POLITECNICO DI TORINO

Scuola di Dottorato

Area: Ingegneria Industriale
SSD: ING-IND/13, Meccanica Applicata alle Macchine

Ph.D. in Mechanics

Diagnostics of machines and structures: dynamic identification and damage detection

Supervisor:
prof. Luigi Garibaldi

Candidate:
Edoardo Gandino

Cycle XXV

Abstract

This research work deals with damage detection of engineering machines and structures. This topic, developed in particular for bearing diagnostics in the first part of the work, is strictly related to dynamic identification when structures are considered. Thus, subspace-based methods are investigated in the second part of the work, with particular attention to nonlinear system identification.

Changes in operational and environmental conditions for structures (such as air temperature, temperature gradients, humidity, wind, etc.) or machines (such as oil temperature, loads, rotating regimes, etc.) are known to have considerable effects on system features and, consequently, on the reliability of diagnostics. Useful tools for eliminating this influence are provided by a Principal Component Analysis (PCA)-based method for damage detection.

The PCA-based method has been applied in many published works for diagnostics of structures, while in this research work a less investigated bearing diagnostic application is considered. After a detailed description of the test rig, the huge amount of acquired data, on several different damaged bearings, is investigated. Results are useful for giving an overview on how the PCA-based method for damage detection can be applied on a complicated real-life machine.

In general cases of real structures, the application of efficient identification techniques is crucial for correctly exploiting the capabilities of the PCA-based method for damage detection. Moreover, in many cases damage causes a structure that initially behaves in a predominantly linear manner to exhibit nonlinear response: the application of nonlinear system identification methods to the feature-extraction process can also be used as a direct detection of damage. For these reasons, a detailed study of the nonlinear subspace-based identification methods is presented in the second part of this work.

Since the classical data-driven subspace method can in some cases be affected by memory limitation problems, two alternative techniques are developed and

demonstrated on numerical and experimental applications. Moreover, a modal counterpart of the nonlinear subspace identification method is introduced, to extend its relevance also to realistic large engineering structures.
In a conclusive application, two of the main sources of non-stationary dynamics, namely the time-variability and the presence of nonlinearity, are analysed through the analytical and experimental study of a time-varying inertia pendulum, whose dynamics is governed by a nonlinear equation of motion due to its large swinging amplitudes.

Contents

Chapter 1

Introduction

Engineering structures and systems are designed to operate within limits specified by the environment in which they will be used. Reliability has always been an important aspect, but a true picture of the behaviour of a system is not available until it is in-service. The solution of implementing very conservative factors of safety produces systems that are heavy and costly; moreover, because of deterioration over time, they will still damage. Maintenance has, thus, been introduced as an efficient way to assure a satisfactory level of reliability during the useful life of a physical system.

The earliest maintenance technique is basically breakdown maintenance (also called unplanned maintenance, or run-to-failure maintenance), which takes place only at breakdowns. A later technique is time-based preventive maintenance (also called planned maintenance), which sets a periodic interval to perform preventive maintenance regardless of the health status of a physical system [1]. Eventually, preventive maintenance has become a major expense of many industrial companies, since products have become more and more complex while better quality and higher reliability are required. Therefore, more efficient maintenance approaches are being implemented to handle the situation, by taking maintenance actions only when there is evidence of abnormal behaviours of a physical system. The history of maintenance technique development for machine tools is briefly summarised in [2]. Indeed, the history applies to other types of machines and systems as well.

Diagnostics and prognostics are two important aspects in a maintenance program [1]. Diagnostics deals with fault detection, isolation and identification when it occurs. Fault detection is a task to indicate whether something is going wrong in

the monitored system; fault isolation is a task to locate the component that is faulty; and fault identification is a task to determine the nature of the fault when it is detected. Prognostics, which involves the determination of the expected remaining useful life of a component, uses the past and current status of a machine to predict the future status with the ultimate goal of using the machine within a safe buffer period to avoid any catastrophic failure [3]. For a successful prognostic system, the first crucial part involves a correct early diagnosis of the fault, which must have sufficient lead time to enable further monitoring and actions.

The present work deals with diagnostics, while prognostics is not treated. In the following section, some concepts of diagnostics will be introduced. Then, Sections 1.2 and 1.3 will focus on diagnostics of machines and structures, respectively, while Section 1.4 deals with pattern recognition methods for diagnostics.

1.1. Diagnostics

An extremely exhaustive overview of diagnostics has been found in [4]. Some of its main concepts, in particular those relevant for the analyses carried out in the present work, are summarised in this section.

In general, all the engineering disciplines are interested in damage evaluation: however, there are four key multidisciplinary areas for which monitoring and assessing damage are principal concerns [4]:

- Condition Monitoring (CM)
- Structural Health Monitoring (SHM)
- Non-Destructive Evaluation (NDE)
- Statistical Process Control (SPC)

Condition Monitoring (CM) is relevant to rotating and reciprocating machinery, such as used in manufacturing. CM also uses on-line techniques that are often vibration-based and use accelerometers as sensors. Structural Health Monitoring (SHM) is relevant to structures such as aircrafts, buildings and bridges and implies a sensor network that monitors the behaviour of the structure on-line. Typical sensors are optical fibres, electrical resistance strain gauges or acoustic

devices; however, all the techniques presented in this work are vibration-based so that accelerometers are used as sensors. Non-Destructive Evaluation (NDE) is usually carried out off-line after the damage has been located using on-line sensors. In some exceptions, NDE is used as a monitoring tool for, for example, pressure vessels and rails. NDE is therefore primarily used for characterisation and as a severity check when there is *a priori* knowledge of the location of the damage. Typical techniques include ultrasound, thermography and shearography. Statistical Process Control (SPC) is process-based rather than structure-based and uses a variety of sensors to monitor changes in the process.
In detail, CM and SHM are addressed in Sections 1.2 and 1.3, for giving an overview of how damage evaluation can be performed for different systems such as rotating machinery or structures. NDE and SPC are not treated in the present work.
In the following, a requisite providing a unified approach to damage evaluation across all the engineering disciplines is given: a precise definition of what constitutes a fault, damage and defect. This unambiguous definition is a primary consideration in developing an intelligent fault detection system. The second step makes use of a hierarchical damage identification scheme.

1.1.1. Definition of fault

All materials and hence all structures contain defects at the nano/microstructural level: the difficulty is to decide when a structure is “damaged”. Because of slight compositional and process variability, materials may have slightly different microstructures and possibly varying numbers or shapes of inclusions, voids and other defects. It is clearly seen that this type of “fault” should not be considered as damage. For the structure in-service, the most common form of damage evolution will be under dynamic load. Here, damage evolves from microcracks and results in changes in the material properties.
A discrimination between when a structure is merely damaged and when a structure contains a fault should be defined: in [4] coherent definitions of faults, damage and defects are established:

- A **fault** is when the structure can no longer operate satisfactorily. If one defines the quality of a structure or system as its fitness for purpose or its ability to meet customer or user requirements, it suffices to define a fault as a change in the system that produces an unacceptable reduction in quality.

- **Damage** is when the structure is no longer operating in its ideal condition but can still function satisfactorily, i.e. in a sub-optimal manner.
- A **defect** is inherent in the material and statistically all materials will contain some unknown amount of defects, this means that the structure can operate at its design condition even if the constituent materials contain defects.

A hierarchical relationship can be developed from the above definitions: defects lead to damage and damage leads to faults. Using this idea, a damage tolerant structure can be designed: it is necessary to decide when the structure is no longer operating in a satisfactory manner. This means that a strict definition of fault has to be given, for example the stiffness of the structure has deteriorated beyond a certain level.

Another consideration is the level of damage tolerance required. An helicopter gearbox is a good example of damage tolerant system: in aeronautics, the monitoring of gearboxes for faults is standard practice. Based on the vibration response of the gearbox, threshold levels are set and, once these are exceeded, a fault is detected and the gearbox is taken out-of-service.

In conclusion, a fault is defined as a change in the condition of the structure, producing an unacceptable quality reduction. By implication, such a change will be evident. Thus, fault detection actually means detecting the damage that will, if not corrected, lead to a fault. Damage detection is just part of a larger problem: damage identification. Before addressing this issue in Section 1.1.3, a brief description of the data acquisition process is given.

1.1.2. Data acquisition

Data acquisition is a process of collecting and storing useful data (information) from targeted physical systems for the purpose of damage identification.

The measurements related to the health condition/state of the physical system are very versatile [1]. It can be vibration data, acoustic data, oil analysis data, temperature, pressure, moisture, humidity, weather or environment data, etc. Various sensors, such as micro-sensors, ultrasonic sensors, acoustic emission sensors, etc., have been designed to collect different types of data [5]. With the rapid development of computer and advanced sensor technologies, data acquisition facilities and technologies have become more powerful and less

expensive. Different techniques for multiple sensor data fusion are also discussed in [1].
After data acquisition, but before processing, data cleaning ensures, or at least increases the chance, that clean (error-free) data are used for further analysis and modelling. Data errors are caused by many factors including the human factor. For monitoring data, data errors may be caused by sensor faults: in this case, sensor fault isolation [6] is the right way to go.

1.1.3. Damage identification

A monitoring system must have the objective of accumulating *sufficient* information about the damage, for taking appropriate remedial action. The system should be restored to high-quality operation or at least safety must be ensured.
The identification problem can be thought as a hierarchical structure, composed of five levels [4].

1. **Detection**: the method gives a qualitative indication that damage might be present in the structure.
2. **Localisation**: the method gives information about the probable position of the damage.
3. **Classification**: the method gives information about the type of damage.
4. **Assessment**: the method gives an estimate of the extent of the damage.
5. **Prediction**: the method offers information about the safety of the system, i.e. estimates a residual life.

In addition, Level 5 needs an understanding of the physics of the damage, i.e. characterisation. Level 1 can also be considered as distinguished from the others, since it can be performed with no prior knowledge of the behaviour of the system when damaged. One of the strategies for addressing damage identification, based on a pattern recognition (see Section 1.4) technique, will be described and applied in Chapters from 2 to 5. It is as general as possible, in the sense that it can be applied to many different types of system, provided that useful (i.e. damage-sensitive) measured data are exploited.

1.2. Rotating machines

Gearbox plays a crucial role in industrial applications, and the condition degradation monitoring of the gearbox is important for its design and maintenance. Rolling element bearing condition monitoring has received considerable attention for many years because the majority of problems in rotating machines are caused by faulty bearings. For this reason, in this section some of the main concepts about bearing diagnostics are summarised. However, gears are also a fundamental part of a gearbox, so that gear condition monitoring is also an important topic. Although gear diagnostics is not treated in this work, some issues are addressed in Section 1.2.3.

Rolling element bearings represent one of the most prevalent and critical components in a majority of machines. There is considerable interest in diagnostics and prognostics of rolling element bearings based on vibration analysis and signal processing, because the major economic benefit from such monitoring comes from being able to predict with reasonable certainty the likely minimum lead time before breakdown.

Some parts of a short history of bearing diagnostics, published in [7], are reported. One of the earliest papers on bearing diagnostics was by Balderston [8] of Boeing in 1969. He recognised that the signals generated by bearing faults were primarily to be found in the high frequency region of resonances excited by the internal impacts, and investigated the natural frequencies of bearing rings and rolling elements, which were often to be found in the response vibrations. Braun [9] made a fundamental analysis of synchronous averaging in 1975, and the basic technique was also applied to bearing signals [10]. This appears to be one of the first references to the fact that bearing signals are not completely periodic, with a random variation in period. Braun made an analysis of the effects of jitter (of the synchronising signal) and likened this to the random spacing of bearing response impulses. This model was much later shown to be incorrect, even though it can give satisfactory results in some situations. At around that time, the “high frequency resonance technique” (HFRT), later called “envelope analysis”, was developed, with the original aim of shifting the frequency analysis from the very high range of resonant carrier frequencies, to the much lower range of the fault frequencies, so that they could be analysed with good resolution [7]. This concept of demodulating high frequency resonant responses led to the development of a number of bearing diagnostic methods, where the demodulated frequency was the

resonance of the transducer itself. Systems including acoustic emission (AE) transducers, with frequency ranges from 50 kHz to 1 MHz, were also introduced at that time: they can often be effective in improving the signal/noise ratio of bearing signal to background noise.

Morever, it is recommended in [7] to choose the appropriate resonance frequency for demodulation in each case. There has long been a discussion on how to choose the optimum bandwidth for the demodulation associated with envelope analysis. For example, prior to the development of the spectral kurtosis based methods described in Section 1.2.2, the best approach was to demodulate the band with the biggest dB change from the original condition, although this does require having reference signals with the bearings in good condition.

In this work, the classical approach based on envelope analysis is not exploited in the experimental application. However, it is summarised in Sections 1.2.1 and 1.2.2 for completeness. A completely different approach to bearing diagnostics is investigated in this work and applied to a specific bearing test rig, in Chapters from 2 to 5. This approach is based on the mathematical tool of Principal Component Analysis (PCA), which is part of a more general set of methods that can be found in the literature as statistical methods based on pattern recognition. These rely on training a pattern recognition system with typical signals representing the different classes to be distinguished. These methods are very general and can be applied to different systems, such as machines or structures. However, they require large amounts of data for the training, and it is very rare that sufficient data can be acquired by experiencing actual faults in practice. Moreover, most published results are not non-dimensionalised and would only apply to a particular bearing on a particular machine for which the system was trained. It is likely that some of these problems will be overcome by fault simulation in the future. For a detailed discussion of methods based on pattern recognition, the reader is referred to [11], while a brief description is given in Section 1.4.

An extremely large amount of case histories, in which condition monitoring is performed by means of envelope analysis or pattern recognition, can be found in literature. Among them, three applications are presented in [7]: (1) a helicopter gearbox test rig, which was run to failure under heavy load; (2) a bearing test rig, on which bearings are tested to failure; (3) a radar tower driving system, consisting of a motor, a gearbox and a spur pinion/ring-gear combination. In [3] a bladed disk test rig, designed to develop models and techniques for monitoring the health of turbomachine blades, is studied. A low speed machinery fault simulator

is presented in [12]: this test rig enables modelling of bearing and gearbox faults under different loading conditions. A fatigue test of an automobile transmission gearbox has been performed in [13]. In [14] a stand-by bearing test rig is analysed, with a spindle driven by a variable speed motor.

1.2.1. Envelope approach

Vibration signals from a defective bearing with a localised fault contain a series of impulse responses, which result from the impacts of the defective part(s) with other elements [3]. These impulses are generated almost periodically and their characteristics depend on the location of the defect; that is, whether it is on the inner race, outer race or rolling elements. In practice, the spacings between the impulses vary randomly to a certain extent, due to slip caused by varying load angle, which leads to the smearing of the defect harmonics at higher frequencies (defect frequencies will appear as discrete harmonics of negligible amplitude in the low frequency region but will be smeared in the high frequency region where their amplitude is amplified by correspondence with resonances). This random slip, while small, does give a fundamental change in the character of the signal, and is the reason why diagnostic information is not available from frequency analyses of the raw signal, in particular at low frequencies, due to the low energy at the bearing frequencies and to the masking by strong background noise [7, 15].
In the vicinity of a resonance (high frequency region), this information could be extracted if no random fluctuations existed, but is often not possible with a small amount of random fluctuation, as the harmonics smear into one other [15]. This problem has been solved by frequency analysing the envelope of the response signal (envelope analysis or high frequency resonance technique (HFRT)) obtained by amplitude demodulation [7, 15]. This enveloping is usually applied to a frequency region where the signal-to-noise ratio (SNR) is the highest, for example around a structural resonance frequency excited by the bearing fault. For this reason, the most powerful bearing diagnostic techniques depend on detecting and enhancing the impulsiveness of the signals.
In the following, the approach based on envelope analysis is briefly presented, by first introducing descriptions on how a faulty bearing can be modelled and how bearing signals can be enhanced.

Bearing fault models

Several studies [16 - 19] have been conducted to explain the mechanism of vibration and noise generation in bearings. Bearings act as a source of vibration and noise due to either varying compliance or the presence of defects in them. Radially loaded rolling element bearings generate vibrations even if they are geometrically perfect [20]. However, a significant increase in the vibration level is caused by the presence of a defect, such as surface roughness, waviness, misaligned races and off-size rolling elements. These defects are caused by manufacturing error, improper installation or abrasive wear, while bearing faults are in general due to fatigue.

Fatigue in rolling element bearings is caused by the application of repeated stresses on a finite volume of material and results in the loss of material from the inner race, the outer race or the rolling elements [3]. Bearing faults usually start as small pits or spalls, and give sharp impulses in the early stages covering a very wide frequency range (even in the ultrasonic frequency range to 100 kHz). As the rolling elements strike a local fault on the outer or inner race a shock is introduced that excites high frequency resonances of the whole structure between the bearing and the response transducer. The same happens when a fault on a rolling element strikes either the inner or outer race. As explained in [16], the series of broadband bursts excited by the shocks is further modulated in amplitude by two factors: (1) the strength of the bursts depends on the load borne by the rolling element(s), and this is normally modulated by the rate at which the fault is passing through the load zone; (2) where the fault is moving, the transfer function of the transmission path varies with respect to the fixed positions of response transducers.

In a number of studies [3, 21], the signature of the vibration signal originating from the passage of a rolling element over the spalled area has been reported as being composed of two main parts. The first originates from the entry of the rolling element into the fault, while the second results from the exit of the rolling element as it strikes the trailing edge of the fault. As the size of the fault increases, the separation between the two points, i.e. the time to impact, increases and if the entry and exit events can be successfully extracted from the vibration signal, the size of the fault can be estimated. In particular, [3] explores the idea of enhancing the two events by means of the envelope approach.

However, for some faults such as brinelling, where a race is indented by the rolling elements giving a permanent plastic deformation, the entry and exit events are not so sharp, and the range of frequencies excited not so wide. They would still generally be detected by envelope analysis, however, as stated in [7]. This

reference also reports that cases have been encountered where faults have not been detected while small and the spalls have become extended and smoothed by wear. Although not necessarily generating sharp impacts any more, this type of fault can often be detected by the way in which it modulates other machine signals, such as the gearmesh signal generated by gears supported by the bearings.
The optimum way to analyse a faulty bearing signal depends on the type of fault present. The main difference is between initial small localised faults and extended spalls, in particular if the spalls become smoothed. Both fault types give rise to signals that can be treated as cyclostationary of order n (i.e. its n-th order statistics must be periodic) [7].

Localised faults

Localised defects include cracks and pits on the rolling surfaces. Whenever a local defect on an element interacts with its mating element, abrupt changes in the contact stresses at the interface result which generates a pulse of very short duration [20]. This pulse produces vibration and noise which can be monitored to detect the presence of a defect in the bearing. As stated in [7], the question arises as to the correct way to model the random spacing of the impacts. Good results were obtained in [22], by modelling the vibration signals from localised bearing faults as cyclostationary of order 2. However, the way of modelling the random variation in pulse spacing in [22] was later found to be incorrect, and in [23] a more correct model was proposed.

Extended spalls

The dominant mode of failure of rolling element bearings is spalling of the races or the rolling elements, caused when a fatigue crack begins below the surface of the metal and propagates towards the surface until a piece of metal breaks away to leave a spall [20].
For extended spalls, there will often be an impact as each rolling element exits the spall, and in that case, envelope analysis will often reveal and diagnose the fault and its type [7]. However, there is a tendency for the spalled area to become worn, in which case the impacts might be much smaller than in the early stages. Such extended spalls can still be detected and diagnosed if the bearing is supporting a machine element such as a gear (see Section 1.2.3).

Enhancement of the bearing signals

One of the major sources of masking of the relatively weak bearing signals is discrete frequency “noise” from gears, since such signals are usually quite strong, even in the absence of gear faults. Even in machines other than gearboxes, there will usually be strong discrete frequency components that may contaminate frequency bands where the bearing signal is otherwise dominant. It is usually advantageous therefore to remove such discrete frequency noise before proceeding with bearing diagnostic analysis [7]. A number of methods are available: (1) Linear prediction [24]; (2) Adaptive noise cancellation (ANC) [25]; (3) Self-adaptive noise cancellation (SANC) [26, 27]; (4) Discrete/random separation (DRS) [28]; (5) Time synchronous averaging (TSA) [29, 30].

Even after removal of discrete frequency “noise”, the bearing signal will often be masked in many frequency bands by other noise, and may also be rendered less impulsive than at the source if the individual fault pulses are modified by passage through a transmission path with a long impulse response (IR) [7].

A method known as minimum entropy deconvolution (MED) removes the effect of the transmission path, for enhancing the bearing signal with respect to residual background noise. The MED method is designed to reduce the spread of IR Functions, to obtain signals closer to the original impulses that gave rise to them. The MED method, which was first proposed in [31], was applied to gear diagnostics in [32] and to bearing diagnostics in [21].

A very powerful technique for enhancing the impulsiveness of bearing signals is Specrtal Kurtosis, which deserves a longer description and it is thus presented in Section 1.2.2.

An alternative to spectral kurtosis methods is to make use of wavelets [33]. Many authors have described the use of wavelets for detecting local faults in gears and bearings (see review in [34]). However, much of the literature on the use of wavelets for machine diagnostics does not take account of all the steps presented here when applying the envelope approach, or makes errors in doing so (as stated in [7]).

Envelope analysis

The spectrum of the raw signal often contains little diagnostic information about bearing faults, and over many years it has been established that the benchmark method for bearing diagnostics is envelope analysis, where a signal is bandpass filtered in a high frequency band in which the fault impulses are amplified by

structural resonances [7]. It is then amplitude demodulated to form the envelope signal, whose spectrum contains the desired diagnostic information in terms of both repetition frequency (ballpass frequency or ballspin frequency) as well as modulation by the appropriate frequency at which the fault is passing through the load zone (or moving with respect to the measurement point) [35].
It was shown in [15] that it is preferable to analyse the squared envelope signal rather than the envelope as such. Ref. [15] also showed that even where the power of the masking noise (random or discrete frequency) was up to three times the power of the bearing signal, in the demodulation band, it was still advantageous to analyse the squared envelope. Using spectral kurtosis, it is usually possible to find a spectrum band where the signal/noise ratio of the bearing signal is much higher.

1.2.2. Spectral Kurtosis

From the earliest days of envelope analysis there has been a debate on how to choose the most suitable band for demodulation. This problem has now largely been solved by the use of spectral kurtosis and the kurtogram to find the most impulsive band (after removal of discrete frequency masking).
Spectral Kurtosis (SK) provides a means of determining which frequency bands contain a signal of maximum impulsivity. It is based on the short time Fourier transform (STFT) and gives a measure of the impulsiveness of a signal as a function of frequency. Kurtosis had long been used as a measure of the severity of machine faults. The application of SK to bearing faults was first outlined in [36, 37].

Definition and calculation

The spectral kurtosis extends the concept of the kurtosis, which is a global value, to that of a function of frequency that indicates how the impulsiveness of a signal, if any, is distributed in the frequency domain. The principle is analogous in all respects to the PSD which decomposes the power of a signal vs frequency, except that fourth-order statistics are used instead of second order. This makes the spectral kurtosis a powerful tool for detecting the presence of transients in a signal, even when they are buried in strong additive noise, by indicating in which frequency bands these take place.

The spectral kurtosis of a signal $x(t)$ may be computed [36] from the STFT $X(t, f)$, that is the local Fourier transform at time t obtained by moving a window along the signal. When seen as a function of t, $X(t, f)$ may be interpreted as the complex envelope of signal $x(t)$ bandpass filtered around frequency f and its squared magnitude will then indicate how energy is flowing in that frequency with respect to time. If that frequency band happens to carry pulses, bursts of energy will then appear. This may be simply detected by computing the kurtosis of the complex envelope $X(t, f)$ as follows:

$$K(f) = \frac{E\left[\left|X(t, f)\right|^4\right]}{E\left[\left|X(t, f)\right|^2\right]^2} - 2,$$

with $E[\cdot]$ the time-averaging operator and where the subtraction of 2 is used to enforce $K(f) = 0$ in the case $X(t, f)$ is *complex* Gaussian (instead of 3 for *real* signals).

Because of the high values it takes at those frequencies where an impulsive bearing fault signal is dominant and because of its theoretical nullity where there is stationary noise only, it makes sense to use the spectral kurtosis as a filter function to filter out that part of the signal with the highest level of impulsiveness [37].

The kurtogram

As previously pointed out, the spectral kurtosis, and therefore the optimal filter which can be obtained from it, will critically depend on the choice of the STFT window length or, equivalently stated, on the bandwidth of the band-pass filter that outputs the complex envelope $X(t, f)$. One solution is to display the spectral kurtosis also as a function of the latter parameter, thus giving rise to a two-dimensional representation called ***kurtogram*** [37].

Computation of the kurtogram for all possible combinations of centre frequencies and bandwidths is obviously costly and not convenient for practical purposes. Suboptimal solutions are however conceivable by subdivision of the bandwidths into rational ratios that permit the use of fast multirate processing. The simplest division in this respect is the dyadic one, where bandwidths are iteratively halved (similar in principle to the FFT algorithm).

In [38], an even finer decomposition is proposed, based on a "1/3-binary tree", where each halved-band is further split into 3 other bands, thus producing a frequency resolution in the sequence $1/2$, $1/3$, $1/4$, $1/6$, $1/8$, $1/12$, …, $1/2^{-k-1}$.

1.2.3. Gear diagnostics

In literature, there is plenty of papers about the study of gears, their modelling and the vibrations they produce. Among them, [39, 40] give general procedures for developing characteristic frequencies (including local fault frequencies) in simple and elaborate gearbox systems. A review of practical techniques and procedures employed to quiet gearboxes and transmission units, by solving the gear noise problem, is presented in [41]. The problem of reducing the noise level or assessing the mechanical condition of a gearbox is also important for planetary gears, which are common in aeronautical and industrial powerplants. The vibration spectra of planetary gears commonly exhibit asymmetry of the modulation sidebands around the meshing frequency: some explanation is given in [42]. In [43] a method to detect gear tooth cracks is proposed, by using the instantaneous phase of the demodulated time signal. The following general description of gear faults can be found in [44].

Gears represent a typical component where the wide frequency range of accelerometers is needed. The basic vibration generating mechanism in gears is the "transmission error" (TE), which can be understood as the relative torsional vibration of the two gears, corrected for the gear ratio. The TE can be expressed as a linear relative displacement along the line of action, which is the same for both gears but represents an angular displacement inversely proportional to the number of teeth on each gear.

The TE results from a combination of geometric errors of the tooth profiles and deflections due to tooth loading. Thus, even a gear with perfect involute profiles will have some TE under load. It is thus important to make comparisons of gear vibration spectra under the same load to obtain information about changes in condition.

Gear vibration signals are dominated by two main types of phenomena [45]:

(1) Effects that are the same for each meshing tooth pair, such as the tooth deflection under load and the uniformly distributed part of initial machining errors and/or wear. These manifest themselves at the toothmeshing frequency and its harmonics. Since there is a pure rolling action at the pitch circle and sliding on

either side, tooth wear tends to occur in two patches on each tooth. Wear is thus often first seen as an increase in the second harmonic of the toothmeshing frequency.
(2) Variations between the teeth, which can be localized or distributed more uniformly around the gears. These manifest themselves at other harmonics of the gear rotational speeds, for the gear on which they are located. Localized faults such as cracks and spalls tend to give a wide range of harmonics and sidebands throughout the spectrum, whereas more slowly changing faults such as those due to eccentricity and distortion during heat treatment, tend to give stronger harmonics grouped around zero frequency and as sidebands around the harmonics of toothmesh frequency.
Since even with faults the same geometric shapes always mesh in the same way, the signals produced by gears are basically deterministic, at least as long as the teeth remain in contact [46].
For light load or very large geometric errors the teeth can lose contact and introduce some randomness or chaotic nature to the signals. For condition monitoring it is better for the loading to be sufficient to maintain tooth contact, to ensure that changes in the vibration signals are due to changes in condition [44].

1.3. Structures

The process of implementing a damage identification strategy for aerospace, civil and mechanical engineering infrastructure is referred to as *structural health monitoring* (SHM). This process involves the observation of a structure or mechanical system over time using periodically spaced measurements, the extraction of damage-sensitive features from these measurements and the statistical analysis of these features to determine the current state of system health [47]. For long-term SHM, the output of this process is periodically updated information regarding the ability of the structure to continue to perform its intended function in light of the inevitable aging and damage accumulation resulting from the operational environments. Under an extreme event, such as an earthquake or unanticipated blast loading, SHM is used for rapid condition screening. This screening is intended to provide, in near real-time, reliable

information about system performance during such extreme events and the subsequent integrity of the system.

1.3.1. Nonlinear dynamics applications

In many cases, in the process of SHM strategy, damage causes a structure that initially behaves in a predominantly linear manner to exhibit nonlinear response when subject to its operating environment and loose parts rattling or sliding against one another. The formation of cracks or delaminations that subsequently open and close under operating loads is an example of such damage. Another type of nonlinearity encountered in engineering systems is the bilinear stiffness characteristics exhibited by a metallic structure that yields during severe loading. An example of such damage is the yielding of steel frame civil engineering structures during an earthquake.
In this section, based on the exhaustive paper by Worden et al. [48], the feature selection portion of the SHM process is discussed, together with the application of nonlinear system identification methods to the feature-extraction process. In particular, the second part of the Thesis (Chapters from 6 to 10) is focused on the development and application of subspace-based identification techniques.
This section is not intended to be a comprehensive review of all damage detection methods rooted in nonlinear dynamics. It provides some concepts for approaching the feature-extraction portion of the damage detection process. These features, which are damage-sensitive and based on nonlinear system response, can either be used as a direct diagnosis of damage or as input to statistical damage classifier, such as the Principal Component Analysis (described in Chapter 2).

Common damage-sensitive features and limitations

A damage-sensitive feature is some quantity extracted from the measured system response data that indicates the presence of damage in a structure. Identifying features that can accurately distinguish a damaged structure from an undamaged one is the focus of most SHM technical literature [49]. The feature-extraction process is based on fitting some model, either physics based or data based, to the measured system response data. The parameters of these models or the predictive errors associated with these models then become the damage sensitive features.

Inherent in many feature-selection processes is the fusing of data from multiple sensors and condensation of these data [48]. A common example of data fusion is the extraction of mode shapes from sensor arrays. Similarly, the extraction of resonant frequencies from measured acceleration time histories can be thought of as a data condensation process. Also, various forms of data normalization are employed in the feature-extraction process in an effort to separate changes in the measured response caused by varying operational and environmental conditions from changes caused by damage. The process of forming a frequency response function (FRF) whereby the measured responses are divided by the measured input can be viewed as a data normalization process. In particular, the present work will focus (see Chapters 2 and 3) on the Principal Component Analysis method for damage detection, in order to take into account the influence of the operational and environmental conditions.

The most common features that have been reported in the SHM literature, and that represent a significant amount of data condensation from the actual measured quantities, are resonant frequencies, mode shape vectors and quantities derived from these parameters. These features are identified by fitting a physics-based model, specifically a lumped-parameter modal model, to measured kinematic response time histories, most often absolute acceleration, or spectra of these time histories. Well-developed experimental modal analysis procedures are applied to the measured response time histories or spectra to estimate the system's modal properties [50, 51]. The fitting process is done using data from the structure in some initial and usually assumed undamaged condition, and then is repeated at periodic intervals or after some potentially damaging event triggers the assessment process. Changes in the modal parameters are then used to indicate the presence and location of damage.

The features described above have several issues associated with them that have prevented their use in most "real-world" applications. First, most of these features involve fitting a linear physics-based model to the measured data from both the healthy and potentially damaged structure. Often these models do not have the fidelity to accurately represent boundary conditions and structural component connectivity, which are prime locations for damage accumulation. Also, this process does not take advantage of changes in the system response that are caused by nonlinear effects. As a result, nonlinear effects tend to be smeared through the linear model-fitting process. From a more practical perspective, real-world structures' modal properties have been shown to be sensitive to changing

environmental and operational conditions [52, 53] and such sensitivity can lead to false indications of damage.

Based on these limitations and the observation that many damage scenarios cause a previously linear structure to exhibit nonlinear behaviour, researchers have developed damage-sensitive features that take advantage of the nonlinear response exhibited by a damaged structure. An exhaustive discussion of some indicators of the presence of nonlinearity is given in [48]: the Harmonic or waveform distortion, the coherence function, the probability density function, two simple correlation tests and the Holder exponent are described. A different approach is analysed in this work. As introduced in the following, it is based on nonlinear dynamical systems theory and identification.

Analysis based on nonlinear dynamical systems theory

The main concept is that, if a given type of damage converts a linear system into a nonlinear system, then any observed manifestations of nonlinearity serve to indicate that damage is present. In this section, a model for representing a crack is described [48], based on nonlinear dynamical systems theory. This model is useful for mathematically representing a relationship between damage (and its extent) and nonlinear contribution.

Consider a simply supported beam. In its undamaged state an assumption that the beam can be modelled as a linear system is quite adequate, but consider what happens when a crack is introduced half-way along its length, as shown in Fig. 1.1. When the beam sags, the effects of the crack are negligible because the two faces of the crack come together and the beam behaves as though the crack was not there. When the beam hogs, however, the presence of the damage must affect the beam because the crack opens and the effective cross-sectional area of the beam is reduced. Under these circumstances, an appropriate model of the beam would perhaps be that shown in Fig. 1.2a, which has the general equation of motion:

$$m\ddot{z}(t) + c_v\dot{z}(t) + \kappa z(t) = f(t), \tag{1.1}$$

where

$$\kappa = \begin{cases} k & z < 0 \\ \alpha k & z \geq 0 \end{cases} = \alpha k - (1-\alpha)k \cdot \left(\frac{\text{sign}(z) - 1}{2} \right). \tag{1.2}$$

When the displacement z of the mass m is positive, the stiffness k of the system is reduced by a factor α. The two-valued stiffness produces an overall restoring force F_κ that is bilinear (Fig. 1.2b). This type of model can be applied to a number of mechanical systems in which moving parts make contact with each other at intermittent points in time.

Different levels of damage can be given to the system by varying the value of the stiffness ratio coefficient α, by assuming that $0 < \alpha \leq 1$ according to model (1.2): the undamaged system is represented by $\alpha = 1$ and higher levels of damage lead to a decreasing value of α. For this reason, it is better to consider a parameter that increases as the damage increase: the coefficient $\beta = (1 - \alpha)$ is preferred, with $0 \leq \beta < 1$. In this case, the undamaged system is represented by $\beta = 0$.

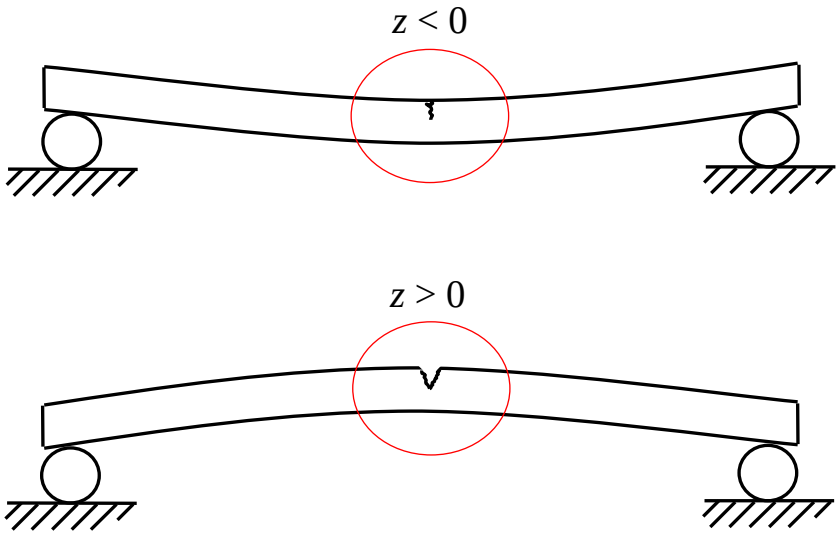
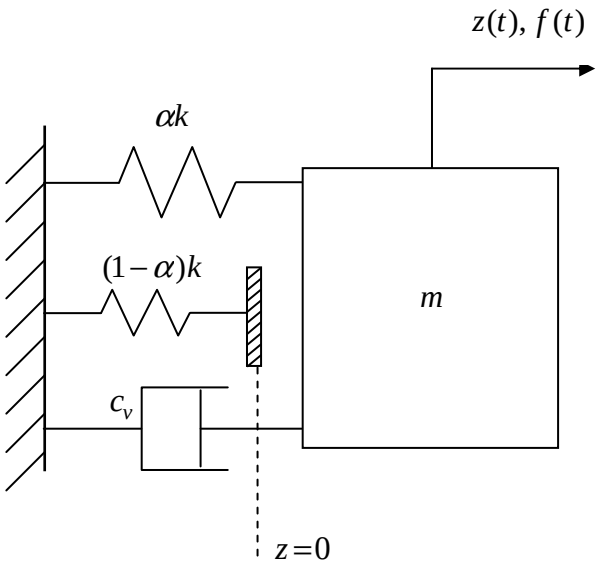


Figure 1.1. A cracked beam under negative and positive deflections [48].

(a)



(b)

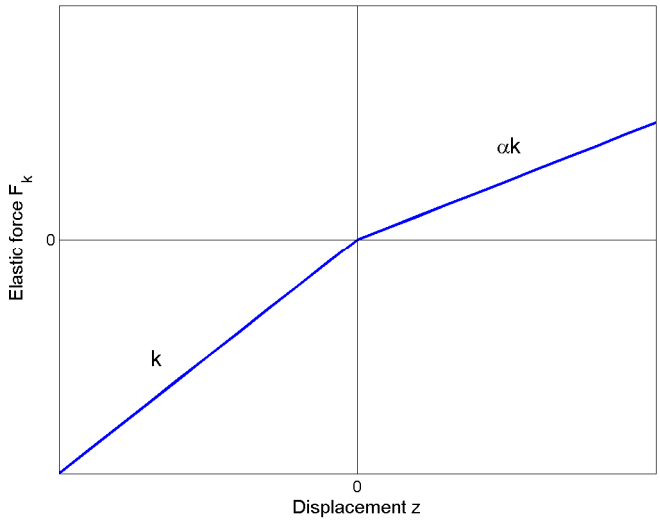


Figure 1.2. (a) Single degree-of-freedom bilinear system. (b) Bilinear elastic force F_κ, with $\alpha = 0.5$.

The question of importance for damage detection purposes is whether the induced bilinearity from a crack can be used to determine the presence of the fault. More specifically, the objective consists of processing the recorded dynamics in order to extract one or more features that detect the crack with a degree of statistical confidence. Further, a deeper level of damage detection is related to obtaining a feature that shows the severity of the damage.
The use of some features proposed in the past, such as the correlation dimension of the attractor and the highest Lyapunov exponent, is discussed in [48], by considering typical nonlinear manifestations such as subharmonic motions and chaotic behaviour. In the present work, a different approach is exploited: nonlinear system identification.

Nonlinear system identification

The nonlinear system identification is a parametric approach which not only allows the immediate characterisation of the nonlinearity, but also yields information on its location. The idea is to obtain estimates of the actual equations of motion of the system of interest using measured time-series data. The presence (and location to an extent) of nonlinearities becomes immediately obvious in this approach, which will be exploited in the second part (Chapters from 6 to 10) of the present work. In fact, this part is focused on subspace-based nonlinear system identification. In the following, a brief state-of-the-art about dynamic identification and subspace methods is given, with some references.
During the last 20 years, there has been a growing interest in subspace-based linear system identification methods. These methods, which process either raw data or correlation matrices, have proven efficient for the identification of the eigenstructure of a linear multivariable system, even without observed inputs [54, 55]. Several variants of stochastic subspace identification exist, but two main categories can be defined, according to the type of data matrices used in the algorithms [56]: data-driven and covariance-driven methods.
In the 90s the control and system identification community carried out an extensive work on data-driven methods. The family of subspace state-space identification (4SID) methods consists of identification techniques which directly estimate a state-space representation [51, 57, 58]. More recently, many researchers started focusing on covariance-driven methods [59 - 61]. A direct estimate of a state-space representation is also obtained, but in a different way.

Nonlinear system identification has been thoroughly investigated in recent years and many efforts have been spent leading to a large number of methods. An exhaustive list of the techniques elaborated to identify the behaviour of nonlinear dynamical systems is hard to write and, moreover, there is no general analysis method that can be applied to all systems in all circumstances. A comprehensive list describing the past and recent developments is given in [62].

One of the established techniques is the Restoring Force Surface (RFS) method [63]: this simple procedure allows a direct identification for single-degree-of-freedom (SDOF) nonlinear systems. There exist in the literature several applications of RFS method to experimental systems: in a recent paper [64], it is applied for the analysis of a nonlinear automotive damper. A similar approach is the Direct Parameter Estimation (DPE) method, which may be applied to multi-degree-of-freedom (MDOF) nonlinear systems: a practical implementation of the procedure is found in [65].

Recent methods are suitable for identification of more complex nonlinear systems, in particular MDOF systems. One of them is the Conditioned Reverse Path (CRP) method [66]: this technique is based on the construction of a hierarchy of uncorrelated response components in the frequency domain, allowing the estimation of the coefficients of the nonlinearities away from the location of the applied excitation. One of the examples of experimental application is given in [67].

More recently, a frequency domain method called Nonlinear Identification through Feedback of the Outputs (NIFO) has been proposed in [68], which has demonstrated [69] some advantages with respect to the CRP, mainly due to the lighter conceptual and computing effort. This method exploits the spatial information and interprets nonlinear forces as unmeasured internal feedback forces.

Starting from the basic idea of NIFO, the Nonlinear Subspace Identification (NSI) method has been developed in [70], showing a higher level of accuracy with respect to NIFO. NSI is a time domain method which exploits the robustness and the high numerical performances of the subspace algorithms. This technique has been deeply investigated in the present work: it is presented, together with some interesting developments and applications, in Chapters from 6 to 10. Several additional references can be found there, referred to the specific arguments dealt with in each of those Chapters.

1.4. Pattern recognition

The idea of **pattern recognition** (PR) is adopted in many modern approaches to damage identification. Generally speaking, a PR algorithm simply assigns a class label to a sample of measured data. The appropriate class labels would encode damage type, location and extent. In order to carry out the higher levels of identification using PR, it will almost certainly be necessary to construct examples of data corresponding to each class [4]. Each possible fault class should usually have a *training set* of measurement vectors that are associated uniquely with it. This type of learning algorithm in which the diagnostic is trained by showing it the desired label for each data set is called **supervised learning**.

The main drawback of supervised learning is that every possible damage situation should be known and data should be available, for training the algorithm with the class labels, from modelling or experiment. The complexity of systems may cause problems in modelling. Moreover, the damage itself may be difficult to model and it may also make the system dynamically nonlinear: a typical structural example is an opening-closing fatigue crack. For experiment, it is simply not possible to damage a real system for accumulating data from all possible damage configurations.

An alternative is **unsupervised learning**, which can only be applied for detection (Level 1 of Section 1.1.3). The techniques are often referred to as *novelty detection* or *anomaly detection* methods [71, 72]. In this case, diagnostics is established by using only training data from the normal operating condition of the system. Any significant deviation from training class is identified as a departure from normal condition, i.e. as acquired damage. An important remark is that only significant deviations should be detected: some criteria must be applied in order to distinguish between statistical fluctuations in the data (for example, measurement noise) and a real deviation from normality. An appropriate approach is *Statistical Pattern Recognition* (SPR).

Another important observation is that there may be variations in the normal conditions that are not statistical: the characteristics of the system may vary with changing environmental conditions, and this must be considered when designing the monitoring system [4]. If the data used to characterise the normal operating condition does not span the whole range of operational and environmental conditions observed in practice, it is likely to signal novelty when a previously unseen condition occurs. A fault will incorrectly be diagnosed.

A technique for preventing this situation to occur, by taking into account all the practical conditions, is presented and applied in Chapters from 2 to 5: it is based on Principal Component Analysis.
Another pattern recognition technique that will be briefly described in Section 1.4.4 is quite different, since it is not based on vibrations. It is the Acoustic Emission method.

1.4.1. Component analysis

Component analysis is a technique of multivariate statistical analysis that can linearly or nonlinearly transform an original set of variables into a substantially smaller set of variables. It can be viewed as a classical method for dimensionality reduction. This technique has been widely applied to virtually every substantive area including cluster analysis, visualization of high-dimensionality data, regression, data compression and pattern recognition.
The Principal Component Analysis (PCA) technique is deeply investigated in Chapter 2, with a discussion (in Section 2.2.5) about other methods based on component analysis. A more detailed description of some extensions of PCA, such as the Kernel-PCA (KPCA) or Local-PCA, can be found in [13]. This reference also introduces a collection of time-domain and frequency-domain statistical features, which should effectively reflect the machine status. PCA is exploited in the first part of this work (Chapters from 2 to 5) for addressing the problem of machine fault diagnosis. However, this technique can be also applied in structural diagnostics, by adopting the identification methods presented in Chapters from 6 to 9 in order to estimate proper features such as the natural frequencies of a structure.
Several papers reported the success of applying component analysis to machine fault diagnosis, often in combination with other pattern recognition techniques. For example, in [73, 74] the combination of component and Support Vector Machine for induction motor fault diagnosis has successfully been implemented. Fault diagnosis of low speed bearings is presented in [12] using a pattern classification method based on Relevance Vector Machine: in this case component analysis was employed with the aim to support the data preparation process. Another class of detection techniques is based on subspace methods: for example, [13] explores subspace-based gearbox condition monitoring using KPCA.

1.4.2. Vector machine

Vector machine methods are here briefly introduced for completeness, without getting into details.

Support Vector Machine (SVM) is a kind of machine learning technique based on statistical learning theory. The basic idea of applying SVM to pattern classification can be stated as follows [12]: first, map the inputs vectors into one features space, possible in higher space, either linearly or nonlinearly, which is relevant with the kernel function. Then, within the feature space from the first step, seek an optimized linear division, that is, construct a hyperplane which separates two classes. However, this technique can also be extended to multi-class classification. SVM training seeks a global optimized solution and avoid over-fitting, so it has the ability to deal with a large number of features. A complete description about SVM is available in [75].

A more recent method is the Relevance Vector Machine (RVM), that uses Bayesian inference to obtain parsimonious solutions for regression and classification. The RVM has an identical functional form to the SVM, but provides probabilistic classification. Interested readers are suggested to refer to [76].

1.4.3. Subspace methods

Subspace methods can be exploited in different applications such as, in particular, diagnostics of machines and structures. These methods can be considered as part of the pattern recognition framework, since they are applied to condition-based maintenance consisting in the early detection of slight deviations with respect to a characterisation of the system in usual working conditions.

When applied to machine diagnostics, subspace methods have recently been more investigated by researchers and were effectively used in pattern recognition. Subspace methods have the good merits of combining feature extraction and pattern classification into one single step. In the method, data in the original pattern space are projected onto a low-dimensional feature subspace extracted by the redundancy reduction techniques, such as PCA, Independent Component Analysis (ICA) and KPCA. For example, in [13] the KPCA technique was chosen to construct the nonlinear subspace for gearbox condition monitoring.

When structures are considered, in many applications the problem of fault detection is solved by investigating changes in the eigenstructure of a linear dynamical system. Several fault detection algorithms, based on subspace-based identification methods and statistical process techniques, are described for example in [77, 78]. Extensions to damage localisation can be found in [79, 80].
Subspace methods are introduced in Chapter 6 and some new techniques are presented in Chapters from 7 to 9 (with an application in Chapter 10).

1.4.4. The acoustic emission method

The Acoustic Emission (AE) method is a high frequency analysis technique which was initially developed as a non-destructive testing (NDT) tool to detect crack growth in materials and structures. The AE technique can be found in a wide area of applications such as structural health monitoring, machine tool monitoring, tribological and wear process monitoring, gear defects monitoring and bearing fault monitoring.
For example, [81] overviews the modern applications of AE technique for monitoring damage in a variety of structures, and the new approaches that have enabled the successful application of the technique, leading to automated crack detection.
The utility of advanced signal processing algorithms and pattern recognition techniques for bearing acoustic emission to achieve early detection of bearing defects is established in [14]. During the bearing operation, bursts of acoustic emissions result from the passage of the defect through the roller and raceway contacts. Defects at different locations of a bearing will have characteristic frequencies at which bursts are generated. Therefore, the signal of a damaged bearing consists of periodic bursts of AE. The signal is usually considered to be amplitude modulated at the characteristic defect frequency. In the end, modulation of the AE by the characteristic defect frequency makes it possible to detect the presence of a defect and diagnose in what part of the bearing the defect appears. The AE method should have an earlier detection capability than is achieved with vibration; moreover, AE is also found to be a better signal than vibrations when the transducers have to be remotely placed from the bearing. A comprehensive and critical review on the application of Acoustic Emission Technology to condition monitoring and diagnostics of rotating machinery is given in [82].

Chapter 2

Principal Component Analysis

Principal Component Analysis (PCA) is one of the most valuable results from applied linear algebra, widely used in all forms of analysis because it is a simple, non-parametric method of extracting relevant information from confusing data sets. A simplified structure often underlies a complex data set: PCA provides a way for reducing it to a lower dimension to reveal this hidden structure, with simple computational issues.

The goal of this chapter is to provide both an intuitive feel for PCA and a thorough discussion of this topic. The mathematical concepts introduced in Sections 2.1 and 2.2 are taken from the excellent tutorial by Shlens [83]. In Section 2.3, particular attention is given on how PCA can be applied for damage detection.

2.1. Motivation

Experimenters are often trying to understand some phenomena by measuring various quantities (e.g. spectra, velocities, voltages, etc.) in a system. Unfortunately, the data appear clouded, unclear and even redundant, so this fundamental obstacle does not allow to figure out what is happening.

A simple example from physics is given to provide an intuitive explanation: the motion of an ideal spring is studied, as shown in Fig. 2.1. The system consists of a

ball of mass m attached to a massless, frictionless spring. The spring is stretched by releasing the ball and it oscillates indefinitely along the x-axis at a set frequency.

This is a standard problem in which the motion along the x direction is solved by an explicit function of time, so the underlying dynamics can be expressed as a function of a single variable x.

However, suppose to ignore which axes and dimensions are important to measure. Thus, the position of the ball is measured in a three-dimensional space, by placing three movie cameras around the system. At every time instant, an image indicating a two-dimensional projection of the position of the ball is recorded by each camera. Supposing the experimenter does not even know what are the *real* x, y and z axes, so three camera axes $\{a,b,c\}$ are chosen, at some arbitrary angles with respect to the system. Moreover, the angles between the measurements might not even be 90°. The cameras record at a set sampling frequency for several minutes. The big question remains how a simple equation of x can be obtained from this data set.

This is what happens in real world. Often, experimenters do not know which measurements best reflect the dynamics of a system. Furthermore, the recorded dimensions are sometimes more than the needed. Also, the real-world problem of noise is always handled. In the spring example this means that air, imperfect cameras or even friction (in case of a not ideal spring) have to be dealt with. To obfuscate the dynamics further, noise contaminates the data set.

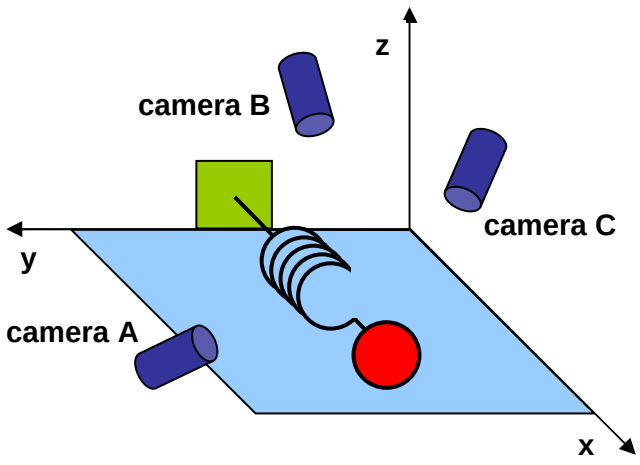


Figure 2.1. A diagram of the spring example [83].

The spring example is used as reference in next Section 2.2, when abstract concepts are introduced. By the end of next section, the question of how to systematically extract x using PCA will be clearly addressed.

2.1. What is behind: mathematics

This section focuses on building a solid intuition for how principal component analysis works; furthermore, it illustrates this knowledge by deriving the mathematics behind PCA [83], from simple intuitions.

2.2.1. Change of basis

The goal of principal component analysis is to compute the most meaningful *basis* to re-express a noisy data set. With this new basis, a hidden structure is supposed to be revealed by filtering out the noise. In other words, the goal of PCA is to discern dimensions to determine which dynamics are important, which are just redundant and which are just noise.

Data can now be defined: every time sample (or experimental trial) is treated as an individual sample in the data set. At each time sample, a set of data consisting of multiple measurements is recorded. Each sample Y is an n-dimensional vector, where n is the number of measurement types (also called features). Equivalently, every sample is a vector that lies in an n-dimensional *vector space* spanned by some orthonormal basis (e.g. the naïve basis $\{e_1 = (1,0,\ldots,0),\ldots,e_n = (0,0,\ldots,1)\}$).

With the assumption of linearity, PCA consists of best re-expressing the data as a linear combination of its original basis vectors. Let $Y \in \mathbf{R}^{n\times N}$ be the original data set, where each column is a single sample (N is the number of samples). Let $X \in \mathbf{R}^{n\times N}$ be another matrix related by a linear transformation $P \in \mathbf{R}^{n\times n}$. Y is the original recorded data set and X is a re-representation of that data set:

$$PY = X \,. \tag{2.1}$$

An interpretation of (2.1), which represents a change of basis, can be seen by writing out the explicit dot products of PY :

$$X = PY = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} \begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix} = \begin{bmatrix} p_1 \cdot y_1 & \cdots & p_1 \cdot y_N \\ \vdots & \ddots & \vdots \\ p_n \cdot y_1 & \cdots & p_n \cdot y_N \end{bmatrix}. \quad (2.2)$$

It can be recognised from (2.2) that the j-th coefficient of each column x_i of X is a projection of the column y_i of Y on to the j-th row of P. Therefore, the rows $\{p_1, \ldots, p_n\}$ of P are indeed a new set of basis vectors for representing the columns of Y.

This new set of basis will become the ***principal components*** of Y. A good choice of basis P depends on what matrix X is wanted to exhibit, as seen in next sections.

2.2.2. Variance

For best expressing the data or, equivalently, for interpreting confused data, three potential interferences can be distinguished: noise, rotation and redundancy. The spring example of Section 2.1 is exploited for better understanding these concepts.

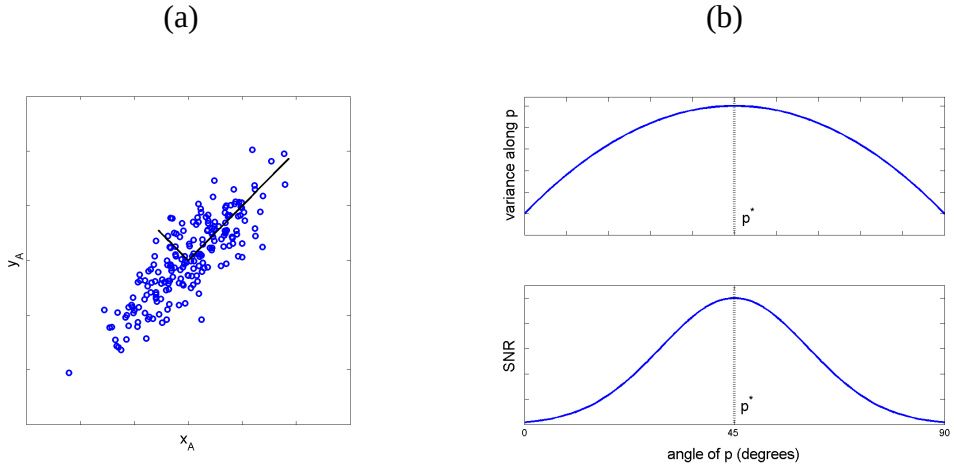


Figure 2.2. (a) Simulated data of (x_A, y_A) for camera A. (b) An optimal p^*, where the variance and SNR are maximized, is found by rotating the axes.

Noise and rotation

No information about a system can be extracted from any data set, if measurement noise is not low, independently from the applied analysis technique. No absolute scale for noise exists, but rather all noise is measured relative to the measurement. A common measure is the following ratio of variances σ^2, called *signal-to-noise ratio (SNR)*:

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}. \tag{2.3}$$

A high SNR (>>1) indicates high precision data, while a low SNR indicates noise contamined data.

Referring to the spring example, suppose that all data from camera A are plotted in Fig. 2.2a: noise consists of any value deviating from straight-line motion. Each line in the diagram represents the variances due to the signal and noise. The SNR is the ratio of the two lengths, measuring how large the cloud is. By considering reasonably good measurements, the dynamics of interest are *assumed* to be contained in the directions with largest variances (and presumably highest SNR) in the vector space of measurements.

Maximizing the variance (and the SNR) consists of finding the appropriate rotation of the naïve basis (Fig. 2.2b). This corresponds to finding the direction p^* that falls along the direction of best-fit line for the data cloud of Fig. 2.2a. Thus, the direction of motion of the spring (for the 2-D case) would be revealed by rotating the naïve basis to lie parallel to p^*. A generalisation will be given in Section 2.2.3.

Redundancy

Fig. 2.2 introduces an additional confounding factor in the data: redundancy. This issue can be clearly demonstrated through the spring example, in which the same dynamic information is recorded by multiple sensors. Fig. 2.3 shows a range of possible plots between two arbitrary measurement types r_1 and r_2. Fig. 2.3a depicts two recordings (for example, $(x_A, humidity)$ [83]) with no apparent relationship: in other words, r_1 is entirely uncorrelated with r_2. Since in this case r_1 cannot be predicted from r_2, they are statistically independent. On the other side, Fig. 2.3c depicts highly correlated recordings: for example it can be a plot of

(x_A, x_B) if cameras A and B are very nearby. Clearly in this case r_1 can be obtained from r_2 (or vice versa) using the best-fit line. In this way, with a reduced number of sensor recordings, a more concise expression of data would be extracted. This is the main idea behind dimensional reduction.

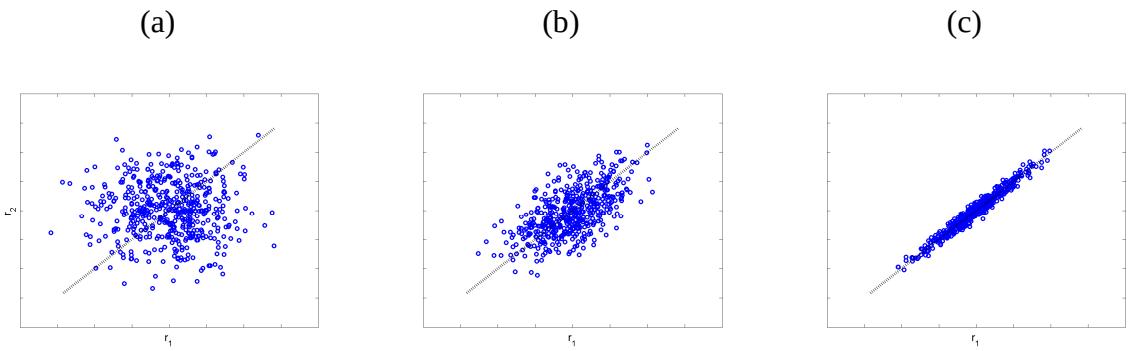


Figure 2.3. Representation of possible redundancies in data from the two separate recordings r_1 and r_2. From (a) to (c): from low to high redundancies. The best-fit line $r_2 = kr_1$ is indicated by the dashed line.

2.2.3. Covariance matrix

In this section the previously described concepts are generalised to arbitrarily higher dimensions. Consider two sets of measurements with zero means (because the means have been subtracted off or are zero):

$$A = \{a_1, a_2, \ldots, a_N\}, \qquad B = \{b_1, b_2, \ldots, b_N\}.$$

The variances of A and B are defined as:

$$\sigma_A^2 = E[a_i a_i], \qquad \sigma_B^2 = E[b_i b_i],$$

where the expectation $E[\cdot]$ is the average over N variables.

The ***covariance*** between A and B is a straight-forward generalisation:

$$\sigma_{AB}^2 = E[a_i b_i].$$

The covariance measures the degree of the linear relationship between two variables. A large (or, on the contrary, small) value indicates high (low)

redundancy. For example, in the case of Figs. 2.2a and 2.3c the covariances are large.

By converting A and B into the corresponding row vectors,

$$a = [a_1 \quad a_2 \quad \dots \quad a_N], \qquad b = [b_1 \quad b_2 \quad \dots \quad b_N],$$

the covariance may be expressed as a dot product matrix computation

$$\sigma_{ab}^2 = \frac{1}{n-1} ab^T,$$

where $1/n-1$ is a constant for normalisation providing an unbiased estimation.

Finally, a generalisation from two vectors to an arbitrary number can be performed by defining a new matrix $Y \in \mathbf{R}^{n \times N}$:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Each row y_i corresponds to all measurements of a particular type. Each column of Y corresponds to a set of measurements from one particular trial. A definition for the ***covariance matrix*** C_Y can now be given:

$$C_Y = \frac{1}{n-1} YY^T. \tag{2.4}$$

Specifically, the ij-th element of C_Y is the dot product between the vector of the i-th measurement type with the vector of the j-th measurement type.

Some properties of the square symmetric matrix $C_Y \in \mathbf{R}^{n \times n}$ can be summarised, reflecting the noise and redundancy in the measurements [83]:

- The diagonal terms are the *variance* of particular measurement types. Large (small) values correspond to interesting dynamics (or noise).
- The off-diagonal terms are the *covariance* between measurement types. Large (small) values correspond to high (low) redundancy.

Diagonalisation

If matrix C_Y could be manipulated, the objectives should be: (1) to minimize redundancy, measured by covariance, and (2) to maximize the signal, measured

by variance. Such an optimised matrix (defined as C_X) must be diagonal, since all its off-diagonal terms should be zero.

PCA applies a method for diagonalising C_X, by assuming that P is an orthonormal matrix and that the directions with the largest variances are the most "important". In the simple example of Fig. 2.2b, P acts as a generalised rotation to align a basis with the axis having maximum variance. In multiple dimensions this could be performed by a simple linear algebra algorithm. The resulting ordered set of row vectors p_i of P are the ***principal components***.

This method allows for judging the "importance" of each principal direction. Namely, the variances associated with each direction p_i quantify how "principal" each direction is [83]. Each basis vector p_i can thus be rank-ordered according to the corresponding variances: the concept of dimensional reduction is appearing again.

2.2.4. Solving PCA

Two solutions to PCA are derived using linear algebra; some of its concepts and theorems are not treated in details. The reader is referred to [84] for them.

Given a data set Y, the goal consists of finding some orthonormal matrix P where $X = PY$ such that $C_X \equiv \frac{1}{n-1} XX^T$ is diagonalised. The rows of P are the *principal components* of Y.

Eigenvectors of covariance

Matrix C_X is rewritten in terms of P:

$$C_X = \frac{1}{n-1} XX^T = \frac{1}{n-1}(PY)(PY)^T = \frac{1}{n-1} P(YY^T)P^T = \frac{1}{n-1} PAP^T, \tag{2.5}$$

where the new matrix $A = YY^T$ is symmetric.

The keypoint is to recognise that a symmetric matrix is diagonalised by an orthogonal matrix of its eigenvectors:

$$A = VDV^T, \tag{2.6}$$

where D is a diagonal matrix and V is a matrix of eigenvectors of A, arranged as columns.

Matrix P can be selected such that each row p_i is an eigenvector of YY^T. By this selection, $P \equiv V^T$ and substituting into (2.6) $A = P^T DP$ is obtained. Moreover, the inverse of an orthogonal matrix is its transpose, so $P^{-1} = P^T$ and the evaluation (2.5) of C_X can be concluded:

$$C_X = \frac{1}{n-1} PAP^T = \frac{1}{n-1} P(P^T DP)P^T = \frac{1}{n-1} PP^{-1} DPP^{-1} = \frac{1}{n-1} D. \tag{2.7}$$

It is evident that the choice of P diagonalises C_X and these matrices summarise the results of PCA [83]:

- The principal components of Y are the eigenvectors of YY^T, or the rows of P.
- The i-th diagonal value of C_X is the variance of Y along p_i.

Singular Value Decomposition

Another algebraic solution for PCA is derived and, in the process, it is found that PCA is closely related to Singular Value Decomposition (SVD), which is a more general method of understanding change of basis.

The decomposition is written in its final form as follows [85, 86]:

$$A = U\Sigma V^T, \tag{2.8}$$

stating that any arbitrary matrix $A \in \mathbf{R}^{N \times n}$ can be converted to an orthogonal matrix $U \in \mathbf{R}^{N \times N}$, a diagonal matrix $\Sigma \in \mathbf{R}^{N \times n}$ and another orthogonal matrix $V \in \mathbf{R}^{n \times n}$.

By manipulating (2.8),

$$U^T A = \Sigma V^T \equiv Z.$$

Hence, U^T is a *change of basis* from A to Z: U^T is termed the *column space* of A since it transforms column vectors, meaning that it is a basis that spans the columns of A.

By symmetry to SVD (2.8),

$$V^T A^T = U^T \Sigma \equiv Z.$$

The rows of V^T (or the columns of V) are an orthonormal basis for transforming A^T into Z. It follows that V is an orthonormal basis spanning the *row space* of A.

With some computations PCA can be demonstrated to fall within the framework of SVD, so that the two methods are related. By returning to the original $n \times N$ data matrix Y, a new $N \times n$ matrix X can be defined:

$$X = \frac{1}{\sqrt{n-1}} Y^T,$$

where each column of X has zero mean. The definition of X becomes clear by analysing $X^T X$:

$$X^T X = \left(\frac{1}{\sqrt{n-1}} Y^T \right)^T \left(\frac{1}{\sqrt{n-1}} Y^T \right) = \frac{1}{n-1} Y Y^T = C_Y.$$

By construction $X^T X$ is equal to the covariance matrix of Y. From previous subsection, the principal components of Y are the eigenvectors of C_Y. By computing the SVD of X as in (2.8), the columns of matrix V contain the eigenvectors of $X^T X = C_Y$. Therefore, the columns of V are the principal components of Y.

As a final interpretation, observe that V spans the row space of X. Therefore, it must also span the column space of $X^T = \frac{1}{\sqrt{n-1}} Y$. The conclusion is that finding the principal components amounts to finding an orthonormal basis that spans the *column space* of Y [83].

2.2.5. Discussion

Dimensional reduction

The main benefit of PCA is that the variances C_X associated with the principal components can be examined. Large values of variances are associated with the first $m < n$ principal components and often a sudden jump to very small values can be seen. The conclusion is that most interesting dynamics occur only in the first m dimensions.

This process, named *dimensional reduction*, of throwing out the less important axes can help in revealing hidden, simplified dynamics in high dimensional data.

Limits and extensions

PCA is a non-parametric analysis. This can be viewed as a weakness, if some features of the structure of a system are known a-priori. These assumptions should be incorporated into a parametric algorithm.
As an example, consider the motion along a circumference [83]: PCA finds two orthonormal principal components, but this answer is not optimal. By recognising that the phase contains all dynamic information, the appropriate parametric algorithm is to first convert the data to the appropriately centered polar coordinates and then compute PCA.
This parametric algorithm is termed *kernel PCA*. The procedure is parametric because the user must incorporate prior knowledge of the structure in the selection of the kernel [87].
Sometimes the assumptions may be too restrictive. For example, there may be situations in which principal components need not to be orthogonal, or the distributions along each dimension are not needed to be Gaussian. This less constrained set of problems is not trivial and has been solved adequately by means of *Independent Component Analysis* (ICA) [88].

Local PCA for nonlinear cases

PCA is limited by its linearity and may sometimes be too simple for dealing with real-world data especially when the relations among variables are nonlinear. Thus, nonlinear generalisations of PCA have emerged, as for instance vector quantization principal component analysis (VQPCA) [89].
VQPCA is a local implementation of PCA that involves a two-step procedure: a clustering of the data space into several disjoint regions and the estimation of the principal axes within each region. From a numerical point of view, this leads to piecewise application of PCA in each local region.
The application of VQPCA for damage diagnosis has been studied in [90], in which a close look at the choice of the distortion function used in data clustering leads to new clustering strategies.

2.3. PCA for damage detection

PCA is a multi-variate statistical method which is very useful for eliminating environmental effects in damage detection. Changes in environmental conditions for structures (such as air temperature, temperature gradients, humidity, wind, etc.) or machines (such as oil temperature, loads, rotating regimes, etc.) are known to have considerable effects on signal features. Most of the time, environmental variables are not measured but their effects are merely observed from the variation of the measured features [53].
The PCA-based method has been successfully used not only for time-invariant systems, but also for time-varying ones, as a bridge with crossing loads [91].

2.3.1. Methodology

The n-dimensional vector y_k denotes a set of signal features identified at time t_k, ($k = 1, \ldots, N$) with N the number of samplings. All the samples are collected in a matrix $Y \in \mathbf{R}^{n \times N}$. For example, if the natural frequencies are chosen as features, n represents the number of selected modes. PCA provides the following linear mapping of data from the original dimension n to a lower dimension m:

$$X = PY, \tag{2.9}$$

where $X \in \mathbf{R}^{m \times N}$ is called the scores matrix and $P \in \mathbf{R}^{m \times n}$ the loading matrix. The dimension m may be interpreted as the physical order of the system which can be here related to the number of combined environmental factors that affect the features. By such dimensional reduction, the system is forced to learn the inherent variables driving changes of the features and to capture the embedded relation between the environmental factors and the features [53].

Selecting the number of components

As seen in Section 2.2.4, matrix P may be calculated by extracting the main m eigenvectors of the covariance matrix of Y. Alternatively, a more practical method is to perform the SVD of the covariance matrix of the features:

$$YY^T = U\Sigma^2 U^T , \tag{2.10}$$

with

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix},$$

where U is an orthonormal matrix, the columns of which define the principal components and form a subspace spanning the data. By definition, $\Sigma_1 = diag(\sigma_1, \sigma_2, \ldots, \sigma_m)$ and $\Sigma_2 = diag(\sigma_{m+1}, \sigma_{m+2}, \ldots, \sigma_n)$ and the singular values are written in decreasing order

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m >> \sigma_{m+1} \geq \cdots \geq \sigma_n \rightarrow 0 .$$

In most practical situations the elements of Σ_2 are assumed to be small with respect to the elements of Σ_1, but they are not equal to zero due to the effect of noise (see Section 2.2.2) and/or the presence of nonlinear effects.

The reason of this dimensional reduction, as seen in Section 2.2.5, is that among all the possible environmental factors, only m of them have a strong influence on the features and have to be considered. Mathematically, the features approximately remain in the hyperplane defined by the m adopted principal components. However, the selection of an appropriate dimension m is not so critical as it appears [53]. Since the relative change of this hyperplane from the reference to the current states is considered, it is possible to obtain stable monitoring results with different values of the order m. In some practical situations temperature is the only significant environmental factor (so $m = 1$), while where the number of factors is not known a priori or it is difficult to find by observing the singular values, a verification by choosing a series of order m may be considered.

Novelty detection technique

From the decomposition in (2.10), the first m columns of U may be used to build matrix P in (2.9), which is exploited to project the measured features into the environmental-factor characterised space. Some information is lost when performing this projection. This loss can be assessed by re-mapping the projected data back to the original space:

$$\hat{Y} = P^T X = P^T P Y . \tag{2.11}$$

The residual error matrix R is estimated as

$$R = Y - \hat{Y} . \tag{2.12}$$

From the residual error vector R_k obtained at time t_k, the Novelty Index (NI) [92] can be defined by using different types of norm, such as the Euclidean norm:

$$NI_k^E = \|R_k\|_2 \tag{2.13}$$

or the Mahalanobis norm

$$NI_k^M = \sqrt{R_k{}^T C_Y R_k} , \tag{2.14}$$

where $C_Y = \frac{1}{n-1} YY^T$ is the covariance matrix of the features. If the Euclidean or Mahalanobis indices are further assumed to be normally distributed, statistical analysis may be performed. The following quantities can be defined for the prediction in the reference state by constructing an X-bar control chart [93]:

Mean value: $\overline{NI}$, Standard deviation: σ,

Threshold value: $Th = \overline{NI} + \alpha\sigma$,

where coefficient α is taken equal to 3, which corresponds to a confidence interval of 99.7% with the assumption of a normal distribution.

After performing this "training" procedure for the reference (which is definitely healthy) state, a "test" procedure is carried out by evaluating N_c current states: equations from (2.11) to (2.14) are applied to a current feature matrix $Y_c \in \mathbf{R}^{n \times N_c}$ in place of the reference feature matrix Y and the resulting Novelty Indexes are compared to the threshold value in order to investigate if damage has occurred.

Outlier statistics allows counting how many prediction errors p_e (represented in percentage) overpass the threshold: this percentage can be used as an indicator of the damage detection. In the absence of damage, the features corresponding to the current data should lie in the hyperplane spanned by the features of the reference state. The outlier statistics value of the current data should thus remain at the same level as for the reference data and $p_e = 0$ %. Conversely, damages should cause a departure of the features from the original hyperplane, and the outlier statistics of the damaged state increase significantly: in the damaged case the prediction errror should be $p_e = 100$ %.

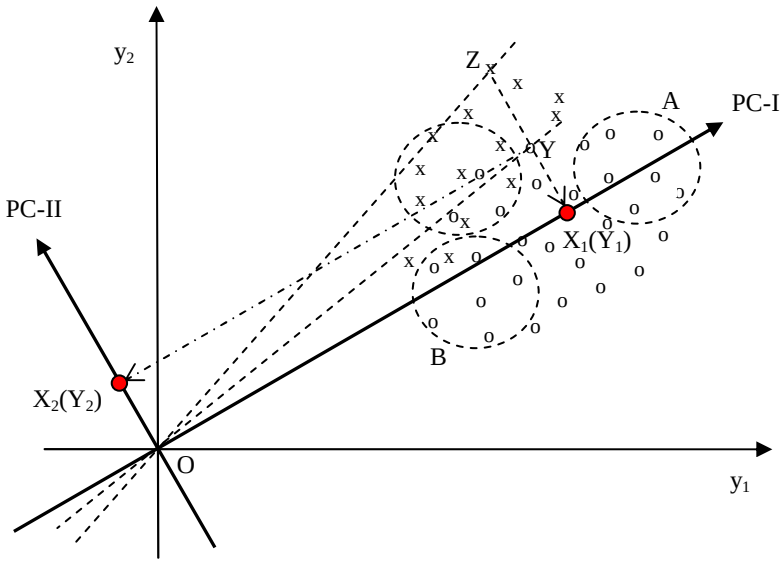


Figure 2.4. Geometric interpretation [53].

In addition to the outlier statistics, the ratio $\overline{NI}_c/\overline{NI}_r$ (c and r denote, respectively, the current and the reference states) may also be used as a quantitative indicator of the damage level [53]. While $\overline{NI}_c/\overline{NI}_r \to 1$ means no damage, a relatively large ratio $\overline{NI}_c/\overline{NI}_r$ corresponds to the opposite situation.

2.3.2. Geometric interpretation

To illustrate the method, an exhaustive geometric interpretation is given in [53]: a two dimensional case, i.e. when two features y_1 and y_2 are considered, is presented since it is the easiest to depict in figures.

In Fig. 2.4, the reference features are represented by circles: they are distributed around their geometric centre. It is assumed that the dispersion of the features is mainly due to environmental variations. The application of PCA to this data set gives two principal components, namely PC-I and PC-II. The first one is associated with the highest singular value and is responsible for the greatest variation of the features; so it corresponds to the main environmental factor (or a combined effect of several factors). By adopting the terminology introduced in Section 2.2, PC-I is associated to the most important dynamics of the system. PC-II represents the effect of secondary factors.

Point Y is selected as an example. It is first projected into the 1D space spanned by PC-I, according to (2.9). The result is a scalar equal to the length of segment OX_1. This data point is then re-mapped into the original 2D space, resulting in point Y_1: the corresponding residual error is $Y\text{-}Y_1$. By adopting the Euclidean norm as in (2.13), the Novelty Index associated to point Y is given by the length of segment Y_1Y, thus $NI_Y^E = \|Y - Y_1\|$.

Suppose now the current set of features (represented by crosses in Fig. 2.4) is obtained from the damaged state. Then, the environmental factors influence the new features in a different way, in comparison with the reference features, although the reference data may be intersected by some of the current data as shown in Fig. 2.4. Consider point Z, chosen in such a way that its re-mapping results again in point Y_1. By processing it as done for Y, the residual error $Z\text{-}Y_1$ leads to a $NI_Z^E = \|Z - Y_1\|$ that increases significantly with respect to NI_Y^E. In such a comparison between healthy and damaged states, the effect of the environmental factors has been approximately eliminated.

In the classical PCA approach as described in Section 2.2, a data normalisation procedure is generally required, leading to variables with zero mean and unitary standard deviation:

$$y_k^* = \frac{(y_k - \bar{y})}{\sigma_y},$$

where $\bar{y}$ and σ_y are, respectively, the mean and standard deviation of each data set. In [53] it is pointed out that such normalisation should be avoided in the present case where PCA is exploited for damage detection. To illustrate this, consider again the two data sets represented in Fig. 2.4. If the mean value of each data set is removed, Fig. 2.5a shows that the features corresponding to the damaged state are merged with those corresponding to the reference state, so that damage may no longer be detected. The problem may be avoided if the damaged-state data are normalised by removing always the mean value of the reference data rather than its own mean value, as shown in Fig. 2.5b.

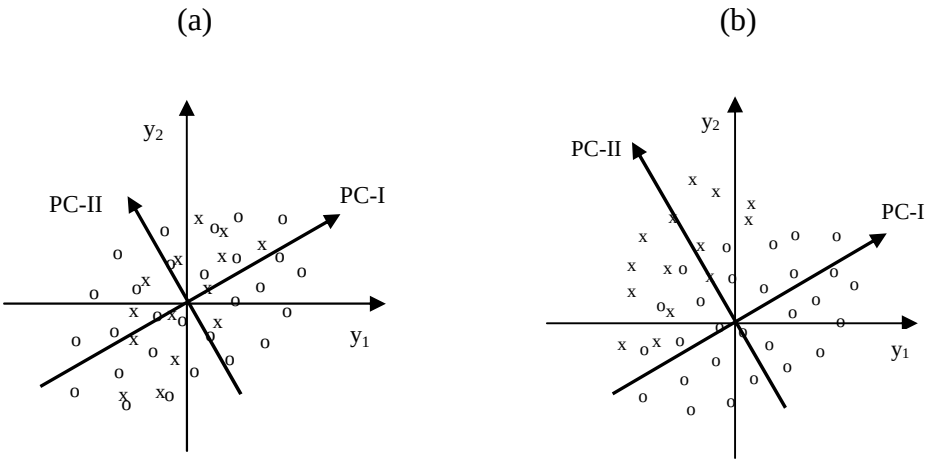


Figure 2.5. Geometric interpretation of PCA with data normalisation [53]: (a) eliminating mean value of each set; (b) eliminating mean value of reference data for both sets.

By using the non-normalised features, it may be demonstrated that, in linear (or weakly nonlinear) cases, the PCA-based detection method is robust even when features are identified in a limited range of environmental variations [53]. This is due to the fact that the PC-I calculated from data set A in Fig. 2.4 is approximately the same as the one calculated from data set B or from the full reference data set. In addition, a part of data from damaged state (e.g., data set C in Fig. 2.4) may be compared with any data set of reference to issue a damage alarm. If data normalisation is adopted as in Fig. 2.5b, this robustness is lost, and one needs to identify the reference features for the whole set of data representative enough of all the environmental variations.

In summary, without data normalisation, it may not be necessary to measure the features in reference state on the full range of environmental variations. It may not be even necessary to care about the environmental conditions in which the features are measured. This is the main reason why the PCA-based method is useful for damage detection. However, it should be kept in mind that the assumption of linear behaviour is very stringent: it is well known that completely linear cases are seldom in practical applications. Therefore, no clear conclusions may be obtained when features are available only from limited environmental variations. A preliminary study of the system, on the full range of environmental variations, should be carried out when possible, to check the range of validity of the linearity assumption. An example is given in Chapter 5.

Chapter 3

Numerical application of PCA for damage detection

In this chapter a simple numerical example is proposed to illustrate the PCA-based method for damage detection under varying environmental conditions, as described in Chapter 2.

3.1. Five degrees of freedom system

The five degrees of freedom system shown in Fig. 3.1 is considered. A damage is introduced between masses 1 and 2, as indicated by the red region in the figure, in the form of a stiffness reduction of 10%. The five masses have the same value of $m = 2$ kg, while the system is assumed to be made of two materials. The stiffness of both materials is considered as temperature dependent, as shown in Fig. 3.2.

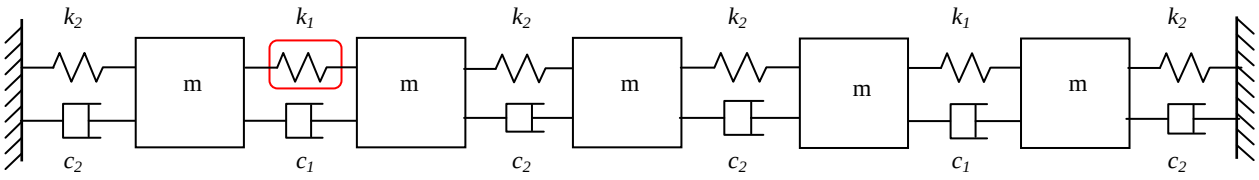
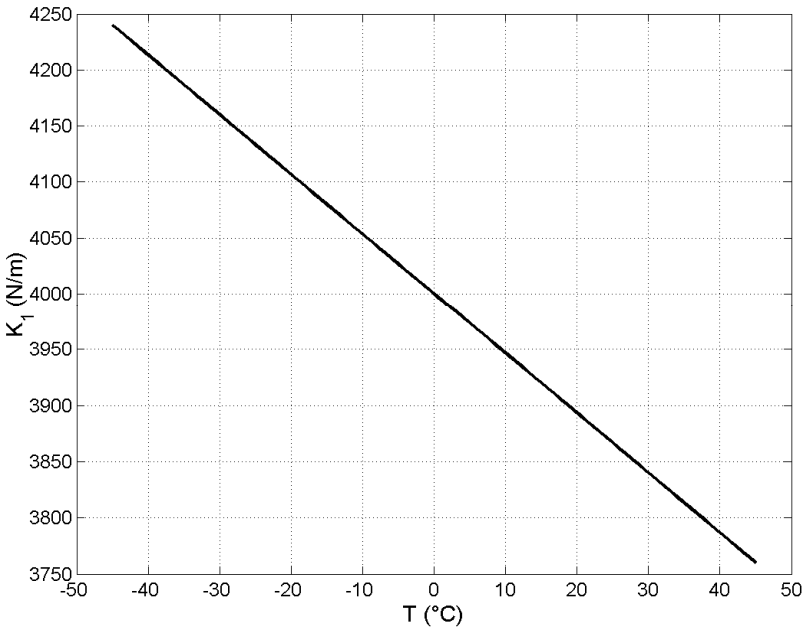


Figure 3.1. Representation of the five degrees of freedom system. Damage is located in the red region in the form of a stiffness reduction of 10%.

(a)



(b)

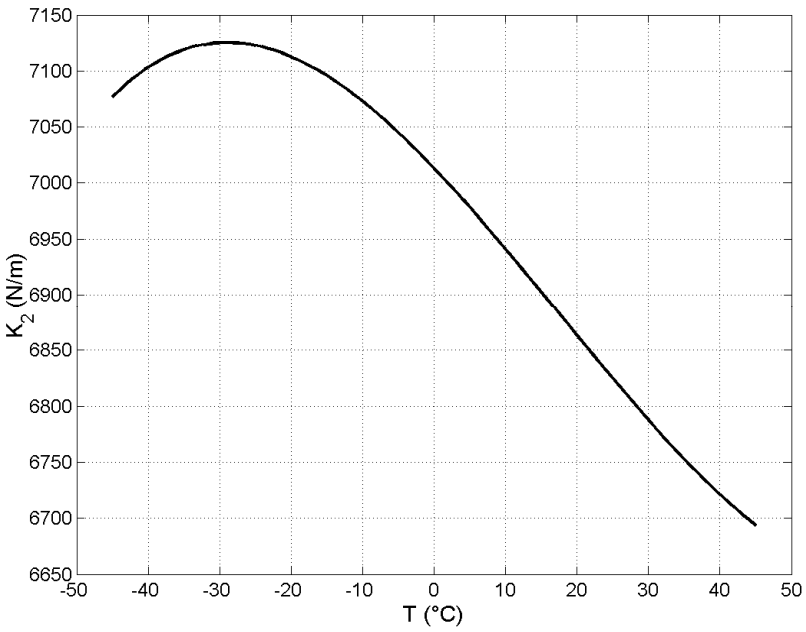


Figure 3.2. Assumed stiffness of the two materials versus temperature: (a) k_1; (b) k_2.

Observe that all the parameters associated to this example are not assumed to be quantitatively related to physics: the aim is to illustrate the capabilities of the PCA-based method from a qualitatively point of view. Even if they are useless in this application, two values of damping $c_1 = 2.8$ Ns/m and $c_2 = 4.8$ Ns/m are also defined for completeness.

The (non-normalised) natural frequencies of the 5 modes of the system are considered as the vibration features: their computation is performed every 0.3 °C in the temperature interval assumed for each of the presented cases. Since temperature is the environmental factor under the action of which the system is monitored, it is assumed to be unknown and unmeasured.

Observe that for real structures the natural frequencies can be computed by performing an identification procedure starting from measurements of input and/or output data. This issue is fundamental for applying PCA in structural health monitoring and some identification methods will be addressed in Chapters from 6 to 9.

3.2. Quasi-linear case

In this section the focus is on the temperature interval from 0 to 30 °C. Fig. 3.3 shows the evolution of the natural frequencies f_i $(i = 2,\ldots,5)$ as a function of f_1. The frequencies are scaled with respect to their mean values, for better visualisation. Fig. 3.3 reveals the presence of a weak nonlinearity due to the existence of two different materials and, in particular, to the nonlinear characteristic of k_2 versus temperature. Since the nonlinearity is weak, the assumption of linearity is made in order to apply the concepts of Section 2.3.

It may be observed that both temperature variations and structural damage are responsible for changes of the natural frequencies, as shown in Fig. 3.4 for modes 3 and 4. In absence of a precise correlation between the variations of the temperature and of the features, since the former are not measured, a simple comparison between the features identified in different environmental conditions does not lead to a clear diagnostics of possible damages [53]. This can be seen in Fig. 3.4 by noticing, for example, that the same natural frequencies can be associated to two conditions: the reference (healthy) system at higher temperatures or the damaged system at lower temperatures.

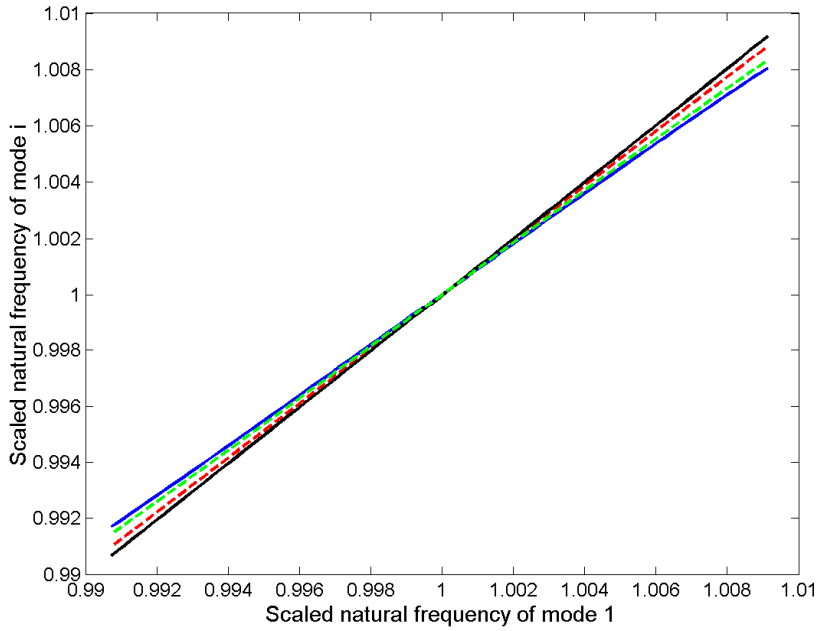


Figure 3.3. Diagram showing weak nonlinearity: evolution of the scaled natural frequencies f_i $(i = 2,\ldots,5)$ as a function of f_1.

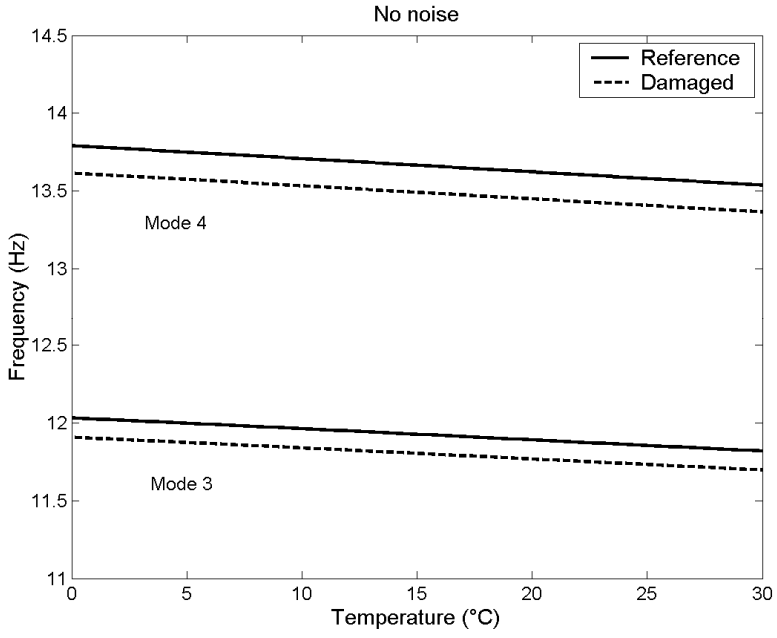


Figure 3.4. Natural frequencies of modes 3 and 4 varying with temperature, for the reference and the damaged case, with no added noise.

3.2.1. Effect of noise

Fig. 3.4 has been obtained by representing the exact values of the natural frequencies, i.e. without accounting for the presence of random effects. To do so, in this section the vibration features are perturbed by adding a certain amount of noise: by referring to the comments about Fig. 3.4, it will be shown that more noise leads to a less clear diagnostics of possible damages.

A high amount of noise, obtained by adding to the features a Gaussian random vector with a standard deviation of 1% of their exact value, is considered in Fig. 3.5. The reference and the damaged conditions merge together and the natural frequencies associated to them are hardly separated: an incorrect diagnostics is expected. This is demonstrated by Fig. 3.6, in which only 2 data acquisitions out of 100 are identified as damaged (and they are very close to the threshold anyway) and the ratio $\overline{NI}_c / \overline{NI}_r$ is very close to 1.

A low amount of noise, obtained by adding to the features a Gaussian random vector with a standard deviation of 0.1% of their exact value, is considered in Fig. 3.7. A clear separation between the reference and the damaged conditions is observable, so that damage is expected to be detected by the PCA-based procedure. This is shown in Fig. 3.8a: the complete set of 100 data acquisitions is correctly identified as damaged, with a large distance from the threshold demonstrated by a high ratio $\overline{NI}_c / \overline{NI}_r$.

In general cases of real structures, the concept of noise is related to errors coming from the performed identification procedures: a low amount of noise is associated to a high level of agreement when estimating the natural frequencies. Again, as stated in Section 3.1, the application of efficient identification techniques is crucial for correctly exploiting the capabilities of the PCA-based method for damage detection: more and more accurate estimates of the natural frequencies lead the method to be more and more sensitive to damage extent. For this reason, a detailed study of the subspace-based identification methods will be addressed in Chapters from 6 to 9.

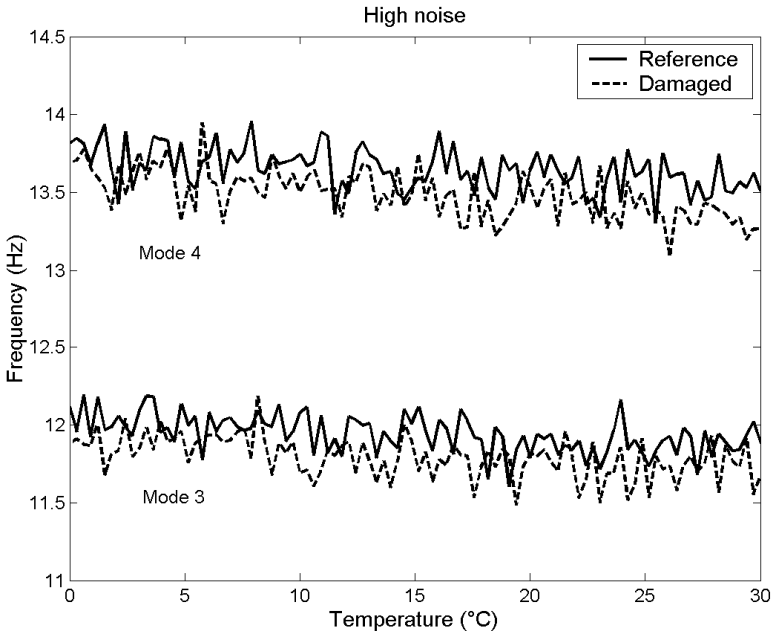


Figure 3.5. Natural frequencies of modes 3 and 4 varying with temperature, for the reference and the damaged case, with a high level (1%) of added noise.

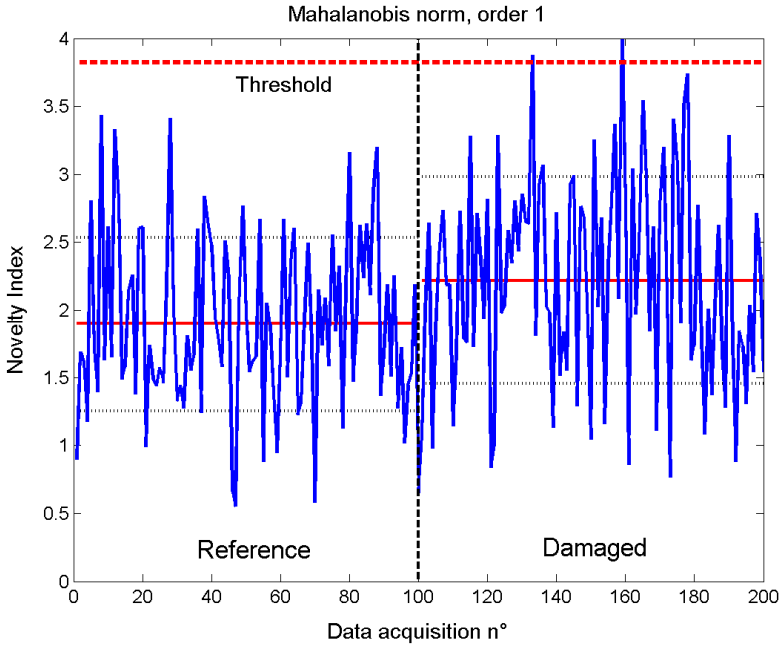


Figure 3.6. Incorrect damage detection in case of a high level (1%) of added noise. The reference data are on the left of the vertical dashed line, while data from the damaged system are on the right. The threshold is represented by the horizontal dashed red line.

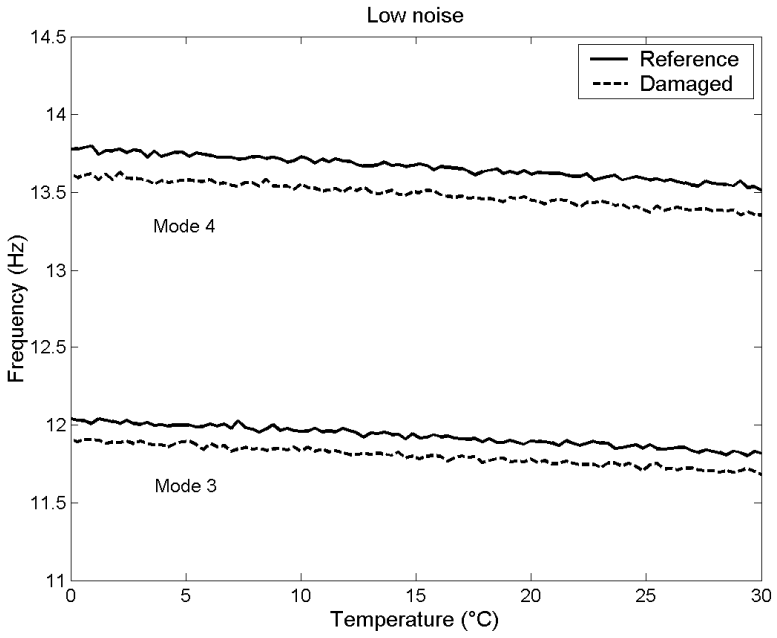


Figure 3.7. Natural frequencies of modes 3 and 4 varying with temperature, for the reference and the damaged case, with a low level (0.1%) of added noise.

3.2.2. Results

The 0.1% amount of noise is retained for illustrating all the results of the present and the following sections. In order to assess the efficiency of the PCA-based damage detection method, two objectives are investigated and the results are presented in the following:

- The method should be able to detect existing damages independently of the environmental conditions at which the identification of the vibration features is performed. In Fig. 3.8a the complete set of 100 data acquisitions is correctly identified as damaged: temperature effects have been removed by the method. Moreover, since non-normalised features are used, Fig. 3.8b is helpful to demonstrate what stated in Section 2.3.2: the PCA-based detection method is robust even when features are identified in a limited range of environmental variations. In fact, in Fig. 3.8b the data acquisition on the damaged structure has been realised at a lower temperature range (0 to 15 °C) than the measurements on the healthy structure (15 to 30 °C). Due to temperature effects, the difference between the features corresponding to the healthy and the damaged structure, respectively, could have been blurred [53]. However,

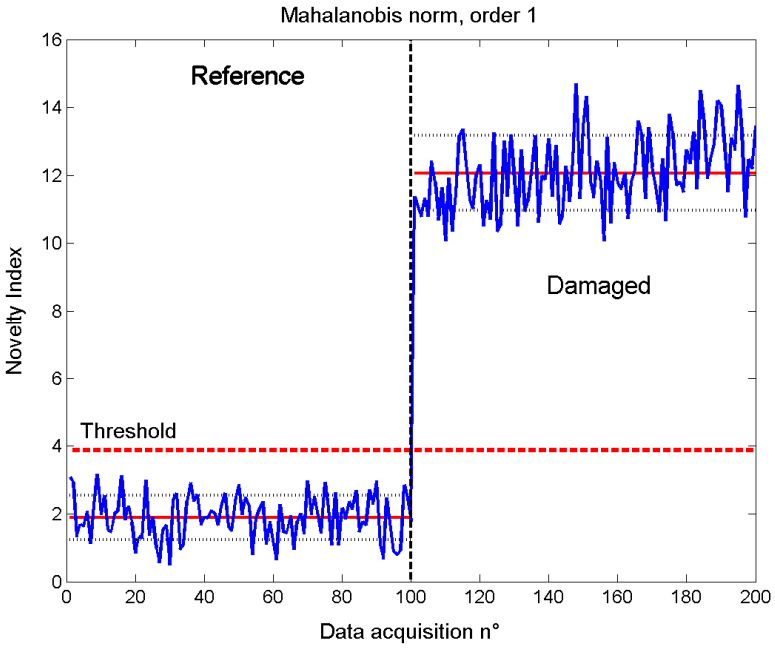
as indicated in Fig. 3.8b, this has not been an obstacle for the method, which clearly detects the presence of damage.

In the end, Figs. 3.8a and 3.8b can be compared. The parts related to reference data are similar, in terms of mean value $\overline{NI}_r$ and threshold: this is due to the application of PCA in combination with the Mahalanobis norm (2.14), which involves the covariance matrix of the features C_y. Referring to the parts related to the damaged system, their NIs both show a slight increasing trend: this is due to the weak nonlinearity of the features. Moreover, the mean value $\overline{NI}_c$ in Fig. 3.8b is a bit larger than the one in Fig. 3.8a. Again, the reason is weak nonlinearity. The principal component (just one, since $m=1$) calculated from a data subset in Fig. 3.8b is not exactly the same (as expected in a linear case) as the one calculated from the full reference data set in Fig. 3.8a: the effect is seen in the NIs of data from damaged state.

- No alarm should be issued if no damage occurs even when measurements are performed under different environmental conditions. A second set of 100 healthy data is computed to be monitored by the PCA-based method. In Fig. 3.9a the complete set is correctly identified as healthy: only 1 data acquisition out of 100 is identified as damaged (but it is very close to the threshold anyway) and the ratio $\overline{NI}_c/\overline{NI}_r$ is very close to 1. In Fig. 3.9b the data acquisition on the monitored structure has been realised at a lower temperature range (0 to 15 °C) than the measurements on the reference structure (15 to 30 °C): the false-positive verification is correct, with only 2 small false alarms.

In this example, the main environmental variable is the temperature: then, a single principal component was retained in the calculation (i.e. $m=1$). However, it should be observed that no significant difference can be seen in the results by considering more principal components in the analysis. For instance, Figs. 3.10 and 3.11 display the results obtained by using $m=3$ principal components. In particular, Fig. 3.10 is the corresponding of Fig. 3.8 and the same comments hold; Fig. 3.11 is the corresponding of Fig. 3.9, the same comments hold but with no false alarms in Fig. 3.11a and only 1 small false alarm in Fig. 3.11b.

(a)



(b)

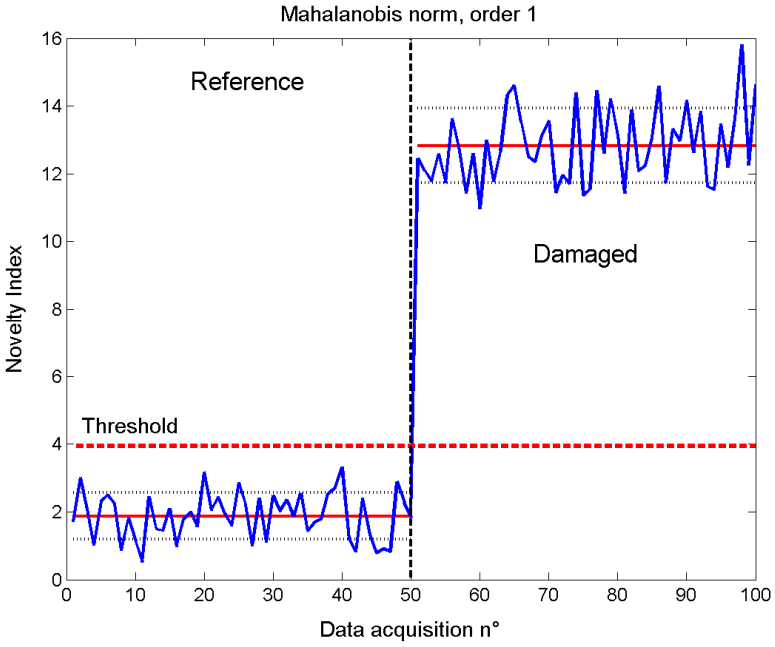
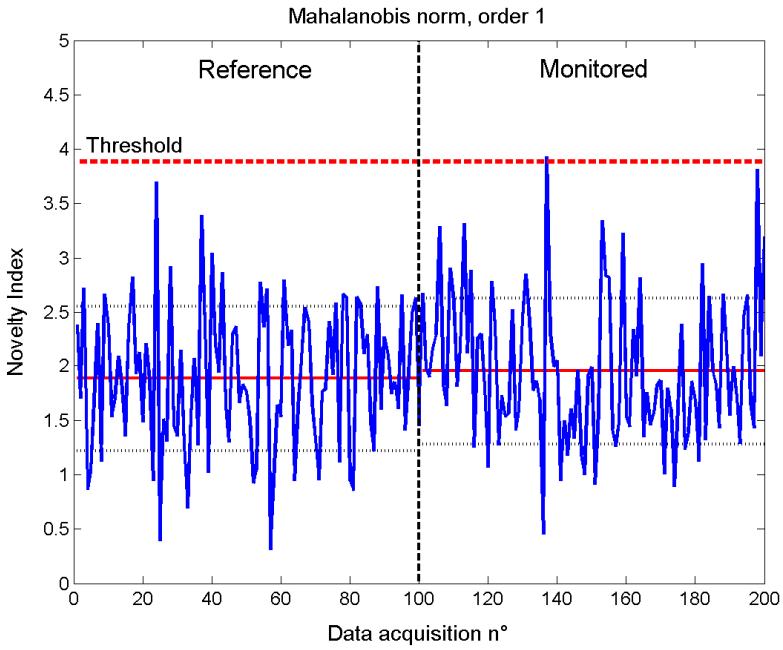


Figure 3.8. (a) Damage detection using full range (from 0 to 30 °C) data of reference and damaged system. (b) Damage detection using two sets of data at different temperatures: reference from 15 to 30 °C and damaged from 0 to 15 °C.

(a)



(b)

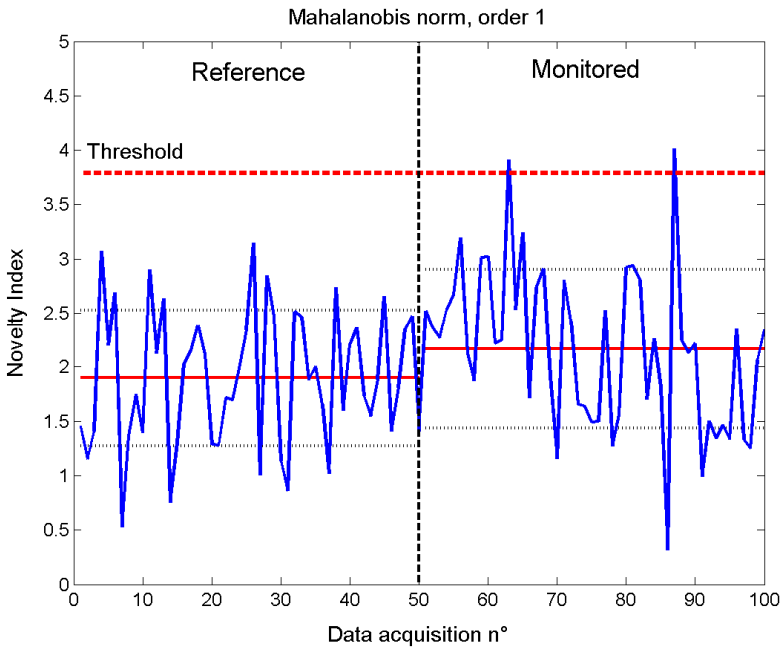
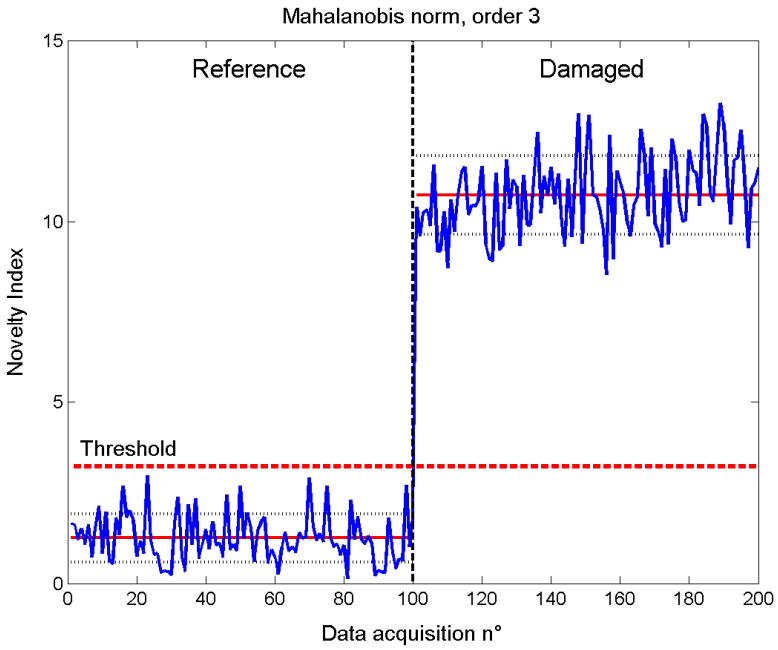


Figure 3.9. (a) False-positive verification using full range (from 0 to 30 °C) data of reference and monitored system. (b) False-positive verification using two sets of data at different temperatures: reference from 15 to 30 °C and monitored from 0 to 15 °C.

(a)



(b)

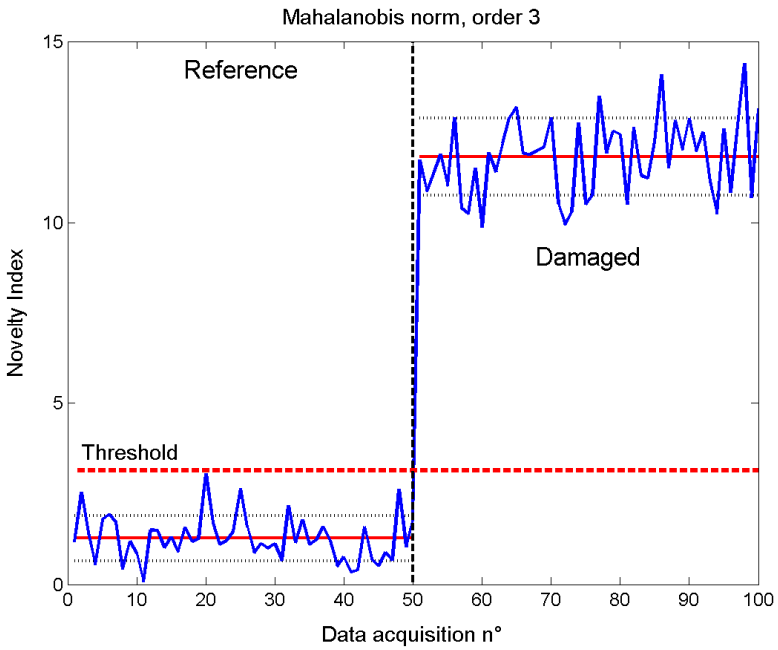
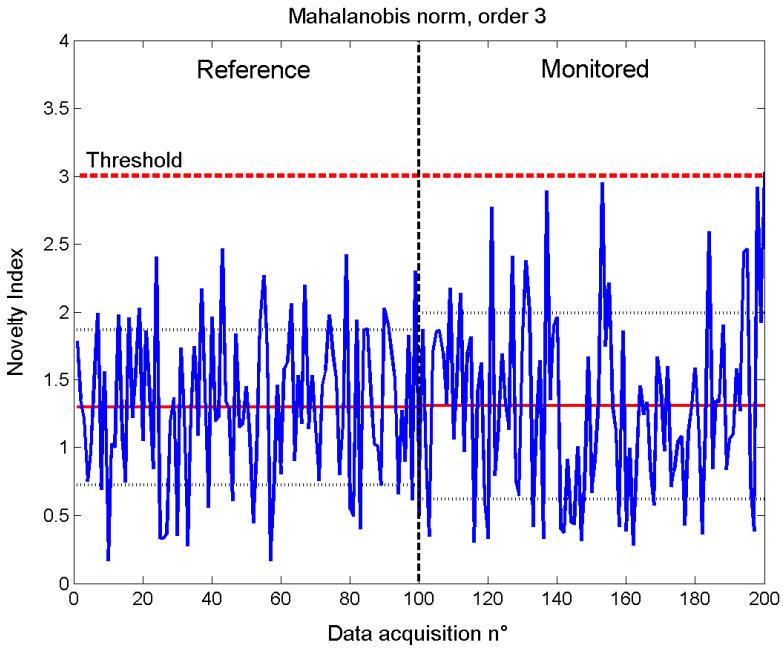


Figure 3.10. $m = 3$. (a) Damage detection using full range (from 0 to 30 °C) data of reference and damaged system. (b) Damage detection using two sets of data at different temperatures: reference from 15 to 30 °C and damaged from 0 to 15 °C.

(a)



(b)

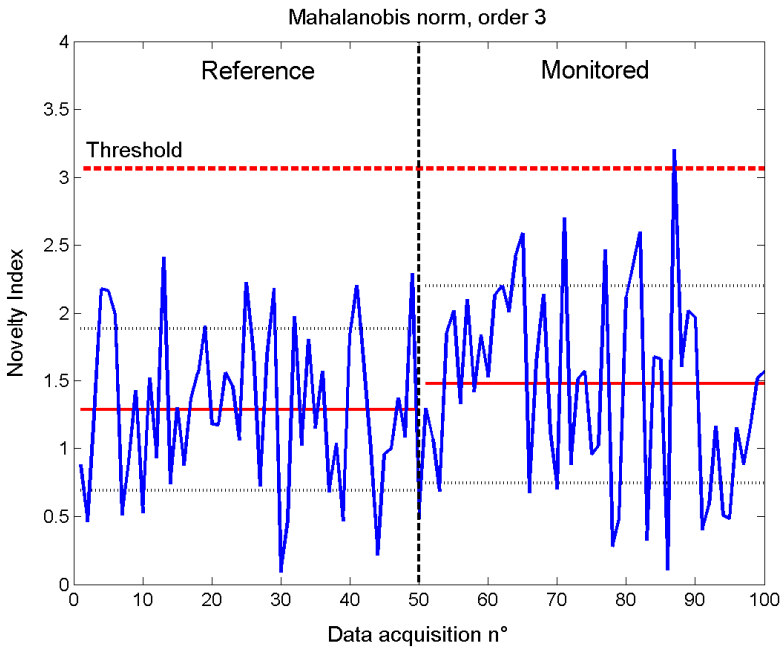


Figure 3.11. $m = 3$. (a) False-positive verification using full range (from 0 to 30 °C) data of reference and monitored system. (b) False-positive verification using two sets of data at different temperatures: reference from 15 to 30 °C and monitored from 0 to 15 °C.

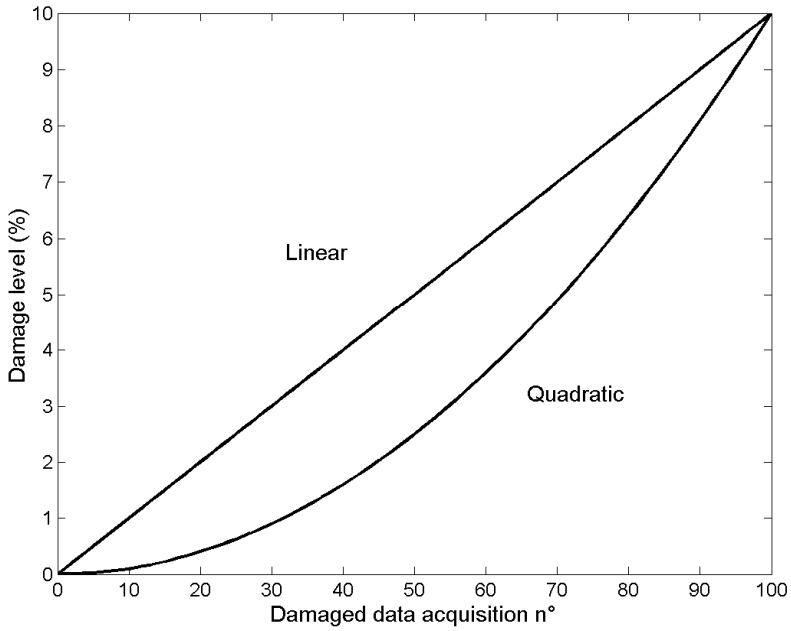


Figure 3.12. Evolution of the damage level percentage α: linear and quadratic representation.

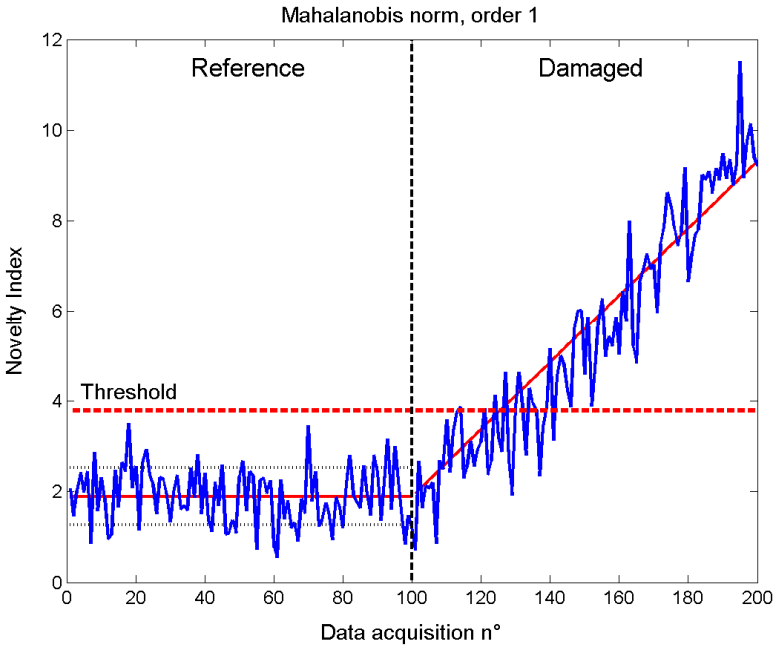
3.3. Damage evolution

The sensitivity of the PCA-based method to damage extent is investigated in this section. The considered temperature interval is from 0 to 30 °C as in previous section, but now each feature acquisition is computed with a temperature chosen randomly from a Uniform distribution $\mathcal{U}$ (0, 30). The aim is demonstrating the capability of the PCA-based method to detect an increasing damage extent independently of any temperature characteristic over time, i.e. to uncouple the time evolutions of damage and temperature.

The damage location is between masses 1 and 2, as in Fig. 3.1, but in this case the damage level is expressed in percentage by a function of time $\alpha(t)$. The damaged stiffness can be represented as $\bar{k}$ defined by the following:

$$\bar{k}(t,T) = \left(1 - \frac{\alpha(t)}{100}\right) k_1(T),$$

(a)



(b)

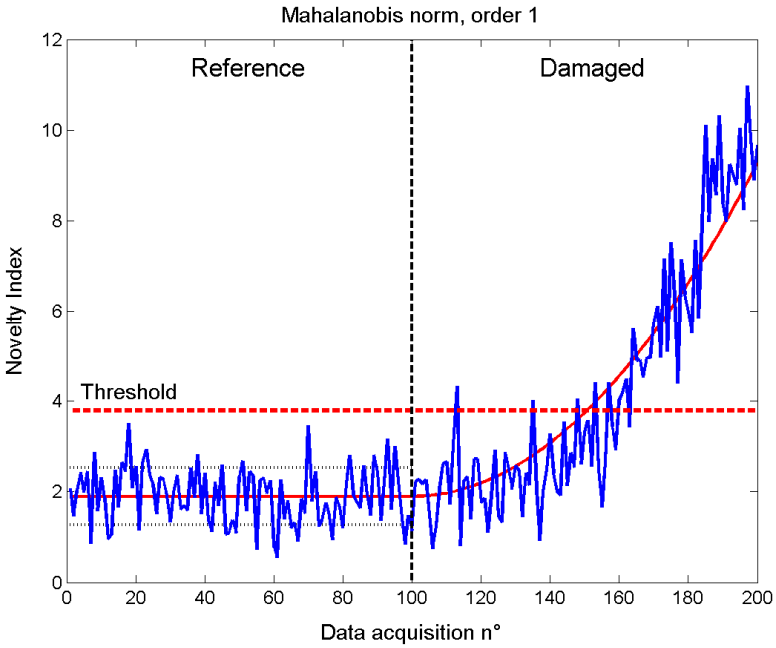


Figure 3.13. Damage detection using the complete data set of reference and damaged system. (a) Linear damage evolution. (b) Quadratic damage evolution.

where $T \sim \mathcal{U}$ (0, 30) °C for each data acquisition. Two characteristics for representing the damage level are considered, as shown in Fig. 3.12: a linear and a quadratic function for increasing damage from 0 to 10%. Observe that the data acquisition number is adopted instead of time, since the study is qualitative and they can be related by specifying some feature sampling frequency. The reference data set is assumed to be the same for both cases.

The capability of the PCA-based method to detect the evolution of damage extent is shown in Fig. 3.13. A qualitative evolution line is traced (in red) for the data sets of the damaged system, from $\overline{NI}_r$ to the last values of NI, for both linear (Fig. 3.13a) and quadratic (Fig. 3.13b) cases: the NIs computed for the damaged cases are in agreement with these evolution lines.

A further quantitative consideration can be done by comparing Fig. 3.13a and Fig. 3.13b, since the reference states are the same. In the linear case, the threshold is crossed after about 35 damaged data acquisitions (Fig. 3.13a), corresponding in Fig. 3.12 to a damage level of about 3.5%. On the contrary, from Fig. 3.12 such a damage level should be reached after about 60 damaged data acquisitions in the quadratic case. This is exactly what can be observed in Fig. 3.13b: the threshold is crossed after about 60 damaged data acquisitions.

3.4. Nonlinear case

The robustness of the method (still with linearity assumptions) is investigated in a slightly stronger nonlinear case, by extending the temperature range (Fig. 3.2) from -45 to 45 °C. The strong nonlinearity is shown in Fig. 3.14, which represents the evolution of the natural frequencies f_i $(i = 2,\ldots,5)$ as a function of f_1. The frequencies are scaled with respect to their mean values, for better visualisation.

The results of the damage detection are given in Figs. 3.15a and 3.15b. In Fig. 3.15a the data acquisition on the damaged structure has been realised at a lower temperature range (-45 to 0 °C) than the measurements on the healthy structure (0 to 45 °C): in this case data contained in region A, corresponding to the high nonlinear effects, are not identified as damaged. On the contrary, when the features start again to behave in a quasi-linear way (as in the reference region), the damage can be detected. A similar situation occurs in Fig. 3.15b, where the complete set of 300 data acquisitions is considered: in this case both the reference

and the damaged data sets include data influenced by the nonlinearity in regions B and C, respectively. Nonlinearity causes outliers in region B and it is also the reason why the damage detection fails in region C. Moreover, in Fig. 3.15 the NIs computed for the damaged case show a trend that can incorrectly be interpreted as a damage evolution: this is again due to the stronger nonlinear nature of the features.

Another main problem when applying the PCA-based method to a nonlinear case occurs in the false-positive verification, as shown in Fig. 3.16a. When measurements are only available for two different temperature intervals (reference from 0 to 45 °C and monitored from -45 and 0 °C) a false or unclear detection is possible. The opposite of what happens in Fig. 3.15a occurs: in Fig. 3.16a, in fact, healthy data corresponding to the high nonlinear effects are incorrectly detected as damage. On the contrary, when the features start again to behave in a quasi-linear way (as in the reference region), the NIs return under the threshold value.

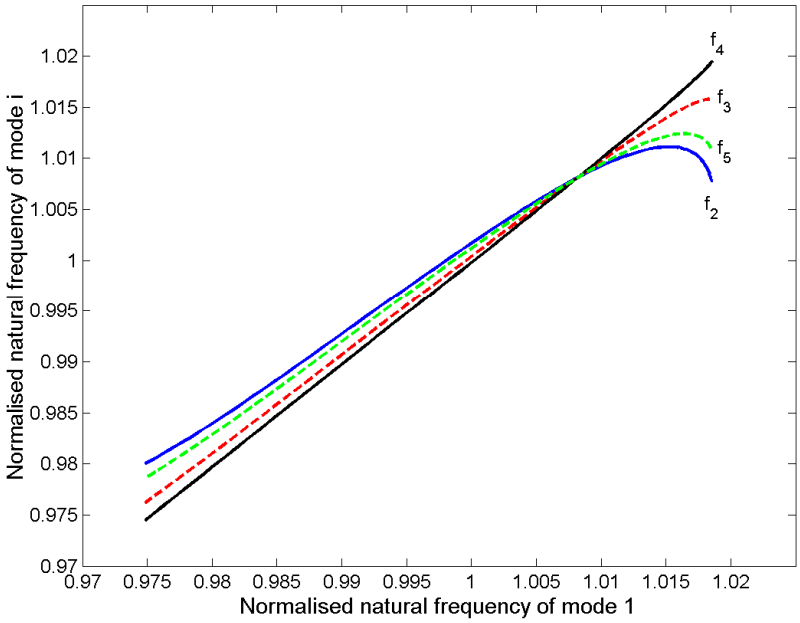
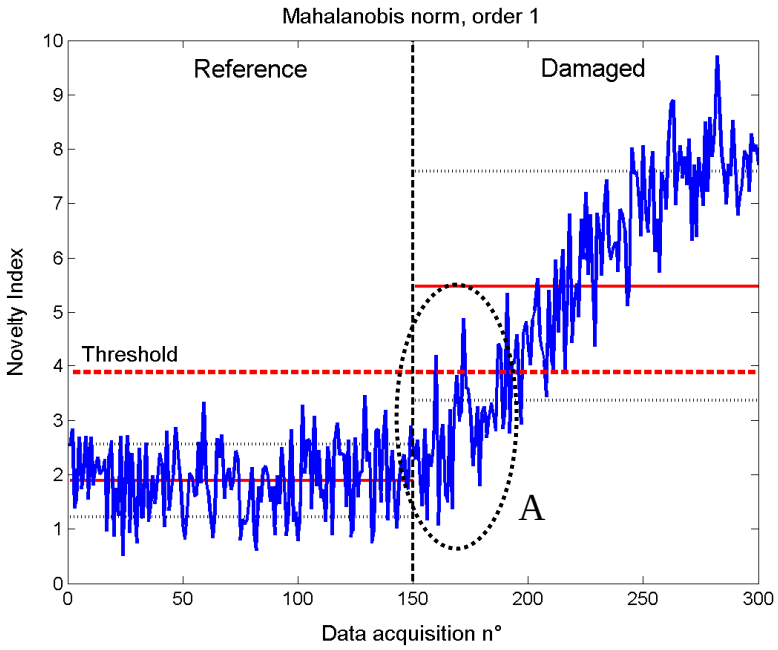


Figure 3.14. Diagram showing stronger nonlinearity: evolution of the scaled natural frequencies f_i $(i = 2,\ldots,5)$ as a function of f_1.

(a)



(b)

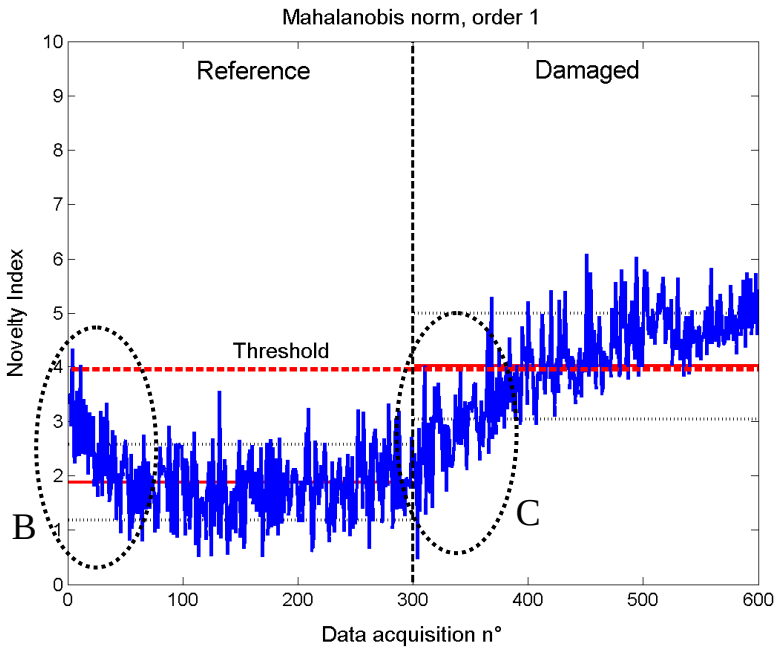
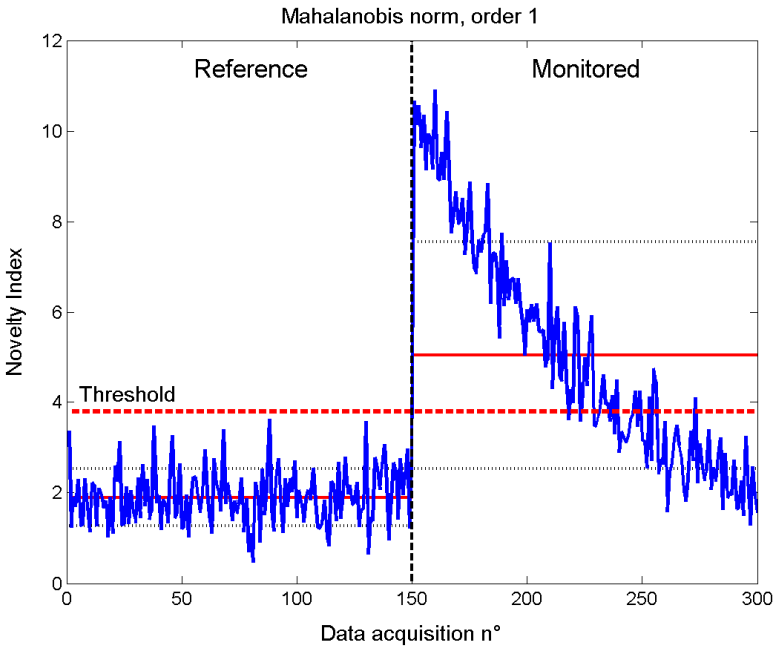


Figure 3.15. (a) Damage detection using two sets of data at different temperatures: reference from 0 to 45 °C and damaged from -45 to 0 °C. (b) Damage detection using full range (from -45 to 45 °C) data of reference and damaged system.

(a)



(b)

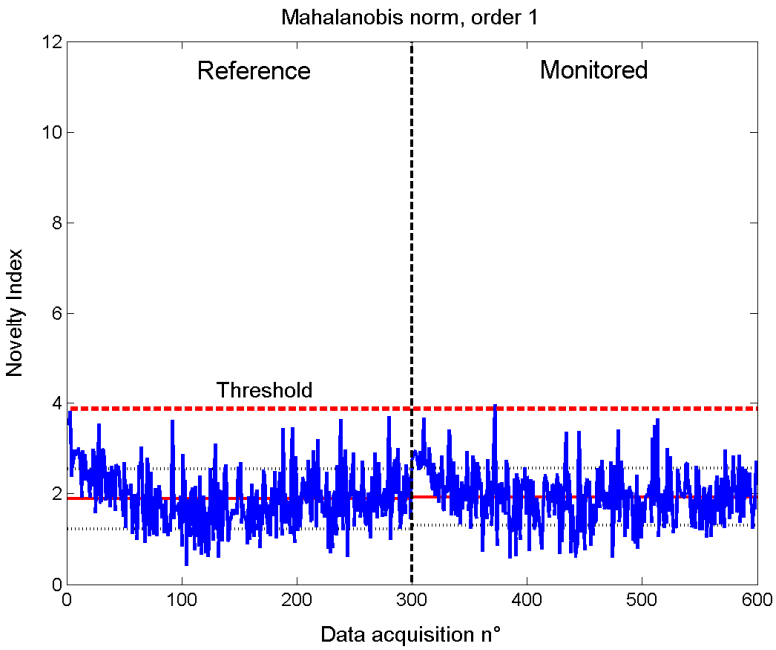


Figure 3.16. (a) False-positive verification using two sets of data at different temperatures: reference from 0 to 45 °C and monitored from -45 to 0 °C. (b) False-positive verification using full range (from -45 to 45 °C) data of reference and monitored system.

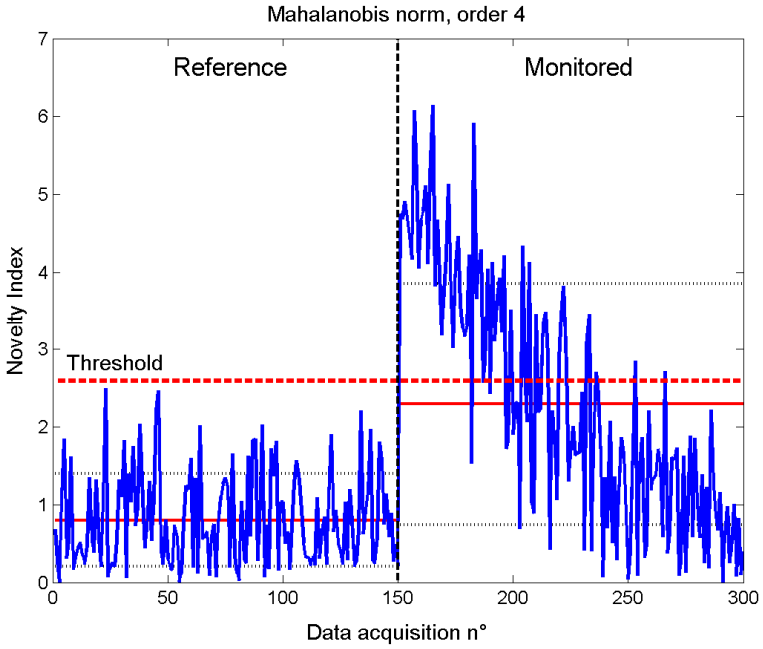


Figure 3.17. $m = 4$. False-positive verification using two sets of data at different temperatures: reference from 0 to 45 °C and monitored from -45 to 0 °C.

Two strategies can be exploited in order to overcome or alleviate this drawback. The first consists in considering the complete set of 300 data acquisitions, as shown in Fig. 3.16b. Nonlinear effects are still observable but are now merged into a larger data set of quasi-linear data: only 1 data acquisition out of 300 is identified as damaged (but it is very close to the threshold anyway) and the ratio $\overline{NI}_c / \overline{NI}_r$ is very close to 1. With the second strategy, the situation is the same as in Fig. 3.16a: the two temperature intervals are kept, but four principal components are used instead of one. As shown in Fig. 3.17, this is a more conservative choice leading to a reduction of false-positive manifestations, even if the drawback has not completely been overcome as with the application of the first strategy.

In conclusion: since the exact number of environmental factors may be unknown in practice, it seems appropriate and safe to try the method with several increasing model orders [53]; since linearity cannot generally be assumed, it is safer to apply the method by considering a full range data set, including all values of environmental conditions that can occur in practice.

A procedure for extending the PCA-based damage detection method to nonlinear cases has been studied in [90]. The method involves a two-step procedure, namely

a clustering of the data space into several regions and then the application of PCA in each local region. The application of local PCA allows performing a piecewise linearization of the nonlinear problem.

Chapter 4

Experimental application: the bearing test rig

In this chapter the experimental application of PCA in bearing damage detection is introduced. The test rig, set up in the Laboratory of the Department of Mechanical and Aerospace Engineering of Politecnico di Torino, is at first described in detail. Then, the instrumentation is presented and a complete description of the (huge) amount of performed tests is given.
Next Chapter 5 will deal with the PCA analysis of the acquired data, by showing all the obtained results.

4.1. Description

The bearing test rig has been conceived to carry out an exhaustive experimental campaign on bearings with different damage levels in controlled laboratory conditions. Since the test rig involves only three bearings and a rotating shaft, the objective is to minimize spurious signals due to the presence of many mechanical elements and in particular meshing gears or other vibrating elements.
The test rig (a detail is shown in Fig. 4.1, in a SolidWorks rendering) has been designed in collaboration with Avio S.p.A. and it has been conceived for simulating the bearings operative conditions as similar to those observed on real

gearboxes. For this reason, the selected elements are derived from those mounted on one of the real Avio gearboxes.

Referring to Fig. 4.1, bearings 1 and 2 have an internal diameter $\Phi = 25$ mm. The rotating shaft has been properly modified, by substituting the central transmission gear with the housing of bearing 3, to which a radial load is applied. Bearing 3 has an internal diameter $\Phi = 40$ mm and a different number of rolling elements with respect to bearings 1 and 2. Moreover, it is overdimensioned and it can be stressed with a load that is about twice the load allowed for bearings 1 and 2: the reason is to avoid early wear or damage, since this bearing is not considered for the tests.

The oil pump and the lubrication system have been lent by Avio for these specific tests. The whole test rig is shown in Fig. 4.2, in which an electric spindle and a load cell can be seen.

The electric spindle is provided with its fixing support, its power supply and its connections (in white, in the background of Fig. 4.2) with the cooling circuit. The spindle has been properly selected in order to reach a rotating speed of 30000 rpm (or 500 revolutions per second, “Hz”), as requested by the project proposal. For the purposes of the tests, the rotating speeds can range from 6000 to 30000 rpm (100 to 500 Hz).

The load cell measures the radial force acting on bearing 3. The springs (in green) for imposing the load on the bearing can be seen in the foreground of Fig. 4.2. The calibration characteristic of the load cell is shown in Fig. 4.3.

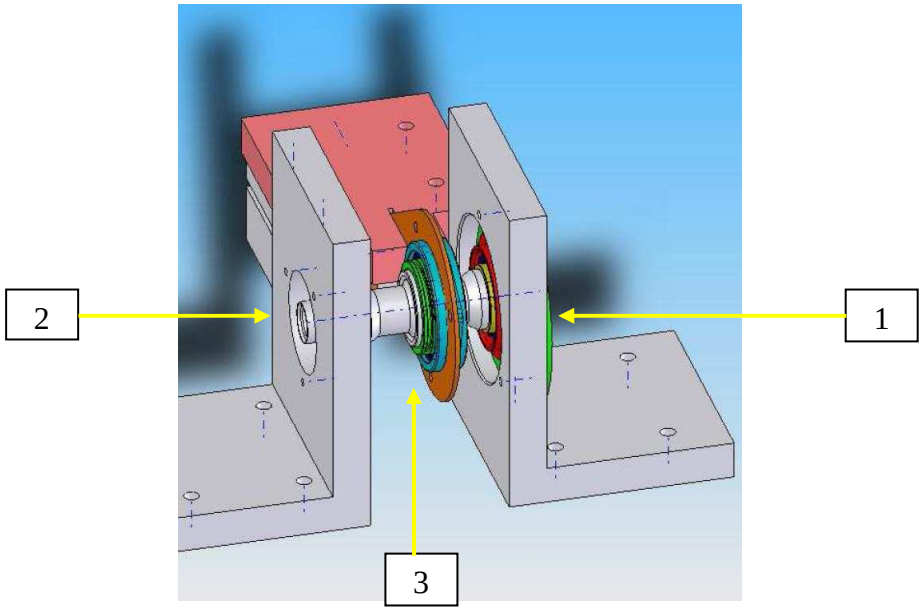


Figure 4.1. A view of the test rig, in a SolidWorks rendering.

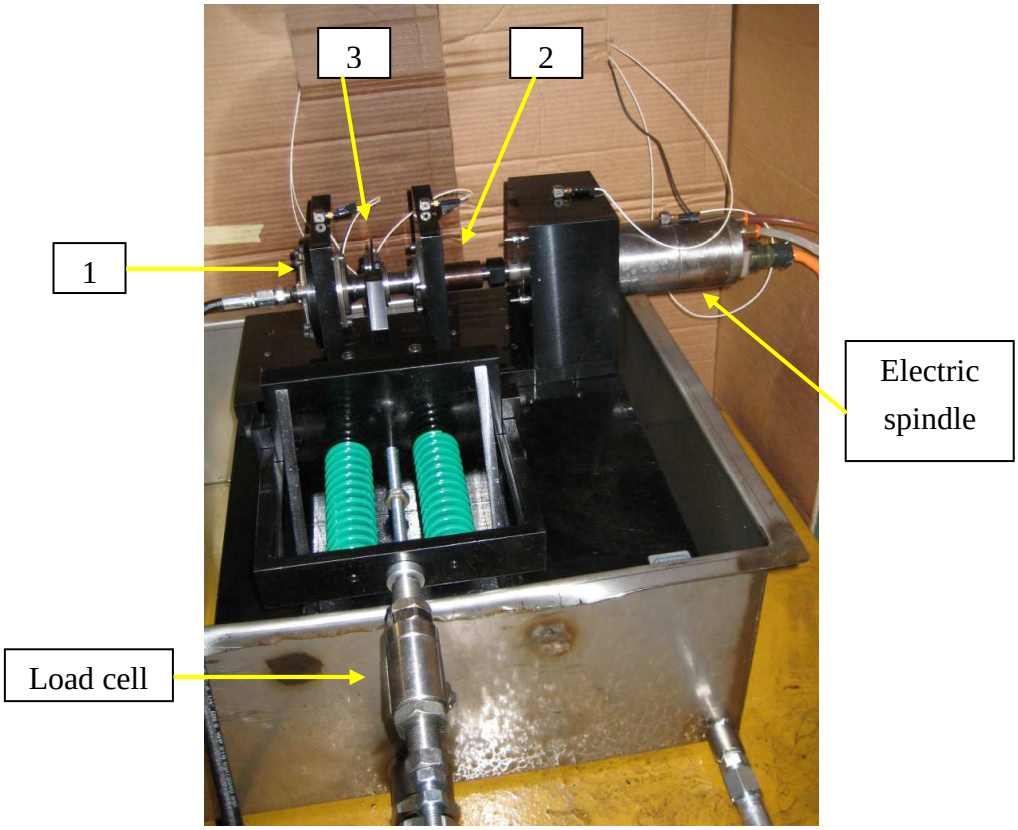


Figure 4.2. The complete test rig.

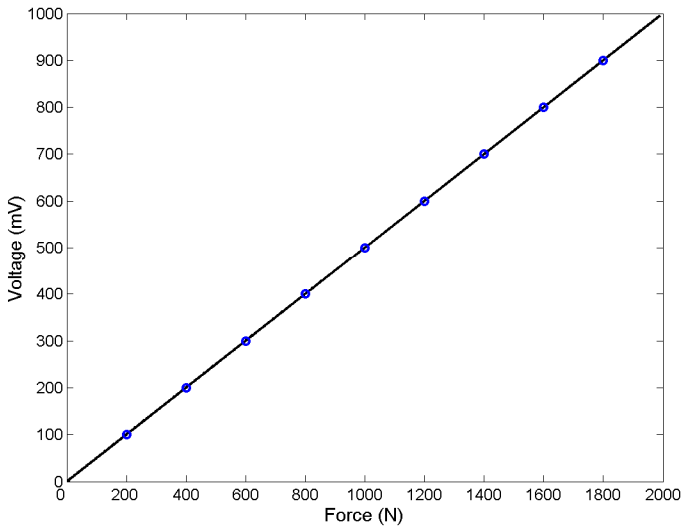


Figure 4.3. Characteristic of the load cell**.**

4.2. Instrumentation

4.2.1. Sensors

The test rig has been equipped with some triaxial accelerometers (3 or 4, depending on the setup as described in the following), strictly fixed to the structure by grub screws. The characteristics of the accelerometers are given in Table 4.1, while their axes orientation is shown in Fig. 4.4: the x, y and z axis correspond, respectively, to the axial, radial and tangential direction. The sensors are connected to a Oros OR38 data acquisition system.
In order to record some significant temperature values, two thermocouples are placed in the oil basin and in proximity of the external ring of bearing 3, respectively.
Two similar setups have been designed for the tests.

- **Setup #1** (Fig. 4.5), with 4 accelerometers: on bearing 1 support (channels 1-2-3), on bearing 2 support (channels 4-5-6), on bearing 3 support (channels 7-8-9) and on the motor support (channels 10-11-12).
- **Setup #2** (Fig. 4.6), with 3 accelerometers: on bearing 3 support (channels 1-2-3), on bearing 2 support (channels 4-5-6) and on bearing 1 support (channels 7-8-9). Moreover, in this setup an electrical resistor can be used to heat the oil properly and control (up to a certain level) its temperature.

4.2.2. Damaged bearings

Some different damaged bearings were available, supplied by SKF by keeping the specifications requested by Avio. A comprehensive list of damages type and extent is given in Table 4.2. These bearings can be mounted in housing number 1 (Figs. 4.1 and 4.2).

Table 4.1. Characteristics of the accelerometers**.**

Producer	Model	Full scale	Nominal sensitivity	Resonance
Kistler	8763A500, triaxial	500 g	10 mV / g	~55 kHz

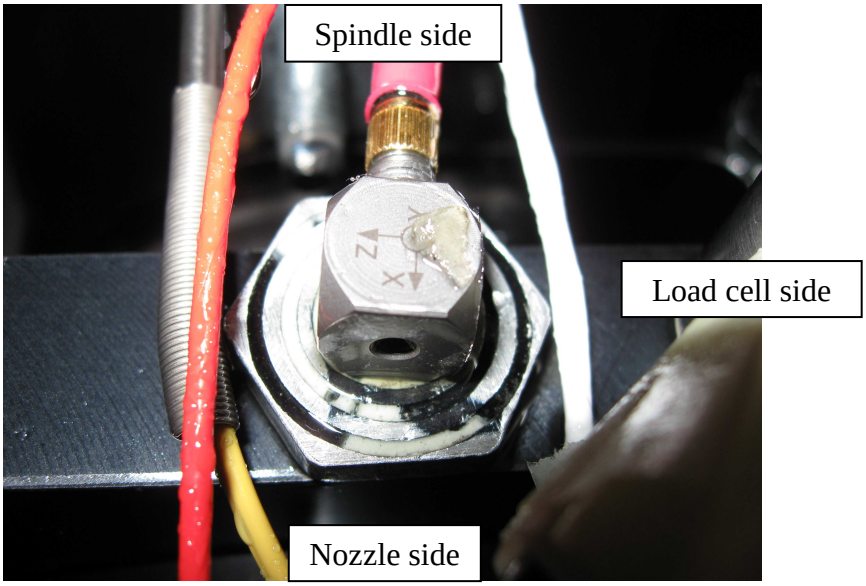


Figure 4.4. Axes orientation.

Table 4.2. List of damages.

Denomination	Type	Extent (in microns)
0A	No damage (“Healthy”)	
1A	Inner ring indentation	450
2A	Inner ring indentation	250
3A	Inner ring indentation	150
4A	Rolling element indentation	450
5A	Rolling element indentation	250
6A	Rolling element indentation	150

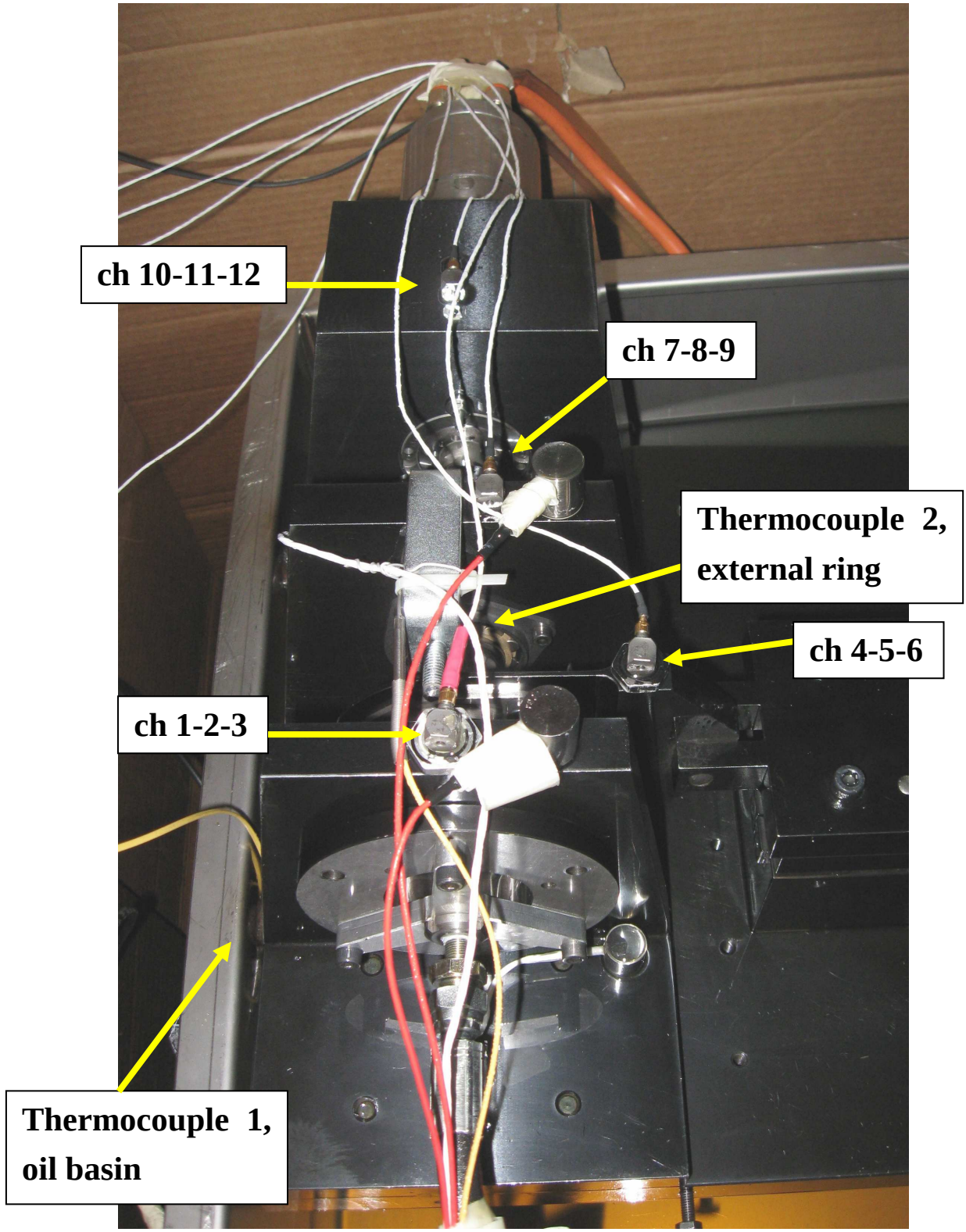


Figure 4.5. Setup #1.

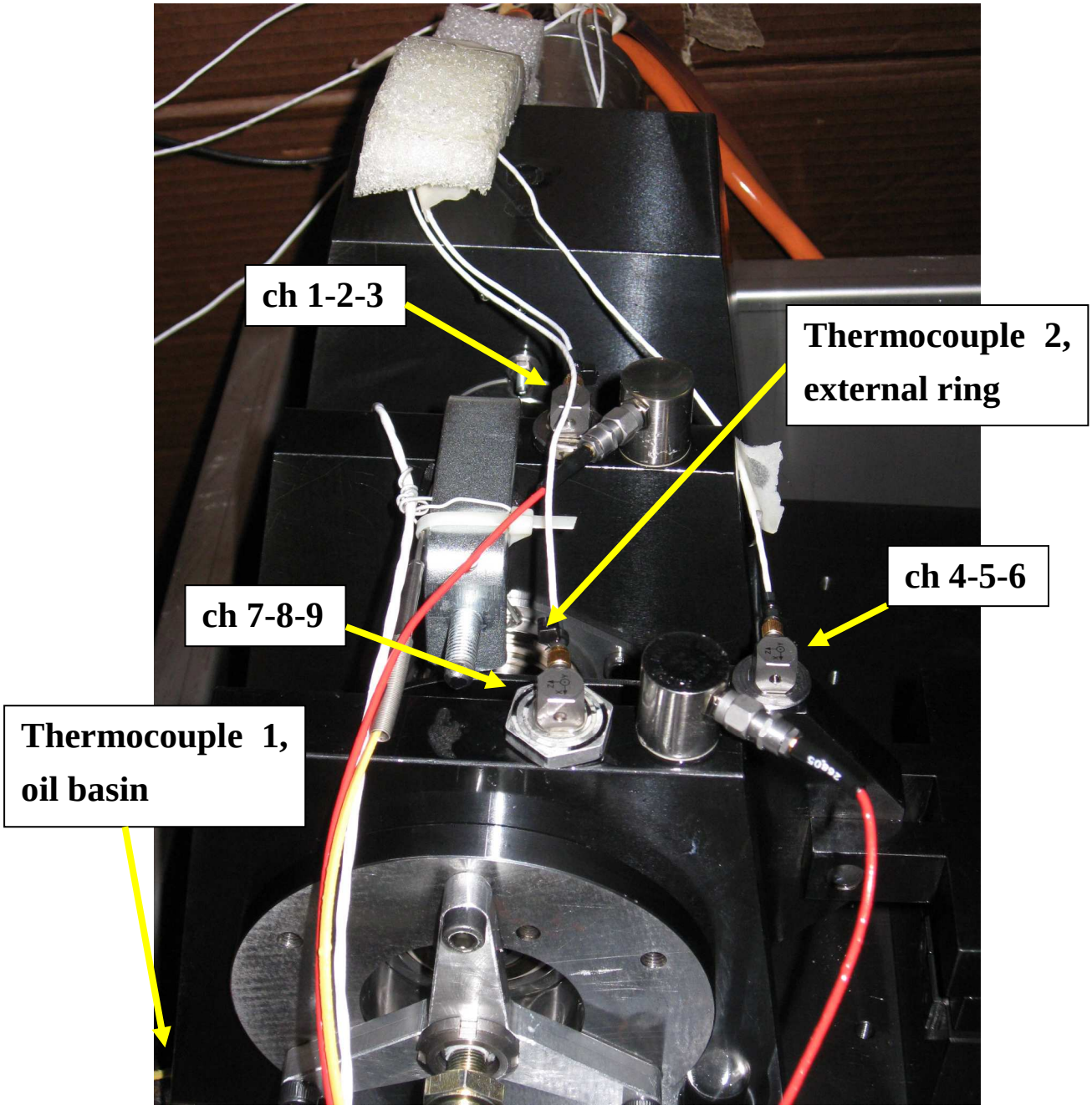


Figure 4.6. Setup #2.

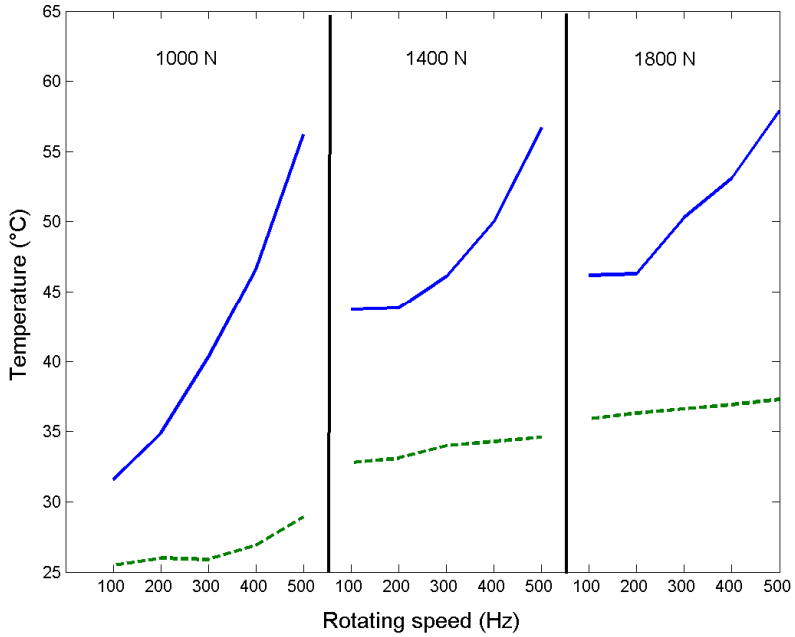
4.3. Data collection

The test campaign covered a period of almost two years, from September 2009 to July 2011. During the tests, both setups have been adopted to acquire a very large data set through some different types of measurements. The complete data set is useful for giving an overview on how the PCA method can be applied on such a complicated real-life system. The PCA analysis will be presented in Chapter 5, by following the conceptual baseline presented in Chapter 3 for a simple structural numerical application.
In the following, a description of the five different types of test is given. Observe that all time series have been recorded at a sampling frequency of 102400 Hz, for 8 seconds.

Test #1

Measurements have been acquired on Setup #1.
The seven types of bearing listed in Table 4.2 have been mounted in sequence on the test rig and, for each of them, the tests have been performed by crossing five values of rotating speed (100, 200, 300, 400 and 500 Hz) with three values of loading (1000, 1400 and 1800 N). A total number of 15 acquisitions per bearing type was finally obtained.
For each type of mounted bearing, the tests have been carried out in close succession (about 30 minutes were needed for the 15 acquisitions), so that the temperatures (oil and bearing) were ***not*** expected to reach stabilisation. This is shown in Figs. 4.7a and 4.7b for the 2A and 6A bearings, respectively: for a fixed value of loading, the five rotating speeds were tested in sequence, from 100 to 500 HZ; then, the motor was switched off for changing the applied load and switched on again, for next tests. Observe that the bearing temperature decreases when the motor is switched off, while the oil temperature is still slightly increasing since the oil circulation system is not switched off.

(a)



(b)

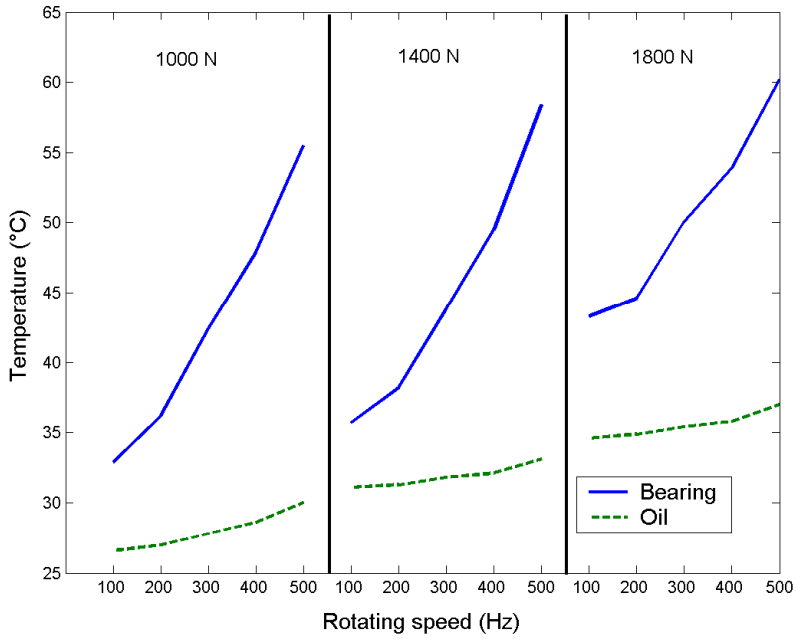


Figure 4.7. Test #1. Bearing and oil temperatures, for each configuration of loading and rotating speed: (a) bearing 2A; (b) bearing 6A. The black line indicates when the motor is switched off, for changing the applied load.

Test #2

Measurements have been acquired on Setup #1, just after the end of Test #1 described above.
Only the bearing 4A (Table 4.2) has been considered for this endurance test. The initial rolling element indentation of bearing 4A is shown in Fig. 4.8. During the test, the operating conditions were constant, with a rotating speed of 300 Hz and a loading of 1800 N.
A total number of 268 measurements has been acquired, in slightly different ways, for a total amount of almost 200 operating hours:

- Measures from 1 to 101: the first measure of the day was recorded 30 minutes after switching the rig on. A measurement every 30 minutes was then acquired. In this way, the first measurements of the day were taken ***without*** reaching a complete rig heating (Fig. 4.9a).
After about 50 operating hours, the bearing has been unmounted and its rolling element indentation has been inspected: no significant differencies have been seen with respect to what observed in Fig. 4.8.

- Measures from 102 to 183: the first measure of the day was recorded 90 minutes after switching the rig on. A measurement every 30 minutes was then acquired. In this way, an almost complete rig heating was reached (Fig. 4.9b).
After about 50 operating hours, the bearing has been unmounted and its rolling element indentation has been inspected again: no significant differencies have been seen with respect to what observed in Fig. 4.8.

- Measures from 184 to 268: the first measure of the day was recorded 90 minutes after switching the rig on (reaching an almost complete rig healing). A measurement every 60 minutes was then acquired. This part lasted for about 95 operating hours. Again, no significant damage increment was seen from a final inspection of the rolling element.

(a)

(b)

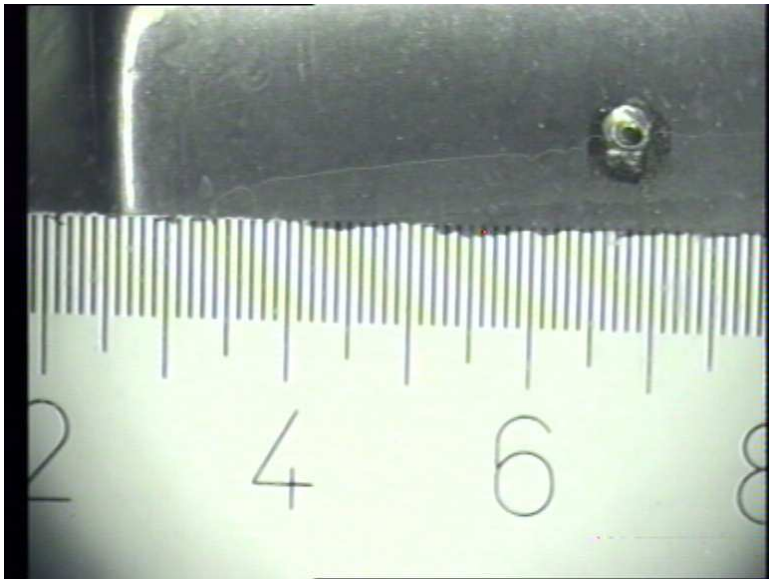
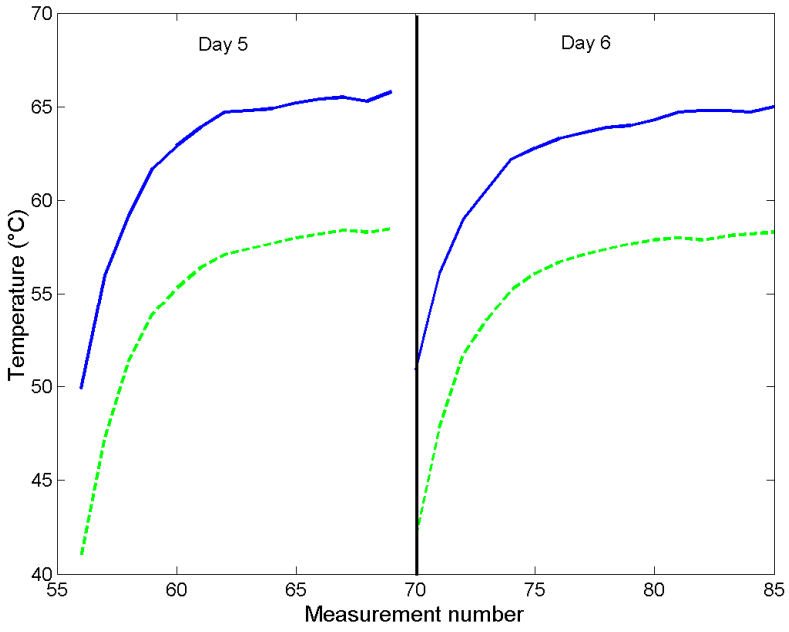


Figure 4.8. (a) The initial rolling element indentation of bearing 4A. (b) Magnification: the scale is represented in tenths of millimeters (the values 2, 4, 6 and 8 are in millimeters).

(a)



(b)

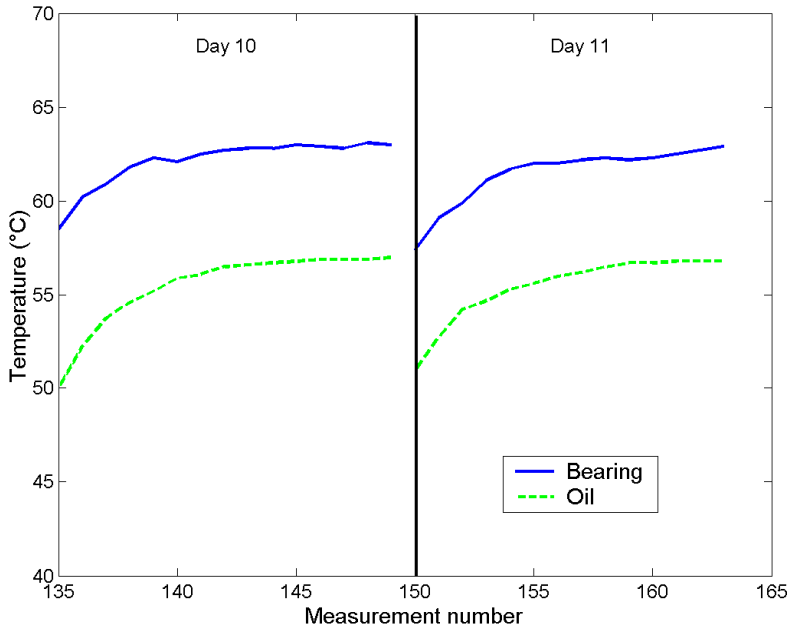
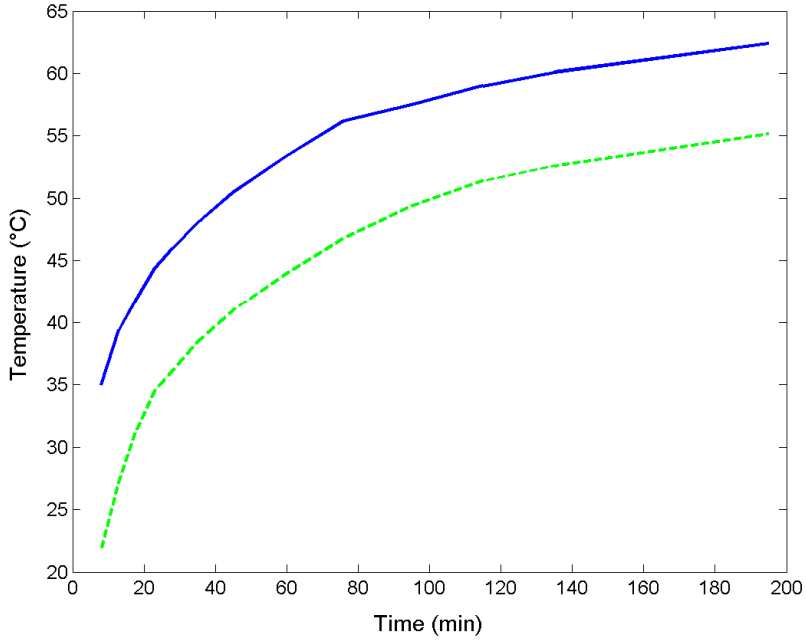


Figure 4.9. Test #2. Bearing and oil temperatures: (a) the first measurements of the day were taken without reaching a complete rig healing; (b) an almost complete rig healing was reached.

(a)



(b)

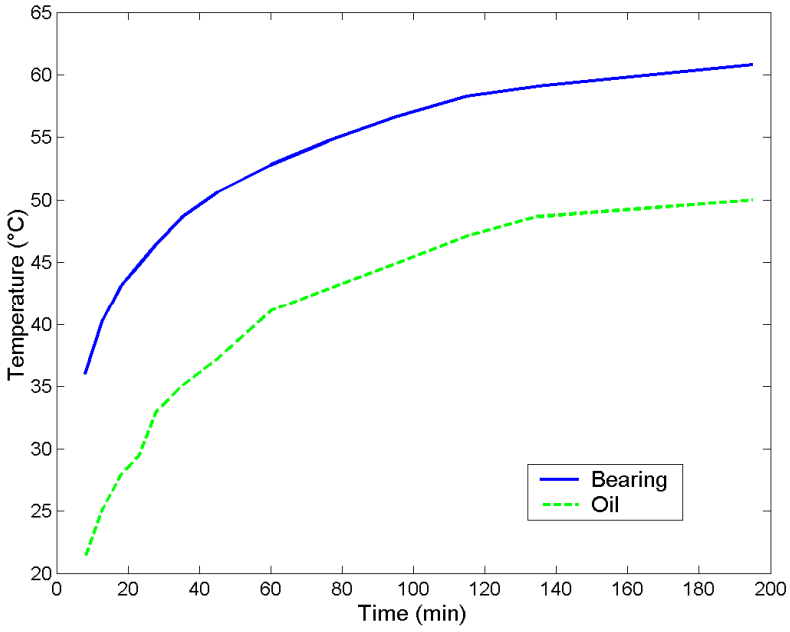


Figure 4.10. Test #3. Bearing and oil temperatures: (a) first succession, 2009 October 19th; (b) second succession, 2009 November 27th.

Test #3

Measurements have been acquired on Setup #1.
The bearing 4A (Table 4.2) has been considered, with the same operational conditions (a rotating speed of 300 Hz and a loading of 1800 N) of Test #2. In this case, the measures have been recorded in close succession in order to follow the bearing temperature increasing, from switching the rig on at about 20 °C.
Two successions of 13 measurements were separately acquired (on the 19th of October and the 27th of November, 2009). For each succession, the measured temperatures are shown in Figs. 4.10a and 4.10b, respectively, against the time elapsed from switching the rig on. They are very similar, but in the second sequence the oil temperature is about 5 °C lower than in the first sequence. There is no particular explanation about that, except for the change of date causing a cooler outside temperature that may induce a longer oil heating time.

Test #4

Measurements have been acquired on Setup #2.
The undamaged bearing 0A has been considered, with a constant loading of 1800 N. Three values of rotating speed (200, 300 and 400 Hz) are tested. For each of them, by using the electrical heater, the measurements have been recorded in four steady-state values of the oil temperature (45, 60, 75 and 85 °C, within a 7% margin). An average number of 20 measures was acquired for each combined condition and the test was repeated twice, for a total number of about 480 measurements. The temperatures for the case with a rotating speed of 200 Hz are shown in Fig. 4.11: the use of the heater allows the system to reach high temperatures. Moreover, they are very similar (as expected) when comparing Day 1 and Day 2.

Test #5

Measurements have been acquired on Setup #2.
The undamaged bearing 0A has been considered at first. The tests have been performed by crossing three values of rotating speed (100, 175 and 200 Hz) with three values of loading (1400, 1600 and 1800 N). For each configuration, 40 measures per day were acquired and a single test was repeated over two days, for a total number of 720 acquisitions.

The electrical heater has been exploited in order to reduce the time needed by the system to reach temperature stabilisation. Observe that different stable temperatures correspond to different configurations, due to the changes in operating conditions (rotating speed in particular). In other words, there is not a global stable temperature for the system, but each configuration reaches an own stable value. This is shown, for example, in Figs. 4.12a and 4.12b. Different stable values of temperature can be seen for different rotating speeds and a fixed loading (Fig. 4.12a). On the contrary, when the rotating speed is fixed (Fig. 4.12b) no significant differencies can be observed for the selected different loadings.

However, the differences are slight (within 10%) and, in addition, no other way of fixing a universal stable value for temperatures could be practically adopted. So these cases will be considered "at a fixed (constant) temperature", when applying PCA in Chapter 5.

The same tests have been repeated by considering the damaged bearing 4A (Table 4.2). In this case each configuration was tested only once, so that a total number of 360 measurements was acquired.

The same considerations about temperatures hold (Figs. 4.13a and 4.13b). Moreover, no particular differencies can be seen by comparing the temperatures of bearing 0A and 4A: this will be useful for damage detection in Chapter 5.

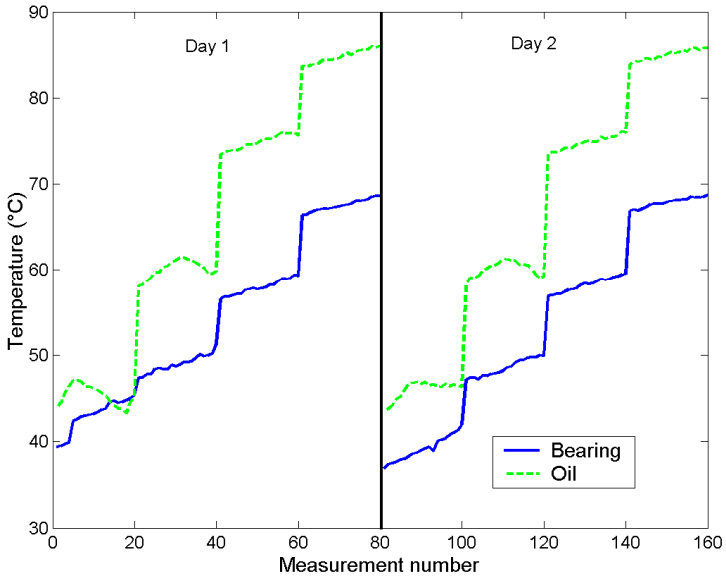
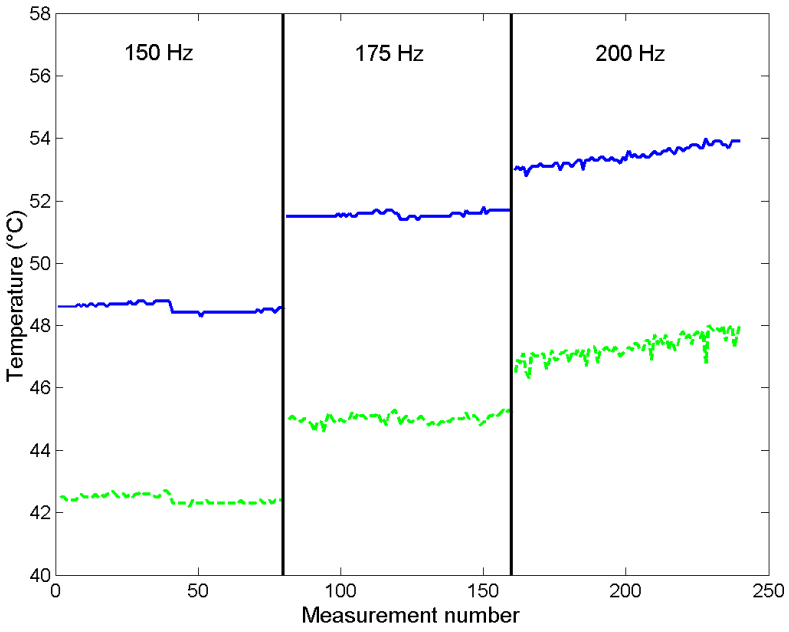


Figure 4.11. Test #4. Bearing and oil temperatures, for the case with a rotating speed of 200 Hz. The four steady-state values of the oil temperature can be recognised, for each day.

(a)



(b)

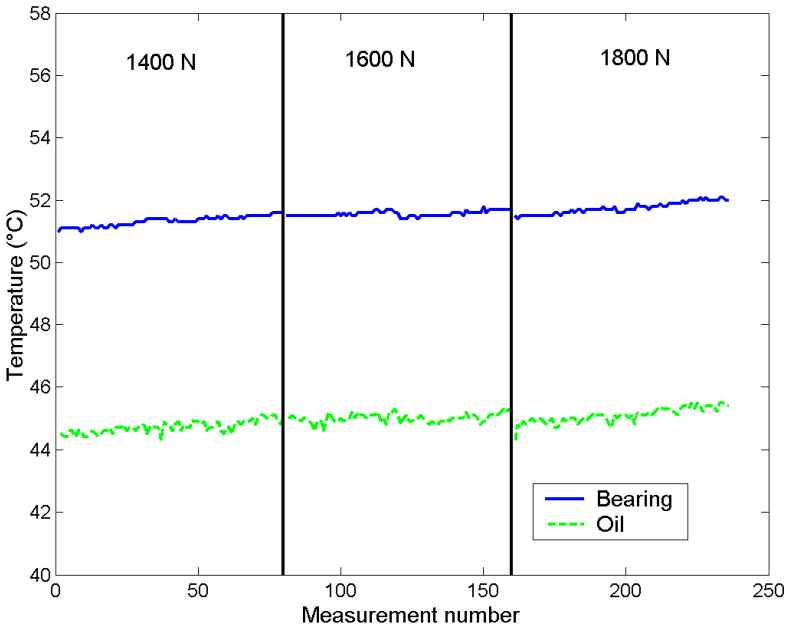
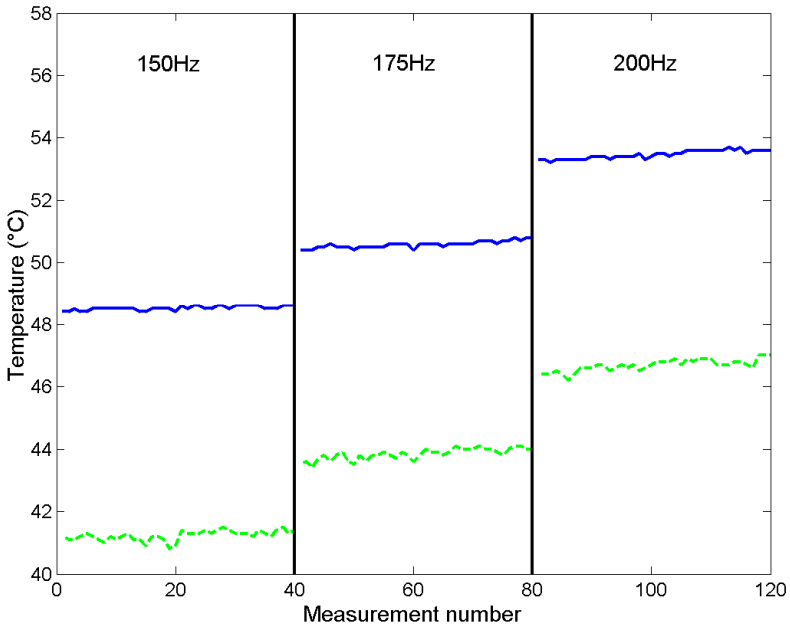


Figure 4.12. Test #5, bearing 0A. Bearing and oil temperatures: (a) fixed loading of 1600 N; (b) fixed rotating speed of 175 Hz.

(a)



(b)

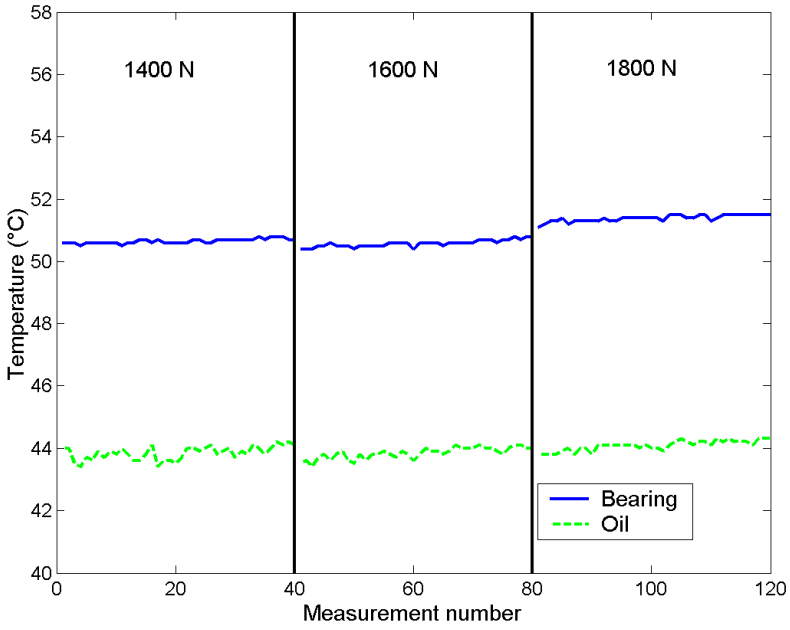


Figure 4.13. Test #5, bearing 4A. Bearing and oil temperatures: (a) fixed loading of 1600 N; (b) fixed rotating speed of 175 Hz.

Chapter 5

Experimental application: PCA-based bearing diagnostics

In this chapter the experimental application of PCA in bearing damage detection is shown through the obtained results (see also [94]). The test rig, set up in the Laboratory of the Department of Mechanical and Aerospace Engineering of Politecnico di Torino, is described in previous Chapter 4.
At first, the motivation for exploiting PCA is illustrated: the features are shown to be dependent on the operational and environmental conditions. Then, the obtained results are described and three objectives of diagnostics are investigated: damage detection (and false-positive verification), damage extent and localisation.

5.1. Motivation: operational and environmental conditions

Starting from the methodology introduced in Section 2.3.1, the n-dimensional set of signal features defining the $n \times N$ matrix Y in (2.9) has to be defined (N is the number of measurements).
The selected parameter, for each of the measurement channels as defined in Section 4.2.1, is the root mean square (RMS):

$$RMS_i = \sqrt{\frac{1}{s}\sum_{k=1}^{s} x_{i,k}^2}\,, \qquad \text{for } i = 1,\ldots,n \tag{5.1}$$

where s is the number of recorded samples for each measurement and $x_{i,k}$ is the k-th sample of the i-th channel.

This statistical parameter has been found to be sensitive to damage and easy to compute. Moreover, the operational and environmental conditions influence the RMS in such a way that it satisfies the linearity (or quasi-linearity) assumption for applying the PCA method, as demonstrated afterwards. Any other damage-sensitive parameter can be used, in both the time or frequency domains [13], provided that the (quasi-) linearity condition is guaranteed.

The number of features n can be 9 or 12, depending on which Setup (see Section 4.2.1) is considered. However, channels 4, 5 and 6 are not so good because they correspond to the sensor placed in proximity of bearing 2, on which the load is applied. This produces high levels of noise which may mask the damage-sensitivity capability of these channels, which are not considered in some cases.

In the following, the dependence of the RMS from each of the measurable operational and environmental conditions of the test rig is investigated in detail, together with the quasi-linearity demonstration that allows the PCA to be applied.

5.1.1. Rotating speed

In order to investigate if both rotating speed variations and damage are responsible for changes of the RMS, Test #5 is considered in Fig. 5.1, with a fixed loading value of 1800 N. In particular channel 8 is shown, but similar results can be obtained for all channels. All measured values are depicted with circles; the solid and dashed lines represent linear regressions that can be used to observe a quasi-linear dependence between RMS and rotating speed. This is a sufficient condition to assert the quasi-linearity among all the features, as demonstrated in the next Fig. 5.2a, but it is not necessary. In fact, the quasi-linearity must be checked in a feature versus feature diagram, while a feature versus condition relationship can also be nonlinear, as reported in Section 5.1.3.

In absence of a precise correlation between the variations of the rotating speed and of the RMS, since the former are supposed to be unknown, a simple comparison between the features identified in different operational conditions does not lead to a clear diagnostics of possible damages. This can be seen in Fig. 5.1 by noticing,

for example, that the same RMS values can be associated to two conditions: the reference (healthy) system at higher rotating speeds or the damaged system at lower rotating speeds.

Fig. 5.2a (Test #5) shows the evolution of RMS_i (for $i=3$ and $i=9$) as a function of RMS_2, for the healthy bearing 0A (see Table 4.2). The RMS values are not scaled, for better visualisation. Fig. 5.2a confirms the absence of a clear nonlinear relationship among features, so in this case the assumption of linearity can be made in order to apply PCA as in Section 2.3.

However, a nonlinear behaviour is observable for some channels. For example, a slight inaccuracy in considering a linear relationship among RMS values can be illustrated in a feature versus feature diagram: Fig. 5.2b shows the evolution of RMS_8 as a function of RMS_2, for the healthy bearing 0A. A quadratic behaviour can be observed: the application of linear PCA may lead to uncorrect results, as demonstrated in Section 3.4. In particular, in this case false alarms may be issued, as shown later on in Section 5.2.2.

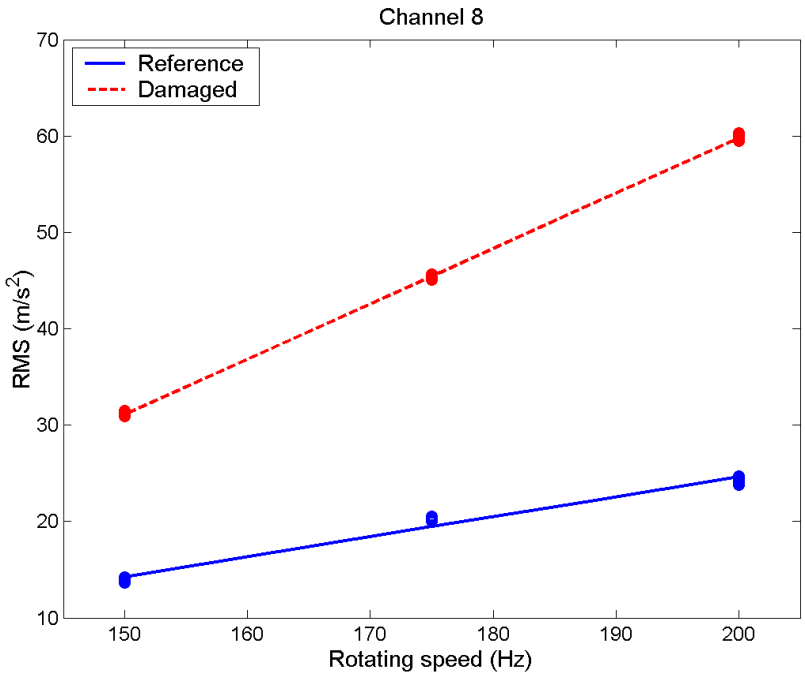
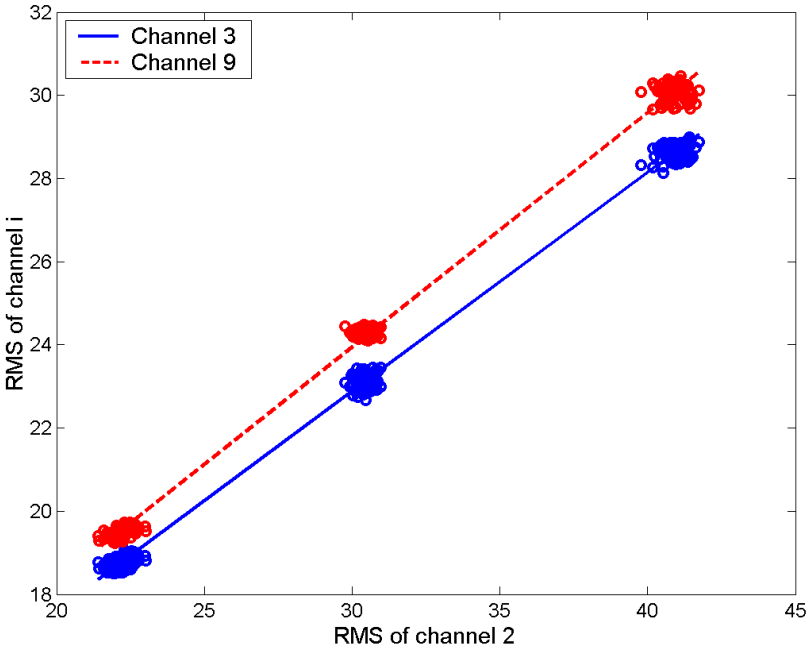


Figure 5.1. Test #5, fixed loading. RMS of channel 8 varying with rotating speed, for the reference and the damaged case. All measured values are depicted with circles; the solid and dashed lines represent linear regressions.

(a)



(b)

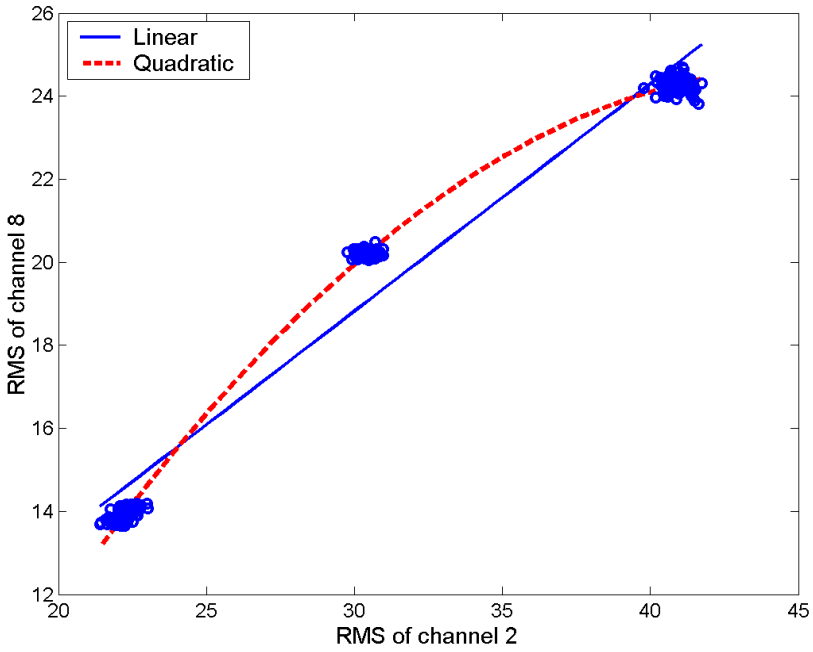


Figure 5.2. Test #5, fixed loading, healthy bearing 0A. (a) Diagram showing the evolution of RMS_i (for $i = 3$ and $i = 9$) as a function of RMS_2. (b) Evolution of RMS_8 as a function of RMS_2, with a linear and a quadratic fitting.

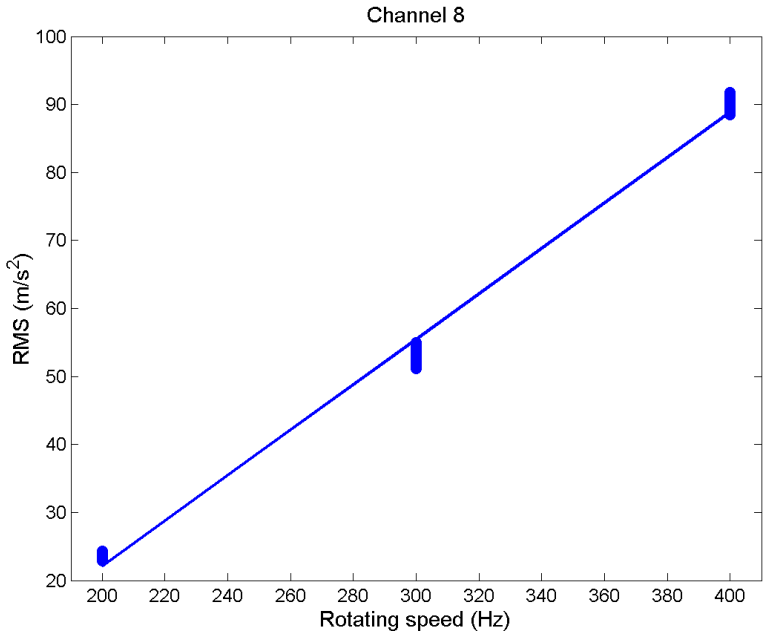


Figure 5.3. Test #4, fixed oil temperature. RMS of channel 8 varying with rotating speed, for the healthy bearing 0A. All measured values are depicted with circles; the solid line represents a linear regression.

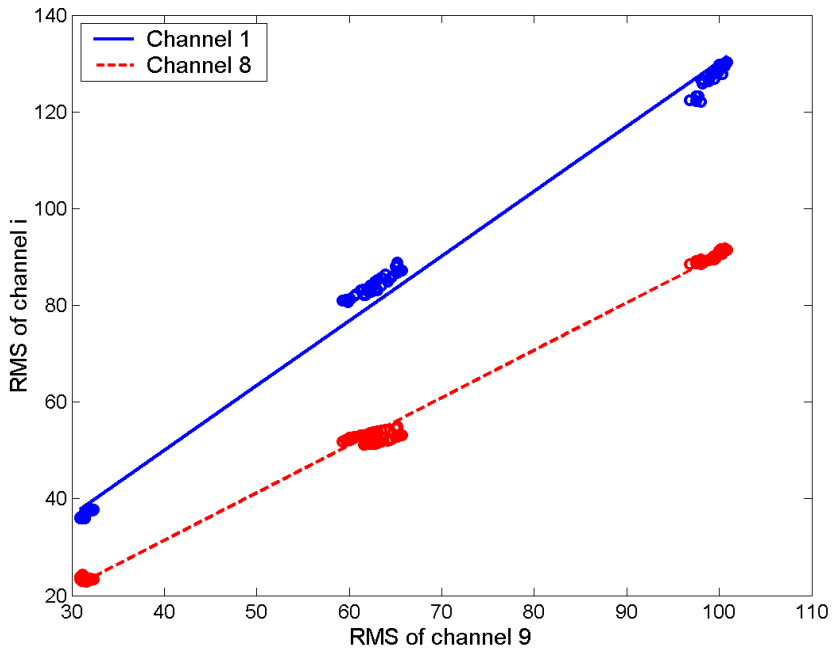


Figure 5.4. Test #4, fixed oil temperature. Diagram showing the evolution of RMS_i (for $i = 1$ and $i = 8$) as a function of RMS_9, for the healthy bearing 0A.

The same remarks can be obtained by considering another test, as shown in Figs. 5.3 and 5.4 for Test #4, with a fixed oil temperature of 45 °C. Since the damaged bearing was not mounted during the Test#4, a comparison is not possible. However, for this test higher rotating speeds were considered, to extend the validity of the linear assumption. In Fig. 5.3 channel 8 is shown, while Fig. 5.4 (Test #5) illustrates the evolution of RMS_i (for $i=1$ and $i=8$) as a function of RMS_9.

In conclusion, in case of a dependence from the rotating speed the assumption of linearity can be made in order to apply PCA, but this assumption may be too strong when trying to apply PCA to different subsets of data, i.e. in a limited range of operational variations (as explained in Section 2.3.2).

5.1.2. Applied load

The same analysis as in Section 5.1.1 can be carried out by investigating the RMS changes due to the applied load. Test #5 is considered in Fig. 5.5, with a fixed rotating speed of 200 Hz. In particular channel 8 is shown, but similar results can be obtained for all channels. All measured values are depicted with circles; the solid and dashed lines represent linear regressions that can be used to observe that the RMS can be considered as constant over the tested values of load. This means that the parameter is not influenced by loading, at least for the values used during the tests. In this case, with fixed temperature and rotating speed, PCA is not needed. However, its application is not expected to alter the results.

Fig. 5.5 also demonstrates that the RMS parameter, without any influence by the operational and environmental conditions, is sensitive to damage: the RMS of the damaged bearing are well separated from those of the healthy bearing.

Fig. 5.6 (Test #5) shows the RMS_i (for $i=3$ and $i=7$) as a function of RMS_2, for the healthy bearing 0A (see Table 4.2). The RMS values are not scaled, for better visualisation. Fig. 5.6 confirms the absence of any relationship among the features: they are placed around a fixed point in the space, without showing a linear behaviour as in Figs. 5.2a and 5.4.

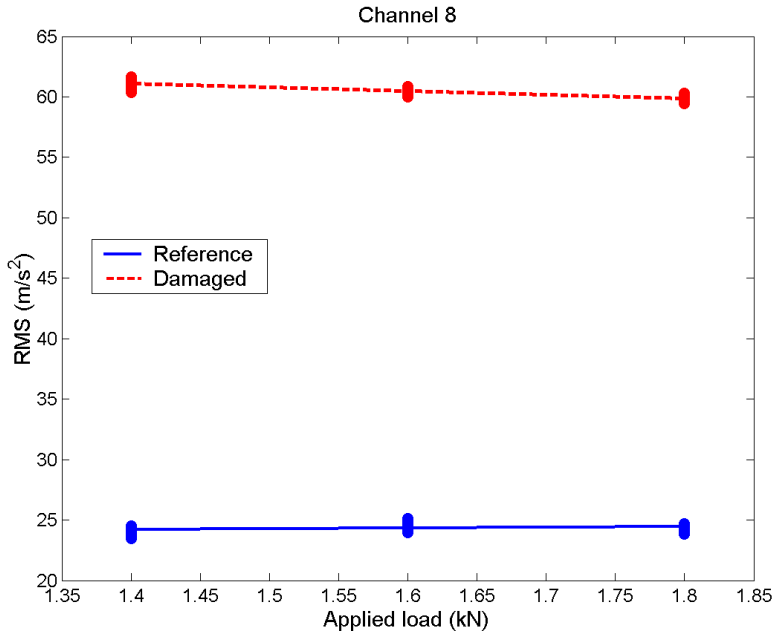


Figure 5.5. Test #5, fixed rotating speed. RMS of channel 8 in function of the applied load, for the reference and the damaged case. All measured values are depicted with circles; the solid and dashed lines represent linear regressions.

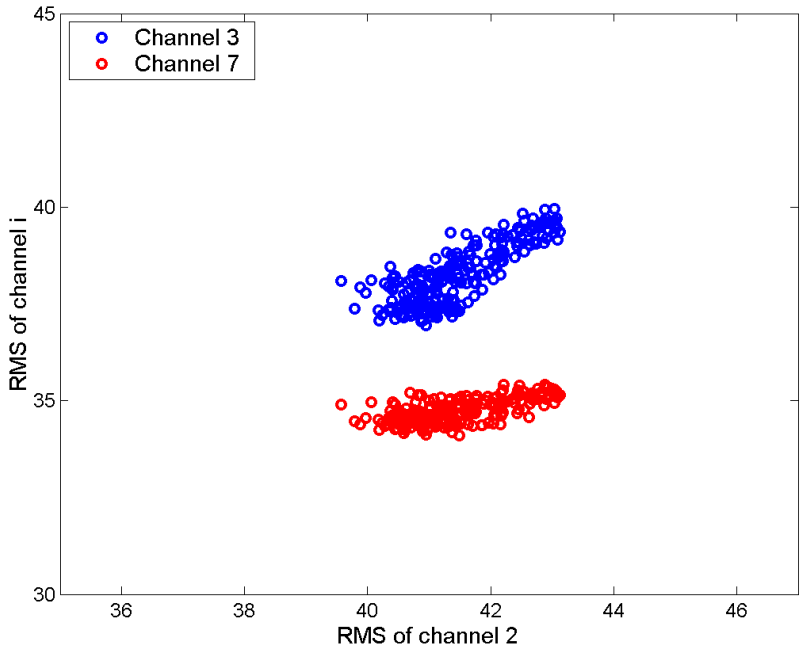


Figure 5.6. Test #5, fixed rotating speed. Diagram showing the RMS_i (for $i = 3$ and $i = 7$) as a function of RMS_2, for the healthy bearing 0A.

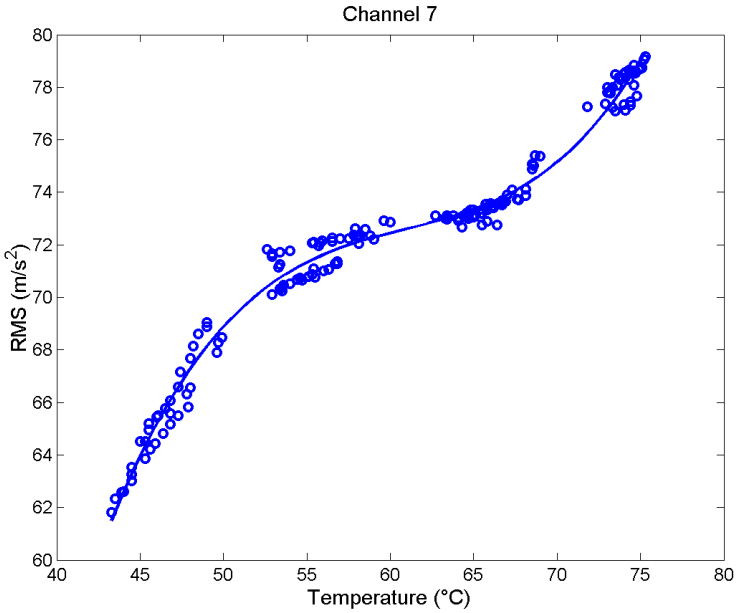


Figure 5.7. Test #4, fixed rotating speed. RMS of channel 7 varying with temperature, for the healthy bearing 0A. All measured values are depicted with circles; the solid line represents a polynomial fitting.

5.1.3. Temperature

The last condition to be investigated involves the RMS changes due to temperature. Since two values of temperature (oil and bearing 3, see Section 4.2.1) are available, but these values have very similar behaviours (see Figs. from 4.9 to 4.13), only one is retained in this Chapter: the bearing 3 temperature, as a reference of the whole system temperature.

Test #4 (Setup #2, bearing 0A) is considered in Fig. 5.7, with a fixed rotating speed of 300 Hz. In particular channel 7 is shown, but similar results can be obtained for all channels. All measured values are depicted with circles; the solid line represents a polynomial fitting that can be used to observe a nonlinear dependence between RMS and temperature. Since the damaged bearing was not mounted during the Test#4, a direct comparison is not possible. However, a similar behaviour can be shown in Fig. 5.8 by considering Test #3 (Setup #1, bearing 4A, first succession of measurements). Channel 1 is shown in this case, as a counterpart (but not exacly located in the same place) of channel 7 of Setup #2 (see Section 4.2.1): almost the same nonlinear dependence can be observed, even

if a quantitative comparison with the healthy case of Fig. 5.7 can not be performed since two different setups are involved.

The results of Figs. 5.7 and 5.8 may drive to the conclusion that PCA can not be applied in the case of temperature dependence, due to the nonlinear characteristic of RMS versus temperature: this is uncorrect. In fact, as stated in Section 5.1.1, the (quasi-)linearity must be checked in a feature versus feature diagram: the quasi-linearity among features is demonstrated in Figs. 5.9a and 5.10.

Fig. 5.9a (Test #4) shows the evolution of RMS_i (for $i = 2$ and $i = 9$) as a function of RMS_7, for the healthy bearing 0A. Fig. 5.10 (Test #3) shows the evolution of RMS_i (for $i = 3$ and $i = 8$) as a function of RMS_1, for the damaged bearing 4A.

However, as seen in Section 5.1.1, a nonlinear behaviour is observable for some channels. For example, an inaccuracy in considering a linear relationship among RMS values can be illustrated in a feature versus feature diagram: Fig. 5.9b shows the evolution of RMS_1 as a function of RMS_7, for the healthy bearing 0A. A nonlinear behaviour can be observed: the application of linear PCA may lead to uncorrect results, as demonstrated in Section 3.4. In particular, in this case false alarms may be issued, as shown later on in Section 5.2.2.

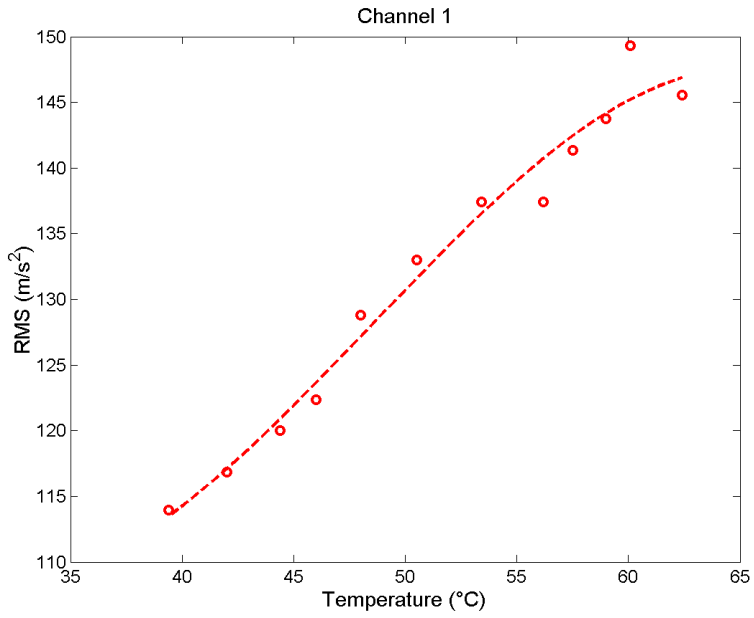
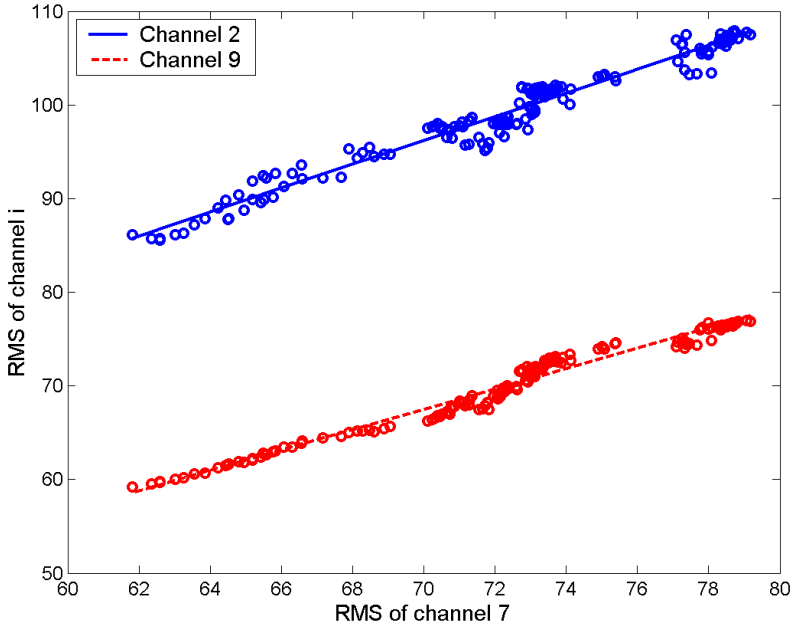


Figure 5.8. Test #3, first succession of measurements. RMS of channel 1 varying with temperature, for the damaged bearing 4A. All measured values are depicted with circles; the dashed line represents a polynomial fitting.

(a)



(b)

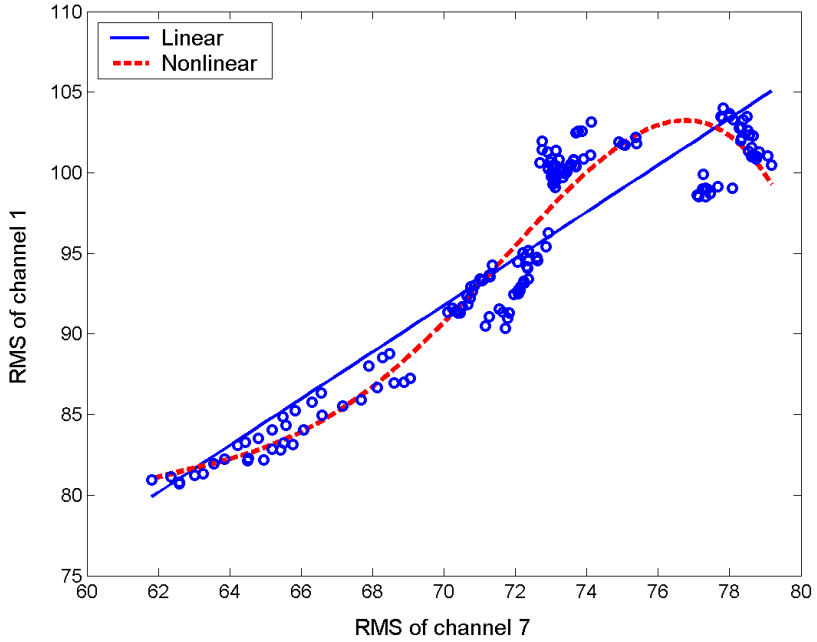


Figure 5.9. Test #4, fixed rotating speed, healthy bearing 0A. (a) Diagram showing the evolution of RMS_i (for $i = 2$ and $i = 9$) as a function of RMS_7. (b) Evolution of RMS_1 as a function of RMS_7, with a linear and a nonlinear fitting.

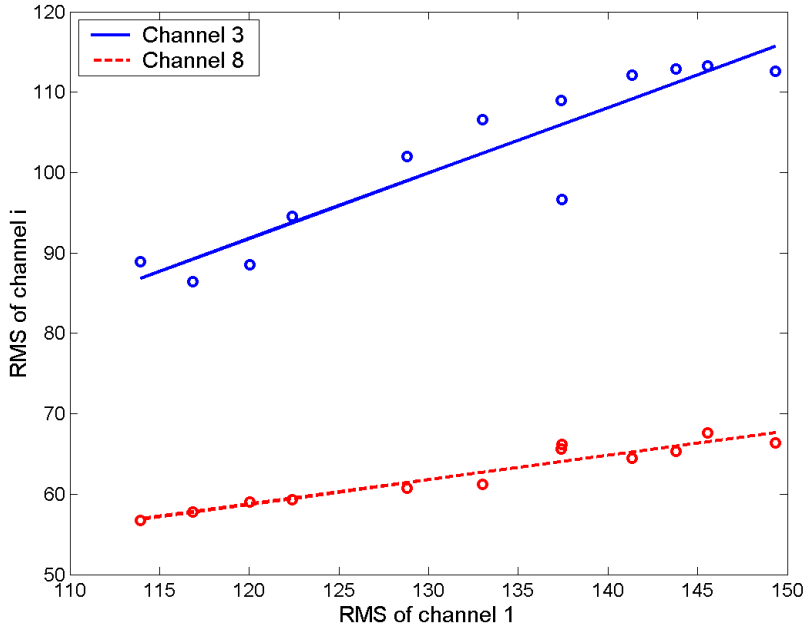


Figure 5.10. Test #3, first succession of measurements. Diagram showing the evolution of RMS_i (for $i = 3$ and $i = 8$) as a function of RMS_1, for the damaged bearing 4A.

In conclusion, also in case of temperature dependence the assumption of linearity can be made in order to apply PCA, but this assumption may be too strong when trying to apply PCA to different subsets of data, i.e. in a limited range of environmental variations (as explained in Section 2.3.2).

5.2. Results

After the detailed description of the motivation for exploiting PCA, this section focuses on the results of the application of the damage detection procedure introduced in Section 2.3. The results are presented in four parts, involving all the tests described in Section 4.3: damage detection; false-positive verification; a technique for damage localisation; an attempt of damage extent evaluation.
For each of the presented cases, a specific number m of principal components is selected. However, it should be observed that no significant difference can be seen in the results by considering a different number of principal components in the analysis.

5.2.1. Damage detection

The method should be able to detect existing damages independently of the operational and environmental conditions at which data are measured. In the following, some of the tests described in Section 4.3 are studied and results are presented.

Test #5

Data collected at the fixed loading of 1800 N are considered and analysed in Fig. 5.11, with $m = 3$ principal components. Fig. 5.11a demonstrates the correct damage detection using full range (150, 175 and 200 Hz) data of reference and damaged system. Moreover, since non-normalised features are used, Fig. 5.11b is helpful to show that the PCA-based detection is robust even when data are measured in a limited range of operational variations. In fact, in Fig. 5.11b the data acquisition on the damaged structure has been realised at lower values (150 and 175 Hz) of roating speed than the measurements of reference (175 and 200 Hz). Due to the effects of rotating speed on RMS, as seen in Fig. 5.1, there may be no difference between the RMS corresponding to the healthy and damaged structure: however, as depicted in Fig. 5.11b, the method clearly detects the presence of damage.

Similar comments emerge when observing Fig. 5.12, in which all loadings are considered, with $m = 4$ principal components. The correct damage detection using full range (150, 175 and 200 Hz) data of reference and damaged system is shown in Fig. 5.12a. Damage is also correctly detected in Fig. 5.12b, in which the data acquisition on the damaged structure has been realised at lower values (150 and 175 Hz) of roating speed than the measurements of reference (175 and 200 Hz).

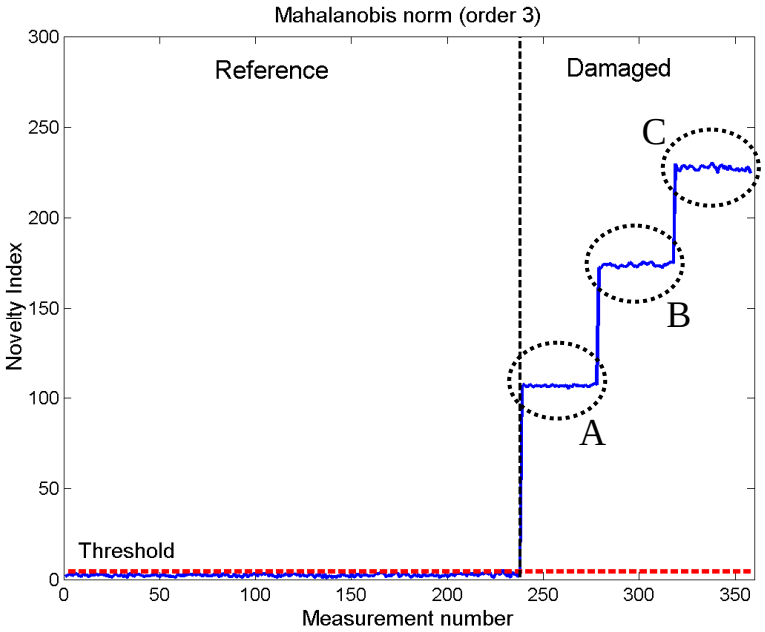
In Figs. 5.11 and 5.12 the effects due to the different operational conditions have been removed by the method: the NIs of reference data are very smooth and do not show any depencence on rotating speed or applied load. However, these effects have not completely been removed on the damaged data: this can be observed in both figures by focusing on regions A, B and C, which correspond to 150, 175 and 200 Hz, respectively. An explanation is given in Fig. 5.13a, which depicts the evolution of RMS_9 as a function of RMS_2, for the reference and the damaged system. Fig. 5.13a shows a different “slope” between the reference and the damaged case: damaged data at 150 Hz are much closer to the blue linear

regression of reference data, while damaged data at 200 Hz are farther. This leads to the “slope” formed by the NIs of data corresponding to regions A, B and C in Figs. 5.11 and 5.12. Note that this concept is not so evident when considering different values of applied load.
Moreover, what is clear for channels close to the damage (such as channel 9 in Fig. 5.13a) may not be observed when considering farther channels. This is the case of Fig. 5.13b, which depicts the evolution of RMS_3 as a function of RMS_2, for the reference and the damaged system. In this case, any difference in “slope” cannot be seen and even a damage detection is not possible. This remark about the damage detection capabilities of a single channel (or group of channels, i.e. a sensor) will be useful for introducing a damage localisation technique in Section 5.2.3.

Test #3

In Fig. 5.14 damaged data are those of Test #3, for the bearing 4A, at fixed values of rotating speed (300 Hz) and applied load (1800 N): the measures have been recorded in close succession in order to follow the bearing temperature increasing. Since for this test a reference case in the same environmental conditions has not been measured, other data from the same setup have been selected as reference. Then, reference data are those of Test #1, for the healthy bearing 0A, at all rotating speeds and applied loads.
Damage detection of Fig. 5.14 is correct. However, reference data of Test #1 have not been acquired in an optimum way for carrying out a PCA-based detection, since this was not the main objective at that time. Data have been measured in close succession, so that the temperatures (oil and bearing) were ***not*** expected to reach stabilisation. The consequence is shown in Fig. 5.15: the temperature at which each reference measurement is acquired is strictly dependent on the assumed values of rotating speed and loading. In other words, reference data from Test #1 do not cover the entire domain of possible temperatures: for each couple of operational conditions (rotating speed and applied load), temperature is single-valued. For this reason, reference data are not enough to perform the PCA-based detection when data are measured in a limited range of environmental variations: temperature is the only condition that changes in damaged data of Test #3.

(a)



(b)

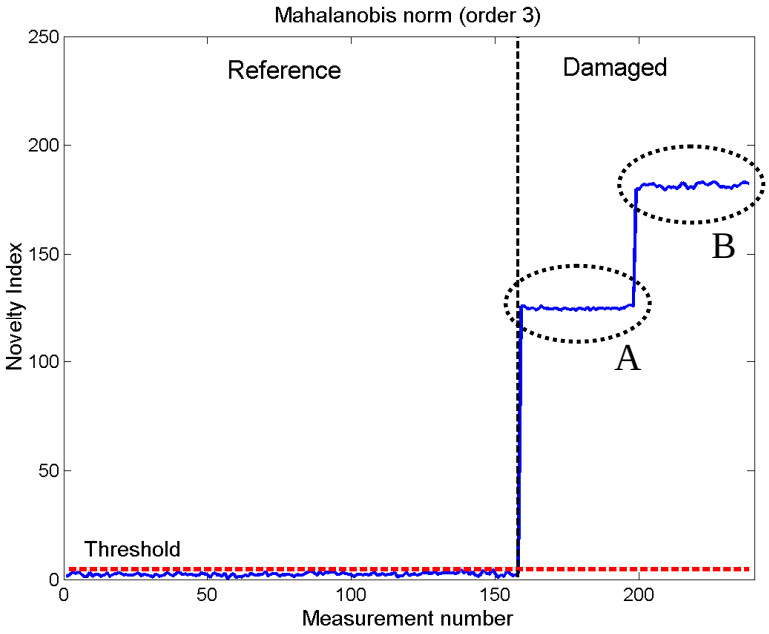
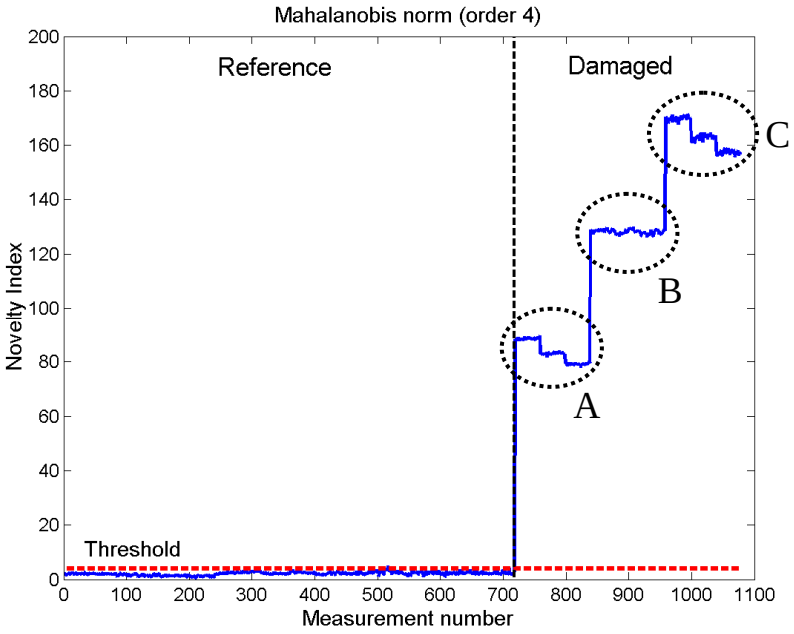


Figure 5.11. Test #5, fixed loading. (a) Damage detection using full range (150, 175 and 200 Hz) data of reference and damaged system. (b) Damage detection using two sets of data at different rotating speeds: reference at higher values (175 and 200 Hz) and damaged at lower values (150 and 175 Hz).

(a)



(b)

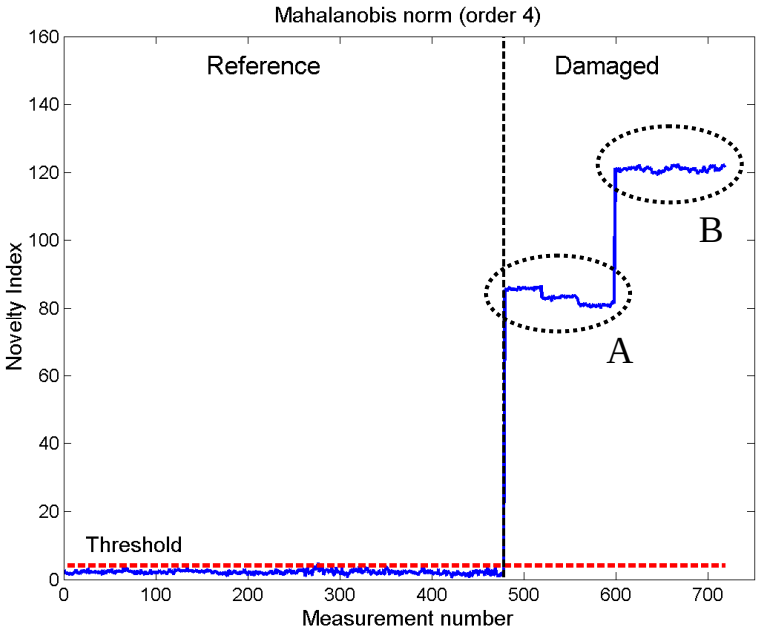
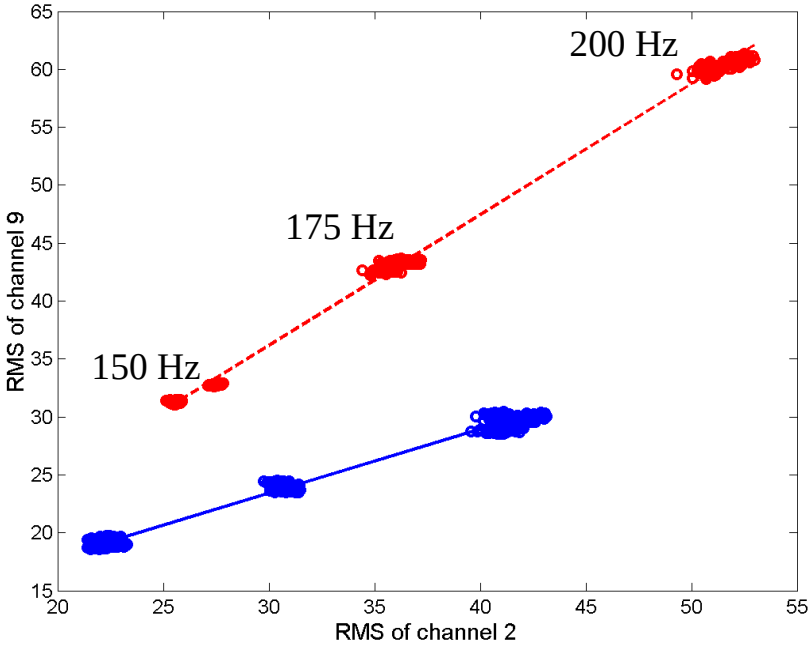


Figure 5.12. Test #5, all loadings. (a) Damage detection using full range (150, 175 and 200 Hz) data of reference and damaged system. (b) Damage detection using two sets of data at different rotating speeds: reference at higher values (175 and 200 Hz) and damaged at lower values (150 and 175 Hz).

(a)



(b)

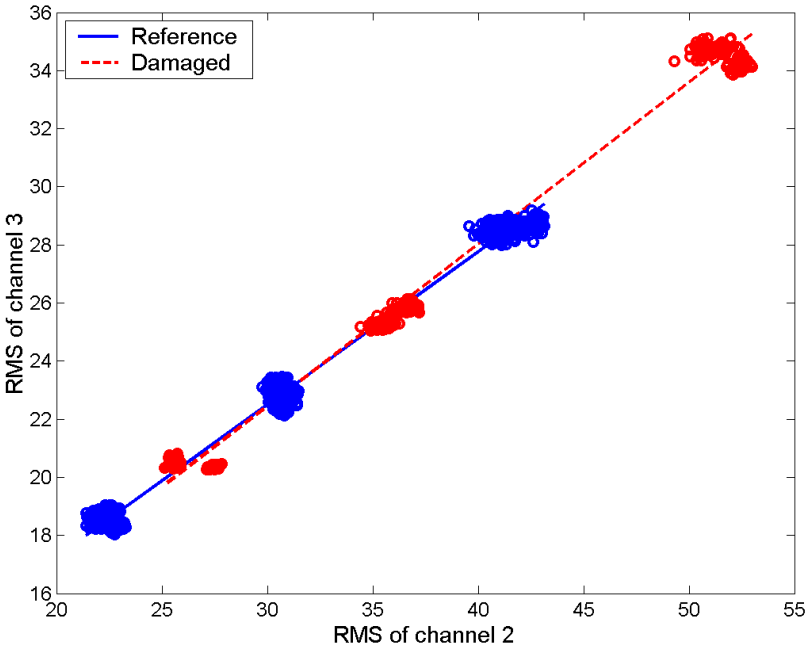


Figure 5.13. Test #5, full range data of reference and damaged system. (a) Diagram showing the evolution of RMS_9 as a function of RMS_2. (b) Evolution of RMS_3 as a function of RMS_2.

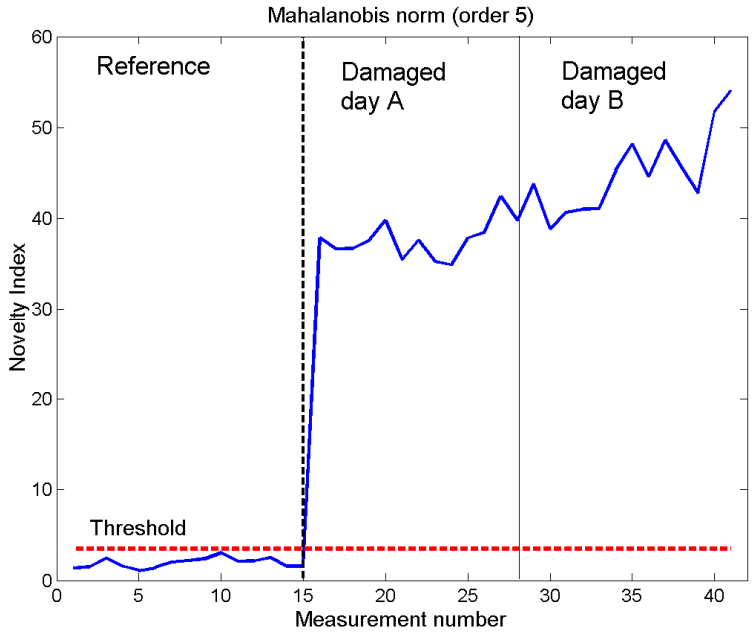


Figure 5.14. Damage detection. Reference data are those of Test #1, for the healthy bearing 0A, at all rotating speeds and applied loads. Damaged data are those of Test #3, for the bearing 4A, at fixed values of rotating speed (300 Hz) and applied load (1800 N).

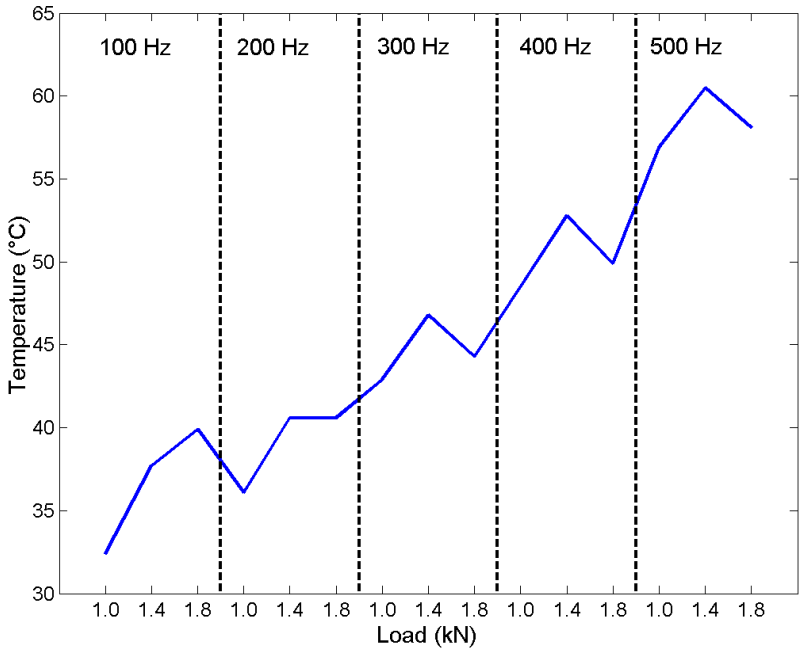


Figure 5.15. Test #1, healthy bearing 0A. Temperatures of bearing 3, at all values of rotating speed and applied load.

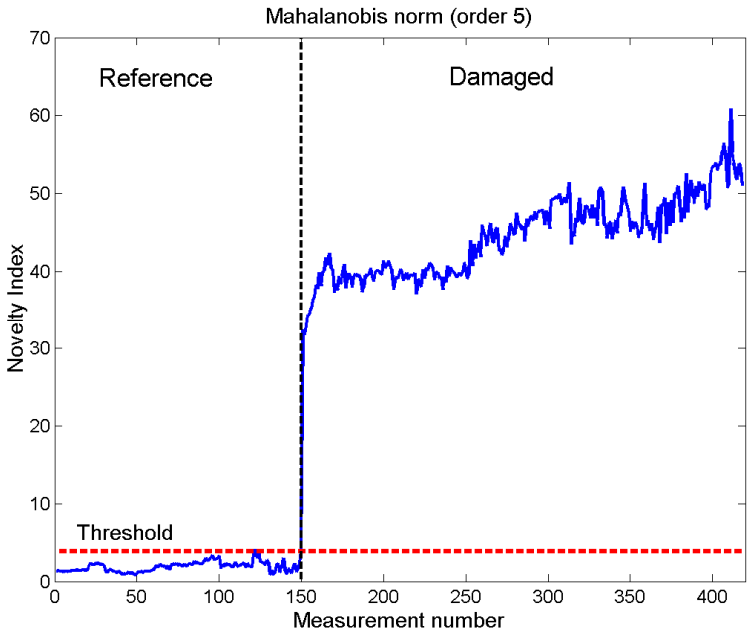


Figure 5.16. Damage detection. Reference data are those of Test #1 (enlarged), for the healthy bearing 0A, at all rotating speeds and applied loads. Damaged data are those of Test #2, for the bearing 4A, at fixed values of rotating speed (300 Hz) and applied load (1800 N).

Test #2

In Fig. 5.16 damaged data are those of Test #2, for the bearing 4A, at fixed values of rotating speed (300 Hz) and applied load (1800 N): the endurance test can be monitored in order to observe any possible damage increase.

This case is very similar to the analysis of Test #3: a reference case in the same environmental conditions has not been measured. As above, reference data are those of Test #1, for the healthy bearing 0A, at all rotating speeds and applied loads. But reference data are few, with respect to the huge amount of damaged data, so a data "enlargement" has been considered to have more data and a more robust statistical confidence. In detail, each acquired reference data has been split into 10 shorter (0.8 seconds) time histories, for which the RMS values have been computed. Observe, however, that this procedure does not solve the problem of having a single-valued temperature, as seen in Fig. 5.15: each group of 10 time sub-histories shares the same temperature value.

Damage detection of Fig. 5.16 is correct. Moreover, an increasing trend in the NIs of damaged data can be observed: this may be due to an increase in the damage extent. Unfortunately, when the bearing has been unmounted and its rolling element indentation has been inspected, no significant differencies have been seen with respect to the original damage of Fig. 4.8. The possibility of a slight damage increment during the endurance test, which can not be clearly detected through a rough visual inspection, should not be excluded. However, the main reason of the increasing trend in the NIs is probably due to temperature, as explained above: reference data are not completely suitable for interpreting the (correct) damage detection as a damage extent evaluation. An attempt of addressing the issue that concerns damage extent evaluation will be made in Section 5.2.4.

5.2.2. False-positive verification

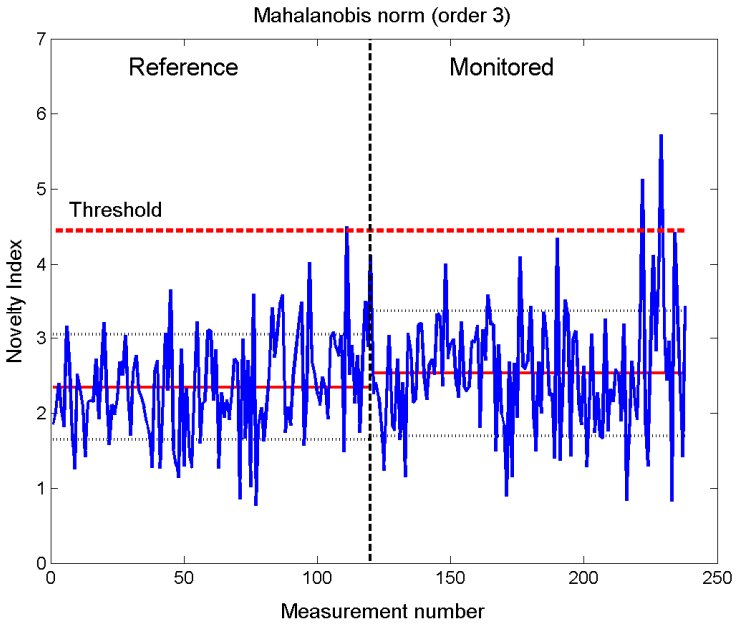
The method should be able to issue no alarm if no damage occurs even when measurements are performed under different operational and environmental conditions. In the following, some of the tests described in Section 4.3 are studied and results are presented.

Test #5

Healthy data (bearing 0A) collected at the fixed loading of 1800 N are considered and analysed in Fig. 5.17, with $m = 3$ principal components. Fig. 5.17a demonstrates the correct false-positive verification using full range (150, 175 and 200 Hz) data of reference and monitored system: only 2 data acquisitions out of 120 are identified as damaged (but they are very close to the threshold anyway) and the ratio $\overline{NI}_c / \overline{NI}_r$ is very close to 1.

Fig. 5.17b shows what happens when measurements are only available for two sets of operational conditions: reference at higher values (175 and 200 Hz) and monitored at lower values (175 and 150 Hz) of rotating speed. Monitored data at 175 Hz are correctly placed under the threshold value, since data at 175 Hz are included in the reference data set. On the contrary, a clearly false detection occurs for monitored data at 150 Hz, since this operational condition is excluded from the reference data set.

(a)



(b)

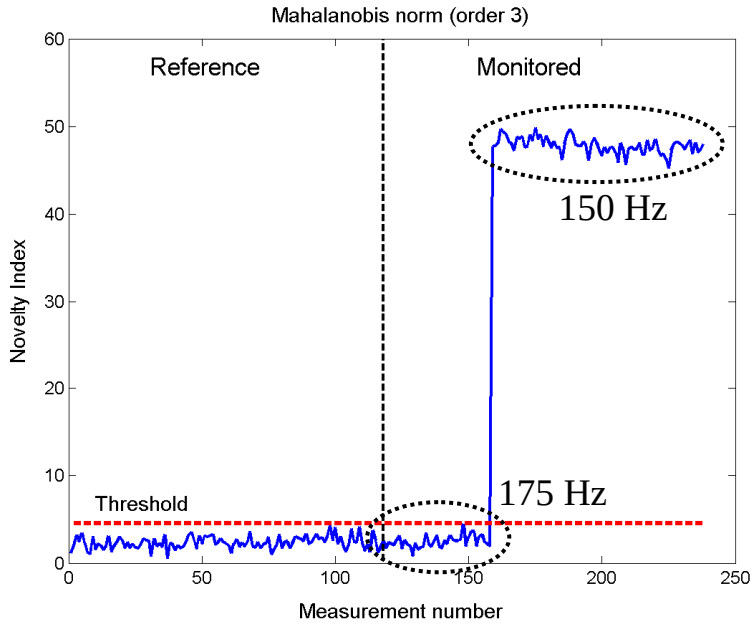
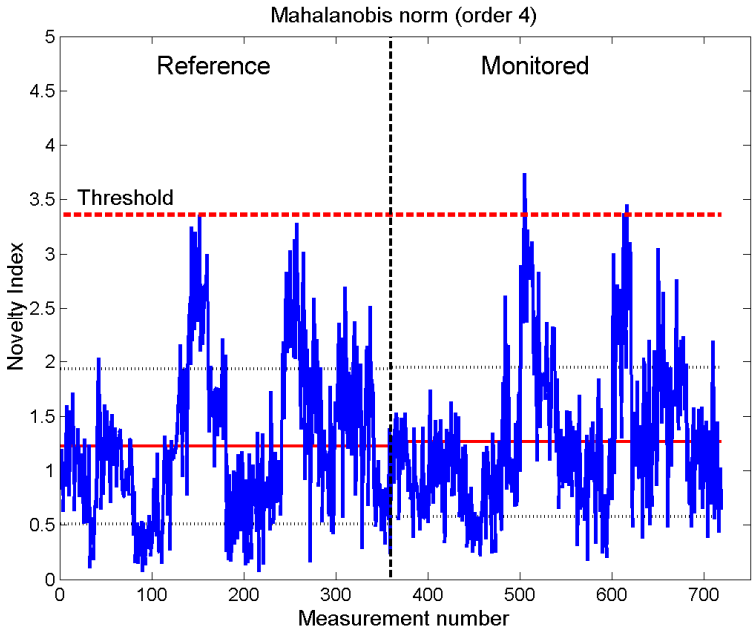


Figure 5.17. Test #5, fixed loading. (a) False-positive verification using full range data (150, 175 and 200 Hz) of reference and monitored system. (b) False-positive verification using two sets of data at different rotating speeds: reference at higher values (175 and 200 Hz) and monitored at lower values (175 and 150 Hz).

(a)



(b)

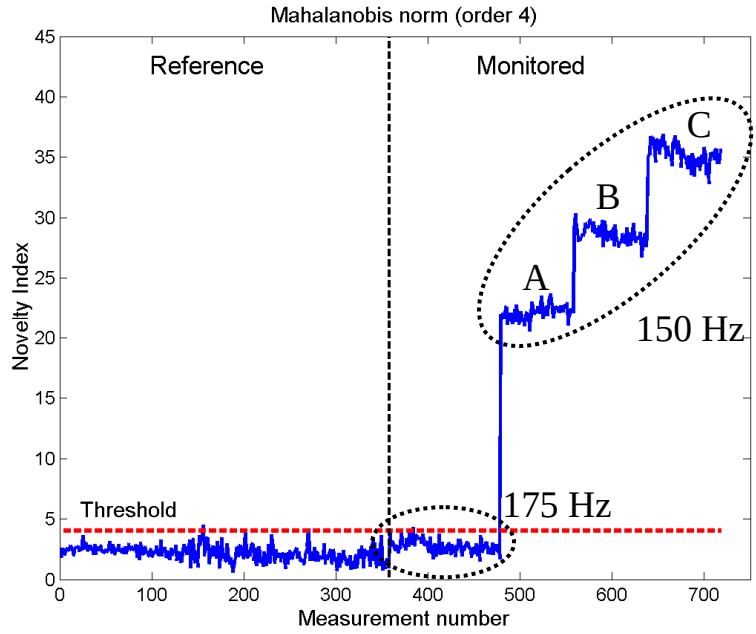
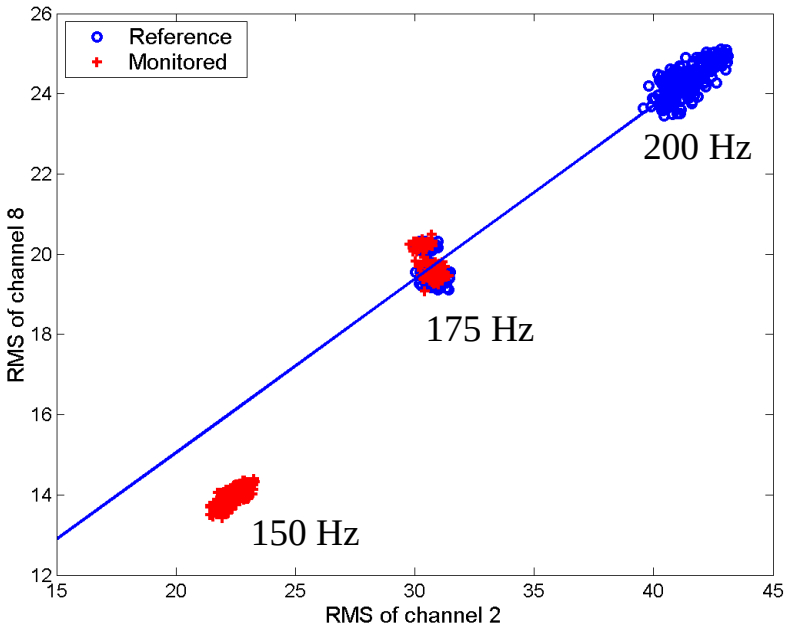


Figure 5.18. Test #5, all loadings. (a) False-positive verification using full range (150, 175 and 200 Hz) data of reference and monitored system. (b) False-positive verification using two sets of data at different rotating speeds: reference at higher values (175 and 200 Hz) and monitored at lower values (175 and 150 Hz).

(a)



(b)

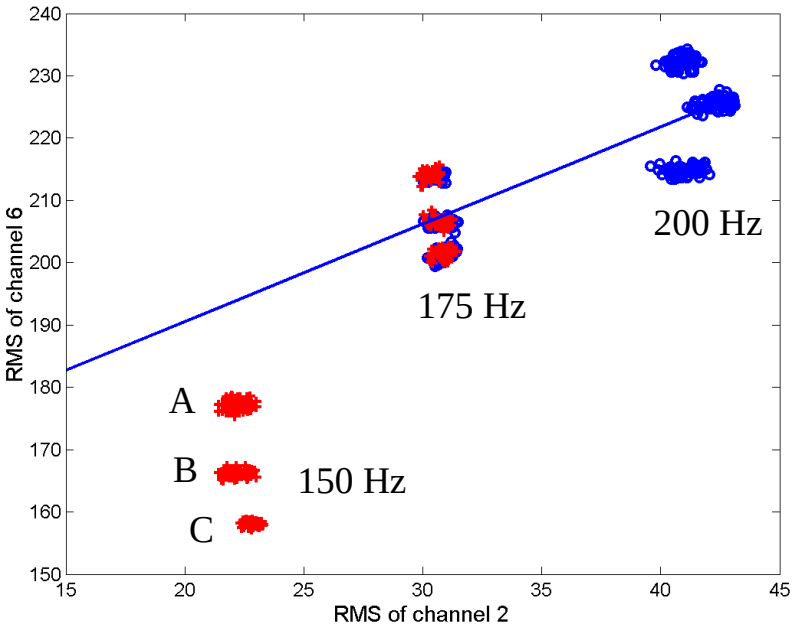


Figure 5.19. Test #5, all loadings. Two sets of data at different rotating speeds: reference at higher values (175 and 200 Hz) and monitored at lower values (150 and 175 Hz). (a) Diagram showing the evolution of RMS_8 as a function of RMS_2. (b) Evolution of RMS_6 as a function of RMS_2.

Similar comments emerge when observing Fig. 5.18, in which all loadings are considered, with $m = 4$ principal components. The correct false-positive verification using full range (150, 175 and 200 Hz) data of reference and monitored system is shown in Fig. 5.18a. In Fig. 5.18b measurements are only available for two sets of operational conditions: monitored data at 175 Hz are correctly placed under the threshold value, while a clearly false detection occurs for monitored data at 150 Hz.

An explanation to the wrong results of Figs. 5.17b and 5.18b is given by observing Fig. 19 (all loadings are considered): since the features are not perfectly linear, the PCA-based detection method is not robust when features are identified in a limited range of operational variations. Fig. 5.19a shows the evolution of RMS_8 as a function of RMS_2: in this simple 2D case, the straight blue line corresponds to the principal component of the reference data set. It is easy to see that monitored data at 175 Hz are on this line, while monitored data at 150 Hz are separated from this line and are consequently detected as damaged.

A similar behaviour can be seen in Fig. 5.19b, showing the evolution of RMS_6 as a function of RMS_2. Moreover, three regions of data can be distinguished at 150 Hz: regions A, B and C corresponding to the three values of applied load, 1400, 1600 and 1800 N, respectively. Each region has a different distance from the straight blue line and this corresponds to the differencies that can be observed in Fig. 5.18b.

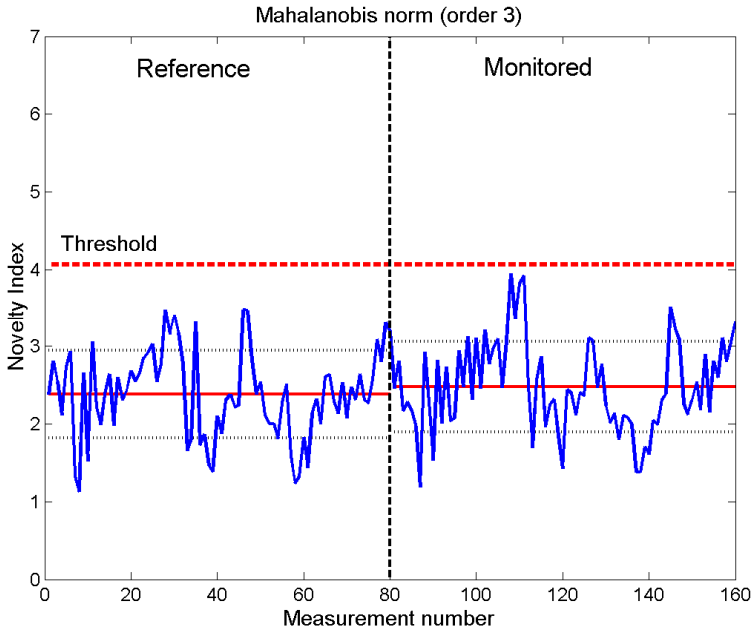
In conclusion, in those cases in which linearity among features is not perfectly guaranteed, the only way for preventing false detections to occur consists in using full range data of reference and monitored system, as performed in Figs. 5.17a and 5.18a.

Test #4

By using the electrical heater, the measurements of Test #4 have been recorded in four steady-state values of the oil temperature (45, 60, 75 and 85 °C, within a 7% margin).

Healthy data (bearing 0A) collected at the fixed rotating speed of 300 Hz are considered and analysed in Fig. 5.20, with $m = 3$ principal components. Fig. 5.20a demonstrates the correct false-positive verification using full range (all temperatures) data of reference and monitored system: all the 80 data acquisitions are identified as healthy and the ratio $\overline{NI}_c / \overline{NI}_r$ is very close to 1.

(a)



(b)

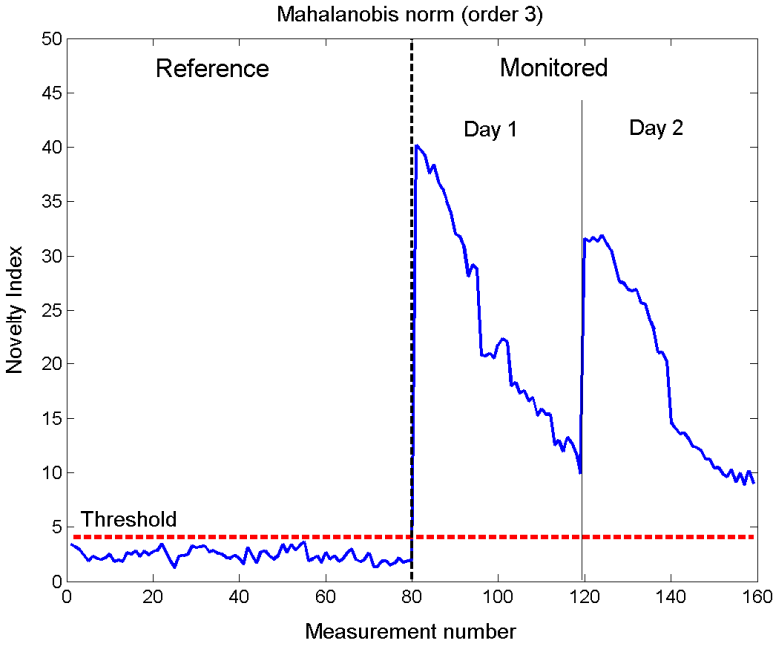
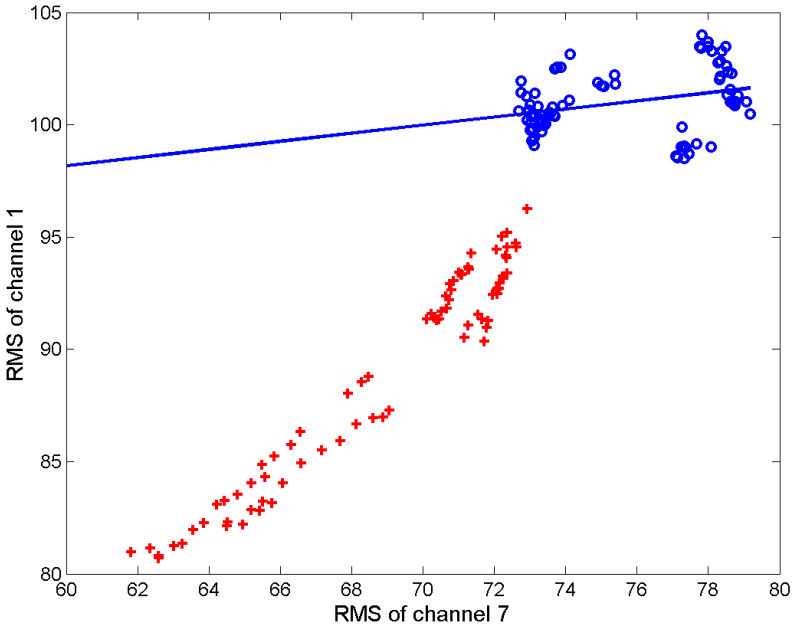


Figure 5.20. Test #4, fixed rotating speed. (a) False-positive verification using full range (all temperatures) data of reference and damaged system. (b) False-positive verification using two sets of data at different temperatures: reference at higher values ($T > 60$ °C) and monitored at lower values ($T < 60$ °C).

(a)



(b)

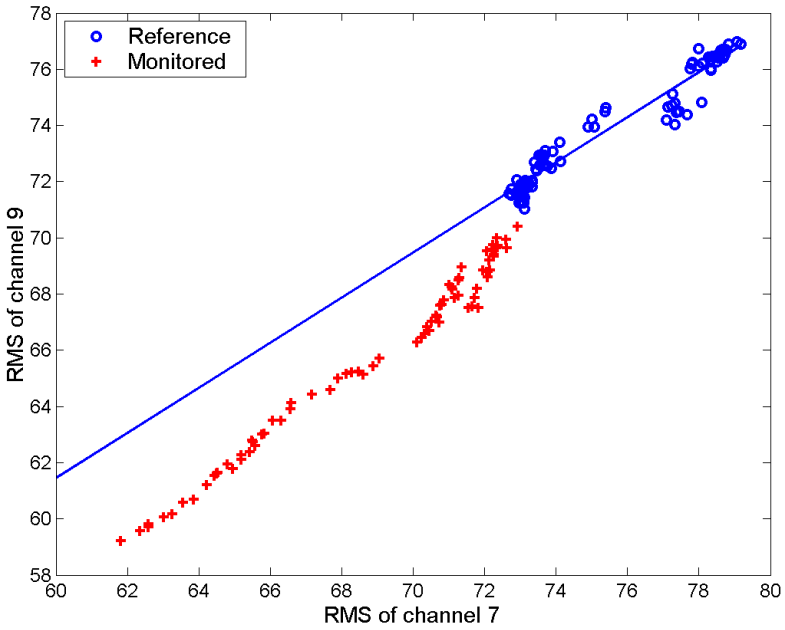


Figure 5.21. Test #4, fixed rotating speed. Two sets of data at different temperatures: reference at higher values ($T > 60\,°C$) and monitored at lower values ($T < 60\,°C$). (a) Diagram showing the evolution of RMS_1 as a function of RMS_7. (b) Evolution of RMS_9 as a function of RMS_7.

Fig. 5.20b shows what happens when measurements are only available for two sets of data at different temperatures: reference at higher values ($T > 60\,°C$) and monitored at lower values ($T < 60\,°C$). A false detection occurs for all monitored data: as seen for Test #5, the reason is nonlinearity, with the difference that in this case nonlinearity is due to temperature variations instead of rotating speed variations. Since the features are not perfectly linear, the PCA-based detection method is not robust when features are identified in a limited range of environmental variations.

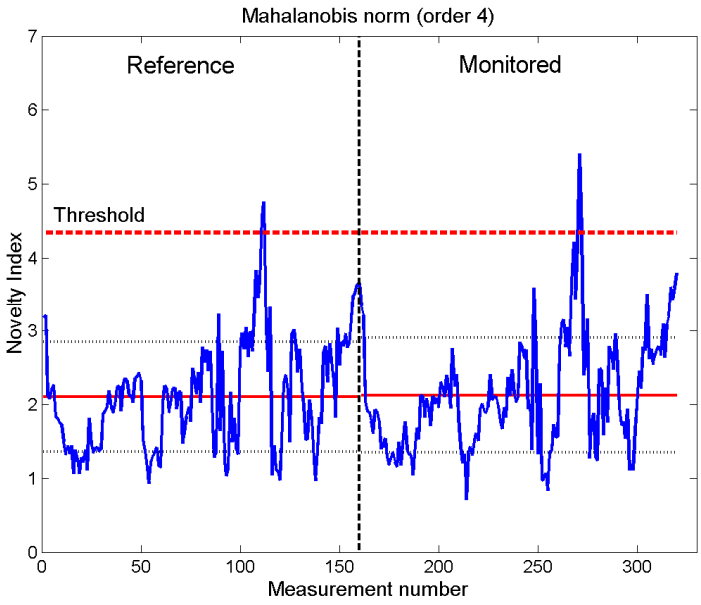
An explanation to the wrong results of Fig. 5.20b is given by observing Fig. 5.21: in a simple 2D case, the straight blue line corresponds to the principal component of the reference data set. All monitored data are separated from this line and are consequently detected as damaged. Moreover, Fig. 5.21 also explains the decreasing trend of the monitored-data NIs in Fig. 5.20b: as their temperature gets closer to the "splitting" value of 60 °C, the distance between the monitored data and the straight blue line is reduced. As a consequence, the NI of these monitored data decreases and they are more hardly detected as damaged.

As seen for Test #5, the same conclusion can be drawn: in those cases in which linearity among features is not perfectly guaranteed, the only way for preventing false detections to occur consists in using full range data of reference and monitored system, as performed in Fig. 5.20a.

The study is extended by adding rotating speed variations to those caused by temperature. Healthy data (bearing 0A) collected at two values (200 and 300 Hz) of rotating speed are considered and analysed in Fig. 5.22, with $m = 4$ principal components. Fig. 5.22a demonstrates the correct false-positive verification using full range (all temperatures) data of reference and monitored system: only 1 data acquisition out of 160 is identified as damaged (but it is very close to the threshold anyway) and the ratio $\overline{NI}_c / \overline{NI}_r$ is very close to 1.

Fig. 5.22b shows what happens when measurements are only available for two sets of data at different temperatures: reference at higher values ($T > 55\,°C$) and monitored at lower values ($T < 55\,°C$). In this case the majority of monitored data are correctly identified as healthy and those identified as damaged are very close to the threshold anyway; moreover, the ratio $\overline{NI}_c / \overline{NI}_r$ is is not far from being equal to 1.

(a)



(b)

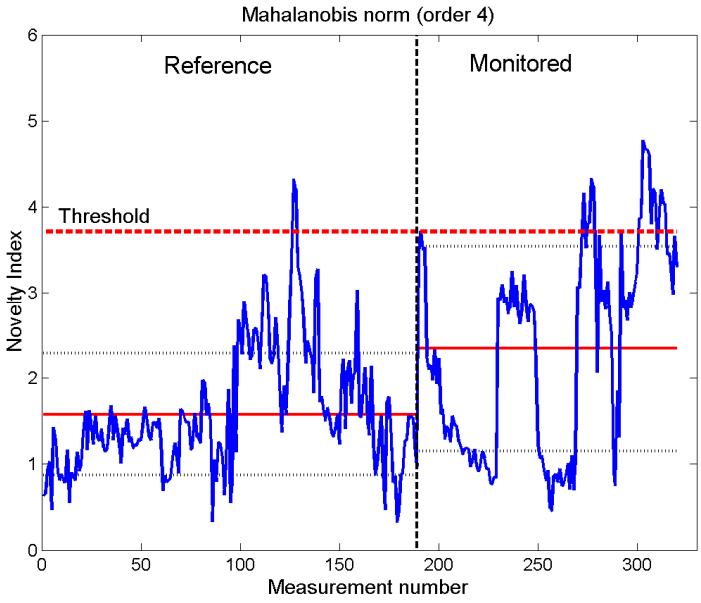
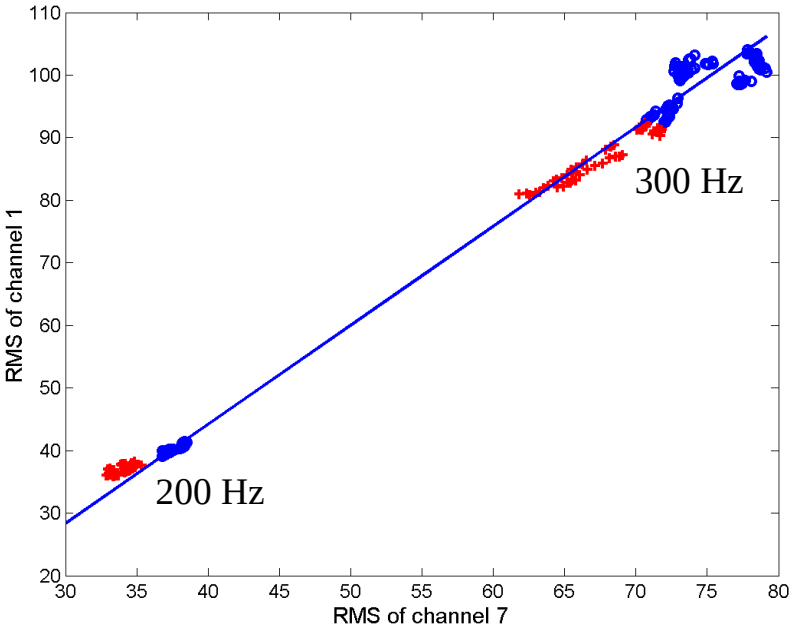


Figure 5.22. Test #4, two rotating speeds (200 and 300 Hz). (a) False-positive verification using full range data (all temperatures) of reference and damaged system. (b) False-positive verification using two sets of data at different temperatures: reference at higher values ($T > 55\,°\text{C}$) and monitored at lower values ($T < 55\,°\text{C}$).

(a)



(b)

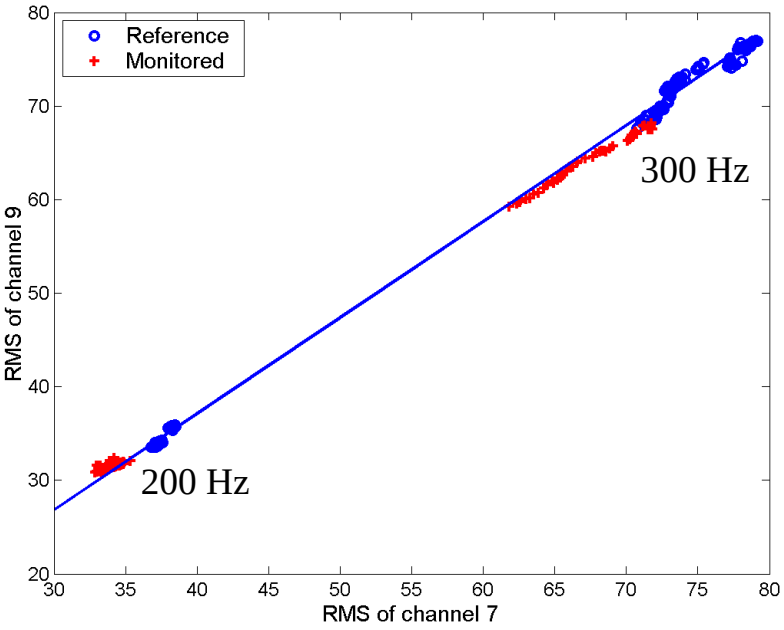


Figure 5.23. Test #4, two rotating speeds (200 and 300 Hz). Two sets of data at different temperatures: reference at higher values ($T > 55\,°\text{C}$) and monitored at lower values ($T < 55\,°\text{C}$). (a) Diagram showing the evolution of RMS_1 as a function of RMS_7. (b) Evolution of RMS_9 as a function of RMS_7.

The situation depicted in Fig. 5.22b is clearly different from the one in Fig. 5.20b: this can be explained by observing Fig. 5.23. In this simple 2D case, the straight blue line corresponds to the principal component of the reference data set, which is composed by data acquired at two rotating speeds (200 Hz and 300 Hz). All monitored data are close to this line and cannot consequently be detected as damaged. In this situation, the feature variations due to temperature are "masked" by the more significative variations due to rotating speed. In other words, temperature variations are local, with respect to the global variations caused by rotating speed, and they are considered by the PCA-based method as a kind of "noise". Since the global variations due to rotating speed are (approximately) linear, the false-positive verification of Fig. 5.22b is correct.

5.2.3. Damage localisation

In this section a simple damage localisation technique is introduced and applied to some of the tests described in Section 4.3. Observe that this procedure is proper for the present case, in which the PCA-based method is applied to bearing diagnostics.

Once a damage has been detected, the starting point is the estimate of the residual error matrix, given by (2.12). The j-th row of R represents the residual error vector associated to the j-th signal feature (measurement channel). Then, a group of channels (denoted by G) can be selected and consequently a new residual error matrix R^G is defined by gathering the rows of R such that the j-th channel belongs to G. For example, G can be composed by the three channels of a single sensor, or by some channels which are placed in the same zone, close or far from the monitored bearing.

The next step consists in defining a local Novelty Index NI_G, by using one of the norms in (2.13) or (2.14), for each group of channels. Then, a threshold value can be computed as in Section 2.3.1:

$$Th_G = \overline{NI_G} + \alpha\sigma_G .$$

In this way, each group of channels can be evaluated in order to determine which is the most sensitive to damage: this is expected to be strictly related to its distance from the damaged bearing. Such an evaluation can be performed by defining a Relative Distance from Threshold (RDT) for each group:

$$RDT_G = \frac{NI_G - Th_G}{Th_G}. \tag{5.2}$$

RDTs from different groups can be directly compared to investigate which group has the highest value, i.e. is closer to damage location. Another way of exploiting RDTs is to define a new threshold, as done for the NIs: a damage localisation analysis can be performed in the same way of damage detection. It is easy to see from (5.2) that the new threshold is equal to 0 for each group, since it corresponds to the relative distance of the threshold from the threshold itself.

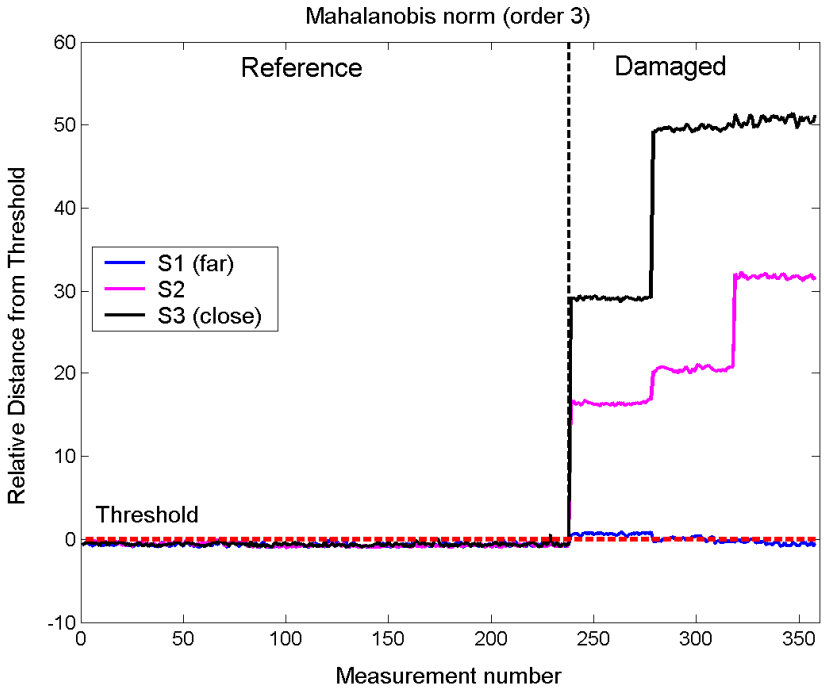
Test #5

The damage localisation technique is applied to data from Test #5. Three groups of channels are considered, corresponding to the 3 accelerometers of Setup #2: group S1 (channels 1-2-3) is far from damage; group S2 (channels 4-5-6) is mid-placed; group S3 (channels 7-8-9) is close to damage. Equation (5.2) is exploited to obtain the results shown in Fig. 5.24.
Data collected at the fixed loading of 1800 N are considered in Fig. 5.24a, with $m = 3$ principal components, while all loadings are considered in Fig. 5.24b, with $m = 4$ principal components. In both cases, the RDTs of damaged data are correctly detecting group S3 as closer to damage. Moreover, the figures demonstrate that RDT depends on the distance of a group of channels from the damaged bearing.

Test #2

The damage localisation technique is applied to data from Test #2. As remarked in Section 5.2.1, for this test a reference case in the same environmental conditions has not been measured: other data from the same setup have been selected as reference. Then, reference data are those of Test #1, for the healthy bearing 0A, at all rotating speeds and applied loads. Moreover, since reference data are few with respect to the huge amount of damaged data, a data "enlargement" has been considered to have more data and a more robust statistical confidence. In detail, each acquired reference data has been split into 10 shorter (0.8 seconds) time histories, for which the RMS values have been computed.

(a)



(b)

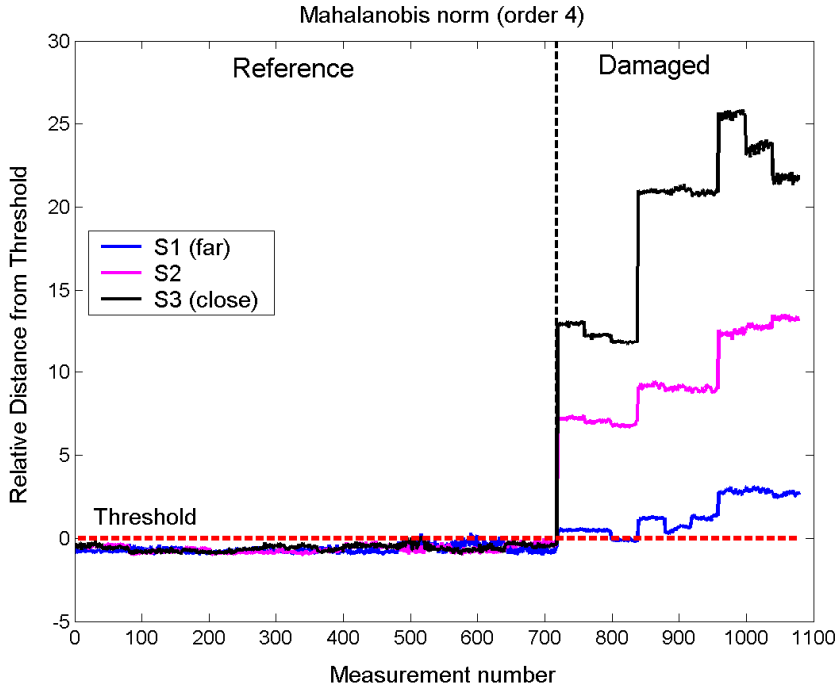


Figure 5.24. Test #5. Damage localisation using full range data (150, 175 and 200 Hz) of reference and damaged system. (a) Fixed loading. (b) All loadings.

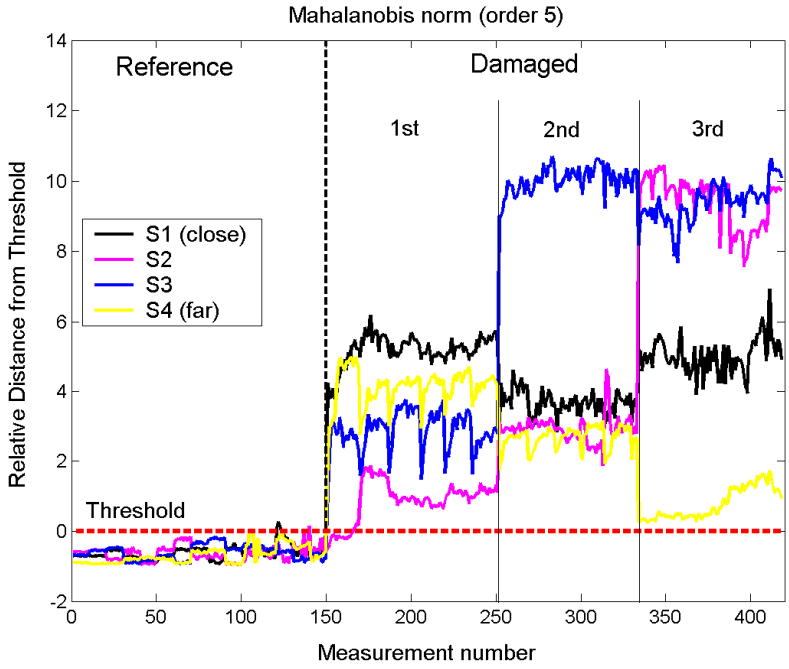


Figure 5.25. Incorrect damage localisation. Reference data are those of Test #1 (enlarged), for the healthy bearing 0A, at all rotating speeds and applied loads. Damaged data are those of Test #2, for the bearing 4A, at fixed values of rotating speed (300 Hz) and applied load (1800 N).

Three groups of channels are considered, corresponding to the 4 accelerometers of Setup #1: group S1 (channels 1-2-3) is the closest to damage; groups S2 (channels 4-5-6) and S3 (channels 7-8-9) are mid-placed; group S4 (channels 10-11-12) is the farthest from damage. Equation (5.2) is exploited to obtain the results shown in Fig. 5.25. The results are quite confused, revealing a failure in applying the presented damage localisation technique.

Some reasons may be found to explain such an incorrect result. The first is related to the different conditions at which reference data are acquired, as already described in Section 5.2.1. Reference data have been measured in close succession, so that the temperatures (oil and bearing) were ***not*** expected to reach stabilisation. Then, reference data are not enough to perform the PCA-based detection when data are measured in a limited range of environmental variations: temperature is the only condition that changes in damaged data of Test #2. The second reason is related to the mounting/unmounting operations that have been applied to the damaged bearing, in order to inspect its rolling element indentation. These operations can be clearly recognised in Fig. 5.25, by splitting the damaged

data into three parts, corresponding to the three successive bearing mountings on the test rig. Large changes can be seen in the results, from one mounting to another. In this way, different mountings can be seen as a new operational condition that can influence the results. This may be due to the different conditions (i.e. screw tightenings) applied to the system when mounting the damaged bearing: such a dependence, for example, has been demonstrated for a cantilever beam in [95].

5.2.4. Damage extent evaluation

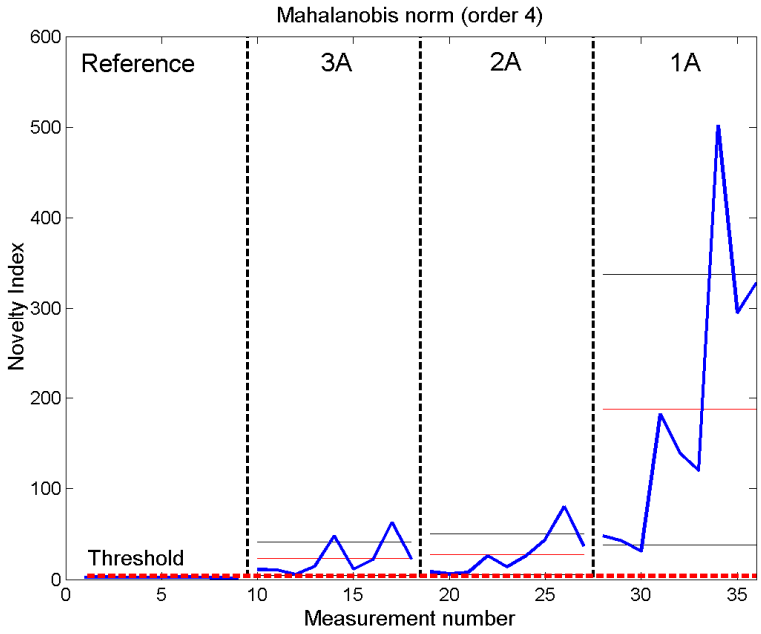
The sensitivity of the PCA-based method to damage extent is investigated in this section. The aim is demonstrating the capability of the method to detect an increasing damage extent independently of any operational and environmental condition.

The damage extent evaluation is applied to data from Test #1, in which all the seven types of bearing listed in Table 4.2 have been mounted on the test rig.

At first, all loadings are considered. Fig. 5.26 shows damage detection by using three rotating speeds (200, 300 and 400 Hz) of reference and damaged system, for increasing damage extents (150, 250 and 450 microns). Figs. 5.26a and 5.26b represent the investigation of the inner ring indentation and the rolling element indentation, respectively. Some dependence on the damage extent can be observed in both cases, even if some influence of the operational conditions is still present. This is due to the few number of reference data, since the tests have been carried out in close succession.

As in previous sections, the data sets (for both the reference and damaged states) have been “enlarged” to have more data and a more robust statistical confidence. In detail, each acquired reference data has been split into 10 shorter (0.8 seconds) time histories, for which the RMS values have been computed. The results are presented in Fig. 5.27, but they are very similar to those of Fig. 5.26, as expected: some dependence on the damage extent can be observed, but some influence of the operational conditions is still present. In conclusion, with a few number of reference data, the influence of all the conditions (rotating speed and applied load, with some additional uncertainty over temperature, see Fig. 5.15) cannot be eliminated by the PCA-based method.

(a)



(b)

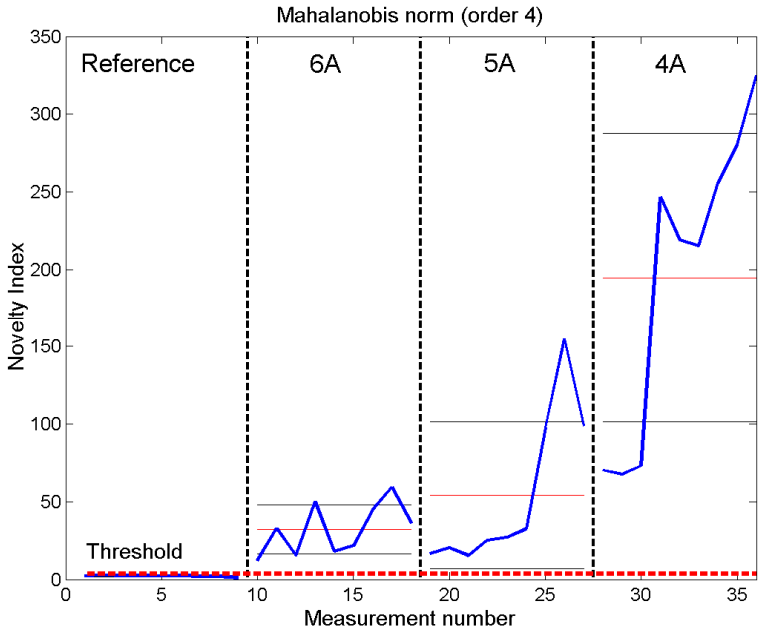
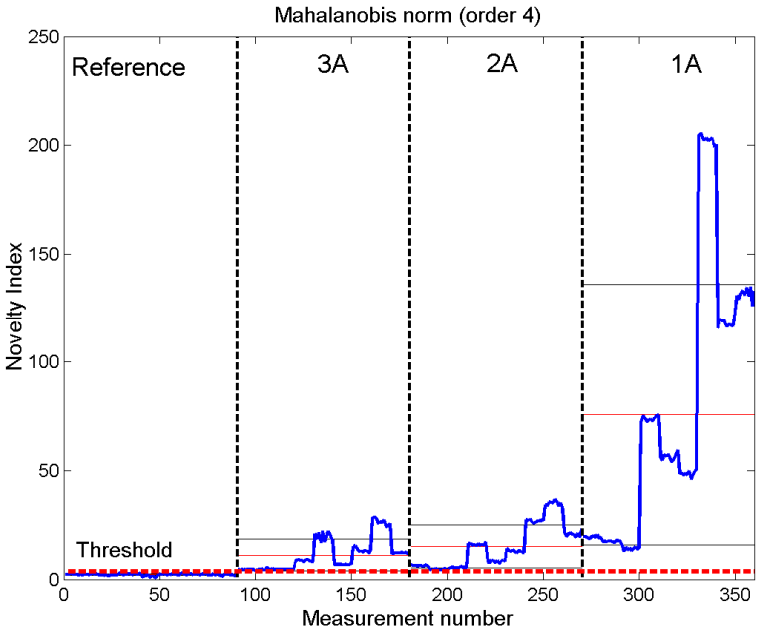


Figure 5.26. Test #1, all loadings. Damage detection using three rotating speeds (200, 300 and 400 Hz) of reference and damaged system, for increasing damage extents (150, 250 and 450 microns). (a) Inner ring indentation. (b) Rolling element indentation.

(a)



(b)

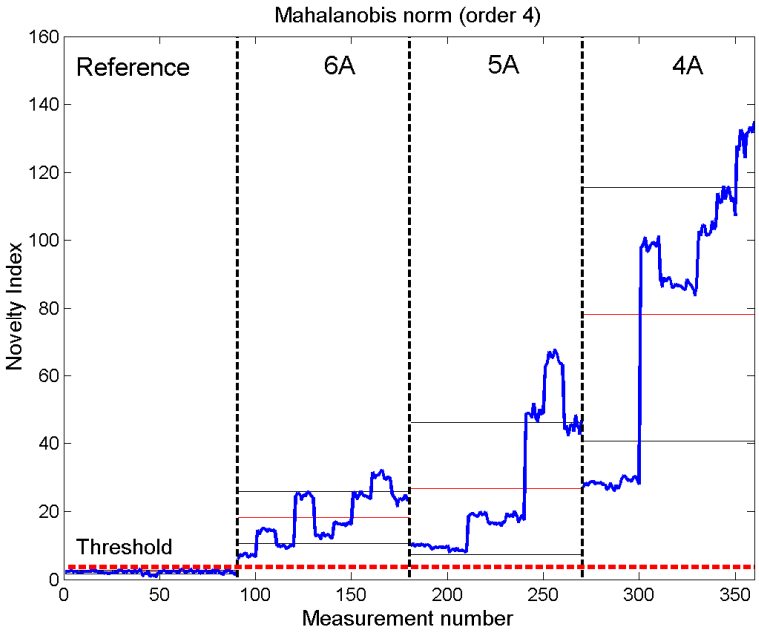


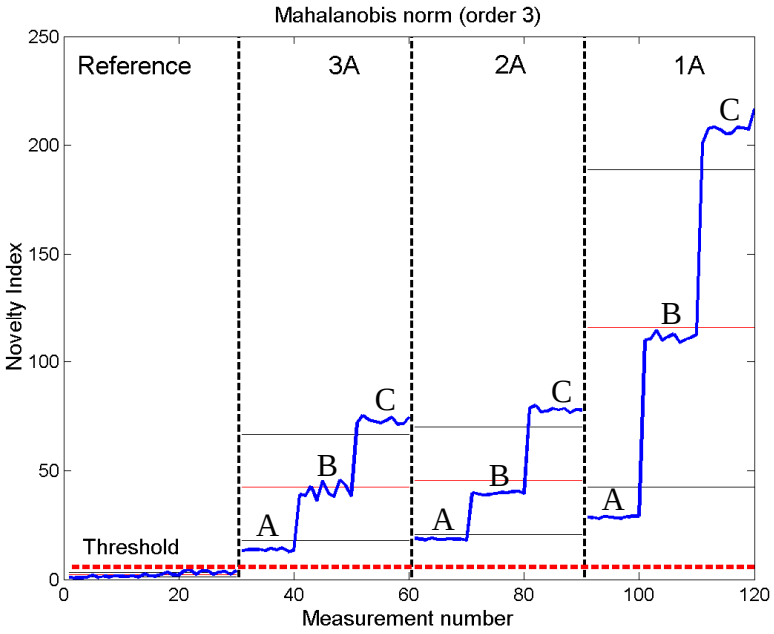
Figure 5.27. Test #1 (enlarged), all loadings. Damage detection using three rotating speeds (200, 300 and 400 Hz) of reference and damaged system, for increasing damage extents (150, 250 and 450 microns). (a) Inner ring indentation. (b) Rolling element indentation.

Then, some restrictions on the operational conditions are applied for obtaining Figs. 5.28 and 5.29. In Fig. 5.28 a fixed loading of 1400 N is considered, at three rotating speeds (200, 300 and 400 Hz): results are more regular and a dependence on the damage extent can be observed. In this case, however, the influence of the rotating speed has not been removed by the method: among the damaged data (and for each damage extent) the three regions A, B and C can be distinguished, corresponding to 200, 300 and 400 Hz, respectively. In this situation an accurate damage extent evaluation is not possible, since a major damage at a lower rotating speed may be confused with a minor damage at a higher rotating speed. This is the case, for example, of Fig. 5.28b: the NIs of bearing 4A, at 200 Hz (region A), are similar to those of bearing 6A, at 400 Hz (region C).

In Fig. 5.29 a fixed rotating speed of 400 Hz is considered, at three applied loads (1000, 1400 and 1800 N). By fixing the rotating speed, the best results are obtained, in terms of damage extent evaluation. The three damage extents are correctly identified and well-separated, in particular for the rolling element indentation (Fig. 5.29b). Some dependence on the applied load can still be observed, especially in Fig. 5.29a, but its effects on the results are slight with respect to those due to the rotating speed variation (Fig. 5.28).

Unfortunately, the problem of having a few number of data, for each type of mounted bearing, affects all the results contained in this section. As already remarked, data of Test #1 have not been acquired in an optimum way for carrying out a PCA-based detection, since this was not the main objective at that time. This is the reason why this section have to be considered only as an “attempt” of applying the PCA-based method for damage extent evaluation. A more suited analysis for this issue should be carried out by performing a more accurate test for each of the seven types of bearing, as done in Test #5 for bearing 4A. Future works and developments will be oriented in this direction.

(a)



(b)

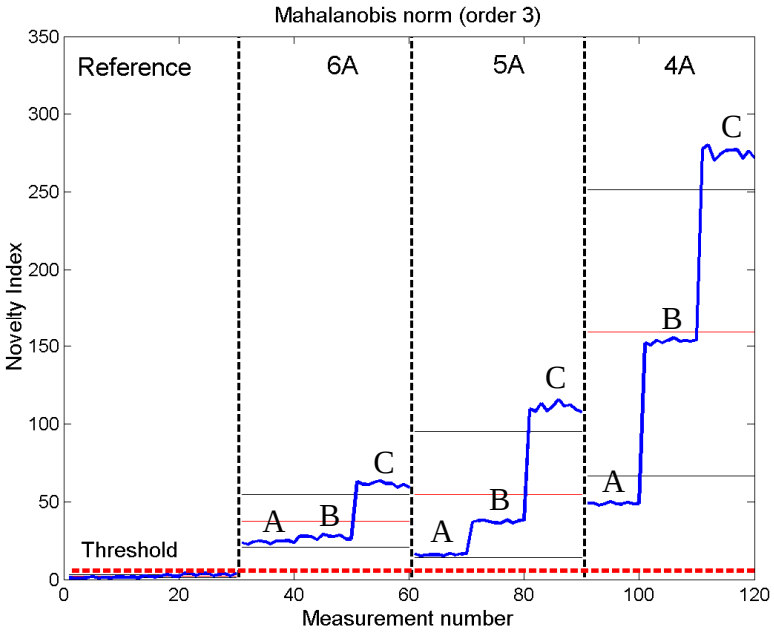
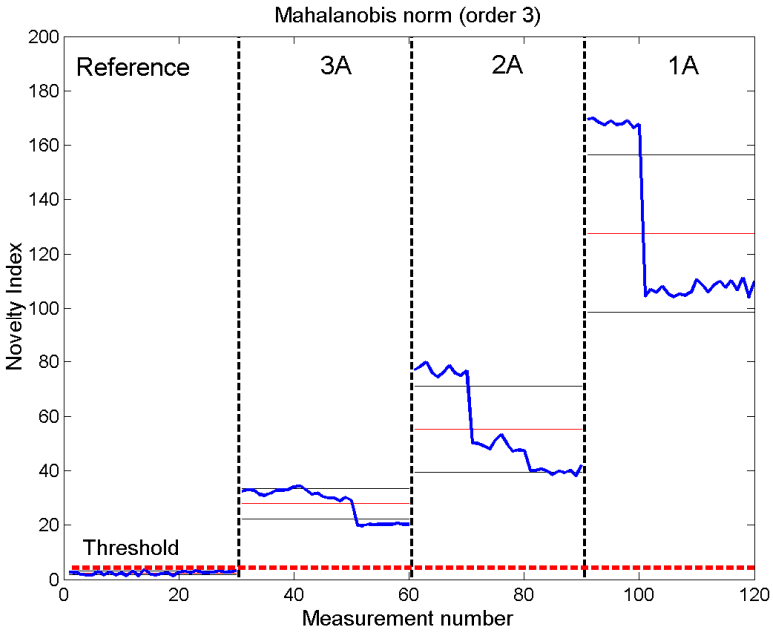


Figure 5.28. Test #1 (enlarged), fixed loading. Damage detection using three rotating speeds (200, 300 and 400 Hz) of reference and damaged system, for increasing damage extents (150, 250 and 450 microns). (a) Inner ring indentation. (b) Rolling element indentation.

(a)



(b)

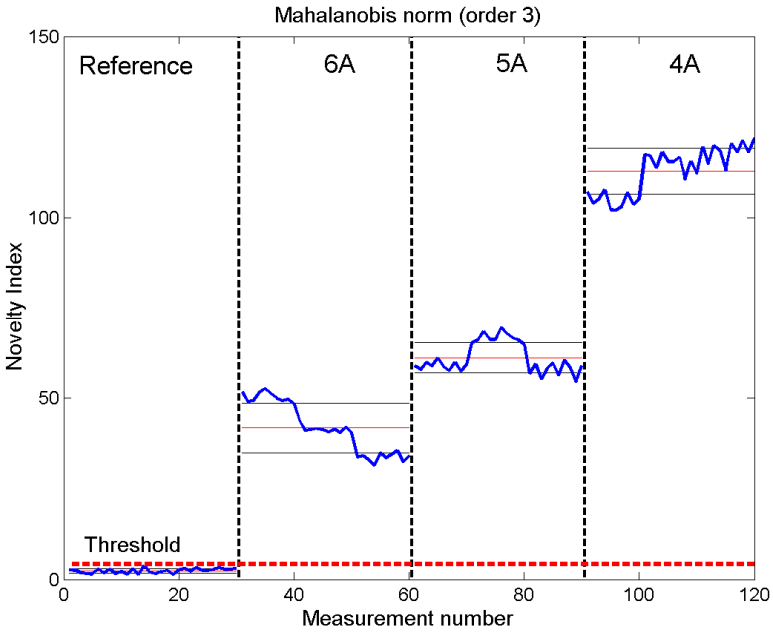


Figure 5.29. Test #1 (enlarged), fixed rotating speed. Damage detection using all loadings (1000, 1400 and 1800 N) of reference and damaged system, for increasing damage extents (150, 250 and 450 microns). (a) Inner ring indentation. (b) Rolling element indentation.

Chapter 6

Subspace identification

In this chapter the state-space representation of linear and nonlinear systems is presented, together with some useful basic properties. In the second part of the chapter the identification procedure is described. Firstly, the state-space matrices are estimated through the application of a subspace method; particular attention is given to the problem concerning computational memory limitations. Then, the estimated matrices are exploited for applying a technique suited for nonlinear system identification. In the end, a procedure for the identification of linear time-varying (LTV) systems is briefly introduced.

6.1. System modelling and properties

6.1.1. Equation of motion

The equation of motion of a linear time-invariant dynamical system, with N degrees of freedom and with lumped parameters, can be described by means of a linear operator $S[z(t)]$ as follows:

$$S[z(t)] = M\ddot{z}(t) + C_v\dot{z}(t) + Kz(t) = f(t), \tag{6.1}$$

where M, C_v $K \in \mathbf{R}^{N\times N}$ are the mass, viscous damping and stiffness matrices respectively, $z(t) \in \mathbf{R}^{N\times 1}$ is the generalised displacement vector and $f(t) \in \mathbf{R}^{N\times 1}$ the generalised force vector, at time t.

In case the dynamical system contains some nonlinearities, that can be described by lumped nonlinear springs and dampers, the equation of motion is expressed from (6.1) through the additional nonlinear operator $NL[z(t), \dot{z}(t)]$:

$$S[z(t)] + NL[z(t), \dot{z}(t)] = S[z(t)] + \sum_{j=1}^{h} \mu_j L_j g_j(t) = f(t). \tag{6.2}$$

The nonlinear operator is expressed as the sum of h components; each of the nonlinear components depends on the nonlinear function $g_j(t) \in \mathbf{R}$, which specifies the class of the nonlinearity (e.g., Coulomb friction, clearance, quadratic damping, etc.), and on a scalar nonlinear coefficient μ_j. The vector $L_j \in \mathbf{R}^{N\times 1}$, whose entries may assume the values 1, -1 or 0, is related to the location of the nonlinear element: it specifies the degrees of freedom joint by the j-th nonlinear component, and the sign of the term appearing in the equation of motion (6.2).

The characterisation of nonlinear terms should be a prerequisite for the implementation of the method proposed in next section (and also for other methods). However, this method shows the capability of dealing simultaneously with several functions $g_j(t)$, in order to select only those providing a significant contribution to the nonlinear term, as can be seen in [96].

By moving the nonlinear term to the right-hand side of (6.2),

$$M\ddot{z}(t) + C_v \dot{z}(t) + Kz(t) = f(t) - \sum_{j=1}^{h} \mu_j L_j g_j(t) = f(t) + f_{NL}(t), \tag{6.3}$$

the original system may be viewed as subjected to the external forces $f(t)$ and the internal feedback forces due to nonlinearities $f_{NL}(t)$.

This concept, already used in [68] to derive the NIFO frequency domain method, is also on the basis of the present time domain identification method.

6.1.2. State-space model

The commonly adopted model for handling and solving the identification problem is not the one described by (6.3), but the so-called state-space model

$$\dot{x}(t) = A_c x(t) + B_c u(t)\,, \tag{6.4}$$

where:

- $x(t) = \begin{bmatrix} z(t) \\ \dot{z}(t) \end{bmatrix}$ is the state vector, $x(t) \in \mathbf{R}^{n\times 1}$

- $A_c = \begin{bmatrix} 0_{N\times N} & I_{N\times N} \\ -M^{-1}K & -M^{-1}C_v \end{bmatrix}$ is the system matrix, $A_c \in \mathbf{R}^{n\times n}$

- $B_c = \begin{bmatrix} 0_{N\times N} & 0_{N\times 1} & \dots & 0_{N\times 1} \\ M^{-1} & M^{-1}\mu_1 L_1 & \dots & M^{-1}\mu_h L_h \end{bmatrix}$ is the input matrix, $B_c \in \mathbf{R}^{n\times \ell}$

- $u(t) = \begin{bmatrix} f(t) \\ -g_1(t) \\ \vdots \\ -g_h(t) \end{bmatrix}$ is the input vector, $u(t) \in \mathbf{R}^{\ell\times 1}$

- ℓ is the number of inputs, $\ell \le (N+h)$. The unforced degrees of freedom, i.e. those for which $f = 0 \;\; \forall t$, are not included in the input vector; this clarifies the use of the previous inequality.

- n is the model order, $n = 2N$.

Since in applied mechanics some observable quantities are needed for reference, (6.4) is associated to a second equation related to the definition of the output vector $y(t) \in \mathbf{R}^{q\times 1}$:

$$y(t) = Cx(t) + Du(t)\,, \tag{6.5}$$

where, by assuming that the measurements concern displacements only,

$C = \begin{bmatrix} I_{N\times N} & 0_{N\times N} \end{bmatrix}$ is the output matrix, $C \in \mathbf{R}^{q\times n}$

$D = \begin{bmatrix} 0_{N\times N} & 0_{N\times 1} & \dots & 0_{N\times 1} \end{bmatrix}$ is the direct feedthrough matrix, $D \in \mathbf{R}^{q\times \ell}$.

The following ***continuous state-space model*** can then be defined:

$$\begin{cases} \dot{x}(t) = A_c x(t) + B_c u(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{6.6}$$

Discrete model

The continuous model may be converted into a discrete state-space model assuming zero-order hold for the input u. Premultiplying the first of (6.6) by the exponential matrix $e^{-A_c t}$ one obtains:

$$\mathrm{e}^{-A_c t}\dot{x}(t) - \mathrm{e}^{-A_c t}A_c x(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{-A_c t}x(t)\right) = \mathrm{e}^{-A_c t}B_c u(t)\,, \tag{6.7}$$

and integrating

$$x(t) = \mathrm{e}^{A_c t}x(0) + \int_0^t \mathrm{e}^{A_c(t-\tau)}B_c u(\tau)\mathrm{d}\tau\,. \tag{6.8}$$

In order to discretise the above expression a discrete state vector is defined as $x_r = x(r\Delta t)$, Δt being the sampling period and $r \in \mathbf{N}$. By assuming that the input u can be considered as constant during each timestep $t_r = r\Delta t$, the equation (6.8) becomes

$$x_r = \mathrm{e}^{A_c r\Delta t}x(0) + \int_0^{r\Delta t} \mathrm{e}^{A_c(r\Delta t-\tau)}B_c u(\tau)\mathrm{d}\tau\,,$$

and for the next timestep

$$x_{r+1} = \mathrm{e}^{A_c(r+1)\Delta t}x(0) + \int_0^{(r+1)\Delta t} \mathrm{e}^{A_c((r+1)\Delta t-\tau)}B_c u(\tau)\mathrm{d}\tau$$

$$= \mathrm{e}^{A_c\Delta t}\left(\mathrm{e}^{A_c r\Delta t}x(0) + \int_0^{r\Delta t} \mathrm{e}^{A_c(r\Delta t-\tau)}B_c u(\tau)\mathrm{d}\tau\right) + \int_{r\Delta t}^{(r+1)\Delta t} \mathrm{e}^{A_c(r\Delta t+\Delta t-\tau)}B_c u(\tau)\mathrm{d}\tau$$

$$= \mathrm{e}^{A_c\Delta t}x_r + \int_0^{\Delta t} \mathrm{e}^{A_c v}\mathrm{d}v\, B_c u_r\,.$$

The following discrete state-space model is finally obtained:

$$\begin{cases} x_{r+1} = Ax_r + Bu_r \\ y_r = Cx_r + Du_r \end{cases}, \tag{6.9}$$

where

$$A = \mathrm{e}^{A_c \Delta t} \in \mathbf{R}^{n \times n}$$

is the dynamical system matrix,

$$B = (\mathrm{e}^{A_c \Delta t} - I) A_c^{-1} B_c \in \mathbf{R}^{n \times \ell} \tag{6.10}$$

is the input distribution matrix, which represents the linear transformation by which the inputs influence the next state (the expression is valid for nonsingular A_c), $C \in \mathbf{R}^{q \times n}$ is the output distribution matrix, that describes how the internal state is transferred to the measurements y_r, and $D \in \mathbf{R}^{q \times \ell}$ is the matrix defining the algebraic relationships between input and output.

Model with noise

A more accurate model must accout for the unavoidable measurement noises and excitation sources that are not considered in the expression of $f(t)$; these errors are often occurring and can be modelled as further noise components. The following discrete-time deterministic-stochastic state-space model is then obtained:

$$\begin{cases} x_{r+1} = Ax_r + Bu_r + w_r \\ y_r = Cx_r + Du_r + v_r \end{cases}, \tag{6.11}$$

where $w_r \in \mathbf{R}^{n \times 1}$ is called process error and it is due to disturbances and model inaccuracy, while $v_r \in \mathbf{R}^{q \times 1}$ is called measurement error and it is due to sensors imprecision. These are unmeasurable vector signals and they are generally assumed to have zero mean value and to be modelled as white noises having covariance matrices:

$$E\left[\begin{pmatrix} w_\alpha \\ v_\alpha \end{pmatrix} \begin{pmatrix} w_\beta^T & v_\beta^T \end{pmatrix} \right] = \begin{bmatrix} Q & Z \\ Z^T & R \end{bmatrix} \delta_{\alpha\beta},$$

where E is the expected value operator and $\delta_{\alpha\beta}$ is the Kronecker operator.

6.1.3. System properties

Frequency domain

The ***transfer function*** is defined as the function relating the Fourier transforms of the input and output signals. Then, by taking the Fourier transform of (6.9) the frequency domain counterpart of the state-space model is obtained:

$$Y(\omega) = H_E(\omega)U(\omega),$$

where $\mathrm{i} = \sqrt{-1}$ and capital letters define the Fourier transforms, i.e. $Y(\omega) = \mathcal{F}[y(t)]$ and $U(\omega) = \mathcal{F}[u(t)]$. In this way the transfer function of the nonlinear system has been defined as:

$$H_E(\omega) = D + C(\mathrm{i}\omega I - A_c)^{-1} B_c . \tag{6.12}$$

One of the objectives of the identification method proposed in the following consists in extracting the transfer function $H(\omega)$ of the linear system (6.1) which is ***underlying*** to (6.2); by applying the Fourier transform to (6.1):

$$H(\omega) = \left(K + \mathrm{i}\omega C_v - \omega^2 M\right)^{-1}, \tag{6.13}$$

that is also defined as Frequency Response Function (FRF) of the system.

Similarity transformation

It is worth noticing that the state-space matrices in (6.11) can be obtained only within a similarity transformation. So there exists an invertible matrix $T \in \mathbf{R}^{n \times n}$ such that

$$\begin{aligned} A &= T\hat{A}T^{-1},\ B = T\hat{B} \\ C &= \hat{C}T^{-1},\ D = \hat{D} \end{aligned} \tag{6.14}$$

(the symbol ^ denotes estimated matrices). In practice it is not necessary to compute the matrix T (subjected to the choice of the algorithm) since it only establishes similarity relationships without adding some knowledge about new system properties. In fact, it is sufficient to extract only one set of matrices satisfying (6.11) and then moving to a similar system through the relationships (6.14).

For example, it is possible to demonstrate that the matrix defined in (6.12) does not depend upon T. In fact, by taking into account (6.10) and (6.14), one yields

$$\begin{aligned}\hat{B}_c &= \hat{A}_c\left(\hat{A}-I\right)^{-1}\hat{B} = \\ &= T^{-1}A_cT\left(T^{-1}AT-T^{-1}IT\right)^{-1}T^{-1}B = \\ &= T^{-1}A_cTT^{-1}\left(A-I\right)^{-1}TT^{-1}B = \\ &= T^{-1}A_c\left(A-I\right)^{-1}B\end{aligned}$$

and, substituting in (6.12),

$$\begin{aligned}\hat{H}_E(\omega) &= \hat{D}+\hat{C}\left(\mathrm{i}\omega I-\hat{A}_c\right)^{-1}\hat{B}_c = \\ &= D+CTT^{-1}\left(\mathrm{i}\omega I-A_c\right)^{-1}TT^{-1}A_c\left(A-I\right)^{-1}B = \\ &= D+C\left(\mathrm{i}\omega I-A_c\right)^{-1}B_c = \\ &= H_E(\omega)\end{aligned}$$

which is not dependent upon T.

Computation of modal parameters

The underlying linear system modal parameters (natural frequencies and damping factors) can be extracted from the eigenvalues of matrix A or from those of A_c.
The eigenvalues λ of matrix A are defined as ***discrete poles*** of the linear system, while the eigenvalues $\lambda^{(c)}$ of A_c are defined as ***continuous poles***. The relationship between them, due to the properties of the exponential matrix, is

$$\lambda = \exp(\lambda^{(c)}\Delta t), \tag{6.15}$$

where Δt is the sampling period.
By supposing to start from a matrix A that can be diagonalised:

$$A = \Phi\Lambda\Phi^{-1},$$

where $\Lambda = diag(\lambda_j) \in \mathbf{C}^{n\times n}$, $j = 1,...,n$ and $\Phi \in \mathbf{C}^{n\times n}$.
It is important to observe that the eigenvectors of A, constitued by the columns of Φ, coincide with those obtained from the dynamical matrix A_c of the continuous time model:

$$\begin{cases} A_c = \Phi_c\Lambda_c\Phi_c^{-1} \\ \Phi_c = \Phi \end{cases},$$

where $\Lambda_c = diag(\lambda_j^{(c)}) \in \mathbf{C}^{n\times n}$, $j = 1, ..., n$.

The eigenvalues $\lambda_j^{(c)}$ of A_c are obtained through the relationship (6.14):

$$\lambda_j^{(c)} = \frac{1}{\Delta t}\ln(\lambda_j).$$

Since matrices M, C_v and K are real valued, symmetric and positive definite, in case of underdamped modes the eigenvalues $\lambda_j^{(c)}$ can be proved to be complex conjugate pairs. These can be expressed, in case of proportional damping, through the relationship:

$$\lambda_j^{(c)} = -\zeta_j\omega_j \pm \mathrm{i}\omega_j\sqrt{1-\zeta_j^2} \tag{6.16}$$

where ω_j represents the ***natural frequency*** and ζ_j the ***damping factor***, related to the j-th system mode.

In the end it is important to notice that, due to the similarity transformation introduced in (6.14), the eigenvalues of A (and then the extracted modal parameters) are invariant with respect to different state-space bases and can then be directly obtained from matrix $\hat{A}$, which is estimated through the identification process.

6.2. Data-driven subspace method

6.2.1. Description

Among the efforts spent from the '90s in the framework of subspace identification, a theoretical overview must follow two milestones such as the books by Ljung [50] and by Van Overschee and De Moor [51]. The following exhaustive summary is taken from [70].

Starting from input and output vectors, a generic subspace method is capable of estimating (up to within a similarity transformation) the state-space matrices A, B, C and D of system (6.11) and the model order n, which is unknown in most of the applications.

Two main categories can be distinguished in the class of subspace methods: data-driven methods are well-established and are now described, while covariance-driven methods will be accurately exposed in Chapter 8.

A deterministic-stochastic state-space model is given as defined by (6.11), with s measurements of the input and of the output. In the data-driven approach the input data are gathered in a block Hankel matrix:

$$U_{0|2i-1} \overset{\text{def}}{=} \left[\begin{array}{cccc} u_0 & u_1 & \cdots & u_{J-1} \\ u_1 & u_2 & \cdots & u_J \\ \vdots & \vdots & \ddots & \vdots \\ u_{i-1} & u_i & \cdots & u_{i+J-2} \\ \hline u_i & u_{i+1} & \cdots & u_{i+J-1} \\ u_{i+1} & u_{i+2} & \cdots & u_{i+J} \\ \vdots & \vdots & \ddots & \vdots \\ u_{2i-1} & u_{2i} & \cdots & u_{2i+J-2} \end{array}\right] = \left[\begin{array}{c} U_p \\ \hline U_f \end{array}\right] \tag{6.17}$$

The subscript p denotes the "past", the subscript f denotes the "future" and the number of block rows i is a user defined index, that should be large enough with respect to the maximum order of the system to be identified [51]. The number of columns is typically equal to $J = s - 2i + 1$, which implies that all given data are used.

The output block Hankel matrices $Y_{0|2i-1}$, Y_p and Y_f are defined in a similar manner by replacing u with y in (6.17). Both input and output data may be then collected in the block Hankel matrix

$$W_p = \left[\begin{array}{c} U_p \\ \hline Y_p \end{array}\right].$$

System properties may be obtained by means of subspace identification algorithms through geometric manipulation of the row spaces of the above defined matrices. It is possible to decompose the matrix Y_f as linear combinations of the two non-orthogonal matrices U_f and W_p and of the orthogonal complement of U_f and W_p as follows:

$$Y_f = L_{U_f} U_f + L_{W_p} W_p + L_{U_f^{\perp}, W_p^{\perp}} \begin{pmatrix} U_f \\ W_p \end{pmatrix}^{\perp},$$

where the symbol $\perp$ means orthogonal complement. The reader is referred to [51] for the computation of these three terms and for their geometric interpretations.
In particular, for the subspace methods the matrix $L_{W_p} W_p$ is needed, which is called the oblique projection of the row space of Y_f along the row space of U_f on the row space of W_p and is indicated with the symbol $\mathcal{O}_i = Y_f /_{U_f} W_p$. Next step consists in performing the Singular Value Decomposition (SVD) of the following weighted oblique projection:

$$\mathcal{O}_i \Pi_{U_f{}^{\perp}} = U \Sigma V^T = \begin{bmatrix} U_n & U_0 \end{bmatrix} \begin{bmatrix} \Sigma_n & 0 \\ 0 & \Sigma_0 \end{bmatrix} \begin{bmatrix} V_n{}^T \\ V_0{}^T \end{bmatrix}, \tag{6.18}$$

where $\Pi_{U_f{}^{\perp}}$ is the projection on the orthogonal complement of the row space of the U_f matrix.
The model order n is determined by inspecting singular values and accordingly U_n, V_n and Σ_n are determined. By defining the following extended observability matrix

$$\Gamma_i \overset{\text{def}}{=} \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{i-1} \end{bmatrix},$$

an estimate may be obtained as $\hat{\Gamma}_i = U_n \Sigma_n{}^{1/2}$, from which the matrices A and C of system (6.11) can be obtained (the estimation of B and D represents a step of the procedure also discussed in [51]), up to within a similarity transformation.
As a conclusive remark, a wide class of subspace-based algorithms require different left and right user defined weighting matrices for the oblique projection $\mathcal{O}_i$ in (6.18); the choice of the weighting matrices affects only the similarity transformation, but this aspect is not relevant referring to the considerations of Section 6.1.3.

6.2.2. Implementation

A common feature in the implementation of all algorithms concerning the subspace methods is the following QR factorisation of a block Hankel matrix $\mathcal{H} \in \mathbf{R}^{J\times 2(\ell+q)i}$, constructed from all input and output measurements:

$$\mathcal{H} = \frac{1}{\sqrt{J}}\begin{bmatrix} U_{0|2i-1}^T & Y_{0|2i-1}^T \end{bmatrix} = \frac{1}{\sqrt{J}}\begin{bmatrix} U_{0|i-1}^T & U_{i|i}^T & U_{i+1|2i-1}^T & Y_{0|i-1}^T & Y_{i|i}^T & Y_{i+1|2i-1}^T \end{bmatrix} = QR \tag{6.19}$$

with $Q \in \mathbf{R}^{J\times 2(\ell+q)i}$ orthonormal ($QQ^T = I_{2(\ell+q)i}$) and $R \in \mathbf{R}^{2(\ell+q)i\times 2(\ell+q)i}$ upper triangular.

The book [51] includes a series of *Matlab*® functions, which are useful for applying the subspace methods. It is also shown that only the term R of this factorisation is needed in order to finally compute the system matrices:

$$R = \begin{bmatrix} R_{11}^T & R_{21}^T & R_{31}^T & R_{41}^T & R_{51}^T & R_{61}^T \\ 0 & R_{22}^T & R_{32}^T & R_{42}^T & R_{52}^T & R_{62}^T \\ 0 & 0 & R_{33}^T & R_{43}^T & R_{53}^T & R_{63}^T \\ 0 & 0 & 0 & R_{44}^T & R_{54}^T & R_{64}^T \\ 0 & 0 & 0 & 0 & R_{55}^T & R_{65}^T \\ 0 & 0 & 0 & 0 & 0 & R_{66}^T \end{bmatrix}.$$

These few guidelines are enough to define the following concept of memory limitation problems. Further and very exhaustive details about the implementation and the geometric interpretation of the so-obtained submatrices of R are given in [51].

6.2.3. Memory limitation problems

The considerations of this section refer to a *Windows*® operative system; by using a *Mac*®, the performances of the data-driven method (in terms of virtual memory space and data storing) should be different and better in general.

Assuming to work in a *Matlab*® environment, matrix R contained in (6.19) should easily be computed through the standard "qr" function, after constructing the block Hankel matrix $\mathcal{H} \in \mathbf{R}^{J\times 2(\ell+q)i}$.

This procedure is certainly valid and efficient for very small linear systems, because an accurate identification does not require the values of i and J to be so

large to fall into the problem described below (typically $J \sim 10^4$ and i does not exceed some tens). But in some cases, especially when the number of degrees of freedom is large and when a high level of noise implies using a large number of samples to avoid poor estimates, such a problem starts appearing even for linear systems.

Moreover, in order to apply subspace methods to nonlinear systems with satisfactory results, it is necessary to consider as many samples s as possible (so $J \approx s$ should be of the order of 10^5 or 10^6) and in particular to extend the index i to some hundreds, especially in presence of noisy measurements. The consequent problem consists in dealing with a matrix $\mathcal{H}$ which results to be too large to be stored nor factorised.

In fact, *Matlab*® stores data (each matrix element takes 8 Bytes) in the computer virtual memory, a space generally limited to 1.3-1.5 GigaBytes. Even if it is extended up to its maximum, this space cannot exceed 2 GigaBytes at the moment. Moreover, the virtual memory space taken through the QR factorisation is doubled and a small part of the virtual memory is always designed for operative system processes and for the graphical interface of *Matlab*® itself. For further details about *Matlab*® and virtual memory space, visit http://www.mathworks.com

Therefore, it is clear that the data-driven subspace method, which is based on the QR factorisation (6.19) undergoes severe limitations in its applicability, in particular as regards large MDOF systems (increasing q) or systems having many nonlinear terms (increasing ℓ).

For example, suppose that an identification procedure is carried out on a nonlinear system with four degrees of freedom; displacements are measured ($q = 4$) and the system has an external force and a nonlinear term ($\ell = 2$). Assume that matrix $\mathcal{H}$ is factorised by considering $s = 5 \times 10^4$ samples (so $J \approx s$) and by choosing $i = 300$ block rows. Only the storing of matrix $\mathcal{H}$ would need about 1.34 GigaBytes of virtual memory space, while its QR factorisation would take about 2.68 GigaBytes. As it can be directly verified, the space is not enough and *Matlab*® generates an error message.

A conclusive remark is given about trying to perform the QR factorisation outside the *Matlab*® environment, for example in *C*® or in *Fortran*®. Although data storing and managing are more autonomous and the virtual memory space can be slightly extended, the problem described above and the consequent failure in performing the QR factorisation can not be avoided anyway.

6.3. Nonlinear identification

The nonlinear identification procedure is based on the computation of system parameters. They are carried out from matrix $H_E(\omega)$ defined in (6.12), (or from other invariant matrices), once the state space matrices $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$ have been estimated by a subspace method in the time domain. In fact, system parameters (included in M, C_v, K, and μ_j) are contained in the invariant matrix $H_E(\omega)$ here called ***extended*** Frequency Response Function (FRF) matrix, because it also includes nonlinear terms in the Multiple Input Multiple Output (MIMO) model.

The implementation of system parameter computation is introduced in the following, according to whether displacement [70] or acceleration [97] measurements are available.

6.3.1. Measured displacements

Let us denote

$$G = \mathrm{i}\omega I - A_c = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \quad \text{and} \quad P = (\mathrm{i}\omega I - A_c)^{-1} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}. \tag{6.20}$$

Taking into account the definition of matrices appearing in (6.6), matrix $H_E(\omega)$ becomes

$$\begin{aligned} H_E(\omega) &= 0_{N\times(N+h)} + \begin{bmatrix} I_{N\times N} & 0_{N\times N} \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} 0_{N\times N} & 0_{N\times 1} & \dots & 0_{N\times 1} \\ M^{-1} & M^{-1}\mu_1 L_1 & \dots & M^{-1}\mu_h L_h \end{bmatrix} \\ &= P_{12} M^{-1} \begin{bmatrix} I & \mu_1 L_1 & \dots & \mu_h L_h \end{bmatrix}. \end{aligned} \tag{6.21}$$

Recalling the block matrix inversion rule

$$P_{12} = -{G_{11}}^{-1} G_{12} {S_{G_{11}}}^{-1}, \tag{6.22}$$

where

$${S_{G_{11}}}^{-1} = \left(G_{22} - G_{21}{G_{11}}^{-1}G_{12}\right)$$

is called the *Shur complement* of G_{11}, one yields:

$$\begin{aligned} P_{12} &= \frac{1}{\mathrm{i}\omega}\left(M^{-1}C_v + \mathrm{i}\omega I + M^{-1}K\frac{1}{\mathrm{i}\omega}\right)^{-1} = \left(M^{-1}K + \mathrm{i}\omega M^{-1}C_v - \omega^2 I\right)^{-1} \\ &= \left(K + \mathrm{i}\omega C_v - \omega^2 M\right)^{-1} M. \end{aligned} \tag{6.23}$$

Since the underlying linear system ***receptance*** matrix is defined by (6.13), equation (6.21) finally becomes

$$H_E(\omega) = \left[H \quad H\mu_1 L_1 \quad \ldots \quad H\mu_h L_h \right]. \tag{6.24}$$

In the particular case $\omega = 0$:

$$H_E(0) = D - CA_c^{-1}B_c = \left[K^{-1} \quad K^{-1}\mu_1 L_1 \quad \ldots \quad K^{-1}\mu_h L_h \right]. \tag{6.25}$$

This latter expression will be adopted in the applications, more precisely for estimating the h nonlinear coefficients μ_j.

6.3.2. Measured accelerations

The computation of parameters is of interest, in particular for experimental applications, when the available measurements come from accelerometers. This is a significant difference with respect to the case of measured displacements, which are more easily considered for numerical simulations.

In case accelerations are measured, matrices C and D appearing in the state-space model (6.6) are modified and can be written as:

$$C = \left[-M^{-1}K \quad -M^{-1}C_v \right]$$

$$D = \left[M^{-1} \quad M^{-1}\mu_1 L_1 \quad \ldots \quad M^{-1}\mu_h L_h \right]$$

In this way, by using (6.20), matrix $H_E(\omega)$ can be written as:

$$
\begin{aligned}
H_E(\omega) &= \begin{bmatrix} M^{-1} & M^{-1}\mu_1 L_1 & \dots & M^{-1}\mu_h L_h \end{bmatrix} \\
&\quad - \begin{bmatrix} M^{-1}K & M^{-1}C_v \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} 0_{N\times N} & 0_{N\times 1} & \dots & 0_{N\times 1} \\ M^{-1} & M^{-1}\mu_1 L_1 & \dots & M^{-1}\mu_h L_h \end{bmatrix} \\
&= \begin{bmatrix} M^{-1} & M^{-1}\mu_1 L_1 & \dots & M^{-1}\mu_h L_h \end{bmatrix} \\
&\quad - \left(M^{-1}K P_{12} + M^{-1}C_v P_{22} \right) M^{-1} \begin{bmatrix} I & \mu_1 L_1 & \dots & \mu_h L_h \end{bmatrix}.
\end{aligned}
\tag{6.26}
$$

As in the previous case, by using (6.22) the equation (6.23) is obtained. Moreover, in a similar way

$$P_{22} = \mathrm{i}\omega\left(K + \mathrm{i}\omega C_v - \omega^2 M\right)^{-1} M$$

can be derived.

Then equation (6.26) can be rewritten as:

$$
\begin{aligned}
H_E(\omega) &= \left(M^{-1} - M^{-1}\left(K + \mathrm{i}\omega C_v\right)\left(K + \mathrm{i}\omega C_v - \omega^2 M\right)^{-1} \right) \begin{bmatrix} I & \mu_1 L_1 & \dots & \mu_h L_h \end{bmatrix} \\
&= \Big(M^{-1}\left(K + \mathrm{i}\omega C_v - \omega^2 M\right)\left(K + \mathrm{i}\omega C_v - \omega^2 M\right)^{-1} \\
&\quad - M^{-1}\left(K + \mathrm{i}\omega C_v\right)\left(K + \mathrm{i}\omega C_v - \omega^2 M\right)^{-1} \Big) \begin{bmatrix} I & \mu_1 L_1 & \dots & \mu_h L_h \end{bmatrix} \\
&= -\omega^2 \left(K + \mathrm{i}\omega C_v - \omega^2 M\right)^{-1} \begin{bmatrix} I & \mu_1 L_1 & \dots & \mu_h L_h \end{bmatrix}.
\end{aligned}
\tag{6.27}
$$

Since the underlying linear system ***inertance*** matrix is defined as

$$H(\omega) = -\omega^2 \left(K + \mathrm{i}\omega C_v - \omega^2 M\right)^{-1},$$

equation (6.27) finally becomes

$$H_E(\omega) = \begin{bmatrix} H & H\mu_1 L_1 & \dots & H\mu_h L_h \end{bmatrix}.$$

It is worth noticing that the particular value of $H_E(\omega)$ at $\omega = 0$ cannot be considered to estimate the h nonlinear coefficients μ_j, since $H_E(0) = 0$. In such a case, the nonlinear coefficients can be estimated by computing a spectral mean over a frequency band of interest [98].

6.4. Linear Time-Varying systems: the ST-SSI method

In addition to the identification of nonlinear systems described in Section 6.3, a procedure for the identification of linear time-varying systems called Short-Time Stochastic Subspace Identification (ST-SSI) [99, 100] is briefly introduced in the following.

The idea is to divide the signal in many parts and to consider the system as time-invariant in that time interval: the process is called frozen technique.

If the output data are measured at discrete times with a sampling interval Δt and the input is a discrete signal characterised by a zero-order hold between consecutive sample points, the corresponding discrete-time state-space representation of a general linear time-varying system at a time instant $t = r\Delta t$ is:

$$\begin{cases} x(r+1) = A(r)x(r) + B(r)u(r) + w(r) \\ y(r) = C(r)x(r) + D(r)u(r) + v(r) \end{cases}$$

where *A(r)* and *B(r)* are not constant and in general their closed forms are unknown [101]; $x(t)$ is the state vector, $u(t)$ the input vector and $y(t)$ the output vector; $w(t)$ and $v(t)$ are the process and measurement error, respectively.

The frozen technique considers the state matrices as constant during each time step so that

$$\begin{cases} x(r+1) = Ax(r) + Bu(r) + w(r) \\ y(r) = Cx(r) + Du(r) + v(r) \end{cases}$$

The complete time record is splitted into time windows (frozen system), which can be almost completely overlapping except for a sampling period τ (or its multiples), as indicated in Fig. 6.1.

The natural frequencies are extracted by calculating the eigenvalues of the identified matrix A in every window. The length of the window L_f is usually chosen as short as possible, in order to consider a brief time interval and hopefully a time-invariant system. This is the main reason why the data-driven approach (Section 6.2) is preferred with respect to the covariance-driven one (Chapter 8), which needs more samples to obtain accurate results.

For these reasons, the ST-SSI method can be used to analyse non-stationary systems that are regarded as time-invariant in each user-defined short-time interval, providing they change "slowly" with time. The term "slowly" here mainly means that their time variations are by far longer than their dynamics, i.e. the frequency ranges are well apart.

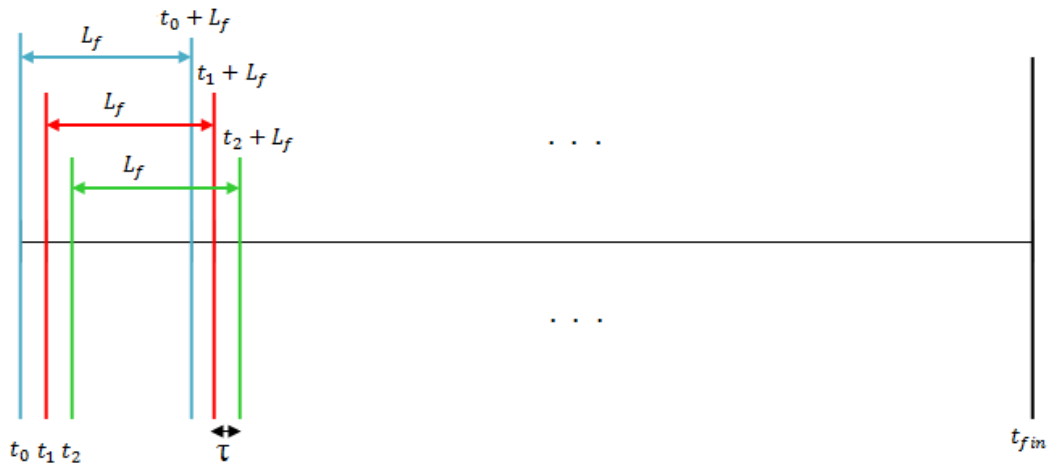


Figure 6.1. Choice of the windows in the ST-SSI method.

Chapter 7

An alternative data-driven implementation

This chapter starts from the memory limitation problems introduced in Section 6.2.3. In order to find a way for overcoming these problems, the NSI method is enforced by the development of a new algorithm to compute the QR factorisation in a *Matlab*® environment, in those cases in which the data matrix is too large to be stored nor factorised. This new algorithm, which exploits some useful features of the Householder transformations, allows the NSI method to reach more accurate results in the parameter estimation [102].

As a global overview of the data-driven subspace method, its limitations and the application of the novel algorithm, the method is applied to an oscillator described by the Duffing equation, with different types of excitation including random forces, which are demonstrated to be very suitable for the identification process. In order to present all the possibilities offered by an accurate method in identifying a nonlinear system and its dynamics, this numerical application also focuses on sudden nonlinear transitions between stable attractors (jumps) caused by nonlinear hysteresis phenomena, which are associated to essentially nonlinear dynamics caused by bifurcations [62].

7.1. Householder transformations

7.1.1. Definition

In this section some concepts, exploited in Section 7.2 to conceive a new useful algorithm to compute the QR factorisation of a matrix, are presented. For a detailed overview of Householder transformations (also known as elementary reflectors), see [85]. In particular, the algorithms presented below are a revised form of those contained in [85, pp. 40-41].
An elementary reflection is a matrix of the form

$$U = I - 2uu^T$$

where u is a vector of length $\| u \|_2 = \sqrt{u^T u} = 1$.
Linear transformations associated to these matrices are also known as ***Householder transformations***. Householder matrices U are symmetric ($U^T = U$), orthogonal ($U^T U = I$) and involutive ($U^2 = I$).
The geometrical interpretation of the transformation $y = Ux$ can now be examined. Let x be a generic nonzero vector, having two components named v and w, so $x = v + w$; v is parallel to vector u while w is orthogonal, so that $v = \alpha \cdot u$ (with α scalar) and $u^T w = 0$. The following can be computed:

$$Ux = Uv + Uw = v - 2uu^T v + w - 2uu^T w = v + w - 2uu^T v.$$

Since $v = \alpha \cdot u$, we have $u(u^T v) = \alpha \cdot u = v$ and

$$Ux = -v + w.$$

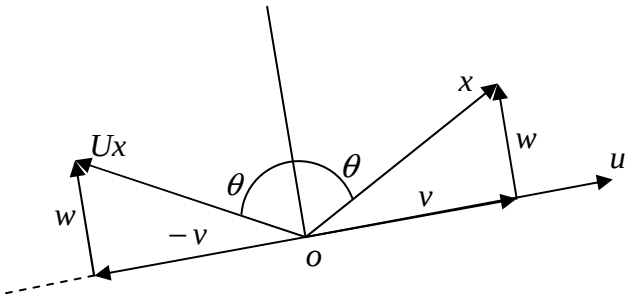


Figure 7.1. Reflection of vector x with respect to the axis orthogonal to vector u and passing through the origin.

Then, from a geometrical point of view matrix U reflects vector x with respect to the axis orthogonal to vector u and passing through the origin, as shown in Fig. 7.1.

Next theorem shows how to exploit elementary reflections for "inserting zeros" into a vector.

Theorem: Given a generic nonzero vector x, the Householder transformation

$$U = I - \beta u u^T \tag{7.1}$$

with $u = x + \sigma \cdot e_1$, $e_1 = [1,0,\ldots,0]^T$, $\sigma = \pm \| x \|_2$ and $\beta = 2/\| u \|_2^2$ yields the following relation:

$$Ux = -\sigma \cdot e_1 . \tag{7.2}$$

The proof is not hard: by first considering vector

$$Ux = \left(I - \beta u u^T\right)x = x - \beta u^T x u$$

and then the quantities (x_1 is the first entry of vector x)

$$u^T x = (x + \sigma \cdot e_1)^T x = (x^T + \sigma \cdot e_1^T)x =$$

$$x^T x + \sigma \cdot e_1^T x = \sigma^2 + \sigma \cdot x_1 = \sigma(\sigma + x_1)$$

and

$$\beta = \frac{2}{u^T u} = \frac{2}{(x + \sigma \cdot e_1)^T (x + \sigma \cdot e_1)} =$$

$$\frac{2}{(x^T x + 2\sigma \cdot x_1 + \sigma^2)} = \frac{1}{\sigma(\sigma + x_1)},$$

we can obtain:

$$Ux = x - u = x - (x + \sigma \cdot e_1) = -\sigma \cdot e_1 .$$

It can be observed that the couple (u, β), formed of $n+1$ real numbers, is sufficient to uniquely determine matrix U, having n^2 elements. Thus, given a vector $x = [\xi_1, \xi_2, \ldots, \xi_n]^T$, an efficient algorithm providing the quantities u (which is overwritten to x) and β (and also σ) can be written:

Algorithm 1

1: $\eta \leftarrow \max\{|\xi_i|,\ i=1,\ldots,n\}$

2: $\sigma \leftarrow 0$

3: **cycle 1**: $i=1,\ldots n$

4: if $|\xi_i| \geq \eta\sqrt{\text{eps}}$ then $\sigma \leftarrow \sigma + (\xi_i/\eta)^2$

5: **end of cycle 1**

6: $\sigma = \text{sgn}(\xi_1)\eta\sqrt{\sigma}$

7: $\xi_1 \leftarrow \xi_1 + \sigma$

8: $\beta \leftarrow 1/(\sigma \cdot \xi_1)$

Note that *eps* stands for the lowest possible machine number, and that this algorithm avoids possible phenomena of *overflow*, *underflow* and numerical cancellation.

The couple (u,β) determined through the above algorithm is sufficient to construct products of the form

$$UA = U[a_1, a_2, \ldots, a_n] = [Ua_1, Ua_2, \ldots, Ua_n], \tag{7.3}$$

in fact we have

$$Ua_i = \left(I - \beta uu^T\right)a_i = a_i - \left(\beta u^T a_i\right)u .$$

This remark is important since matrices U will be adopted for operations like (7.3), or products of the form $AU = (UA^T)^T$.

Given the two vectors $u = [v_1, v_2, \ldots, v_n]^T$ and $a = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T$, and the number β, the substitution of a with vector Ua can be computed in the following way:

Algorithm 2

1: $\tau \leftarrow \beta\sum_{i=1}^{n} v_i\alpha_i$

2: $\alpha_i \leftarrow \alpha_i - \tau \cdot v_i, \quad i = 1,\ldots,n$

Observe that a general product of an n order matrix by a vector requires n^2 flops (floating point operations), while only $2n$ flops are required for the Ua transformation.

7.1.2. Application: the QR factorisation

In this section the action of elementary reflections is generalised, in order to derive useful matrix transformations. In particular, the following problem can initially be faced: given a vector $x = [x_1, x_2, \ldots, x_n]^T$, the possibility of determining an elementary reflection $U_k \equiv U_k^{(n)}$ such that

$$U_k x = [\bar{x}_1, \ldots, \bar{x}_k, 0, \ldots, 0]^T$$

can be investigated.
The answer is positive. In fact, by remembering the characterisation of the elementary reflection $U \equiv U_1^{(n)}$ as defined by the Theorem above, U_k can be built as:

$$U_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & U_1^{(n+1-k)} \end{bmatrix} = I - \beta_k u_k u_k^T$$

$$u_k = \begin{bmatrix} 0 \\ u_k' \end{bmatrix}, \quad u_k' \in R^{n+1-k}, \quad \beta_k = \beta_k' = \frac{2}{\| u_k' \|_2^2},$$

where I_{k-1} is the identity matrix of order $(k-1)$ and $U_1^{(n+1-k)} = I_{n+1-k} - \beta_k' u_k' (u_k')^T$ is the elementary reflection of order $(n+1-k)$, defined by the relationship

$$U_1^{(n+1-k)} [x_k, x_{k+1}, \ldots, x_n]^T = -\sigma_k \cdot e_1.$$

The proposed matrix U_k has also the advantage of not modifying the first $k-1$ components of vector x; this means that only the k-th component of the new vector $U_k x$ is needed. In the end it can be observed that, once the integer k is

fixed, matrix $U_k \in \mathbf{R}^{n\times n}$ is univocally defined by the couple (u_k', β_k'), with $u_k' \in \mathbf{R}^{n+1-k}$; moreover,

$$U_k a = \begin{bmatrix} I_{k-1} & 0 \\ 0 & U_1^{(n+1-k)} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ U_1^{(n+1-k)} a_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 - \beta_k' (u_k')^T a_2 u_k' \end{bmatrix}.$$

The possibility of finding elementary reflections U_k with the properties described above can be exploited for reducing a generic matrix A to the upper triangular form. In fact, given a matrix $A \in \mathbf{R}^{n\times n}$, $n-1$ elementary reflectors $U_1, U_2, \ldots, U_{n-1}$ can be built such that the new matrix

$$U_{n-1} \ldots U_2 U_1 A = Q^T A = R \tag{7.4}$$

is upper triangular; note the orthogonality of

$$Q = [U_{n-1} \ldots U_2 U_1]^T = [U_1 U_2 \ldots U_{n-1}],$$

which is a product of orthogonal matrices. From (7.4), the ***QR factorisation*** is then defined as

$$A = QR.$$

As a final observation, the QR factorisation can be computed even if matrix A is rectangular $m \times n$; in this case $A = QR$ with $Q \in \mathbf{R}^{m\times m}$ and $R \in \mathbf{R}^{m\times n}$, and the factorisation is attained with $r = \min\{m-1, n\}$ elementary reflectors $U_1, U_2, \ldots, U_r$.

7.2. New algorithm

As explained in Section 6.2.3, memory limitation problems affect the QR factorisation contained in the data-driven subspace methods. It is then necessary to conceive a new algorithm to compute the QR factorisation. This algorithm is based on *Matlab*® commands “save” and “load”, which allow to save and load variables directly from the hard disk, and the command “clear”, useful to clean virtual memory.

Moreover, it is observed that the development of this new procedure exploits the particular structure of the matrix $\mathcal{H}$ as defined in (6.19) and the useful features of

Householder transformations: in particular, from now on, the Algorithms 1 and 2 reported in Section 7.1.1 will be considered.
The new algorithm is described in the following and a flow chart representation is given in Fig. 7.2:

1. Load measured data y, representing the q system outputs, and the values of the external force f; compute from these data the vector u of the ℓ system inputs.
2. Choose the number of samples s for the identification procedure, and the number of block rows i; this choice determinates the number of rows and columns of matrix $\mathcal{H}$, respectively $J = s - 2i + 1$ and $d = 2(q + \ell)i$.
3. Start a **Cycle 1,** $k = 1,\ldots d$; define δ as the k-th column of matrix $\mathcal{H}$. δ is constructed by using the input (if it is a column of submatrix $U^T_{0|2i-1}$) or output (if it is a column of submatrix $Y^T_{0|2i-1}$) data, as defined in (6.19).
4. Start a **Cycle 2**, $g = 1,\ldots,k-1$; for each iteration g:

   4.1. "load" from the hard disk vector $Q_g = [v_g,\ldots,v_J]^T$;

   4.2. execute, on part $\tilde{\delta} = [\delta_g,\ldots,\delta_J]^T$ of vector δ, the transformations defined in Alg. 2, also using number β_g; vector $\bar{\delta}$ is obtained;

   4.3. "clear" vector Q_g from virtual memory.

   End of Cycle 2.
5. Subdivide vector $\bar{\delta}$ into two vectors $\gamma = [\bar{\delta}_1,\ldots,\bar{\delta}_{k-1}]^T$ and $\xi = [\bar{\delta}_k,\ldots,\bar{\delta}_J]^T$. Make a copy ψ of vector ξ.
6. Apply Alg. 1 to vector ψ, which becomes the new $Q_k = [v_k,\ldots,v_J]^T$, obtaining also number β_k.
7. Execute, on vector ξ, the transformations defined in Alg. 2, in order to obtain the new vector $\bar{\xi} = [\bar{\xi}_1,0,\ldots,0]$.
8. Attain the k-th column of matrix R, denoted here as R_k:

   8.1. construct vector $\tilde{R} = [\gamma \quad \bar{\xi}]^T \in \mathbf{R}^J$;

8.2. truncate vector $\tilde{R}$, by eliminating all unnecessary zeros and keeping only the first d elements, in order to obtain $R_k \in \mathbf{R}^d$.

9. "save" vectors Q_k and R_k on the hard disk, and "clear" them from the virtual memory.

   End of Cycle 1.

10. Reconstruct matrix R, by loading ("load") the d columns R_k from the hard disk.

At the end of the algorithm, all saved vectors Q_k and R_k (and β also) will be deleted from the hard disk.

Note (referring in particular to step 3 of the above algorithm) that in this way it is not necessary to store the entire matrix $\mathcal{H}$, and the already discussed memory problems can be avoided. It is indeed sufficient to construct and factorise a new column for each iteration k of Cycle 1.

As a final consideration, it should be observed that this new algorithm does not present any limitations about the choice of index i and the number of samples s to be considered in the NSI procedure. The only limitation may be represented by a larger amount of time requested for the computation of matrix R.

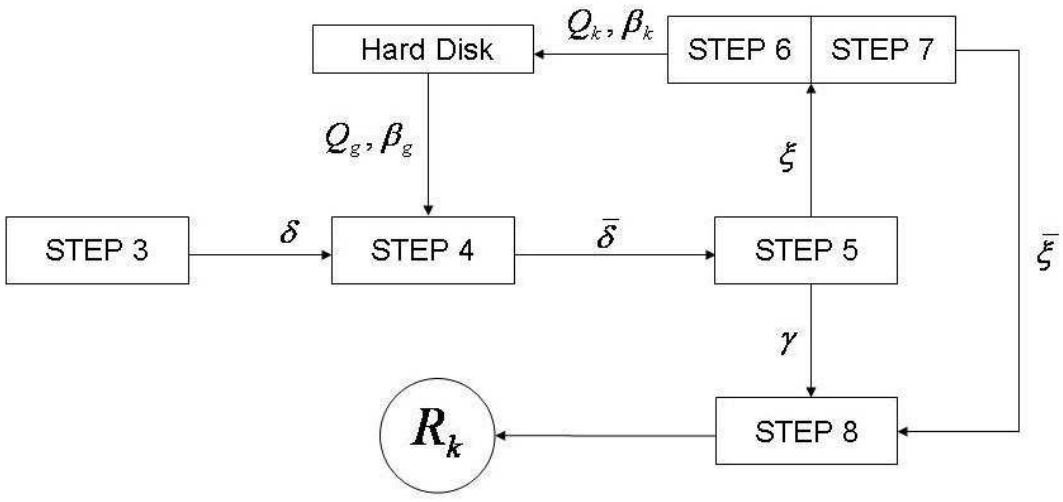


Figure 7.2. Flow chart representation of the new algorithm, from step 3 to step 8.

7.3. Numerical example

In this section the NSI method is applied to a Duffing oscillator, which has been studied for many years as representative of many nonlinear systems [103]. This system can be considered in order to simply describe the sudden transitions between co-existing stable branches of solutions. For this type of system there are frequencies at which the vibration suddenly jumps-up or down, when it is excited harmonically with slowly changing frequency.

One of the main topics about the study of the Duffing oscillator consists in searching for analytical expressions of the jump frequencies and the amplitudes of vibration at these frequencies. For example, in [104, 105] these points are computed by using the harmonic balance method, while in [106] the minimum excitation force required for the jump phenomenon to appear is determined, by using a method based on the elimination theory of polynomials. A more recent paper [107] provides a full set of expressions determined by using the harmonic balance approach, as a link between the earlier analytical work and the later numerical studies.

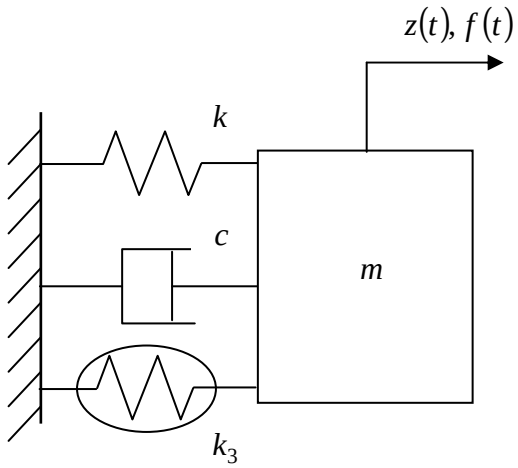


Figure 7.3. The nonlinear system described by the Duffing equation.

Table 7.1. System parameters.

m (kg)	k (N/m)	c (Ns/m)	k_3 (N/m^3)
1.3	800	1.3	1.5×10^6

7.3.1. Identification

Consider the SDOF system with cubic hardening stiffness depicted in Fig. 7.3, whose motion is described by the following Duffing equation

$$m\ddot{z}(t)+c\dot{z}(t)+kz(t)+k_3z^3(t)=f(t) \tag{7.5}$$

with system parameters summarized in Table 7.1. The strength, the type and the location of the nonlinearity are defined respectively by the three scalar quantities $\mu_1=k_3$, $g_1(t)=-z^3(t)$ and obviously $L_1=1$. The system is excited by two different types of force:

- **Case 1**. A linearly varying frequency sweep (of amplitude $A=1$) between 3 and 6 Hz, applied for an upward (Case 1up) and a downward (Case 1down) frequency sweep.

- **Case 2**. A zero-mean Gaussian random input whose r.m.s. is 20 N, selected so that the r.m.s. of the nonlinear force is equal to 67% of the corresponding linear stiffness force.

A fourth order Runge-Kutta numerical integration (with a time step $\Delta t=10^{-3}$ s) of the equation of motion has been performed and a total number of $s=10^5$ samples has been generated (so $t_{fin}=100$ s) and then corrupted by adding a zero-mean Gaussian noise (1% of the r.m.s. value of the output).
The invariant matrix $H_E(\omega)$ can be easily computed for $\omega=0$, as in (6.25):

$$H_E(0)=D-CA_c^{-1}B_c=\begin{bmatrix}0 & 0\end{bmatrix}-\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}-\frac{c}{k} & -\frac{m}{k}\\ 1 & 0\end{bmatrix}\begin{bmatrix}0 & 0\\ \frac{1}{m} & \frac{k_3}{m}\end{bmatrix}=\begin{bmatrix}\frac{1}{k} & \frac{k_3}{k}\end{bmatrix} \tag{7.6}$$

From the eigenvalues of the system matrix A_c it is possible to obtain [55] estimates for the angular frequency ω_n of the undamped system and for the

damping factor ζ, so that all system parameters can be estimated from (7.6) and from the following relationships:

$$\omega_n = \sqrt{\frac{k}{m}} \qquad \text{and} \qquad \zeta = \frac{c}{c_{crit}} = \frac{c}{2\sqrt{km}}. \tag{7.7}$$

It is observed here that in each of the identification procedures performed, the model order $n = 2$ is determined by inspecting a singular value plot (with $i = 60$ block rows), as shown in [70].

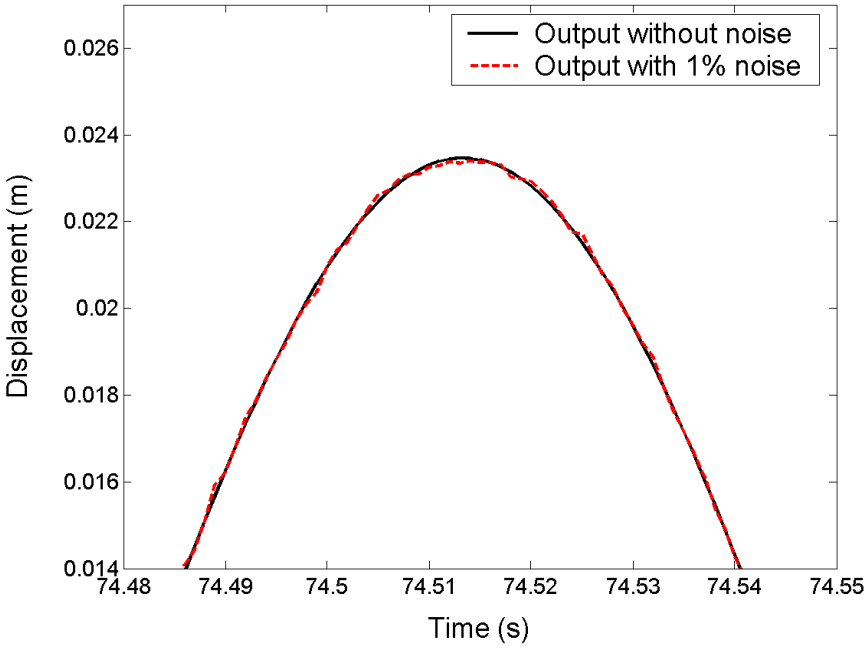
The identification results for all system parameters are presented in Table 7.2: the best estimates are obtained by applying a random input. In fact, for Case 1, it should be observed that the added noise is related to the r.m.s. of the entire time history, which is non-stationary; so, samples corresponding to small displacements are more deeply corrupted by noise and are consequently counterproductive for the identification procedure. This is shown in Fig. 7.4 for Case 1up, in which this concept is more evident because the system reaches higher values of response amplitudes (and then a higher r.m.s. of the time histories).

A slightly better result for Case 1 can be obtained by considering k_3 as depending on ω: for each ω, matrix $H_E(\omega)$ defined in (6.12) simply reduces to a vector $\bar{h}_E$ with two elements as in (7.6), and it is possible to compute $k_3 = \bar{h}_E(2)/\bar{h}_E(1)$.

Table 7.2. Identification results: percentage error (100 • | estimated – actual | / actual).

	m	k	c	k_3
1up) Upward sweep	4.63	4.01	4.04	5.86
1down) Downward sweep	1.71	1.30	2.64	3.97
2) Random	0.13	0.54	0.73	0.73

(a)



(b)

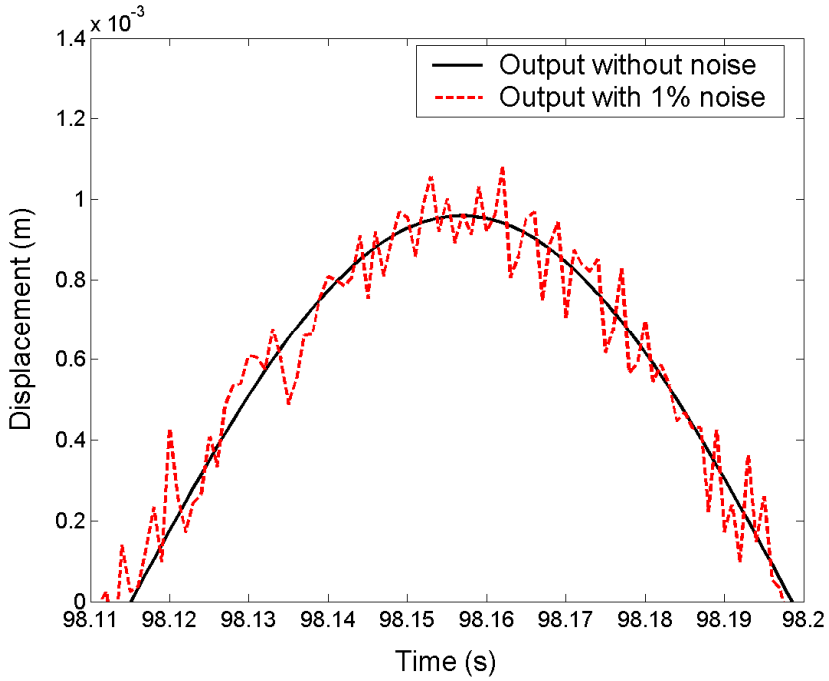


Figure 7.4. Effect of noise corruption for Case 1up. The r.m.s. of the entire time history is 0.0088 m. (a) Magnification just before the jump-down (large amplitudes). (b) Magnification after the jump (small amplitudes).

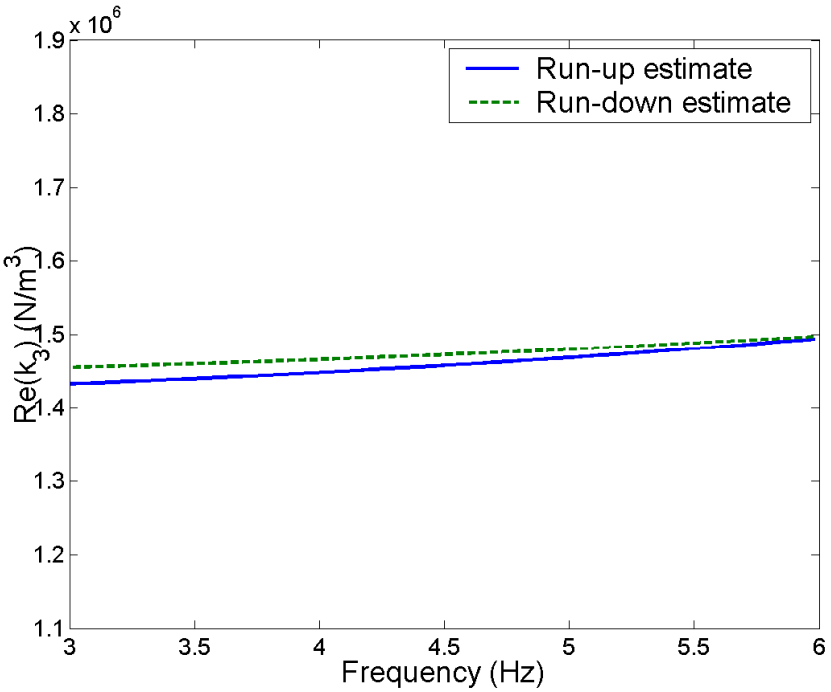
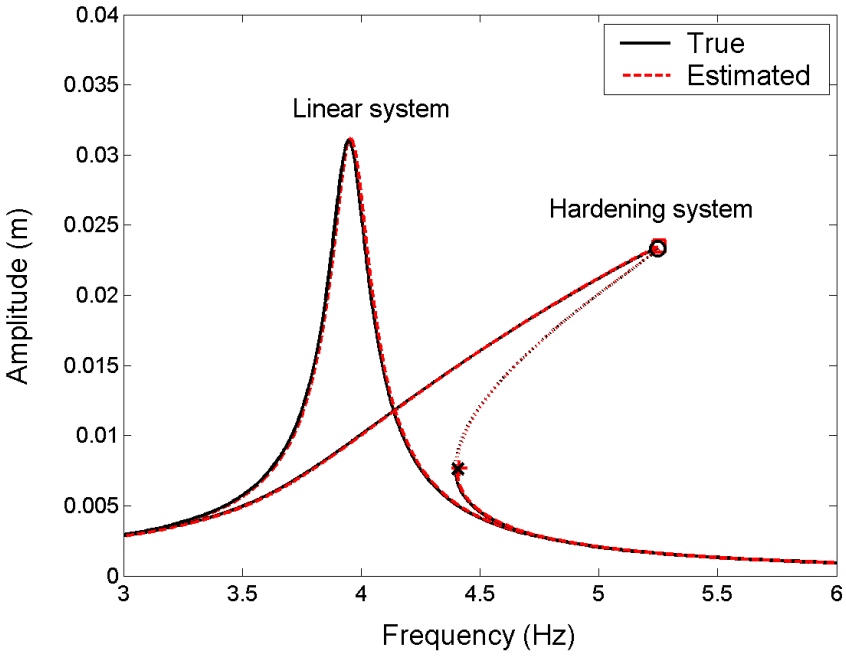


Figure 7.5. Real part of the estimated nonlinear coefficient k_3, in the frequency range considered.

The estimated coefficient of the nonlinear term is frequency dependent and complex, albeit its imaginary part is some orders of magnitude smaller than the real part. A single value can be obtained by performing a spectral mean in the frequency range from 3 to 6 Hz (Fig. 7.5). In this way, the percentage errors related to the k_3 estimates become 2.74 for Case 1up and 1.78 for Case 1down. Note that this procedure is not applicable to get a spectral mean for k, because for $\omega > 0$ vector $\bar{h}_E$ is not defined as in (7.6).

In Fig. 7.6a the true Frequency Response Functions (FRFs) of the nonlinear and underlying linear system are shown in comparison with the NSI estimates, computed from the identified system parameters in Case 2. As a consequence of the results reported in Table 7.2, the curves are almost overlaid: an excellent agreement can be observed, even in estimating the jump-up and jump-down frequencies and responses. The values for the jump-down and the jump-up (Fig. 7.6b) have been obtained from the approximate expressions derived in [107]: the approximation of the true jump is obtained with the real system parameters of Table 7.1, while the approximation of the estimated jump is obtained with the NSI estimates of Case 2.

(a)



(b)

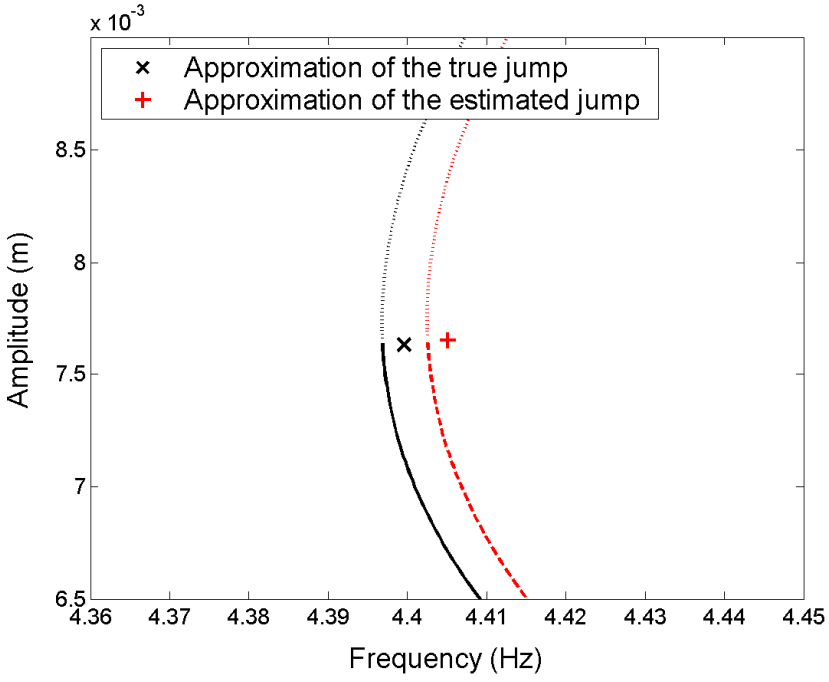


Figure 7.6. (a) Frequency response curves. The crosses and the circles denote the responses at the jump-up and jump-down frequencies, respectively. The dotted lines denote unstable solutions. (b) Magnification near the jump-up.

7.3.2. Output prediction

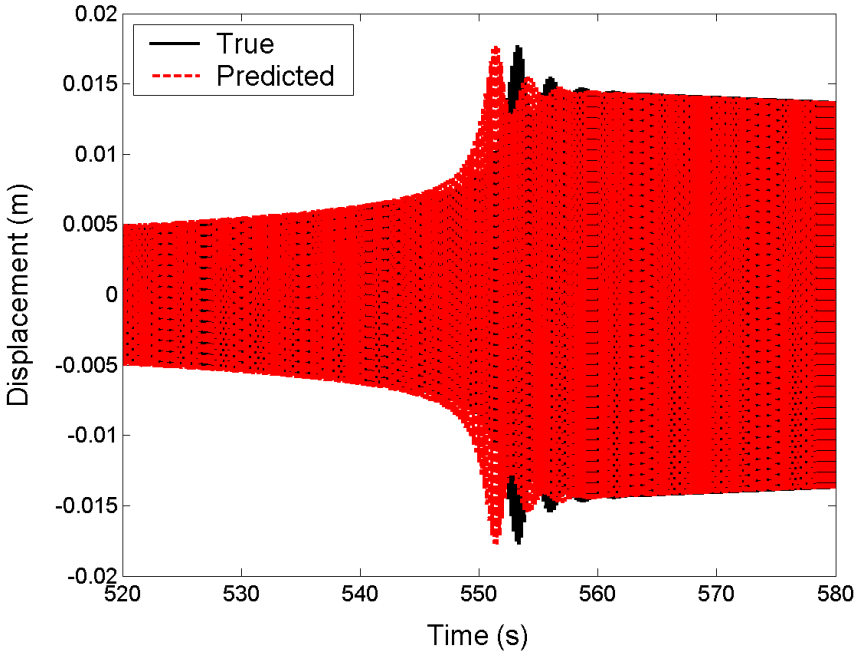
The NSI method presented in Section 6.2 is also attractive for its predictive capability. In fact, once the system matrices A, B, C and D in (6.9) have been estimated, it is possible to predict the system behaviour when it is subject to a different type of excitation.

It is important to remark that recent methods such as CRP [66, 108] and NIFO [68] would require a second step to perform output prediction in a general case of MDOF systems. In fact, these methods only produce estimates of the underlying linear FRFs and of nonlinear coefficients. On the contrary, the NSI capability of predicting the output is intrinsic in its formulation, since a state-space model is used. In other words, system parameter estimation is not strictly necessary and this represents a great advantage of NSI in case of MDOF systems. However, for simplicity's sake, in section a SDOF numerical example is considered so estimating system parameters out of state-space matrices is both possible and easy to perform.

Starting from the best estimates of system parameters, obtained through the Case 2 identification procedure, it is possible to generate new time histories considering the system as excited by the frequency sweeps described in Case 1. Now the numerical integration has been performed for $t_{fin} = 1000\,\mathrm{s}$, in order to have a slower frequency sweep and to obtain a more accurate representation of jump phenomena.

In Fig. 7.7 the results are shown, in terms of a comparison between the true (i.e.: system parameters as in Table 7.1) and the predicted (i.e.: identified system parameters) time histories, for the Case 1down. In Fig. 7.7a it can be observed that the predicted jump-up occurs at a higher frequency (at a lower time instant in the downward sweep), as expected from the FRFs zoom shown in Fig. 7.6b. After the jump-up, this slight shift has no longer effect on the prediction: as shown in Fig. 7.7b, the true and the predicted output are almost overlaid just a few seconds after the jump. Notice the high global level of accuracy of the prediction results, albeit system parameters have been estimated starting from a time history corrupted by measurement noise.

(a)



(b)

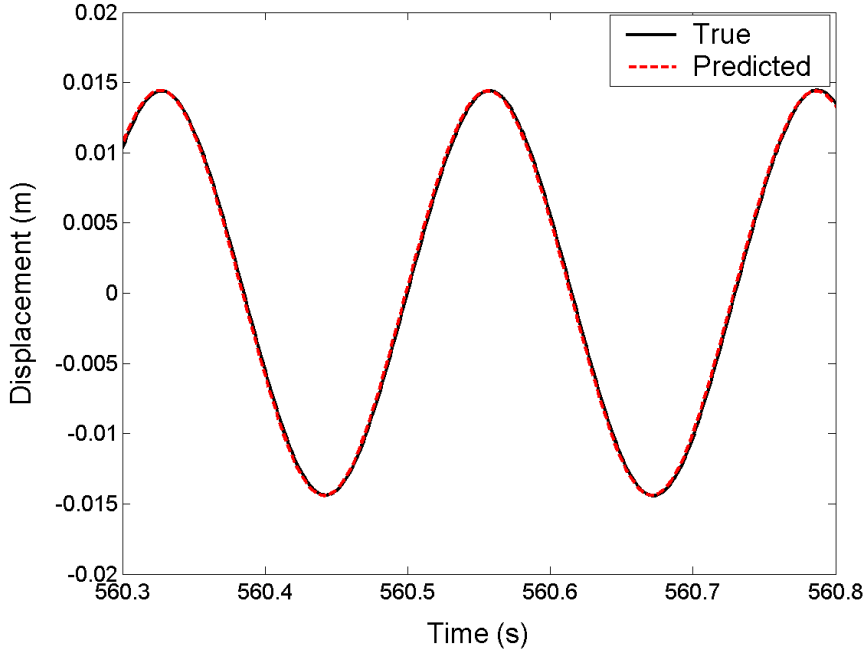


Figure 7.7. Downward prediction. (a) Comparison between true and predicted output, near the jump-up. (b) Magnification just after the jump.

7.3.3. Improved results

In this section the novel algorithm of Section 7.2 is tested and the results of the NSI procedure exploiting it are analysed. Note that the previously adopted $i = 60$ is the maximum index (for the calculator used for the computations) which allows to avoid the memory limitation problems described in Section 6.2.3. In fact, for larger values of i, *Matlab*® goes out of memory and the NSI procedure with the standard "qr" function fails.

The same time histories ($s = 10^5$ samples) as in Section 7.3.1 are considered, and the NSI procedure with the novel algorithm is performed for higher values of the number of block rows i.

Since Table 7.2 shows that the best parameter estimations are obtained in Case 2 (Gaussian random input), the results presented in this section refer only to Case 2. Note also that in all the following tables the results obtained by choosing $i = 60$ are also reported for comparison purposes. For this value of i the results are the same as in Table 7.2, as expected: the novel algorithm does not alter the NSI results, it just proposes a useful way to compute matrix R in those cases in which *Matlab*® produces an "out of memory" message. However it is observed that, when the standard *Matlab*® "qr" function is still applicable, the novel algorithm is about 26 times slower because of its many savings and loadings from the hard disk.

Table 7.3 shows the identification results relative to an output corrupted by 1% of noise: it is clear that the percentage error in the estimates of k and k_3 decreases as i increases. This trend is not so evident for the estimates of m and c: this is due to the fact that these parameters are not directly estimated from matrix $H_E(\omega = 0)$, as k and k_3 in (7.6), but they depend on the estimates of k, ω_n and ζ through the relationships of (7.7); this may cause a sort of error propagation or compensation. This remark is also valid for Tables 7.4 and 7.5.

From Table 7.3 it can also be observed that a value of $i = 60$ is anyway sufficient to obtain an excellent level of accuracy in the estimates, so the application of the new algorithm is not necessary.

The new algorithm appears to be more appealing when the output is corrupted by a higher level of noise: in this case it is necessary to increase the value of i in order to attain acceptable accuracy in the estimates, in particular as regards the nonlinear coefficient k_3.

For this reason, the previously generated output is corrupted by adding a higher percentage of zero-mean Gaussian random noise and the results of the identification procedures are shown in Tables 7.4 and 7.5 for 3% and 5% noise, respectively. It can be observed that the index i required in order to obtain the same level of accuracy increases as the noise percentage increases.

Table 7.3. Identification results (noise 1%): percentage error (100 • | estimated – actual | / actual).

i	m	k	c	k_3
60	0.13	0.54	0.73	0.73
90	0.13	0.33	0.57	0.49
120	0.08	0.13	0.15	0.21
180	0.07	0.11	0.33	0.18

Table 7.4. Identification results (noise 3%): percentage error (100 • | estimated – actual | / actual).

i	m	k	c	k_3
60	0.68	1.87	1.54	2.98
90	0.76	1.37	1.22	2.32
120	0.57	0.74	0.80	1.37
180	0.51	0.66	0.63	1.20

Table 7.5. Identification results (noise 5%): percentage error (100 • | estimated – actual | / actual).

i	m	k	c	k_3
60	1.19	3.08	0.53	6.24
90	1.60	2.62	0.15	5.29
120	1.41	1.84	0.73	3.82
180	1.26	1.61	1.59	3.34

Chapter 8

A complete covariance-driven method

In this chapter a multivariate subspace-based formulation in the time domain for modal parameter identification using covariances is developed, with the aim of proposing a complete input-output covariance-driven identification method applicable in the same way as its well-established data-driven counterpart.

Input-output covariance-driven subspace identification has been less investigated in the past than its output-only counterpart, but some results are available. In the literature, most of the proposed approaches [61] consist in handling different projections of the system onto the subspace generated by the inputs, or onto its orthogonal subspace, for eigenstructure identification only.

In this chapter the main purpose consists of the estimation of the input matrix B and the direct feedthrough matrix D of (6.11), since it is believed that these matrices should also be provided by an input-output identification method in order to be defined as complete. This state-space reconstruction capability, which is fundamental for time response prediction and frequency response analysis, is well-established for data-driven methods as seen in Chapter 6, but it has not been investigated for an input-output covariance-driven method yet.

Moreover, an important issue of structural vibration analysis consists of considering mechanical structures subject to uncontrolled, unmeasured and nonstationary excitation [109]. The concept of nonstationary consistency of subspace-based methods in the identification of the eigenstructure of a linear multivariable system is dealt with, for example, in [110] for nonstationarities in

unobserved disturbances. The same concepts can be applied when nonstationarities are in the known input, the state-space estimates also being consistent.

Two numerical examples are given, to demonstrate the capabilities of the present covariance-driven subspace identification method (a further experimental application can be found in [111]). Some comparisons with the data-driven subspace identification method are also indicated. A simple SDOF system is used to demonstrate the consistency of the proposed method under weak nonstationary input and, consequently, output. A 15 DOFs example is used to show the capability of the covariance-driven method to deal with large and complicated systems, even when a high level of noise implies using a large number of samples to avoid poor estimates. The covariance-driven method, in fact, is not suffering from the memory limitation problems described in Section 6.2.3, such as in the data-driven method.

As a final remark, the presented covariance-driven method has been only demonstrated on linear identification procedures, although on large MDOF systems. Some computational difficulties leading to wrong results have been encountered when trying to apply the method on nonlinear systems. The reasons of this failure (and possibly a solution) are under investigation: this aspect can be a starting point for future research developments.

8.1. Methodology

8.1.1. Output only

The output-only covariance-driven subspace identification is well-established in the literature [54, 59, 112] and it is recalled here in order to introduce some concepts and notations [60] that will also be used in the following sections, where the input-output approach will be presented.

Given a stochastic state-space model with *s* measurements of the output

$$\begin{cases} x_{r+1} = Ax_r + w_r \\ y_r \quad = Cx_r + v_r \end{cases} \tag{8.1}$$

where w_r and v_r are unmeasurable vector signals called process error and measurement error respectively, the subspace identification problem consists in estimating the model order n and the system matrices A and C up to within a similarity transformation.

Assuming w_r and v_r as Gaussian white noise, the stochastic nature of the problem leads to the definition of the covariances between all outputs

$$R_j = E\left[y_{r+j}\, y_r^T \right].$$

Furthermore, the next state-output covariance matrix is defined as

$$G_1 = E\left[x_{r+1}\, y_r^T \right].$$

After defining a block index i, the algorithm starts with the construction of a block Hankel matrix of output covariance matrices, constructed with the q measured outputs:

$$R = \begin{bmatrix} R_1 & R_2 & R_3 & \cdots & R_i \\ R_2 & R_3 & R_4 & \cdots & R_{i+1} \\ R_3 & R_4 & R_5 & \cdots & R_{i+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_i & R_{i+1} & R_{i+2} & \cdots & R_{2i-1} \end{bmatrix}, \qquad R \in \mathbf{R}^{iq \times iq}. \tag{8.2}$$

Note that in this paper square matrices are the best choice found by the authors for representing the block Hankel covariances matrices (as in [59]); in some other papers [61, 112] rectangular matrices are preferred, but the influence of this choice on the results is not significant.

In order to get an implementation of this matrix, consider the following iq-dimensional column vectors, containing the future and past received data, respectively:

$$Y_r^+ = \begin{bmatrix} y_{r+1} \\ y_{r+2} \\ \vdots \\ y_{r+i} \end{bmatrix}, \qquad Y_r^- = \begin{bmatrix} y_r \\ y_{r-1} \\ \vdots \\ y_{r-i+1} \end{bmatrix}. \tag{8.3}$$

Due to the stationarity assumption, Hankel matrix in (8.2) writes:

$$R = E\left[Y_r^+ Y_r^{-T} \right]. \tag{8.4}$$

Given a s-size data sample, it can be deduced from (8.4) that the corresponding empirical block Hankel matrix also writes:

$$\hat{R} = \frac{1}{J} \sum_{r=i}^{s-i} Y_r^+ Y_r^{-T}, \tag{8.5}$$

where $J = s - 2i + 1$ denotes the number of data used in the calculation of the covariances (the symbol $\wedge$ indicates estimated matrices).
The block Hankel matrix in (8.2) decomposes as

$$R = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{i-1} \end{bmatrix} \begin{bmatrix} A^{i-1}G & A^{i-2}G & \cdots & AG & G \end{bmatrix} = \Gamma_i C_i,$$

where $\Gamma_i \in \mathbf{R}^{iq \times n}$ is called the extended observability matrix and $C_i \in \mathbf{R}^{n \times iq}$ the stochastic controllability matrix; n is the system order.
From this equation, it is clear that Γ_i and C_i can be estimated (up to a similarity transform) from $\hat{R}$ by performing a singular value decomposition (SVD):

$$\hat{R} = U \Sigma V^T = \begin{bmatrix} U_n & U_0 \end{bmatrix} \begin{bmatrix} \Sigma_n & 0 \\ 0 & \Sigma_0 \end{bmatrix} \begin{bmatrix} V_n^T \\ V_0^T \end{bmatrix}, \tag{8.6}$$

where matrices U_n, Σ_n and V_n have dimensions $iq \times n$, $n \times n$ and $iq \times n$, respectively. Observe that $\Sigma_0 \to 0$ as the number of samples s increases. The model order n is determined by inspecting singular values. Then an estimate of the extended observability matrix may be obtained as $\Gamma_i = U_n \Sigma_n^{1/2}$; in order to get an estimate of the system matrices A and C, it can be seen that Γ_i has the following shift-invariant structure:

$$\Gamma_i^{(2)} = \Gamma_i^{(1)} A,$$

where submatrices $\Gamma_i^{(1)}$ and $\Gamma_i^{(2)}$ are defined from Γ_i as

$$\Gamma_i = \begin{bmatrix} \Gamma_i^{(1)} \\ CA^{i-1} \end{bmatrix} = \begin{bmatrix} C \\ \Gamma_i^{(2)} \end{bmatrix}.$$

It follows that matrix C can be estimated from the first q rows of Γ_i, while the state matrix A can be estimated as

$$A = \Gamma_i^{(1)\dagger} \Gamma_i^{(2)},$$

where the symbol "†" denotes the Moore-Penrose pseudo-inverse.

8.1.2. Input-output

Input-output covariance-driven subspace identification (CDSI) has been less investigated in the past than its output-only counterpart, but some results are available. In the literature, most of the proposed approaches [61] consist in handling different projections of the system onto the subspace generated by the inputs, or onto its orthogonal subspace, for estimating matrices A and C only. This procedure has been used in this section, but it is believed that a complete input-output identification method should also cater for the estimation of matrices B and D. In the following, a covariance-driven approach for obtaining these matrices is proposed and details for its implementation will also be provided in next section.

Given a deterministic-stochastic state-space model (the same derived in (6.11), which is reported here for completeness), with *s* measurements of the input and of the output:

$$\begin{cases} x_{r+1} = Ax_r + Bu_r + w_r \\ y_r \quad = Cx_r + Du_r + v_r \end{cases} \tag{8.7}$$

where w_r and v_r are unmeasurable vector signals as in (8.1), the subspace identification problem consists in estimating the model order n and the system matrices A, B, C and D up to within a similarity transformation.

Observe that (8.1) can be considered as a particular case of (8.7), by combining $Bu_r + w_r = \tilde{w}_r$ and $Du_r + v_r = \tilde{v}_r$ and regarding these two terms as noises. This assumption is usually adopted when excitations cannot be measured, such as during ambient testing of large complex structures, and the output-only identification is performed.

In (8.7), the stochastic terms w_r and v_r are unknown, but they are assumed to be zero mean, stationary noise vector sequences. This stochastic nature leads to the

definition of the following covariance matrices: the covariance matrix between all outputs

$$R_j = E\left[y_{r+j}\, y_r^T \right],$$

between the states and the outputs

$$G_j = E\left[x_{r+j}\, y_r^T \right],$$

and between all inputs and outputs

$$L_j = E\left[u_{r+j}\, y_r^T \right]. \tag{8.8}$$

The following properties can be easily verified:

$$G_j = A^{j-1} G_1 + \sum_{g=2}^{j} A^{j-g} B L_{g-1},$$

and

$$R_1 = CG_1 + DL_1$$

$$R_2 = CG_2 + DL_2 = CAG_1 + CBL_1 + DL_2$$

$$\vdots$$

$$R_j = CG_j + DL_j = CA^{j-1} G_1 + \sum_{g=2}^{j} CA^{j-g} B L_{g-1} + DL_j. \tag{8.9}$$

Estimating matrices A and C

The algorithm which leads to an estimate of matrices A and C is similar to that introduced in Section 8.1.1, but it needs some preliminary operations in order to consider the contribution deriving from the knowledge of the inputs [60].
After defining a block index i, the algorithm starts by considering the output data as in (8.3), and the following $i\ell$- dimensional column vectors, containing the future and past input data, respectively:

$$U_r^+ = \begin{bmatrix} u_{r+1} \\ u_{r+2} \\ \vdots \\ u_{r+i} \end{bmatrix}, \qquad U_r^- = \begin{bmatrix} u_r \\ u_{r-1} \\ \vdots \\ u_{r-i+1} \end{bmatrix}.$$

To remove the influence of the input in the output data formulation, the projection of the output data on the orthogonal space of the past input data is performed.

The key point of this section consists in performing the orthogonal projection *not before* (for example as in [61]), *but after* the calculation of the covariance matrices: this issue turns out to be very useful in terms of a reduced computational effort and, in particular, of a strongly reduced memory occupation (see the 15DOFs numerical example in Section 8.3.2).

For two random vectors V and Z with zero means and finite variances, the orthogonal projection of V on $sp(Z)$ is defined as

$$V/Z = E\left(VZ^T\right)E\left(ZZ^T\right)^\dagger Z$$

$$V/Z^\perp = V - V/Z\,.$$

In this case, the projection of each output y_r on the past inputs $u_1,\ldots,u_r$ is to be known. Due to the property of contraction of A, it is sufficient to project on the i last inputs $u_{r-i+1},\ldots,u_r$, that means on U_r^-, for i large enough.

For this procedure, the following empirical covariance matrices are needed:

$$C_0^- = E\left[Y_r^- U_r^{-T}\right] = \frac{1}{J}\sum_{r=i}^{s-i} y_r U_r^{-T}\,,$$

$$R^{--} = E\left[U_r^- U_r^{-T}\right] = \frac{1}{J}\sum_{r=i}^{s-i} U_r^- U_r^{-T}\,,$$

where $J = s - 2i + 1$ denotes the number of data used in the calculation of the covariances.

Now, the orthogonalized outputs are given by

$$y_r/U^\perp = y_r - C_0^- R^{--\dagger} U_r^-$$

and, with these new measures, the classical output only identification method as described in Section 8.1.1 can be applied.

Defining the future and past new data as

$$\left(Y/U^{\perp}\right)_r^{+} = Y_r^{+} - C_0^{-} R^{--\dagger} U_r^{-}$$

$$\left(Y/U^{\perp}\right)_r^{-} = Y_r^{-} - C_0^{-} R^{--\dagger} U_r^{-},$$

the new empirical block Hankel matrix W has the same structure as R in (8.2) and is defined as

$$W = \frac{1}{J} \sum_{r=i}^{s-i} \left(Y/U^{\perp}\right)_r^{+} \left(Y/U^{\perp}\right)_r^{-T} . \tag{8.10}$$

So, once this new Hankel matrix is computed, the rest of the method leading to an estimate of matrices A and C is exactly the same as in Section 8.1.1, with W in place of R.

Estimating matrices B and D

Consider now the state space system of (8.7), and suppose that the extended observability matrix Γ_i has successfully been estimated by the procedure described above. By using the covariance matrices between inputs and outputs, defined in (8.8), it is possible to get a new Hankel matrix L having the same structure of matrix R defined in (8.2), so the following block Hankel matrix is introduced:

$$L = \begin{bmatrix} L_1 & L_2 & L_3 & \cdots & L_i \\ L_2 & L_3 & L_4 & \cdots & L_{i+1} \\ L_3 & L_4 & L_5 & \cdots & L_{i+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L_i & L_{i+1} & L_{i+2} & \cdots & L_{2i-1} \end{bmatrix}, \qquad L \in \mathbf{R}^{i\ell \times iq} . \tag{8.11}$$

By manipulating (8.7) in a “covariance” form, and by exploiting the property written in (8.9), it is possible to yield the following important matrix relationship:

$$R = \Gamma_i G + \Delta L , \tag{8.12}$$

where

$$G = \begin{bmatrix} G_1 & G_2 & \cdots & G_i \end{bmatrix} \in \mathbf{R}^{n \times iq}$$

contains the covariances between the states and the outputs.

Δ is a $iq \times i\ell$ lower triangular Toeplitz matrix formed from the first i impulse responses as

$$\Delta = \begin{bmatrix} D & 0 & 0 & \cdots & 0 \\ CB & D & 0 & \cdots & 0 \\ CAB & CB & D & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ CA^{i-2}B & CA^{i-3}B & CA^{i-4}B & \cdots & D \end{bmatrix}. \tag{8.13}$$

Pre-multiplying (8.12) by $\Gamma_i^{\perp}$, in order to exploit the properties of the orthogonal matrix, and then post-multiplying it by $L^{\dagger}$, the following equation is obtained:

$$\Gamma_i^{\perp} R L^{\dagger} = \Gamma_i^{\perp} \Delta . \tag{8.14}$$

A set of linear equations for B and D can be derived from (8.14), by using the estimated $\Gamma_i^{\perp}$ and by examining the structure of Δ in (8.13). This procedure will be described in detail in Section 8.2.

8.2. Implementation

As a remark, note that all the covariance matrices defined in (8.2), (8.11) and in the following can also be computed in a very fast way by using the FFT algorithm [113, 114].

The calculation of covariance matrices has a main drawback: data are squared up, while data-driven methods are implemented as numerically robust square root algorithms. For this reason, data-driven results should be more accurate than covariance-driven ones, even if in practical applications no accuracy differences can be observed when looking at the identified modal parameters [56]. However, the inaccuracy of covariance-driven methods can be reduced by considering very large data sets [109].

Another source of errors leading to poor estimates is noise: when measured signals are heavily corrupted by noise (i.e. with low SNR) and some denoising procedures are required, the quality of estimates can still be improved by considering a larger number of samples.

Estimating matrices A and C

An easy and robust technique for building the empirical block Hankel matrix W defined in (8.10), by performing the orthogonal projection after the computation of covariance matrices, has been found in [60].

Consider the notations introduced above for i, J, Y_r^+, Y_r^-, U_r^+ and U_r^-; moreover, the following collections of vectors are introduced:

$$Y^+ = \begin{bmatrix} Y_i^+ & Y_{i+1}^+ & \cdots & Y_{s-i}^+ \end{bmatrix},$$

$$Y^- = \begin{bmatrix} Y_i^- & Y_{i+1}^- & \cdots & Y_{s-i}^- \end{bmatrix},$$

and in the same way the future and past collections of inputs U^+ and U^-.

An almost equivalent way to remove the influence of the inputs from the outputs is to project the vector Y_r^+ orthogonally on U_r^+ and U_r^-, since for each r

$$\left(Y/U^{\perp}\right)_r^+ \approx (Y/\begin{pmatrix} U_r^+ \\ U_r^- \end{pmatrix}^{\perp})_r^+ .$$

In order to write the projection procedure in a compact form, the following empirical cross-covariance matrices

$$C^{++} = \frac{1}{J}\sum_{r=i}^{s-i} Y_r^+ U_r^{+T}, \qquad C^{+-} = \frac{1}{J}\sum_{r=i}^{s-i} Y_r^+ U_r^{-T},$$

$$C^{-+} = \frac{1}{J}\sum_{r=i}^{s-i} Y_r^- U_r^{+T}, \qquad C^{--} = \frac{1}{J}\sum_{r=i}^{s-i} Y_r^- U_r^{-T}$$

and empirical auto-covariance matrices

$$R^{++} = \frac{1}{J}\sum_{r=i}^{s-i} U_r^+ U_r^{+T}, \qquad R^{+-} = \frac{1}{J}\sum_{r=i}^{s-i} U_r^+ U_r^{-T},$$

$$R^{--} = \frac{1}{J}\sum_{r=i}^{s-i} U_r^- U_r^{-T}, \qquad R^{-+} = \left(R^{+-}\right)^T$$

are introduced.

The new empirical block Hankel matrix of (8.10)

$$W = \frac{1}{J}(Y^+ / \begin{pmatrix} U^+ \\ U^- \end{pmatrix}^\perp) Y^{-T}$$

is simply

$$W = R - \begin{bmatrix} C^{++} & C^{+-} \end{bmatrix} \begin{bmatrix} R^{++} & R^{+-} \\ R^{-+} & R^{--} \end{bmatrix}^\dagger \begin{bmatrix} \left(C^{-+}\right)^T \\ \left(C^{--}\right)^T \end{bmatrix},$$

where R is defined as in (8.2).

Then, the extended observability matrix Γ_i and the stochastic controllability matrix C_i can be estimated by performing the SVD as in (8.6):

$$W = U\Sigma V^T = \begin{bmatrix} U_n & U_0 \end{bmatrix} \begin{bmatrix} \Sigma_n & 0 \\ 0 & \Sigma_0 \end{bmatrix} \begin{bmatrix} V_n^T \\ V_0^T \end{bmatrix}, \tag{8.15}$$

and matrices A and C are estimated the same way as described in Section 8.1.1.

Estimating matrices B and D

Before considering (8.14), it is necessary to define how to write the covariance matrices R and L: matrix R writes as in (8.5), while matrix L is defined as

$$L = E\left[U_r^+ Y_r^{-T} \right] = \frac{1}{J} \sum_{r=i}^{s-i} U_r^+ Y_r^{-T} = \left(C^{-+}\right)^T . \tag{8.16}$$

The last term to be estimated, before proceeding in solving the linear system, is the orthogonal matrix $\Gamma_i^\perp$: this can be easily obtained by exploiting the orthonormality property of matrix U in the SVD of (8.15). It can be shown that a numerically robust estimate is given by

$$\Gamma_i^\perp = U_0^T \tag{8.17}$$

which, in particular, has full rank (equal to $iq - n$). Note that this peculiarity is crucial in order to obtain correct results for the estimates, by keeping a compatibility with the dimensions of B and D. In fact, the following classical definition of left orthogonal matrix

$$\Gamma_i^\perp = I - \Gamma_i \left(\Gamma_i^T \Gamma_i\right)^\dagger \Gamma_i^T \in \mathbf{R}^{iq \times iq}$$

is not applicable in this case, since it is not of full rank.

Once the quantities in (8.16) and (8.17) have been defined, it is possible to determine an estimate for matrices B and D, after partitioning and rearranging of (8.14), using the standard least squares techniques. A similar approach, but applied in a data-driven framework, has been adopted in [115].

An appropriate way of computing B and D exploits the lower triangular structure of matrix Δ, as it can be seen in (8.13), in order to write the following set of $(iq-n)i$ linear equations:

$$F\begin{bmatrix} D \\ B \end{bmatrix} = \widetilde{P}\,.$$

The unknown has dimensions $(q+n)\times\ell$ and can be determined through the pseudo-inverse of F as

$$\begin{bmatrix} D \\ B \end{bmatrix} = F^{\dagger}\widetilde{P}\,.$$

Matrices F and $\widetilde{P}$ have dimensions $(iq-n)i\times(q+n)$ and $(iq-n)i\times\ell$, respectively, and are defined as

$$F = \begin{bmatrix} U_0^T(:,1:q) & U_0^T(:,q+1:iq)\Gamma_i(1:q(i-1),:) \\ U_0^T(:,q+1:2q) & U_0^T(:,2q+1:iq)\Gamma_i(1:q(i-2),:) \\ \vdots & \vdots \\ U_0^T(:,q(i-2)+1:(i-1)q) & U_0^T(:,q(i-1)+1:iq)\Gamma_i(1:q,:) \\ U_0^T(:,q(i-1)+1:iq) & 0 \end{bmatrix}$$

and

$$\widetilde{P} = \begin{bmatrix} P(:,1:\ell) \\ P(:,\ell+1:2\ell) \\ \vdots \\ P(:,\ell(i-2)+1:\ell(i-1)) \\ P(:,\ell(i-1)+1:i\ell) \end{bmatrix}.$$

Matrix $P = U_0^T R L^{\dagger} \in \mathbf{R}^{(iq-n)\times i\ell}$ from the above equation represents the left-hand side of (8.14) and $\widetilde{P}$ is formed by partitions of P.

8.3. Numerical examples

In this section two numerical examples are shown, to demonstrate the capabilities of the proposed covariance-driven subspace identification method (CDSI) to deal with large and complicated systems, even when a high level of noise implies using a large number of samples to avoid poor estimates. Some comparisons with the data-driven subspace identification method (DDSI) described in Section 6.2 are also indicated. By using the same user-defined parameters (number of samples s and block index i), the latter method gives in general better results but in some cases, especially when the number of samples and degrees of freedom is large, the DDSI method can not be performed with values of s and i sufficiently high to obtain good results. This is due to the memory limitation problems described in Section 6.2.3, related to the storing and managing of these large Hankel data matrices. In detail, when working with *Matlab*®, these problems consist of an error message ("out of memory") which may occur in two situations: (1) when the block Hankel matrices are too large to be stored in the virtual memory space; (2) when they can be stored but the QR factorisation expected by the procedure can not be performed since it requires more virtual memory space.

The CDSI method is not suffering from these drawbacks and the user-defined parameters can be increased in order to improve significantly the quality of the results. Furthermore, the results can reach an excellent level, even better than those obtained by applying the DDSI method at its maximum capabilities, at least for eigenstructure identification. This concept will be shown in the second example, concerning a 15 DOFs system.

8.3.1. Single degree of freedom system

Consider a simple SDOF system whose parameters are $m = 1.3$ kg, $c = 2$ N s/m and $k = 800$ N/m. This system can be used to demonstrate the consistency of the proposed CDSI method even under weak nonstationary input and, consequently, output. The concept of nonstationary consistency is dealt with, for example, in [110] for nonstationarities in unobserved disturbances. In this example nonstationarities are in the observed input and output, but the same concepts can be applied.

In order to obtain statistically significant results, 100 Monte Carlo identification procedures have been performed, each with a different excitation specified by the following

$$u(t) = 6\left[\sin\left(2\pi\left(t/P + \Phi\right)\right)\right]\cdot V + Z\,,$$

where

$$V\,,\ Z\ \sim \mathcal{N}(0,\,1); \qquad P \sim \mathcal{U}(25,\,80); \qquad \Phi \sim \mathcal{U}(0,\,1).$$

Note that the notation $\mathcal{N}(\mu,\,\sigma)$ refers to a Normal distribution with mean μ and standard deviation σ, while $\mathcal{U}$ (a, b) refers to a Uniform distribution defined in the interval (a, b).

For each identification procedure, time histories have been obtained through a numerical simulation with a time step $\Delta t = 1.25\times10^{-2}$ s, and a total number of $s = 10^4$ samples has been generated. As an example, one of the realisations of the input and the corresponding output (assuming that the output $y(t)$ is the displacement) is shown in Fig. 8.1. The effect of the measurement noise on the parameter estimation results is investigated by corrupting the output with an additive Gaussian zero-mean noise (5% of the root mean square value).

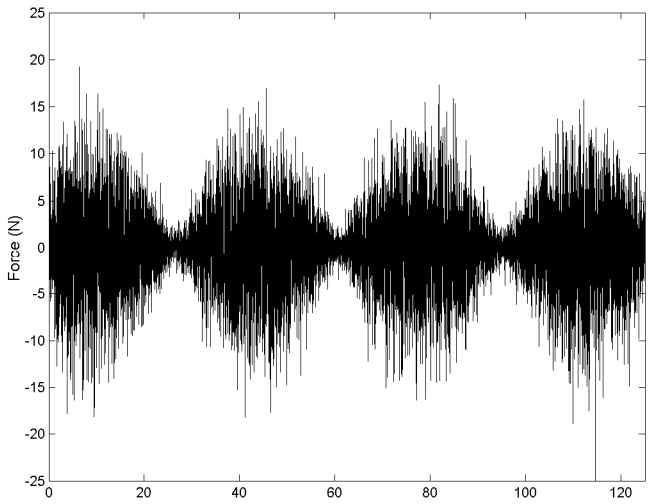
The analysis through the covariance-driven approach described in Sections 8.1 and 8.2 has been carried out by considering $i = 120$ block rows and by selecting the model order $n = 2$. The results are presented in the following, together with those obtained by applying the DDSI method. The mean and the standard deviation (std) of the percentage errors over 100 Monte Carlo estimates of the natural frequency f_n and the damping factor ζ are indicated in Table 8.1.

The results obtained by simply exciting the system with a stationary Gaussian random input (with a root mean square value of 4 N, giving the same levels of displacement as for the nonstationary case) are also presented for comparison purposes. The estimates provided by the CDSI method are worse than those attained by the DDSI method, but they are still very good if considering that 5% noise affects the output. However, it is believed that the quality of the estimates, for the CDSI method in particular, may be further improved by increasing the selected model order, according to a usual procedure based on stabilization diagrams [55, 59]. The quality of nonstationary results is also excellent as compared to the stationary ones, that are supposed to be the best.

In order to evaluate also the quality of the estimation of matrices B and D, a comparison between the actual and the estimated Frequency Response Function is

carried out by computing the matrix defined in (6.12). The curves, as shown in Fig. 8.2, are almost overlapped, with a maximum error of 0.8% around the resonance for the CDSI estimate (the DDSI estimate is better, as for the estimated parameters).

(a)



(b)

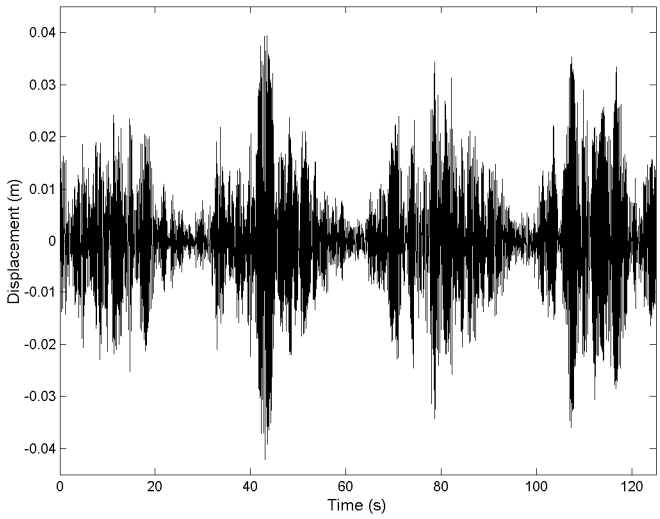


Figure 8.1. One of the realisations of the input (a) and the corresponding output (b).

Table 8.1. Percentage error $\left(100 \cdot \frac{|\text{estimated} - \text{actual}|}{\text{actual}}\right)$ with 100 Monte Carlo experiments with 5% noise. The mean and the std of the estimated parameters are presented for both the CDSI and DDSI methods, and for both the nonstationary (NonSt) and stationary (St) excitations.

	f_n				ζ			
	mean		std		mean		std	
	CDSI	DDSI	CDSI	DDSI	CDSI	DDSI	CDSI	DDSI
NonSt	0.0083	0.0025	0.0062	0.0021	0.2931	0.0895	0.2258	0.0707
St	0.0080	0.0028	0.0058	0.0022	0.2580	0.0833	0.1944	0.0652

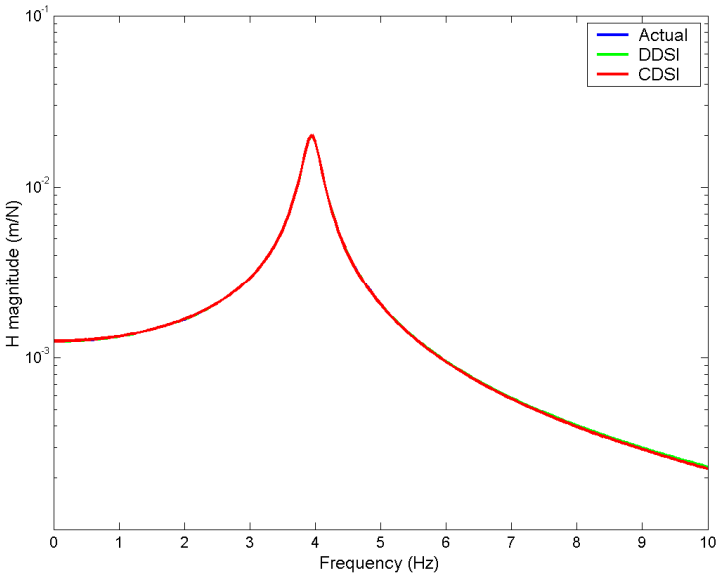


Figure 8.2. FRF of the system: actual (blue line), DDSI estimate (green line) and CDSI estimate (red line).

8.3.2. Fifteen degrees of freedom system

Consider the 15 degrees of freedom system shown in Fig. 8.3. The system, which is the same defined in [116], is excited by a zero-mean Gaussian random input with a root mean square (r.m.s.) value of 120 N, applied only to DOF 6. By assuming an output consisting of all displacements, time histories have been obtained through a numerical simulation with a time step $\Delta t = 2.5\times10^{-4}$ s, and a total number of $s = 3\times10^{4}$ samples has been generated. The effect of the measurement noise on the parameter estimation results is investigated by corrupting the output adding a Gaussian zero-mean noise (5% of the r.m.s. value). A comparison is made by performing the CDSI and the DDSI methods with different choices for the number of samples s and the block index i. A number of samples of $s = 2\times10^{4}$ and $s = 3\times10^{4}$ is selected and for each value of s the block index is selected as $i = 60$ and $i = 120$.

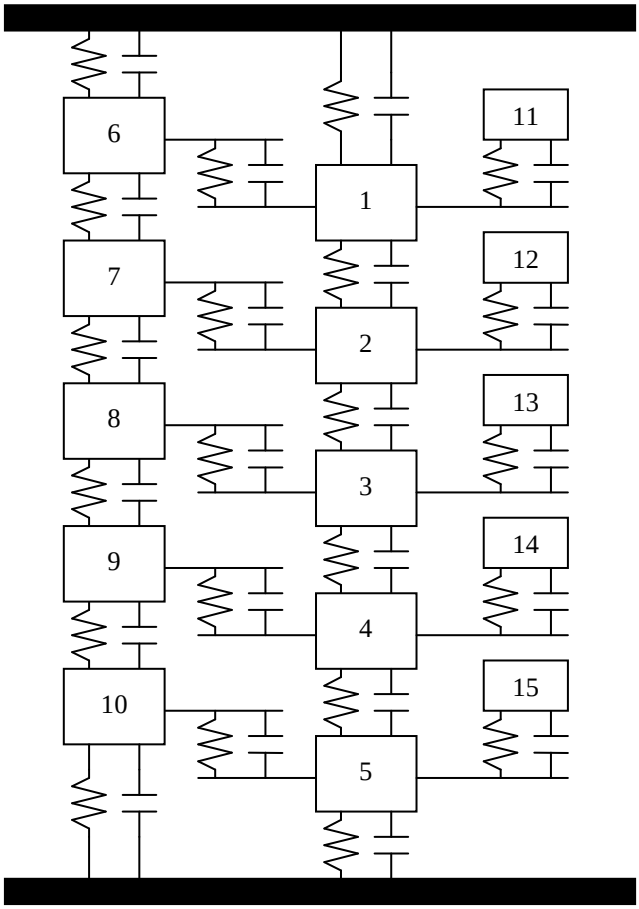


Figure 8.3. Representation of the 15 degrees of freedom system [116].

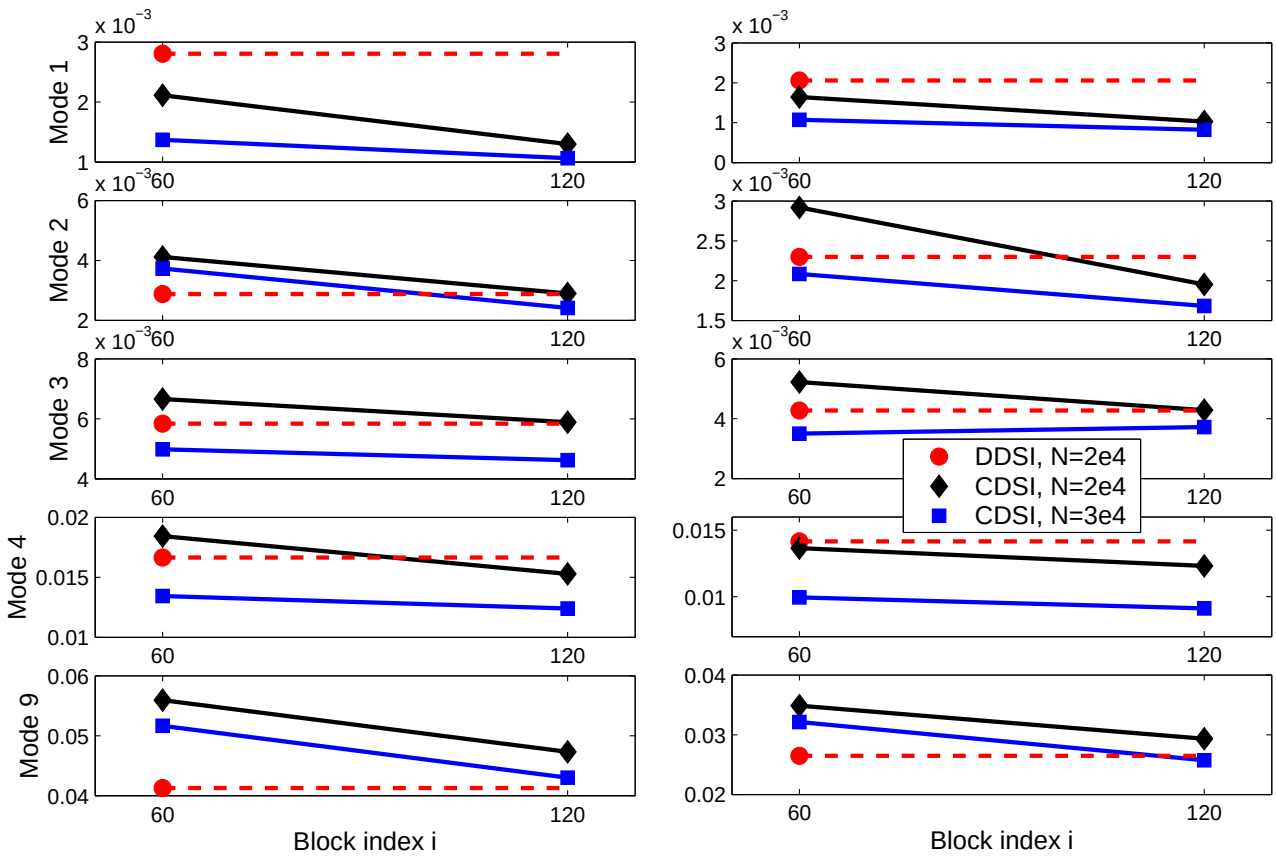


Figure 8.4. Mean (a) and standard deviation (b) of the percentage errors related to 100 Monte Carlo estimates of some of the natural frequencies. Solid lines represent performed identifications, while dashed lines indicate that the DDSI method with $i = 120$ cannot be performed. Thus, the dashed line is reported for comparison purposes only.

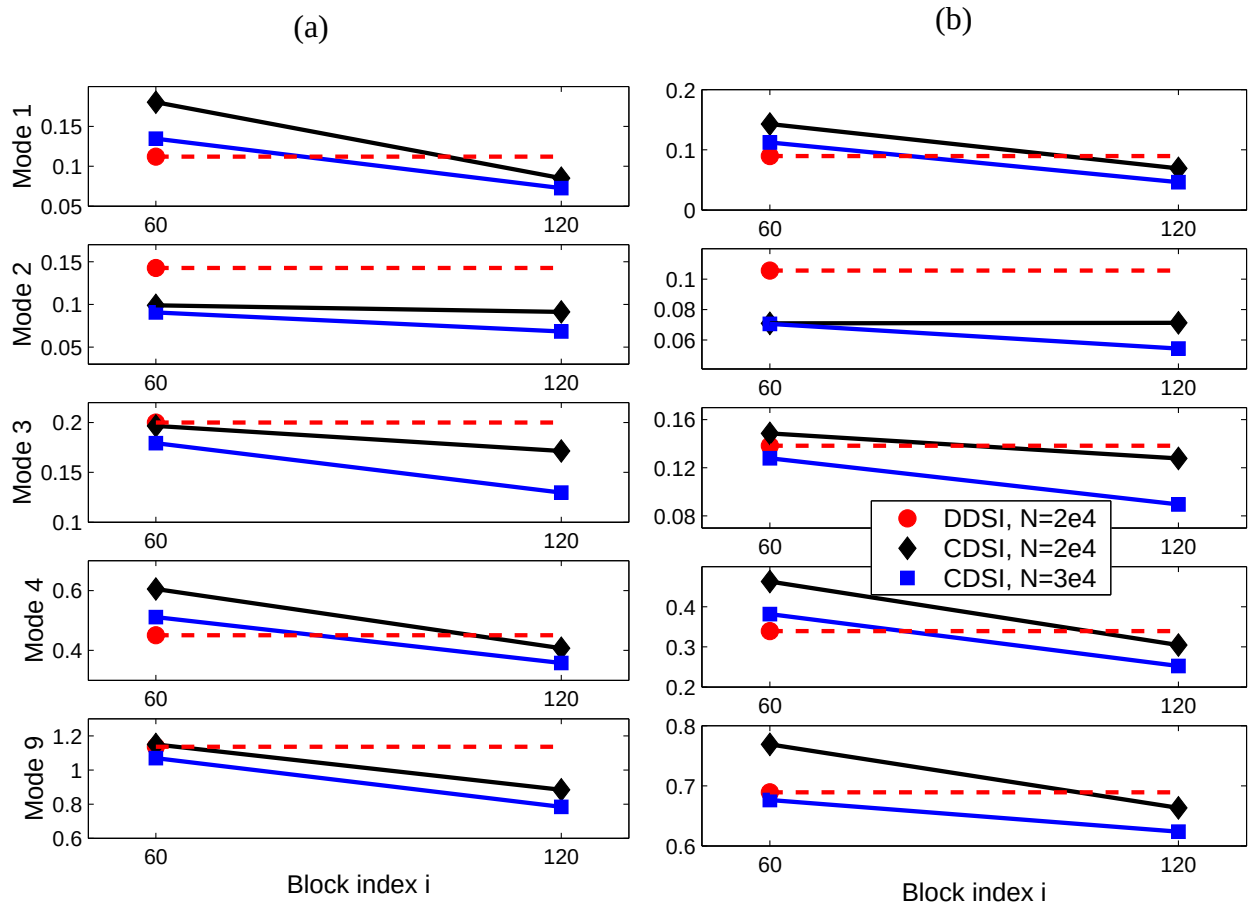


Figure 8.5. Mean (a) and standard deviation (b) of the percentage errors related to 100 Monte Carlo estimates of some of the damping factors. Solid lines represent performed identifications, while dashed lines indicate that the DDSI method with $i = 120$ cannot be performed. Thus, the dashed line is reported for comparison purposes only.

The model order $n = 30$ is fixed and both methods identify 10 out of 15 modes: modes from 1 to 5, from 7 to 10 and mode 13. The natural frequency of mode 6 is very close to that of mode 5, so its identification is difficult in case of noisy measurements like these. In order to improve the number and the quality of identified modes, the model order can be increased through a stabilization diagram, but this is not the goal of this example.

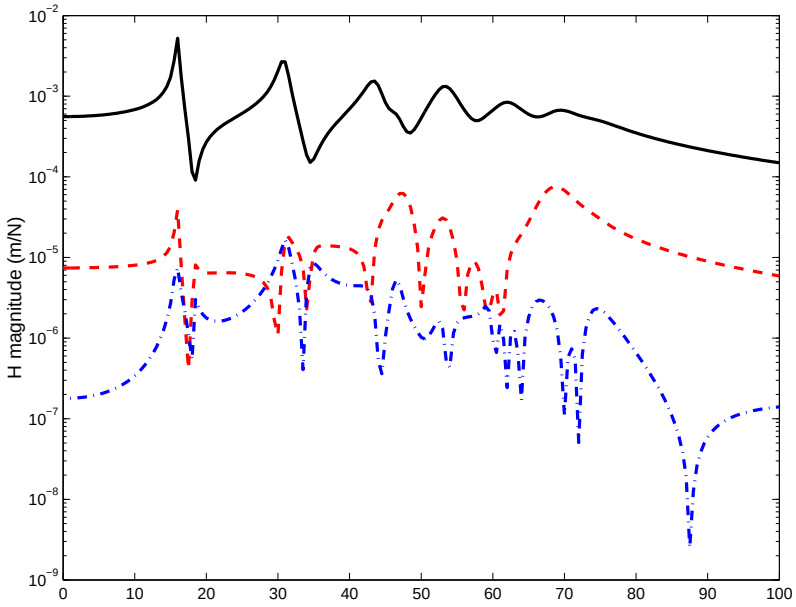
Information about the accuracy of the methods is given in Figs. 8.4 and 8.5, where the mean and the standard deviation (std) of the percentage errors related to 100 Monte Carlo estimates of some natural frequencies f_n (Fig. 8.4) and damping

factors ζ_n (Fig. 8.5) are indicated. Similar results and considerations hold for the other identified modes. In most of the cases the DDSI method cannot be performed because of the memory limitation problems described above: the only applicable choice (the full circle in the figures) is $s = 2\times10^4$ and $i = 60$. For higher values of these parameters the DDSI method is not available (the dashed lines in the figures are indicated for comparison purposes only). The CDSI method is not affected by memory limitations and the quality of the results get better by increasing the values selected for the user-defined parameters, as indicated by the solid lines.

A comparison between the actual and the estimated Frequency Response Functions is carried out by computing the matrix defined in (6.12). This equation involves all the estimated state-space matrices, so this comparison is useful in order to evaluate also the quality of the estimation of matrices B and D. The results are shown in Fig. 8.6 for $H_{6,6}$ and $H_{12,6}$, the same level of accuracy being obtained by all other FRFs. The DDSI estimate is not shown, since it is very close to the CDSI estimate and differences could not be appreciated in the figures. For evaluating the quality of estimates, the residues computed as $|\text{estimated} - \text{actual}|$ are also reported, for both the CDSI and the DDSI methods. An excellent agreement can be observed by the CDSI estimates (obtained with parameters $s = 3\times10^4$ and $i = 120$), since in presence of noise the residues are about two orders of magnitude smaller. Even better results are obtained (with parameters $s = 2\times10^4$ and $i = 60$) by the DDSI method, since the residues are about three orders of magnitude smaller.

In conclusion, this numerical example demonstrates the presented CDSI method, which can be applied as well as the established DDSI method with similar results. For both methods, the quality of results is excellent and they can be both used in practical situations, depending on the size of the data sets that have to be managed. In particular, the CDSI method can handle larger data sets to obtain more accurate estimates of modal parameters such as natural frequencies and damping ratios, while the DDSI method produces better FRF estimates, due to its robustness in estimating matrices B and D.

(a)



(b)

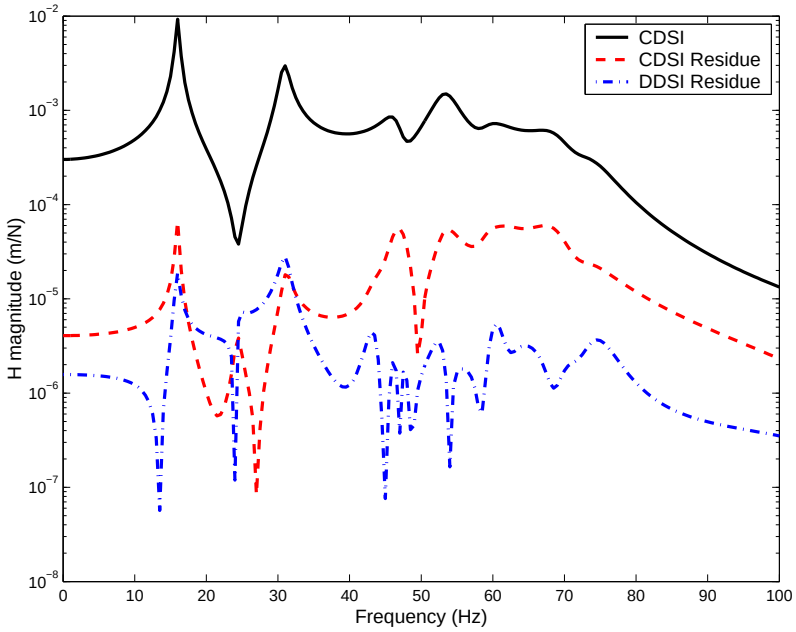


Figure 8.6. Two of the system FRFs: $H_{6,6}$ (a) and $H_{12,6}$ (b). The residues have been computed as $|\text{estimated-actual}|$. The CDSI estimate has been obtained with parameters $s = 3\times10^4$ and $i = 120$; the DDSI estimate has been obtained with parameters $s = 2\times10^4$ and $i = 60$.

Chapter 9

Continuous structures: a modal approach

In practice most structural nonlinearities are distributed and an ideal nonlinear identification method should cater for them as well [117]. Moreover, any nonlinear method is requested to correctly identify the types of nonlinearity present and possibly to quantify the extent of their force contributions. Therefore, the identification of a whole parametric nonlinear model is an important instrument for many purposes. For example, it would allow for treating nonlinearities in possibly damaged structures [118], or attaining improved predictions of vibration response amplitude, which is an issue for accurate long term fatigue estimates.

In order to extend the current method to be applied also on realistic large engineering structures, a model reduction is needed and can be performed by selecting a set of modes that span the dominant dynamics. To this purpose, a modal counterpart of the NSI method has been developed, together with ideas for handling a large complex nonlinear system as simple modal single degree of freedom systems [119].

In this chapter, the modal NSI technique is described and some mathematical details are consolidated through a numerical example. The method is also demonstrated through an experimental application set up in Pescara, Italy, where some tests on reinforced concrete beams have been performed.

9.1. Methodology

9.1.1. NSI method in modal space

Nonlinear modal model

A dynamical system with N degrees of freedom and with lumped nonlinear springs and dampers can be described by the equation of motion (6.3), which is reported here for completeness:

$$M\ddot{z}(t)+C_v\dot{z}(t)+Kz(t)=f(t)+f_{NL}(t), \tag{9.1}$$

where

$$f_{NL}(t)=-\sum_{j=1}^{h}\mu_j L_j g_j(z,\dot{z}). \tag{9.2}$$

The transformation between physical and modal space, after selecting N_R retained modes, is defined by

$$z(t)=\Phi p(t), \tag{9.3}$$

where $p(t)$ is the generalised modal displacement vector and Φ is the $N\times N_R$ modal matrix of the underlying linear system. The modal approach is a form of model order reduction, since the number of retained modes, in general, will be much smaller than the number of physical degrees of freedom, so that $N_R << N$. Substituting equation (9.3) into the equation of motion (9.1) and pre-multiplying by Φ^T yields:

$$\Phi^T M\Phi\ddot{p}(t)+\Phi^T C_v\Phi\dot{p}(t)+\Phi^T K\Phi p(t)=\Phi^T f(t)-\Phi^T\sum_{j=1}^{h}\mu_j L_j g_j(\Phi p,\Phi\dot{p}). \tag{9.4}$$

By using the orthogonality of the modes, equation (9.4) becomes:

$$\overline{M}\ddot{p}(t)+\overline{C}_v\dot{p}(t)+\overline{K}p(t)=\bar{f}(t)+\bar{f}_{NL}(t), \tag{9.5}$$

where

$$\bar{f}_{NL}(t)=-\sum_{j=1}^{\bar{h}}\overline{\mu}_j\overline{g}_j(p,\dot{p}). \tag{9.6}$$

The matrices $\overline{M}, \overline{C}_v$ and $\overline{K}$ are $N_R \times N_R$. The modal mass matrix $\overline{M}$ and the linear modal stiffness matrix $\overline{K}$ are diagonal, while the modal damping matrix $\overline{C}_v$ is diagonal for proportionally damped systems only. $\bar{f}(t)$ is the $N_R \times 1$ applied modal force vector and $\bar{f}_{NL}(t)$ is the $N_R \times 1$ vector of internal feedback modal forces due to nonlinearities. Each of the $\overline{h}$ nonlinear components depends on the scalar modal nonlinear function $\overline{g}_j(p, \dot{p})$, which is related to $\bar{f}_{NL}(t)$ through a $N_R \times 1$ vector of coefficients $\overline{\mu}_j$ (this aspect will be clarified in next section).

Assuming that the measurements concern displacements only, so that $y(t) = p(t)$, the state-space formulation of the equation of motion (9.5), corresponding to a state vector chosen as $x(t) = \begin{bmatrix} p(t)^T & \dot{p}(t)^T \end{bmatrix}^T$ and to an input vector $u(t) = \begin{bmatrix} \bar{f}(t)^T & -\overline{g}_1(t)^T & \dots & -\overline{g}_{\overline{h}}(t)^T \end{bmatrix}^T$, is

$$\begin{bmatrix} \dot{p} \\ \ddot{p} \end{bmatrix} = \begin{bmatrix} 0_{N_R \times N_R} & I_{N_R \times N_R} \\ -\overline{M}^{-1}\overline{K} & -\overline{M}^{-1}\overline{C}_v \end{bmatrix} \begin{bmatrix} p \\ \dot{p} \end{bmatrix} + \begin{bmatrix} 0_{N_R \times N_R} & 0_{N_R \times 1} & \cdots & 0_{N_R \times 1} \\ \overline{M}^{-1} & \overline{M}^{-1}\overline{\mu}_1 & \cdots & \overline{M}^{-1}\overline{\mu}_{\overline{h}} \end{bmatrix} \begin{bmatrix} \bar{f}(t) \\ -\overline{g}_1(t) \\ \vdots \\ -\overline{g}_{\overline{h}}(t) \end{bmatrix} \tag{9.7}$$

$$y = \begin{bmatrix} I_{N_R \times N_R} & 0_{N_R \times N_R} \end{bmatrix} \begin{bmatrix} p \\ \dot{p} \end{bmatrix} + \begin{bmatrix} 0_{N_R \times N_R} & 0_{N_R \times 1} & \cdots & 0_{N_R \times 1} \end{bmatrix} \begin{bmatrix} \bar{f}(t) \\ -\overline{g}_1(t) \\ \vdots \\ -\overline{g}_{\overline{h}}(t) \end{bmatrix} \tag{9.8}$$

By writing the previous equations in a compact form, the following continuous modal model can be derived:

$$\begin{aligned} \dot{x} &= A_c x + B_c u \\ y &= Cx + Du \end{aligned} \tag{9.9}$$

The continuous model of (9.9) may be converted into a discrete modal model and then processed by means of the subspace methods, exaclty the same way as discussed in Chapter 6.

Nonlinearities in modal coordinates

Let's focus on how $\bar{f}_{NL}(t)$ in (9.6) can be obtained from the nonlinear term of equation (9.4). A vector $\bar{L}_j = \Phi^T L_j$, whose entries depend on matrix Φ, can be defined. Moreover, it can be assumed that

$$\sum_{j=1}^{h} g_j(\Phi p, \Phi \dot{p}) = \sum_{j=1}^{\bar{h}} \varphi_j \bar{g}_j(p, \dot{p}), \tag{9.10}$$

where the coefficients φ_j and the number of nonlinear modal components $\bar{h}$ depend on Φ, h, N_R and on the nonlinear functions $g_j(z, \dot{z})$. In (9.10), $\bar{h}$ new nonlinear modal functions $\bar{g}_j(p, \dot{p})$ are derived. Then,

$$\sum_{j=1}^{h} \mu_j \Phi^T L_j g_j(\Phi p, \Phi \dot{p}) = \sum_{j=1}^{\bar{h}} \mu_j \bar{L}_j \varphi_j \bar{g}_j(p, \dot{p}) = \sum_{j=1}^{\bar{h}} \bar{\mu}_j \bar{g}_j(p, \dot{p}). \tag{9.11}$$

A simple example is given, in case of a 3DOFs system with a single nonlinear function $g_1(z) = (z_2 - z_1)^3$, a cubic stiffness between DOFs 1 and 2. In this case, $N = 3$ and $h = 1$.

From (9.3), g_1 can be written as:

$$g_1 = \left(\sum_{\gamma=1}^{N_R} \Phi_{2,\gamma} p_\gamma - \sum_{\gamma=1}^{N_R} \Phi_{1,\gamma} p_\gamma \right)^3 = \left(\sum_{\gamma=1}^{N_R} \left(\Phi_{2,\gamma} - \Phi_{1,\gamma} \right) p_\gamma \right)^3 = \left(\sum_{\gamma=1}^{N_R} \alpha_\gamma p_\gamma \right)^3. \tag{9.12}$$

By considering just the first mode, $N_R = 1$, (9.12) becomes

$$g_1 = \alpha_1^3 p_1^3 = \varphi_1 p_1^3 = \varphi_1 \bar{g}_1(p), \tag{9.13}$$

so $\bar{h} = 1$.

By taking into account $N_R = 2$, (9.12) turns into a more complicated expression:

$$\begin{aligned} g_1 &= (\alpha_1 p_1 + \alpha_2 p_2)^3 = \alpha_1^3 p_1^3 + 3\alpha_1^2 \alpha_2 p_1^2 p_2 + 3\alpha_1 \alpha_2^2 p_1 p_2^2 + \alpha_2^3 p_2^3 = \\ &= \varphi_1 p_1^3 + \varphi_2 p_1^2 p_2 + \varphi_3 p_1 p_2^2 + \varphi_4 p_2^3 = \sum_{j=1}^{\bar{h}} \varphi_j \bar{g}_j(p) \end{aligned} \tag{9.14}$$

and in this case $\bar{h} = 4$.

Written as in (9.14), the transformation from physical to modal coordinates can be seen as an unattractive step, since it can introduce more nonlinear terms and then

increasing difficulties and computational efforts. This is true for numerical applications in which small systems with known lumped parameters are analyzed: the class and characterisation of nonlinear terms are known and their expression in modal coordinates is in general disadvantageous. For these systems, other methods (such as those introduced in Chapters from 6 to 8, or in [68]) in physical coordinates are most suited, although the modal space representation behaves well anyway.

The present modal approach can be much effective when analyzing real, continuous (not lumped-parameter) and complex structures, for which the characterisation of distributed nonlinearities is not known but can be easily approximated with some nonlinear modal functions $\bar{g}_j(p, \dot{p})$.

Estimating the modal model

The modal coordinates $p(t)$ of equation (9.3) will in practice be obtained from the measurements of N_q measured physical coordinates z_q, where usually $N_R < N_q << N$. Then

$$p = \left(\Phi_q^T \Phi_q\right)^{-1} \Phi_q^T z_q, \tag{9.15}$$

where Φ_q is the $N_q \times N_R$ modal matrix corresponding to the measured set of responses [117].

For systems with a significant class of stiffness nonlinearities (such as polynomials, sine and piecewise linear functions) it is possible to obtain nearly linear responses as long as the excitation applied to the system has a sufficiently low amplitude, so the system behaves in an essentially linear manner. A general nonlinear stiffness element whose extension is denoted by x and restoring force by $f(x)$ can be said to be nearly linear at small displacements if

$$\lim_{x \to x_E} \frac{\mathrm{d}f}{\mathrm{d}x} = c, \tag{9.16}$$

where c is a real constant and x_E is an equilibrium point of the system. Applying excitation forces of low amplitude to systems containing such nonlinearities allows for the identification of the modal matrix Φ_q and consequently of the modal model of the underlying linear system [117].

9.1.2. Single degree of freedom approach

Starting from the equation of motion (9.5), the equation for a particular mode for a proportionally damped nonlinear system is in the form of a single degree of freedom (SDOF) system:

$$\bar{m}_\gamma \ddot{p}_\gamma + \bar{c}_\gamma \dot{p}_\gamma + \bar{k}_\gamma p_\gamma = \bar{f}_\gamma(t) + \bar{f}_{NL,\gamma}(t), \tag{9.17}$$

where

$$\bar{f}_{NL,\gamma}(t) = -\sum_{j=1}^{\bar{h}_\gamma} \bar{\mu}_{\gamma,j} \bar{g}_{\gamma,j}(p, \dot{p}). \tag{9.18}$$

p_γ is the γ-th modal displacement, $\bar{m}_\gamma$, $\bar{c}_\gamma$ and $\bar{k}_\gamma$ are the γ-th mode modal mass, damping and stiffness and $\bar{f}_\gamma$ is the applied modal force [117]. The term $\bar{f}_{NL,\gamma}$ refers to the γ-th mode internal feedback modal force and in general is a function of several modal coordinates to allow for nonlinear cross-coupling. The contribution of each of the $\bar{h}_\gamma$ nonlinear terms $\bar{g}_{\gamma,j}(p, \dot{p})$ is defined through a scalar coefficient $\bar{\mu}_{\gamma,j}$. Note that the subscript γ has been introduced here also to denote $\bar{h}_\gamma$ and $\bar{g}_{\gamma,j}(p, \dot{p})$, in order to underline the possibility to choose a suitable number and type for each of the γ-th modal nonlinear terms, since the equations are separately dealt with. This can be useful in particular for real continuous structures, for which the main nonlinear modes can be identified in more detail.
Nonproportional damping would lead to the presence of modal damping coupling terms, so (9.5) should be identified by subspace methods as a whole N_R degrees of freedom system instead of N_R SDOF systems.

Forced response

One of the main advantages of the present modal approach is that it is not required to perform a mode by mode excitation as for example in [117], so multi-exciters are not necessary. By dealing with N_R separate SDOF systems, a single point excitation on the structure is sufficient to obtain N_R modal forces, with also a little gain in terms of testing time. Moreover, as demonstrated in Chapter 7, the NSI method best performs with a zero-mean Gaussian random excitation, but is effective with many types of applied forces.

Consider the equation (9.17) for the γ-th mode, in presence of an applied modal force. Assuming that the measurements concern displacements only, the modal formulation corresponds to a state vector chosen as $x = \begin{bmatrix} p_\gamma & \dot{p}_\gamma \end{bmatrix}^T$ and to an input vector $u = \left[\bar{f}_\gamma(t) \quad -\bar{g}_{\gamma,1}(t) \quad \cdots \quad -\bar{g}_{\gamma,\bar{h}_\gamma}(t) \right]^T$, as the following:

$$\begin{bmatrix} \dot{p} \\ \ddot{p} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\dfrac{\bar{k}_\gamma}{\bar{m}_\gamma} & -\dfrac{\bar{c}_\gamma}{m_\gamma} \end{bmatrix} \begin{bmatrix} p \\ \dot{p} \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \dfrac{1}{\bar{m}_\gamma} & \dfrac{\bar{\mu}_{\gamma,1}}{\bar{m}_\gamma} & \cdots & \dfrac{\bar{\mu}_{\gamma,\bar{h}_\gamma}}{\bar{m}_\gamma} \end{bmatrix} \begin{bmatrix} \bar{f}_\gamma(t) \\ -\bar{g}_{\gamma,1}(t) \\ \vdots \\ -\bar{g}_{\gamma,\bar{h}_\gamma}(t) \end{bmatrix} \tag{9.19}$$

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} p \\ \dot{p} \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \bar{f}_\gamma(t) \\ -\bar{g}_{\gamma,1}(t) \\ \vdots \\ -\bar{g}_{\gamma,\bar{h}_\gamma}(t) \end{bmatrix} \tag{9.20}$$

and matrices A_c, B_c, C and D of (9.9) are consequently defined.

For the identification of system parameters, the computation of matrix $\bar{H}_{\gamma,E}(\omega)$ can be performed as in (6.24).

$$\bar{H}_{\gamma,E}(\omega) = \begin{bmatrix} \bar{H}_{\gamma,\gamma} & \bar{H}_{\gamma,\gamma}\bar{\mu}_{\gamma,1} & \ldots & \bar{H}_{\gamma,\gamma}\bar{\mu}_{\gamma,\bar{h}_\gamma} \end{bmatrix}. \tag{9.21}$$

Once all the N_R SDOF systems have been analyzed, the modal underlying linear system frequency response function (or "modal FRF") can be written as follows:

$$\bar{H}(\omega) = \begin{bmatrix} \bar{H}_{1,1} & & & \\ & \bar{H}_{2,2} & & \\ & & \ddots & \\ & & & \bar{H}_{N_R,N_R} \end{bmatrix}, \tag{9.22}$$

By simple calculations and using the properties of the modal matrix Φ it is possible to derive an expression for the underlying linear system FRF matrix $H(\omega)$ in physical space:

$$H(\omega) = \Phi \overline{H}(\omega) \Phi^T . \tag{9.23}$$

Note that for lumped-parameter systems, in the particular case in which $N_R \equiv N$, the expression (9.23) gives the exact complete FRF matrix $H(\omega)$, even if the force is applied on just one of the N degrees of freedom (different from a node). In general cases in which $N_R << N$, an approximation of the exact $H(\omega)$ is obtained, whose accuracy depends on the number of retained modes.

Consider now the identification of the γ-th mode nonlinear coefficients $\overline{\mu}_{\gamma,j}$. The particular case $\omega = 0$ can be easily computed from (6.25), (9.19) and (9.20):

$$\overline{H}_{\gamma,E}(\omega=0) = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} -\dfrac{\overline{c}_\gamma}{\overline{k}_\gamma} & -\dfrac{\overline{m}_\gamma}{\overline{k}_\gamma} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \dfrac{1}{\overline{m}_\gamma} & \dfrac{\overline{\mu}_{\gamma,1}}{\overline{m}_\gamma} & \cdots & \dfrac{\overline{\mu}_{\gamma,\overline{h}_\gamma}}{\overline{m}_\gamma} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\overline{k}_\gamma} & \dfrac{\overline{\mu}_{\gamma,1}}{\overline{k}_\gamma} & \cdots & \dfrac{\overline{\mu}_{\gamma,\overline{h}_\gamma}}{\overline{k}_\gamma} \end{bmatrix} \tag{9.24}$$

From (9.24), the parameter $\overline{k}_\gamma$ can be estimated from the first entry of vector $\overline{H}_{\gamma,E}(\omega=0)$ and then the $\overline{h}_\gamma$ modal nonlinear coefficients $\overline{\mu}_{\gamma,j}$ can be obtained from the other entries.

Moreover, from the eigenvalues $\lambda_{c,\gamma}$ of the system matrix A_c it is possible to obtain estimates for the eigenfrequencies $f_{n,\gamma}$ of the undamped system and for the damping ratios ζ_γ, as follows:

$$f_{n,\gamma} = \frac{\omega_{n,\gamma}}{2\pi} = \frac{|\lambda_{c,\gamma}|}{2\pi} \quad \text{and} \quad \zeta_\gamma = -\frac{\mathrm{Re}(\lambda_{c,\gamma})}{|\lambda_{c,\gamma}|} \tag{9.25}$$

Then, all modal parameters can be estimated from (9.24), (9.25) and from the following relationships:

$$\omega_{n,\gamma} = \sqrt{\frac{\overline{k}_\gamma}{\overline{m}_\gamma}} \quad \text{and} \quad \zeta_\gamma = \frac{\overline{c}_\gamma}{\overline{c}_{\gamma,crit}} = \frac{\overline{c}_\gamma}{2\sqrt{\overline{k}_\gamma \overline{m}_\gamma}} . \tag{9.26}$$

Free response

In absence of an applied force, a different type of analysis can be performed by considering the system as subject to initial conditions or an impulsive excitation. Although it is not possible to estimate the FRF matrices as in (9.22) and (9.23), and the exact modal nonlinear coefficients $\overline{\mu}_{\gamma,j}$ can not be obtained, a free

response analysis can be useful in order to perform a characterisation of modal nonlinearities, in particular for large structures, when forced tests are often uneasy.

In this case, the input vector of (9.9) is $u = \left[-\bar{g}_{\gamma,1}(t) \quad \cdots \quad -\bar{g}_{\gamma,\bar{h}_\gamma}(t) \right]^T$ and matrix B_c is

$$B_c = \begin{bmatrix} 0 & \cdots & 0 \\ \dfrac{\bar{\mu}_{\gamma,1}}{\bar{m}_\gamma} & \cdots & \dfrac{\bar{\mu}_{\gamma,\bar{h}_\gamma}}{\bar{m}_\gamma} \end{bmatrix}, \tag{9.27}$$

the equation (9.21) turns into

$$\bar{H}_{\gamma,E}(\omega) = \left[\bar{H}_{\gamma,\gamma}\bar{\mu}_{\gamma,1} \quad \cdots \quad \bar{H}_{\gamma,\gamma}\bar{\mu}_{\gamma,\bar{h}_\gamma} \right] \tag{9.28}$$

and the equivalent of (9.24) is

$$\bar{H}_{\gamma,E}(\omega=0) = [0 \quad \cdots \quad 0] - [1 \quad 0] \begin{bmatrix} -\dfrac{\bar{c}_\gamma}{\bar{k}_\gamma} & -\dfrac{\bar{m}_\gamma}{\bar{k}_\gamma} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 \\ \dfrac{\bar{\mu}_{\gamma,1}}{\bar{m}_\gamma} & \cdots & \dfrac{\bar{\mu}_{\gamma,\bar{h}_\gamma}}{\bar{m}_\gamma} \end{bmatrix} = \begin{bmatrix} \dfrac{\bar{\mu}_{\gamma,1}}{\bar{k}_\gamma} & \cdots & \dfrac{\bar{\mu}_{\gamma,\bar{h}_\gamma}}{\bar{k}_\gamma} \end{bmatrix}. \tag{9.29}$$

In case of monotonic nonlinear functions, such as polynomials, these estimated values can be used to evaluate the ratio between the j-th nonlinear modal feedback and the linear stiffness force contribution:

$$\lambda_j = \frac{\bar{\mu}_{\gamma,j}}{\bar{k}_\gamma} \frac{\max\left(\bar{g}_{\gamma,j}(p,\dot{p})\right)}{\max\left(p_\gamma\right)}, \quad \text{for} \quad j = 1, 2, \ldots, \bar{h}_\gamma \tag{9.30}$$

Then, the absolute contribution of each nonlinear term with respect to the whole nonlinear force can be evaluated in percentage as follows:

$$\delta_j = 100 \frac{|\lambda_j|}{\sum_{r=1}^{\bar{h}_\gamma} |\lambda_r|}, \quad \text{for} \quad j = 1, 2, \ldots, h_r \tag{9.31}$$

The investigation of the δ_j leads to the elimination of the minor or negligible nonlinear terms and then to a suitable choice of the $\bar{h}_\gamma$ nonlinear modal functions for further detailed identification procedures.

For example, a broad free response analysis can be performed by considering as much nonlinear terms as permitted by the computational resources; then, those corresponding to the minor values of δ_j can be discarded; eventually, the FRF matrix and the coefficients of the selected $\bar{h}_\gamma$ nonlinear modal functions can be identified with more accuracy by performing a forced response analysis, with less computational effort.

The relationships (9.25) and (9.26) are still useful to estimate the modal parameters, but in this case this is not sufficient because the estimate of $\bar{k}_\gamma$ from (9.24) is not available.

Note that in this case (9.26) are two equations in the three unknowns $\bar{m}_\gamma$, $\bar{c}_\gamma$ and $\bar{k}_\gamma$: depending on the system under analysis, a solution can be estimating the modal mass, or normalizing the parameters with respect to a unitary modal mass.

9.2. Numerical example

In order to illustrate the proposed approach, the same five degrees of freedom system used in [117] is considered: some of the modes are linear while others are nonlinear.

The system is shown in Fig. 9.1: the parameters used are $m = 1$ kg, $c = 4.8$ Ns/m, $k = 4\times10^3$ N/m and $\beta = 5\times10^9$ Nm^{-3}. It has a hardening cubic stiffness nonlinearity between Masses 2 and 4, and it is designed to be symmetric in its linear components so as to yield very simple mode shapes.

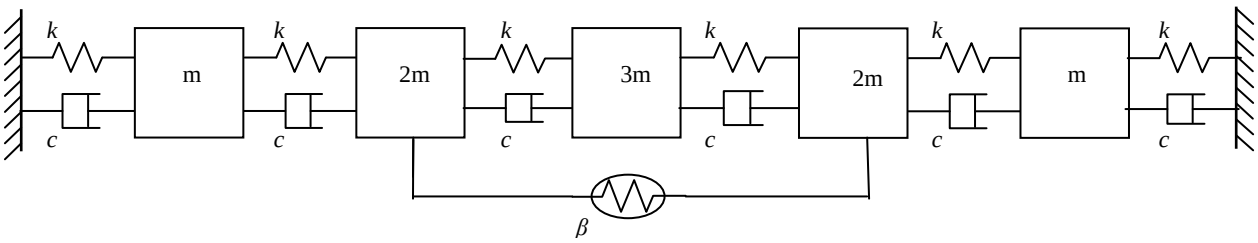


Figure 9.1. Representation of the five degrees of freedom system.

Table 9.1. Natural frequencies and damping ratios of the 5 DOF system

Mode	1	2	3	4	5
Natural frequency (Hz)	3.51	8.01	10.72	15.48	15.63
Damping ratio	0.0132	0.0302	0.0404	0.0584	0.0589

For best describing the capabilities of the proposed method, as explained in prior sections, the case $N_R \equiv N$ is considered, so $N_R = 5$. It is also assumed to know the modal matrix Φ from a previous low level identification.

The natural frequencies and damping ratios are summarized in Table 9.1. It can be seen that modes 4 and 5 are very close in frquency. The normalized mode shapes of the undamped system, corresponding to the modal matrix, are:

$$\Phi = \begin{bmatrix} 0.264 & -0.418 & -0.446 & -0.664 & 0.651 \\ 0.496 & -0.571 & -0.386 & 0.243 & -0.267 \\ 0.607 & 0 & 0.550 & 0 & 0.102 \\ 0.496 & 0.571 & -0.386 & -0.243 & -0.267 \\ 0.264 & 0.418 & -0.446 & 0.664 & 0.651 \end{bmatrix}. \tag{9.32}$$

As pointed out in [117], by writing the equations of motion in modal space it is shown that some of the modes are completely uncoupled while the modal equations for modes 2 and 4 contain nonlinear contributions.

For this system, the elements appearing in the nonlinear term of equation (9.2) are $h = 1,\ \mu_1 = \beta,\ L_j = [0 \quad -1 \quad 0 \quad 1 \quad 0]^T$ and $g_1(z, \dot{z}) = (z_4 - z_2)^3$.

Following the procedure described in (9.12), g_1 can be written as:

$$\begin{aligned} g_1 &= \left(\sum_{\gamma=1}^{N_R} \Phi_{4,\gamma} p_\gamma - \sum_{\gamma=1}^{N_R} \Phi_{2,\gamma} p_\gamma \right)^3 = \left(\sum_{\gamma=1}^{N_R} \left(\Phi_{4,\gamma} - \Phi_{2,\gamma} \right) p_\gamma \right)^3 = \left(\sum_{\gamma=1}^{N_R} \alpha_\gamma p_\gamma \right)^3 = \\ &= \left(\alpha_2 p_2 + \alpha_4 p_4 \right)^3 = \varphi_1 {p_2}^3 + \varphi_2 {p_2}^2 p_4 + \varphi_3 p_2 {p_4}^2 + \varphi_4 {p_4}^3 = \sum_{j=1}^{\bar{h}} \varphi_j \bar{g}_j(p) \end{aligned} \tag{9.33}$$

so the $\bar{h} = 4$ nonlinear modal functions $\bar{g}_j(p, \dot{p})$ are defined.

The vector $\overline{L}_j = \Phi^T L_j = [0 \quad \alpha_2 \quad 0 \quad \alpha_4 \quad 0]^T$ is derived and in the end the $\overline{h}$ vectors of nonlinear coefficients $\overline{\mu}_j$ can be obtained as in (9.11):

$$\overline{\mu}_j = [0 \quad \beta\alpha_2\varphi_j \quad 0 \quad \beta\alpha_4\varphi_j \quad 0]^T, \quad \text{for } j = 1, 2, \ldots, \overline{h} \tag{9.34}$$

Only the coefficients $\overline{\mu}_{2,j}$ and $\overline{\mu}_{4,j}$ are different from zero: this is a further proof that only modes 2 and 4 behave in a nonlinear manner.

The system is excited by a zero-mean Gaussian random input having a root mean square (r.m.s.) value of 15 N, applied only to DOF 4. Time histories have been obtained through a numerical simulation with a time step $\Delta t = 2.5 \times 10^{-4}$ s, and a total number of $s = 1.6 \times 10^4$ samples has been generated.

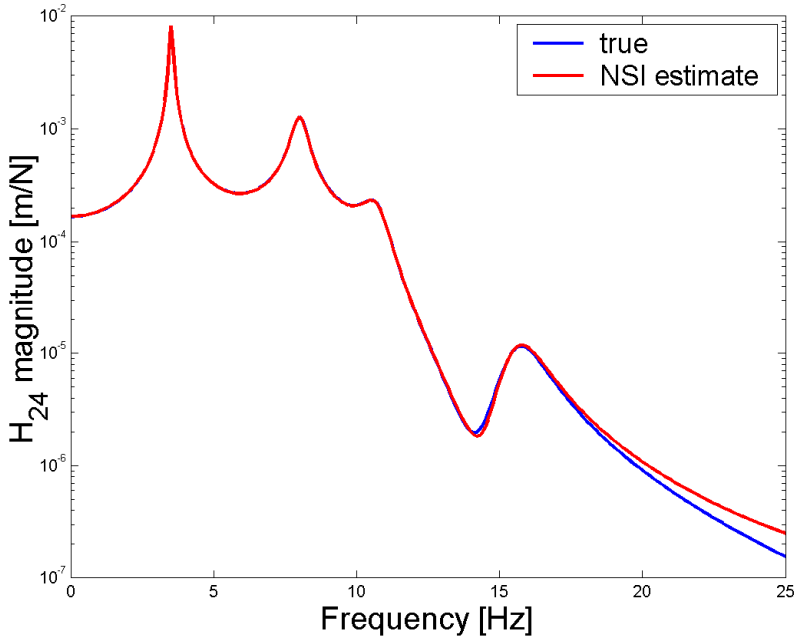
The effect of the measurement noise on the parameter estimation results is investigated by corrupting the output adding a Gaussian zero-mean noise (1% of the r.m.s. value).

The modal coordinates are obtained through the relationship (9.15) and the single degree of freedom approach is applied for each of the N_R retained modes, by performing a subspace identification with $i = 300$ block rows and s measurements. Note that the model for modes 1, 3 and 5 may be written from the low level identification stage, but it is identified here as a linear alternative of the proposed SDOF approach. The results are presented in the following.

The Frequency Response Functions (FRFs) of the underlying linear system in physical space are estimated through the procedure described in (9.23), and some of them are shown in Figs. 9.2 and 9.3; the level of accuracy of the other ones is the same. Since the excitation is applied only at DOF 4, Fig. 9.2 shows the FRFs H_{24} and H_{44}. In this example $N_R \equiv N$ is assumed, so the modal NSI method is able to estimate the complete FRF matrix: this capability is demonstrated in Fig. 9.3, by showing the FRFs H_{11} and H_{35}. In both figures the estimates and the true linear FRFs are almost overlaid: an excellent agreement can be observed.

Information about the accuracy of the modal NSI is given in Table 9.2, where estimates of natural frequencies, damping ratios and linear modal parameters are indicated, as obtained through the relationships (9.24), (9.25) and (9.26).

(a)



(b)

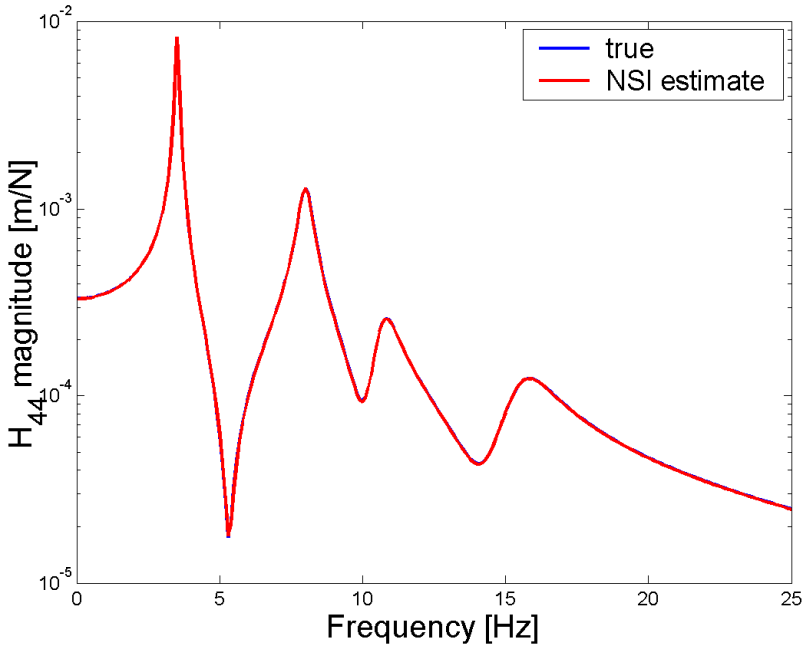
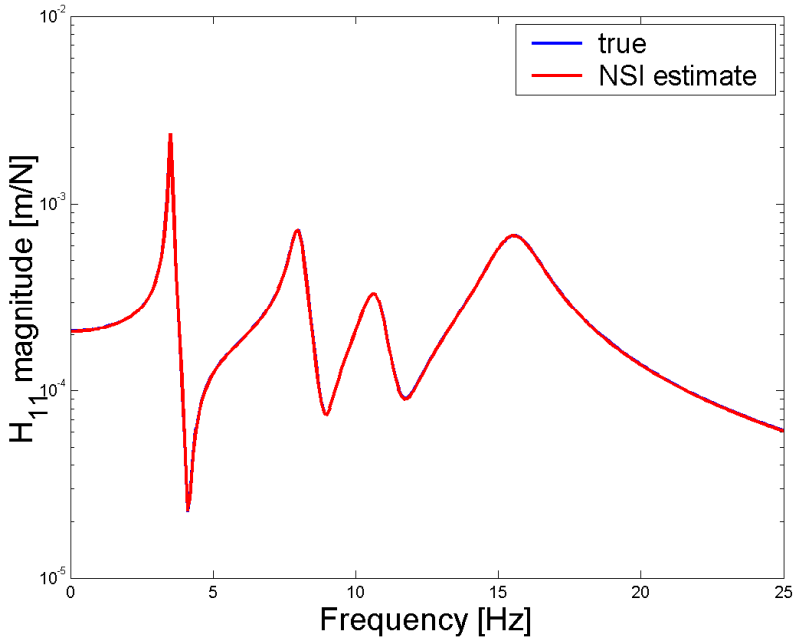


Figure 9.2. Frequency Response Functions H_{24} (a) and H_{44} (b).

(a)



(b)

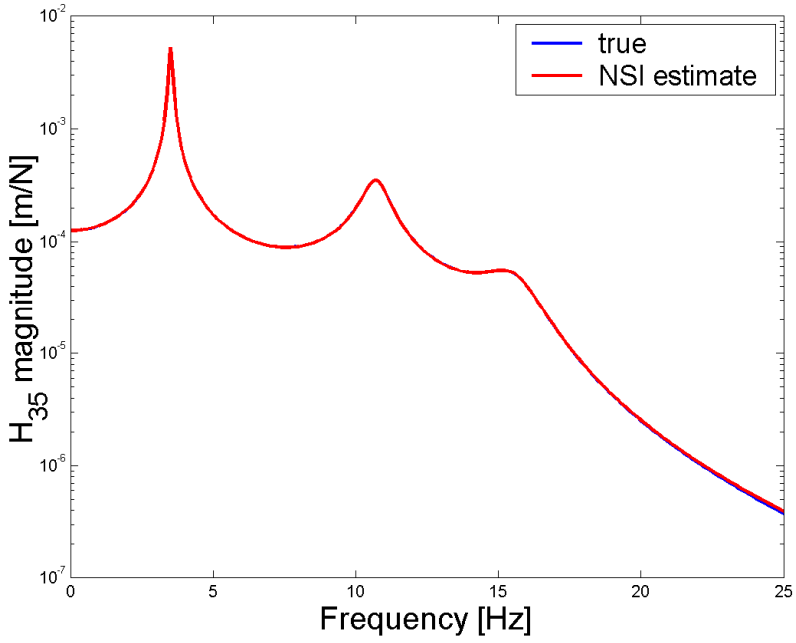


Figure 9.3. Frequency Response Functions H_{11} (a) and H_{35} (b).

Table 9.2. Estimates of natural frequencies, damping ratios and modal parameters: percentage error (100 • | estimated – actual | / actual).

Mode	1	2	3	4	5
Natural frequency	0.0009	0.0229	0.0018	0.0715	0.0042
Damping ratio	0.0883	0.1213	0.1053	1.2214	0.0628
Modal mass	0.0341	0.7444	0.1505	1.1955	0.0461
Modal damping	0.0533	0.5992	0.0468	0.1120	0.0209
Modal stiffness	0.0358	0.6983	0.1541	1.0508	0.0377

The identification results for the nonlinear modal coefficients $\bar{\mu}_{2,j}$ and $\bar{\mu}_{4,j}$ of Mode 2 and 4 are presented, respectively, in Tables 9.3 and 9.4. The estimates are accurate, except for the values obtained for $j = 4$ which seem to have a too high percentage error. This is due to the fact that the contribution of the nonlinear terms related to coefficients $\bar{\mu}_{2,4}$ and $\bar{\mu}_{4,4}$ to the whole nonlinear force is negligible w.r.t. the others, so good estimates are hardly obtained. This can be demonstrated through the application of (9.31), with the following definition of λ_j which is slightly different from (9.30):

$$\lambda_j = \bar{\mu}_{\gamma,j}\,\text{r.m.s.}\left(\bar{g}_{\gamma,j}(p, \dot{p})\right), \quad \text{for} \quad j = 1, 2, \ldots, \bar{h}_\gamma$$

The differences come from the knowledge of the exact estimate of $\bar{\mu}_{\gamma,j}$ (and not of $\bar{\mu}_{\gamma,j} / \bar{k}_\gamma$ as in (9.29)), since the applied force is measured; moreover, since the force is a zero-mean Gaussian random input, the root mean square (r.m.s.) is preferred. Then, the application of (9.31) leads to the following (observe that the contributions are the same for mode 2 and mode 4, due to the system configuration):

$$\delta_1 = 54.4636\,\%, \quad \delta_2 = 32.1937\,\%, \quad \delta_3 = 10.7921\%, \quad \delta_4 = 2.5506\,\%.$$

The value of δ_4 is less than 3%, so the error associated to the estimates of $\bar{\mu}_{2,4}$ and $\bar{\mu}_{4,4}$ is not considerably affecting the estimate of the whole nonlinear modal force.

As a further demonstration, the coefficients $\bar{\mu}_{2,j}$ and $\bar{\mu}_{4,j}$ can be used to obtain the contributions of the 2nd and 4th mode internal feedback modal force as in (9.18): a comparison between the true and estimated nonlinear functions $\bar{f}_{NL,2}$ and $\bar{f}_{NL,4}$ is made, respectively, in Figs. 9.4 and 9.5: an excellent degree of agreement is achieved.

Table 9.3. Nonlinear modal coefficients for Mode 2.

Coefficient	$\beta\alpha_2\varphi_1$	$\beta\alpha_2\varphi_2$	$\beta\alpha_2\varphi_3$	$\beta\alpha_2\varphi_4$
Actual	8.4782×10^{9}	-1.0835×10^{10}	4.6154×10^{9}	-6.5536×10^{8}
Estimate	8.4208×10^{9}	-1.0725×10^{10}	4.4872×10^{9}	-6.0429×10^{8}
Percentage error	0.6778	1.0167	2.7774	7.7916

Table 9.4. Nonlinear modal coefficients for Mode 4.

Coefficient	$\beta\alpha_4\varphi_1$	$\beta\alpha_4\varphi_2$	$\beta\alpha_4\varphi_3$	$\beta\alpha_4\varphi_4$
Actual	-3.6116×10^{9}	4.6154×10^{9}	-1.9661×10^{9}	2.7917×10^{8}
Estimate	-3.5944×10^{9}	4.5779×10^{9}	-1.9330×10^{9}	2.3719×10^{8}
Percentage error	0.4763	0.8115	1.6814	15.0366

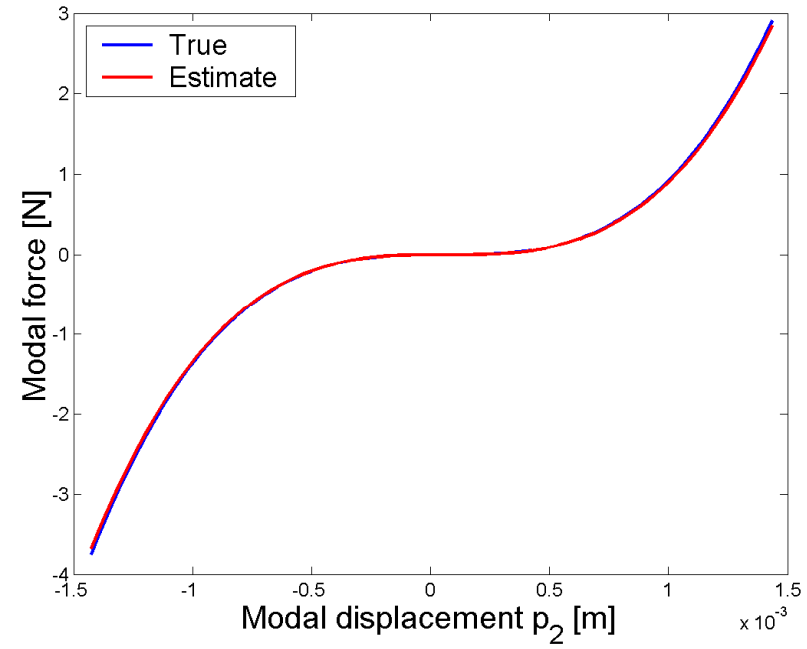


Figure 9.4. 2nd mode internal feedback modal force: true versus estimated contribution. The sign of the force is changed for a better representation.

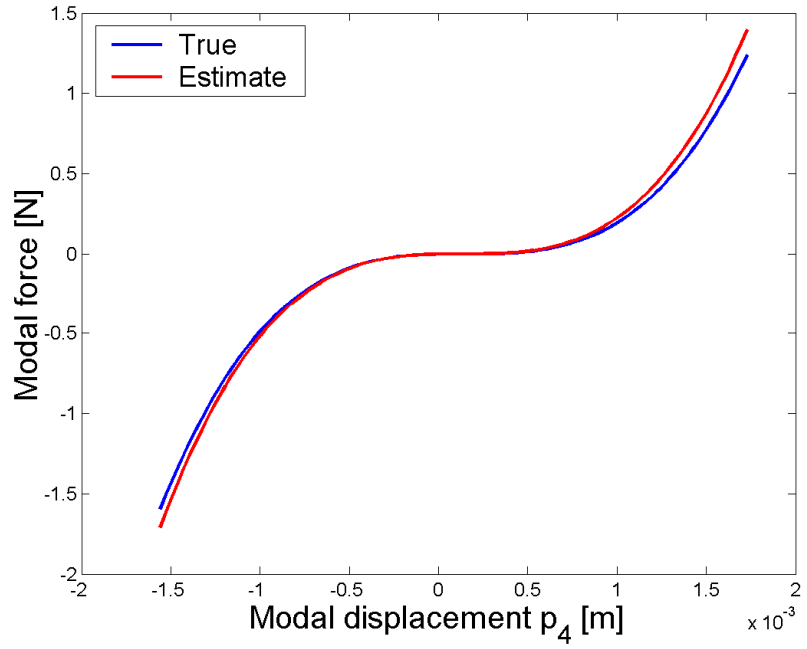


Figure 9.5. 4th mode internal feedback modal force: true versus estimated contribution.

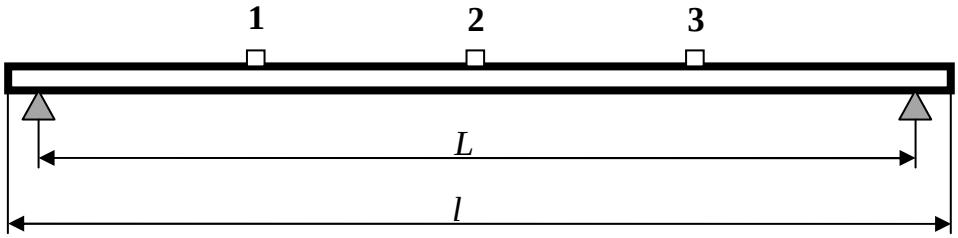


Figure 9.6. Scheme of the beam analyzed in Pescara.

Table 9.5. Characteristics of the beam.

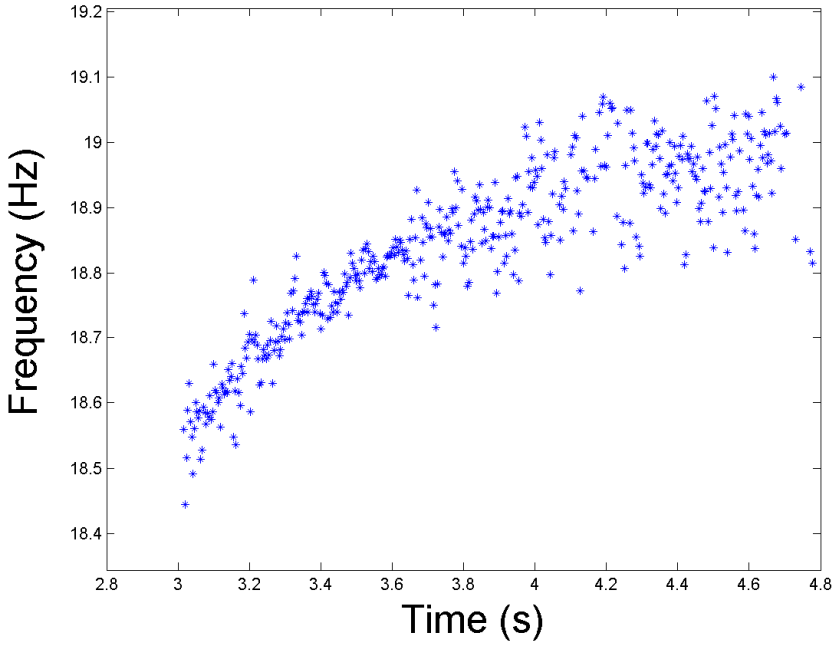
Total length	$l = 3.2$ m
Length between the supports	$L = 3$ m
Area	$A = 0.0072$ m^2
Moment of inertia	$I = 4.32 \cdot 10^{-6}$ m^4
Young modulus	$E = 50.481$ GPa
Density	$\rho = 2597.2$ kg/m^3

9.3. Experimental application

The proposed modal NSI method can be effectively applied to real, continuous structures, as demonstrated by the following experimental application.

In the framework of the project titled "Monitoring and diagnostics of railway bridges by means of the analysis of the dynamic response due to train crossing", financed by Italian Ministry of Research, some experimental tests on reinforced concrete beams have been performed in Pescara, Italy. A detailed description of the tests is given in [120]. In Fig. 9.6, a scheme of the simply supported beam under study (called T4-1, series 4, number 1) is shown, and the characteristics of the beam are presented in Table 9.5 as a general reference.

(a)



(b)

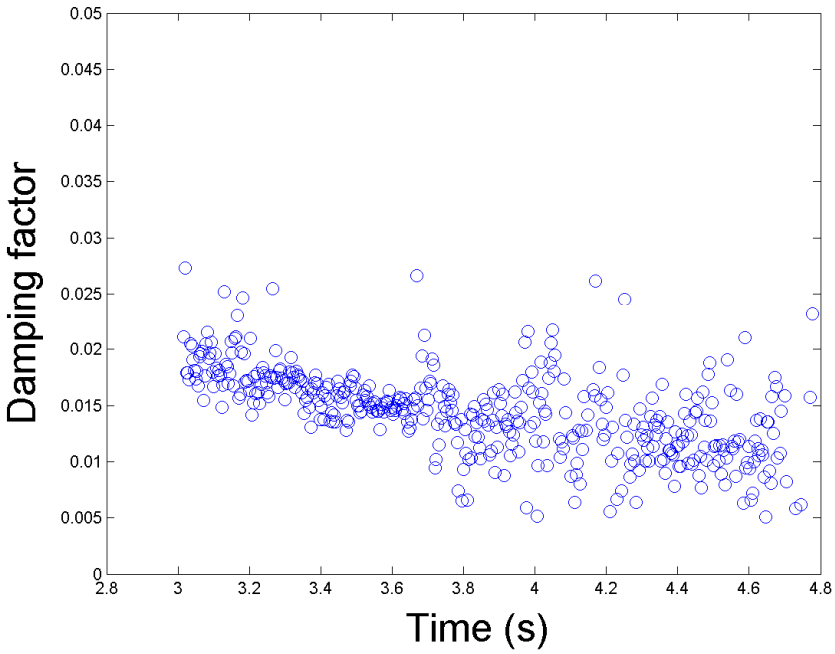


Figure 9.7. Preliminary linear analysis: first natural frequency (a) and damping factor (b) over time, during a decay.

By carrying out a preliminar study through the ST-SSI method (Section 6.4), a softening nonlinear behaviour has been observed as shown in Fig. 9.7a: during a decay, lower values of the identified first natural frequency correspond to higher signal amplitudes, i.e. those in which nonlinear effects are stronger. As time goes by, the beam behaves more and more as a linear system (and noise affects more the signal for low amplitudes): the first natural frequency should asymptotically approach the value assumed for the "underlying linear" system. A similar behaviour can be seen for damping, as shown in Fig. 9.7b. From a physical point of view, such a nonlinear behaviour is hardly explained: it is likely that a material nonlinearity is involved, maybe due to the presence of air bubbles inside the reinforced concrete.

This qualitative demonstration justifies the application of the proposed method, in order to characterise the nonlinear contribution. It is assumed that the "underlying linear" beam can be approximated by a simply supported beam, so the mode shapes of the undamped system can be obtained from the theoretical relationship

$$\phi_\gamma = \sin\left(\pi \frac{x}{L} \gamma\right), \quad \text{for} \quad \gamma = 1, \ldots, N_R \tag{9.35}$$

where L denotes the length of the beam and x is the position over the beam.

Since the contribution of other modes is negligible with respect to the first one [120], $N_R = 1$ is selected and the normalized modal matrix is:

$$\Phi = \begin{bmatrix} 0.5000 \\ 0.7071 \\ 0.5000 \end{bmatrix}. \tag{9.36}$$

The beam has been excited by an impulsive force given by lifting it a few centimetres by one end and then by releasing it, such that it bumped against the support; accelerations in points 1, 2 and 3 (placed respectively at $L/4$, $L/2$ and $3L/4$) have been measured. For the present analysis, displacements $z(t)$ are obtained through a numerical integration and shown in Fig. 9.8. Then, equation (9.15) is applied in order to obtain the modal coordinate $p_1(t)$.

A SDOF equation for the first mode is obtained as in equation (9.17); in order to perform a free response analysis as described previously, a ninth-order odd polynomial stiffness has been attempted, with the following four modal nonlinear terms (the subscript $\gamma = 1$ is omitted from now on):

$$\bar{g}_1(t) = p_1(t)^3,\ \bar{g}_2(t) = p_1(t)^5,\ \bar{g}_3(t) = p_1(t)^7,\ \bar{g}_4(t) = p_1(t)^9 \qquad (9.37)$$

A subspace identification is carried out by considering $s = 6\times10^3$ samples and $i = 150$ block rows. The matrix $\overline{H}_E(\omega = 0)$ is obtained as in (9.29); the absolute contribution of each nonlinear term with respect to the whole nonlinear force is evaluated as in (9.31) and shown in Table 9.6.

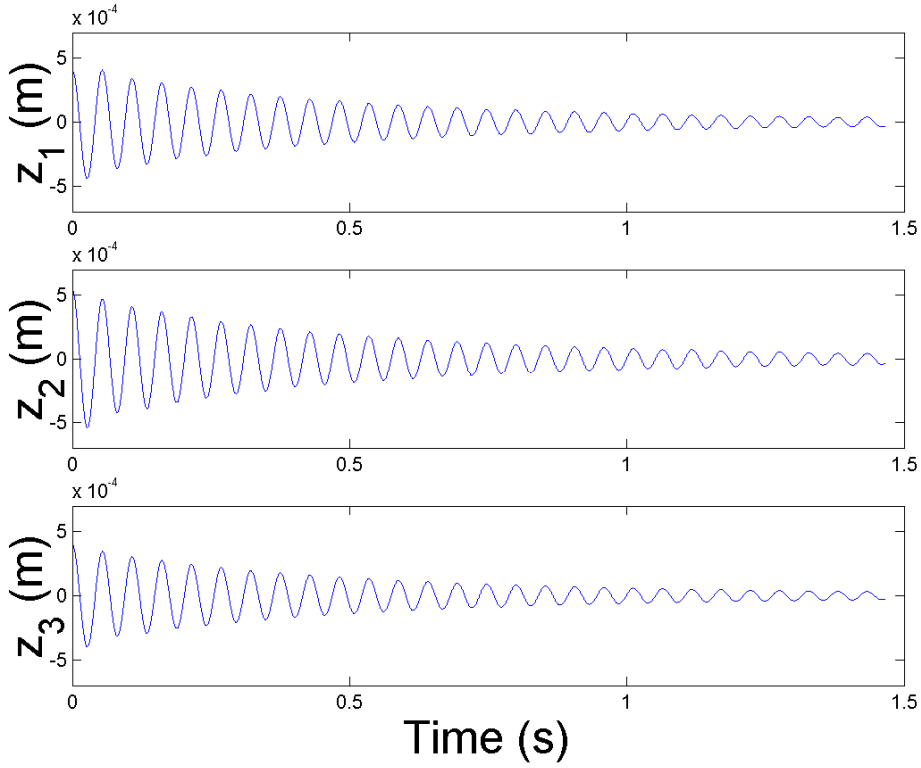


Figure 9.8. Generalised displacement vector $z(t)$ used for the present analysis.

Table 9.6. Nonlinear terms contribution to the nonlinear force, in percentage.

δ_1	δ_2	δ_3	δ_4
23.84	44.07	30.11	1.98

Since δ_4 is negligible with respect to the others, the nonlinear term $\bar{g}_4(t)$ is discarded and a new identification procedure is performed with the remaining three modal nonlinear terms $\bar{g}_1(t)$, $\bar{g}_2(t)$ and $\bar{g}_3(t)$. Estimates for the natural frequency, damping ratio and modal parameters are given in Table 9.7 for both the identification procedures carried out (in case of 4 and 3 nonlinear terms): the modal parameters are computed, by assuming that the modal mass is theoretically equal to $\bar{m} = \frac{\rho AL}{2}$. The results of a linear analysis (0 nonlinear terms included in the model) are also reported to show that a simple linear analysis would lead to wrong estimates, especially for damping and natural frequency, as can be seen by comparing them with Fig. 9.7.

Moreover, both the estimated nonlinear modal stiffness characteristics are shown in Fig. 9.9: the two nonlinear functions are almost identical, this justifying discarding the term $\bar{g}_4(t)$. The estimated modal nonlinear internal force, evaluated at the maximum measured modal displacement, is equal to 8% of the corresponding linear internal stiffness force. This measure of the weight of the nonlinear contribution with respect to the linear one seems not significant, but it is sufficient to considerably reduce the accuracy of the model.

Table 9.7. Estimates for the natural frequency, damping ratio and normalized modal parameters, in case of 4 and 3 modal nonlinear terms. The case with 0 nonlinear terms (linear analysis) is also reported.

Number of nonlinear terms	f_1 (Hz)	ζ_1	$\bar{c}_1$ (Ns/m)	$\bar{k}_1$ (N/m)
4	18.97	0.012	82.9	3.98×10^5
3	18.94	0.013	88.3	3.97×10^5
0	18.74	0.016	106.7	3.89×10^5

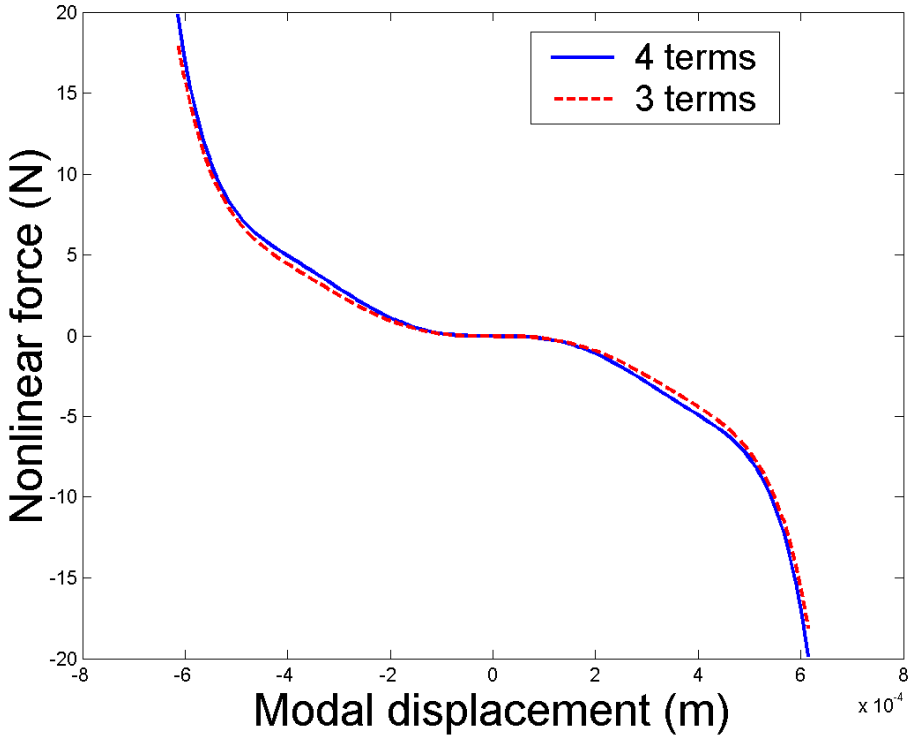


Figure 9.9. Estimates of the nonlinear contribution to the modal stiffness curve. The two nonlinear functions are almost overlaid.

In order to verify the accuracy and validity of the identified model, it is possible to perform a numerical simulation starting from the estimated parameters, by considering the system as subject to known initial conditions. This way, a reconstruction of the output is obtained and it can be compared with the measured modal displacement. In Fig. 9.10 this comparison is shown, together with the output reconstructed by carrying out a classical linear identification [51]. An excellent level of accuracy is observed for the nonlinear reconstruction, while the linear one is inadequate in estimating the amplitudes and the frequency of the system, especially towards the end of the decay as shown in detail in Fig. 9.10.

In the end, it is possible to notice that the identified model can then be used to predict the behavior of the system starting from different initial conditions. For example, a second “lift and release” excitation has been produced for the same beam, with a similar level of response. In Fig. 9.11 the measured modal displacement is compared against the one predicted by considering the previously estimated parameters. A good agreement can be observed.

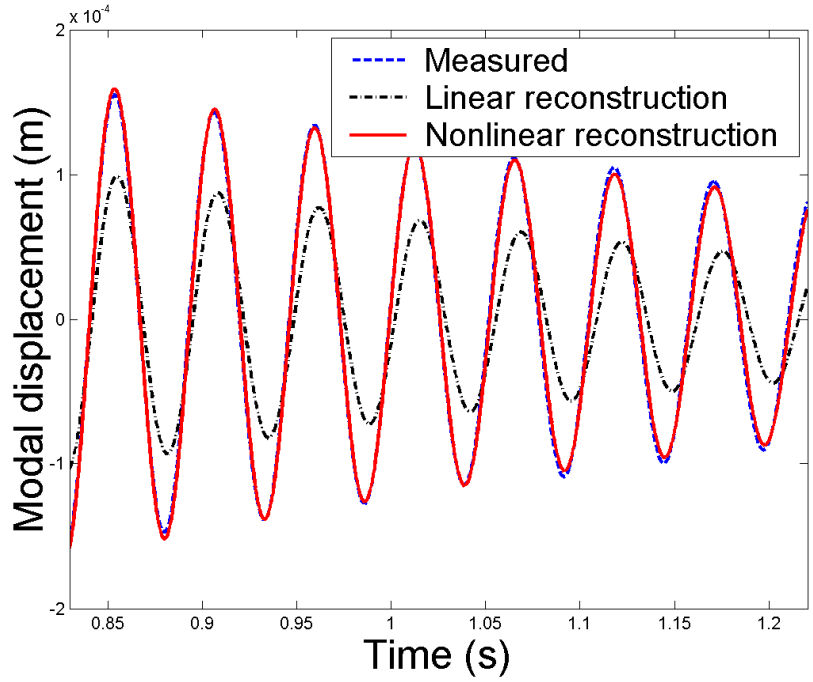


Figure 9.10. Detail of the comparison between measured and reconstructed modal displacement.

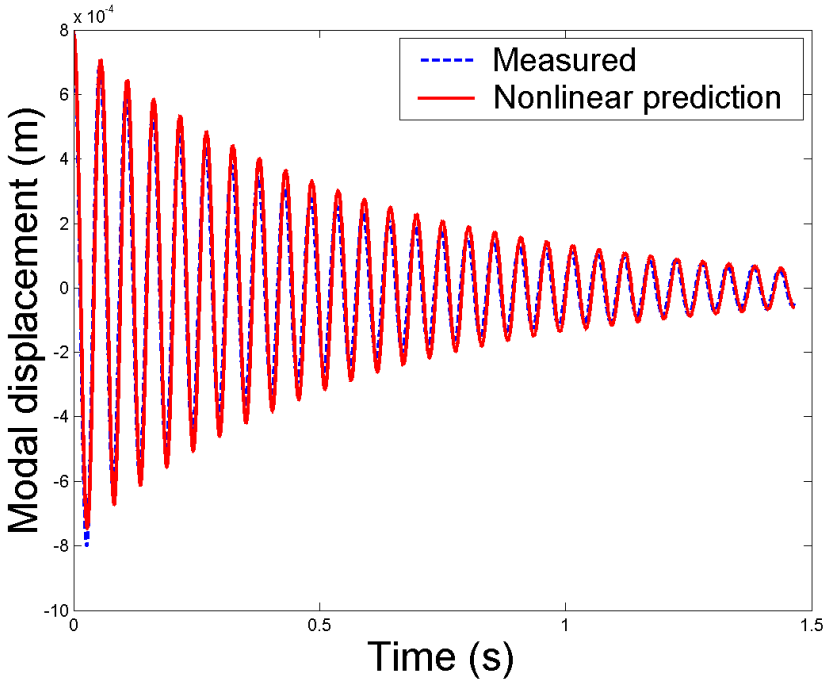


Figure 9.11. Comparison between measured and predicted modal displacement.

Chapter 10

A time-varying inertia pendulum

In this chapter two of the main sources of non-stationary dynamics, namely the time-variability and the presence of nonlinearity, are analysed through the analytical and experimental study of a time-varying inertia pendulum. The pendulum undergoes large swinging amplitudes, so that its equation of motion is definitely nonlinear, and hence becomes a nonlinear time-varying system.

The analysis and simulation [121] of mechanical systems with imposed relative motion of components is challenging: time-varying inertia, created by a part that slides along a rotating member, reveals the Coriolis-type effects present in the system. This relative movement can excite but also reduce the structure's vibration, providing new means or techniques for active attenuation of vibrations. An example of such a technique, in which a mass moving radially is treated as a controller to attenuate the pendulum oscillations, was demonstrated in [122].

The concept of controlling the motion of a system through mass reconfiguration has been examined in [123] using a variable length mathematical pendulum. The control of angular oscillations is accomplished by sliding the end mass towards and away from the pivot. A variable length pendulum has also been considered in [124], where a rigorous qualitative investigation of its equation is carried out without any assumption on small oscillations. The exact and approximate study of the nonlinear pendulum can be found in various recent papers; most of them deal with obtaining analytical approximate expressions for the large-angle pendulum period [125, 126]. Among the few papers devoted to obtaining approximate solutions (the angular displacement as a function of time), [127] derives an accurate expression in terms of elementary functions.

10.1. Experimental set up

The structure under testing is a pendulum with time-varying inertia: a disk on a cart can travel along it through a runner, while the pendulum is swinging. Moreover, this structure cannot be considered simply as a linear time-variant system, since for large swinging amplitudes the equation of motion of the pendulum has to be considered as nonlinear.

10.1.1. Description

An overview of the design of the structure is presented in this section, together with a description of the instrumentation used for acquiring data. The measured characteristics of the considered elements, such as mass and dimensions, are defined in next section, where the equation of motion is introduced.
A CAD model of the whole structure is shown in Fig. 10.1, in which the main supports are observable. The pendulum is constituted by an aluminium runner along which a cart can slide, as represented in Fig. 10.2. The cart has two screw holes for mounting an added mass which can slide on the pendulum, thus varying its inertia.
Moreover, in order to avoid a non optimal clamp between the runner and the shaft due to the large deformability of aluminium, a small plate has been added to the system to enforce the clamp.
The travelling mass is a steel disk, whose motion is regulated by a counterbalancing mass driven by hand without affecting the pendulum swing. This counterweight is connected to the moving mass through a system of pulleys and a cable that can be considered as non-extendable. The complete structure is shown in Fig. 10.3, with the sensors used and described in next section. The main supports, plates, pulleys, bearings and precision shaft are observable.

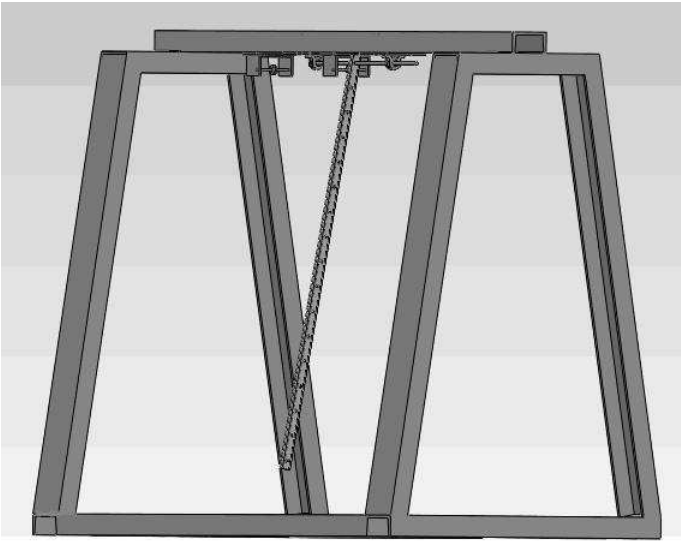


Figure 10.1. CAD model of the whole structure, including the pendulum and its supports.

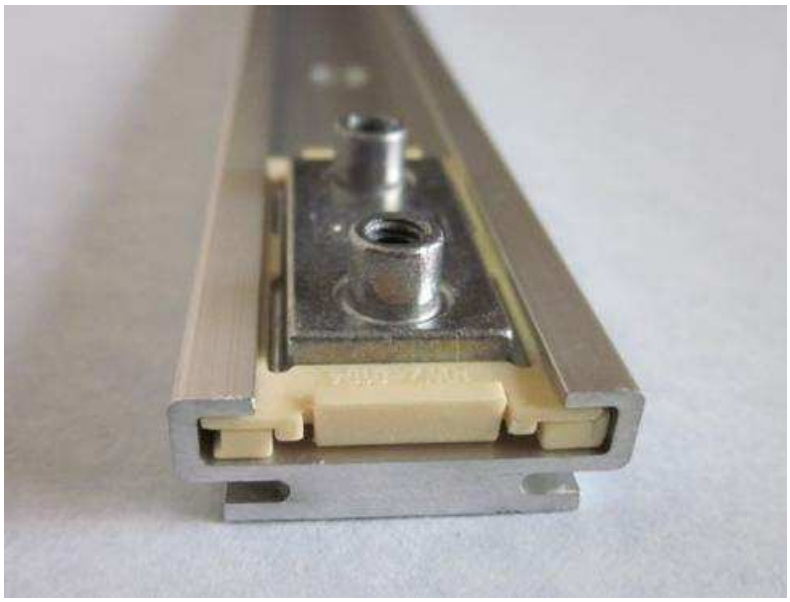


Figure 10.2. Detail of the runner and the cart.

Figure 10.3. Complete structure. At the top, the plate added to the system to enforce the clamp is highlighted. The travelling mass and the counterweight are observable on the left and on the right, respectively.

Instrumentation

The sensors can be seen in Fig. 10.3. A triaxial and four monoaxial accelerometers have been mounted along the beam: information about their masses and positions is given in Table 10.1. Their characteristics are useful for elaborating some considerations about the parameter updating that will be performed in Section 10.2.1.

The triaxial accelerometer is a PCB 356B18 piezoelectric sensor (ICP). This sensor is used to demonstrate some practical considerations in Section 10.1.2. To show typical measurement errors, some data sets have been acquired by adding a capacitive accelerometer to the system, in the same position of the ICP sensor.

Each monoaxial accelerometer is a Brüel&Kjær 4507 B 004 piezoelectric sensor. These sensors are used to measure the transversal vibrations of the pendulum for performing the analysis of the flexural vibrations, which is not contained in this work (it can be found in [128]).

Table 10.1. Characteristics of the accelerometers.

Type	Mass	Distance from the pivot point
Monoaxial	$m_{mono} = 0.0046$ kg	$s_{mono,1} = 0.205$ m $s_{mono,2} = 0.525$ m $s_{mono,3} = 0.750$ m $s_{mono,4} = 0.980$ m
Triaxial	$m_{tri} = 0.0243$ kg	$s_{tri} = 0.930$ m

A direct measure of the angular position of the pendulum is given by a Penny+Giles SRS280 sealed rotary sensor, with an accuracy of ±1% over 100°, connected to the precision shaft.
A Celesco PT1A linear potentiometer, with a maximum extension of 1.2 m, has been connected to the counterweight (see Fig. 10.3). The position of the travelling mass along the runner can be simply obtained from this measure.
All signals have been acquired with a sampling frequency of 256 Hz. The signals have been measured by using an OROS acquisition system, with 32 channels and anti-aliasing filter.

10.1.2. Considerations about accelerometers

In this section some considerations about the signals measured by the accelerometers are proposed, together with comparisons with those acquired directly from the rotary sensor.
In the following, $\theta(t)$ is the output of the potentiometer, which is very accurate at these low frequencies; thus using supplementary sensors is not necessary to describe the dynamics of the SDOF system. However, accelerometers are mounted to give some useful guidelines in case a potentiometer would not be available.

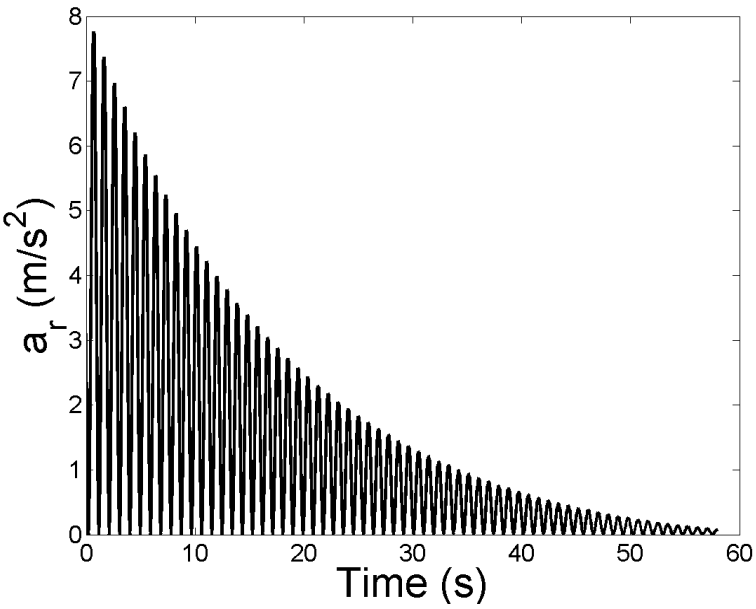


Figure 10.4. Actual value of the radial acceleration.

A piezoelectric sensor (ICP) is not suited to measure the radial acceleration of the pendulum under exam, because it removes the DC component of the output, which is non-null. This is why a capacitive accelerometer was chosen for comparison.

To show this, the two accelerometers have been mounted on the beam in the same position $p_a = 0.93$ m. Let's focus on the radial direction. The signal $\theta(t)$ of the rotary sensor has been numerically differentiated in order to obtain $\dot{\theta}(t)$, which is used for computing the "actual" value of the radial acceleration $a_r(t) = p_a \dot{\theta}(t)^2$, shown in Fig. 10.4.

The signals acquired by the accelerometers are represented in Fig. 10.5: the ICP measurement $\tilde{a}_{r,ICP}(t)$ has zero mean and its value is zero for $\theta = 0$ and $\dot{\theta} = 0$ at the end of time history, while the capacitive sensor output $\tilde{a}_{r,DC}(t)$ is asymmetric and its value tends to g for $\theta = 0$ and $\dot{\theta} = 0$. Clearly, none of the two behaviours can be associated with the actual value of radial acceleration.

Another consideration arises from this figure: the effect of the acceleration of gravity g on the measured signals must be taken into account and removed in order to get the correct value of the radial acceleration. This is due to the fact that the measurement axes of the accelerometer on the pendulum have an orientation that changes considerably over time, while most of dynamics applications do not show such a behaviour.

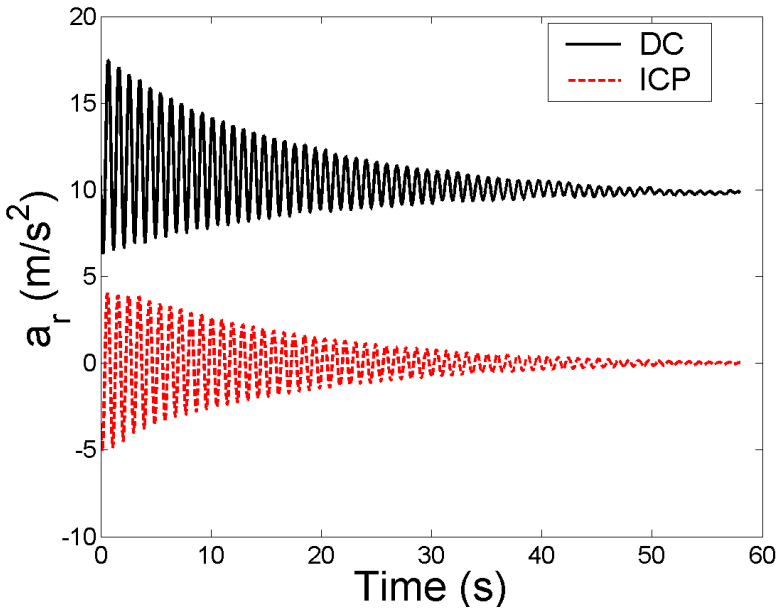


Figure 10.5. Measured accelerations in radial direction, by the DC and the ICP sensor.

A "cleaning" operation can be thought as trivial, but it cannot be performed on signals from classical ICP accelerometers, because of the continuous component removal described above.

Then, we can consider the capacitive sensor signal as corrupted by the presence of g

$$\tilde{a}_{r,DC}(t) = a_r(t) + g\cos\theta(t)\,. \tag{10.1}$$

By cleaning this measurement, we get an estimate (indicated by $\wedge$) of the actual radial acceleration:

$$\hat{a}_{r,DC}(t) = \tilde{a}_{r,DC}(t) - g\cos\theta(t)\,. \tag{10.2}$$

Fig. 10.6 shows a comparison between the actual radial acceleration and this latter estimate: an almost perfect correspondence can be obtained.

A similar approach can be adopted if the tangential direction is considered. By differentiating again the signal $\dot{\theta}(t)$ in order to obtain $\ddot{\theta}(t)$, the "actual" value of the tangential acceleration $a_t(t) = p_a\ddot{\theta}(t)$ has been computed. The DC estimate $\hat{a}_{t,DC}(t)$ can be obtained from the measured signal $\tilde{a}_{t,DC}(t)$ as:

$$\hat{a}_{t,DC}(t) = \tilde{a}_{t,DC}(t) - g\sin\theta(t)\,. \tag{10.3}$$

The comparison between the actual tangential acceleration and the DC estimate is shown in Fig. 10.7: again, a perfect agreement can be observed.

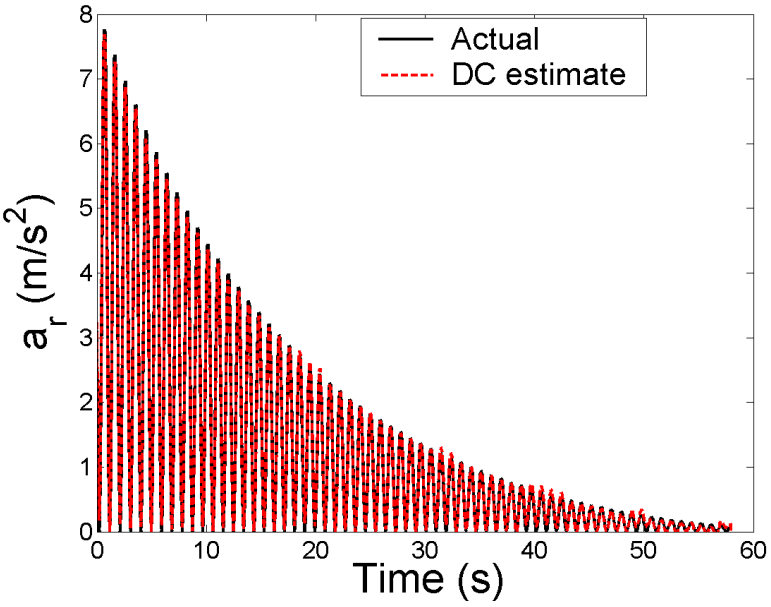


Figure 10.6. Radial direction: comparison between actual acceleration and DC estimate.

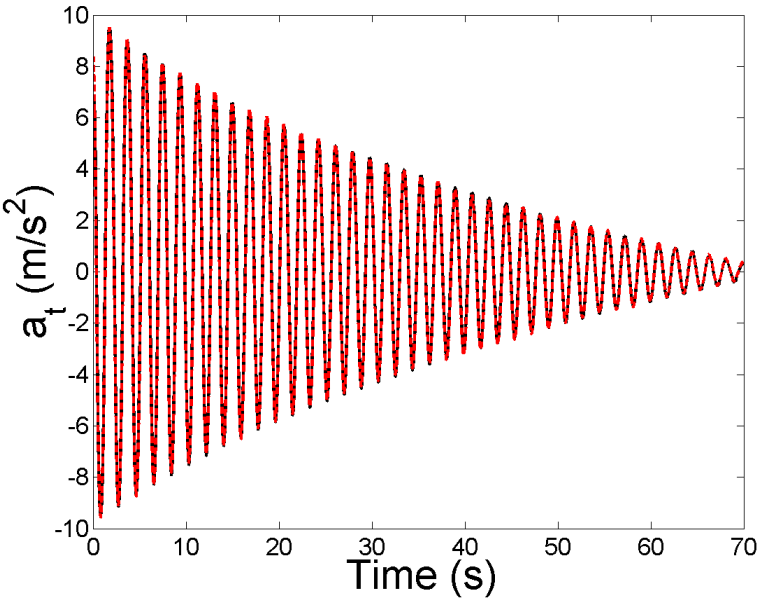


Figure 10.7. Tangential direction: comparison between actual acceleration and DC estimate.

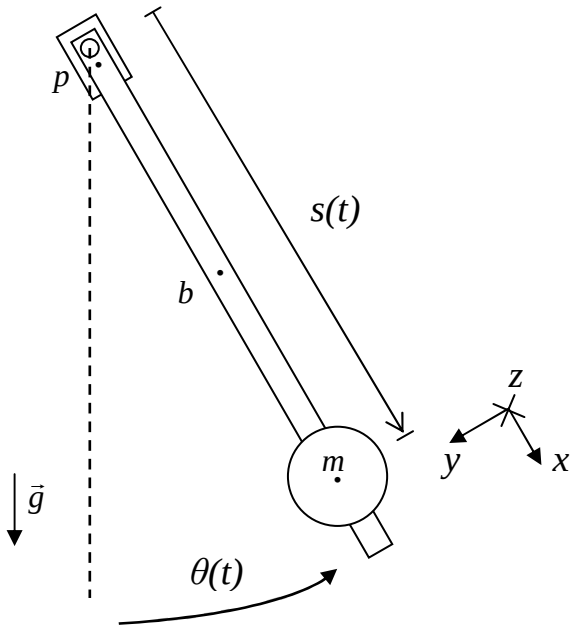


Figure 10.8. Pendulum with the travelling mass.

10.1.3. Equation of motion

The pendulum with a travelling mass is considered as vibrating in two directions: in the flexural direction (the z axis in Fig. 10.8) the system behaves as a linear continuous beam, while in the orthogonal direction y the swinging pendulum can be considered as a simple nonlinear SDOF system, since the flexural stiffness along this direction is very high. In this work only the swinging motion in the orthogonal direction is studied.

Note that the two motions can be considered as uncoupled: flexural vibrations are influenced by the effect of an axial force (tensile positive) due to gravity. For the beam of this experimental application the contribution of gravity, which depends on $\theta(t)$ (Fig. 10.8), can be considered as negligible on the basis of results in [129].

From the rotational equilibrium of the system shown in Fig. 10.8, the equation of motion can be derived as follows:

$$\left(I_{p0} + I_{b0} + I_{m0}(t)\right)\ddot{\theta}(t) + \left(c_v + 2m_m s(t)\dot{s}(t)\right)\dot{\theta}(t) + \left(m_p g d_p + m_b g d_b + m_m g s(t)\right)\sin\theta(t) = I_{tot}(t)\ddot{\theta}(t) + C_{tot}(t)\dot{\theta}(t) + P_{tot}(t)\sin\theta(t) = 0 \tag{10.4}$$

in which the subscripts p, b and m refer to the plate, the beam and the travelling mass, respectively. The angle swept by the pendulum is indicated by $\theta(t)$. Other terms appearing in (10.4) are the position $s(t)$ of the travelling mass, the acceleration of gravity $g = 9.81$ m/s^2 and a viscous damping coefficient c_v.

For each component of the system, the moment of inertia has been computed with respect to the pivot point O as follows:

$$I_{p0} = \frac{1}{12} m_p \left({a_p}^2 + {b_p}^2 \right) + m_p {d_p}^2,$$

$$I_{b0} = \frac{1}{12} m_b {L_b}^2 + m_b {d_b}^2,$$

$$I_{m0} = \frac{1}{2} m_m {r_m}^2 + m_m s(t)^2 = I_m + m_m s(t)^2.$$

The properties of the components have been measured and are reported in Table 10.2. These values are those used for preliminary comparisons between identified and "expected" results in Section 10.2.1.

When the travelling mass is fixed on the beam, then $s(t) = \bar{s}$, $\dot{s}(t) = 0$ and the following restricted forms of (10.4) are considered:

$$\begin{aligned} &\left(I_{p0} + I_{b0} + \bar{I}_{m0} \right) \ddot{\theta}(t) + c_v \dot{\theta}(t) + \left(m_p g d_p + m_b g d_b + m_m g \bar{s} \right) \sin \theta(t) = \\ &\left(I + m_m \bar{s}^2 \right) \ddot{\theta}(t) + c_v \dot{\theta}(t) + \left(P + m_m g \bar{s} \right) \sin \theta(t) = \\ &\bar{I}_{tot} \ddot{\theta}(t) + c_v \dot{\theta}(t) + \bar{P}_{tot} \sin \theta(t) = 0 \end{aligned} \tag{10.5}$$

where $I = I_{p0} + I_{b0} + I_m$ and $P = m_p g d_p + m_b g d_b$.

Table 10.2. Characteristics of the components.

Component	Mass (kg)	Sizes (m)	Centre of mass distance from O (m)
Plate	0.0713	$a_p = 0.044$, $b_p = 0.063$	$d_p = 0.01$
Beam	0.29	$L_b = 1$	$d_b = 0.5$
Travelling Mass	0.5025	$r_m = 0.05$	$s(t)$

Swinging frequency

When the swings are not "small", i.e. the linearization $\sin\theta \cong \theta$ of (10.4) is not possible, the period of oscillation of the pendulum depends on its angular amplitude.

In the following, the behaviour of the undamped pendulum in case of large swings is studied, with fixed mass, in order to achieve an analytical expression of its time-varying frequency [125]. The starting point is Eq. (10.5), with the assumption that $c_v = 0$.

Consider the energy balance for the undamped pendulum (the time dependency is omitted from now on), in which θ_0 stands for the maximum amplitude (note that $\dot{\theta} = 0$ in θ_0):

$$\frac{1}{2}\bar{I}_{tot}\dot{\theta}^2 + \bar{P}_{tot}(1-\cos\theta) = \bar{P}_{tot}(1-\cos\theta_0).$$

By using the trigonometric identity $\cos\theta = 1 - 2\sin^2(\theta/2)$, we obtain

$$\frac{1}{2}\bar{I}_{tot}\dot{\theta}^2 = 2\bar{P}_{tot}\left(\sin^2\frac{\theta_0}{2} - \sin^2\frac{\theta}{2}\right)$$

and then

$$dt = \frac{1}{2}\sqrt{\frac{\bar{I}_{tot}}{\bar{P}_{tot}}}\left(\sin^2\frac{\theta_0}{2} - \sin^2\frac{\theta}{2}\right)^{-\frac{1}{2}} d\theta. \tag{10.6}$$

By integrating both sides of (10.6) between 0 and θ_0, a quarter of the period T is obtained on the left and then:

$$T = 2\sqrt{\frac{\bar{I}_{tot}}{\bar{P}_{tot}}}\int_0^{\theta_0}\left(\sin^2\frac{\theta_0}{2} - \sin^2\frac{\theta}{2}\right)^{-\frac{1}{2}} d\theta. \tag{10.7}$$

We can now apply the following relations (note that T_0 is the period for "small" oscillations)

$$T_0 = 2\pi\sqrt{\frac{\bar{I}_{tot}}{\bar{P}_{tot}}}, \quad \sin\phi = \frac{\sin(\theta/2)}{\sin(\theta_0/2)} \tag{10.8}$$

to rewrite (10.7) as

$$T = \frac{2}{\pi} T_0 K(k), \tag{10.9}$$

where $k = \sin(\theta_0/2)$ and

$$K(k) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{1-k^2 \sin^2 \phi}}, \qquad \text{for } |k| < 1 \quad \text{and} \quad |\theta_0| < \pi, \tag{10.10}$$

is the incomplete elliptic integral of the first kind. This integral can be approximated, for example, by expanding the integrand function in power series or by using an accurate arithmetic-geometric mean [126].

The period T_0 is a function of the mass position $\bar{s}$ and then the analytical expression of the swinging frequency can be written from (10.9) as:

$$\bar{f} = f(\bar{s}) = \frac{\pi}{2} \frac{1}{T_0(\bar{s}) K(k)}. \tag{10.11}$$

In order to derive an analytical representation of frequency for large oscillations in presence of small damping, (10.11) is simply extended by considering the swinging frequency as "instantaneous": for each value of time t the pendulum is seen as a new system having a new maximum amplitude θ_0, which implies a new value of $K(k)$ and consequently a new value of $\bar{f}(t)$ in (10.11).

The meaning of $\theta_0(t)$ is shown in Fig. 10.9: it can be seen as a time-varying maximum amplitude. Note that $\theta_0(t)$ can be computed (for example, by interpolation of maxima/minima) only *a posteriori*, after having full knowledge of the time history of $\theta(t)$.

Then, the new definition of frequency is given as follows:

$$\bar{f}(t) = f(\bar{s}, \theta_0(t)) = \frac{\pi}{2} \frac{1}{T_0(\bar{s}) K(k(t))}, \tag{10.12}$$

where $k(t) = \sin(\theta_0(t)/2)$.

A nonlinear effect can then be observed in (10.12): the large oscillations of the pendulum affect the elliptic integral $K(k(t))$. Moreover, this nonlinear contribution is decreasing in time, since the maximum amplitude is reduced by damping, and the frequency tends to the value $f_0 = 1/T_0(\bar{s})$ assumed for small swings. An example of this latter effect is given in Fig. 10.10, where (10.12) is used to compute the frequency for the mass in a fixed position $\bar{s} = 0.5$ m.

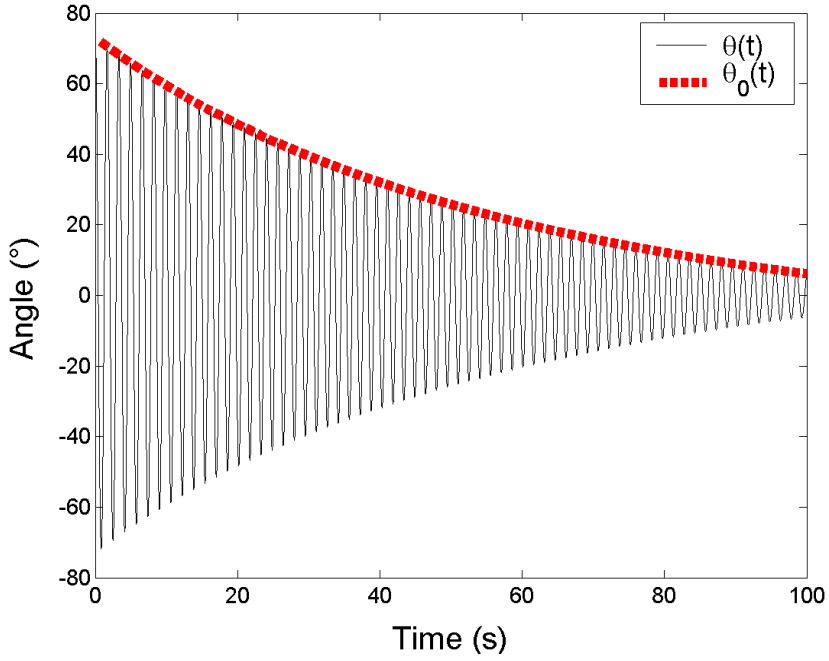


Figure 10.9. Representation of the angle $\theta(t)$ and the maximum amplitude $\theta_0(t)$.

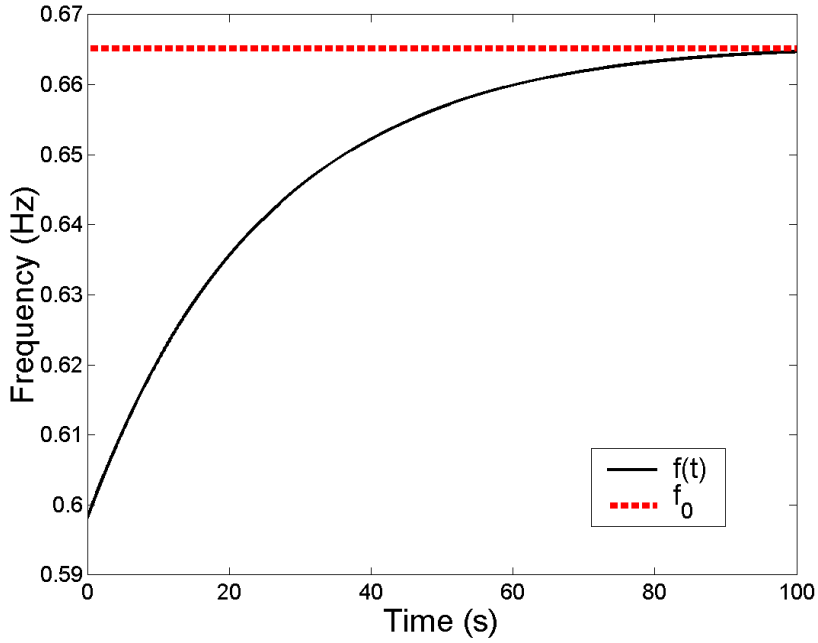


Figure 10.10. Example of frequency of oscillations, with the mass in a fixed position $\bar{s} = 0.5$ m.

The problem of estimating the frequency of a nonlinear system as a function of time, and also of the maximum amplitude of swings, can be handled by using other approaches. As an example, by approximating the term $\sin\theta$ with its Taylor series expansion, a nonlinear system having a general polynomial nonlinearity can be obtained. Then, many different analytical methods can be applied, as described in [130], in order to derive a relationship between the amplitude and frequency of the oscillator. With these techniques, such as the Variational Approach (VA), the Energy Balance Method (EBM) or the Hamiltonian Approach (HA), the same results are attained and can be considered as approximations of (10.12). In the present case, however, the elliptic integral approach is preferable since it is an *ad hoc* procedure for the nonlinear term $\sin\theta$ of the pendulum and it can also be associated to a precise physical meaning of time-varying frequency.

Baseline frequencies for fixed mass positions

Equation (10.12) is useful as an analytical "expectation" that can be adopted for comparisons with the identified results of Section 10.2. In particular, in order to analyze the cases with the travelling mass, a representation of some curves, for fixed values of $\bar{s}$, can be used as a baseline grid. This is shown in Fig. 10.11, in which the frequencies are plotted as a function of the maximum amplitude of swing, for 10 equally spaced mass positions $\bar{s}$. It can be observed that the 10 curves have the same behaviour, as expected. Moreover, if a fixed value of $\bar{\theta}_0$ is considered, the frequencies are not monotonic with the position $\bar{s}$. In fact, they are increasing for values of $\bar{s}$ between 0.95 m and 0.25 m, but then they start decreasing for $\bar{s} = 0.15$ m and 0.05 m. This is due to the values assumed by the characteristics of the components (Table 10.2) and is confirmed by showing in Fig. 10.12 the behaviour of the small-swings frequency f_0 as a function of $\bar{s}$:

$$f_0(\bar{s}) = \frac{1}{2\pi}\sqrt{\frac{P + m_m g\bar{s}}{I + m_m \bar{s}^2}}\,. \tag{10.13}$$

This function has a maximum at about $s = 0.237$ m so that there are values of frequency that can be associated to two different mass positions. This might lead to extra difficulties in interpreting the results.

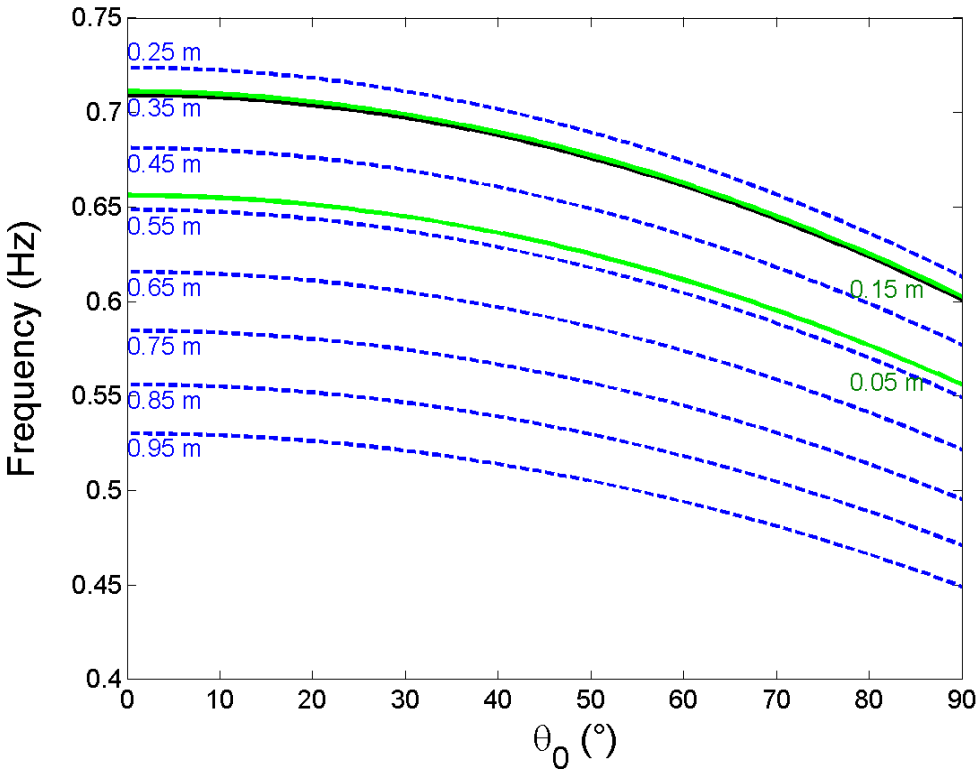


Figure 10.11. Representation of the time-varying frequencies as a function of the maximum amplitude $\theta_0(t)$, for different fixed mass positions $\bar{s}$ (indicated in meters on the figure).

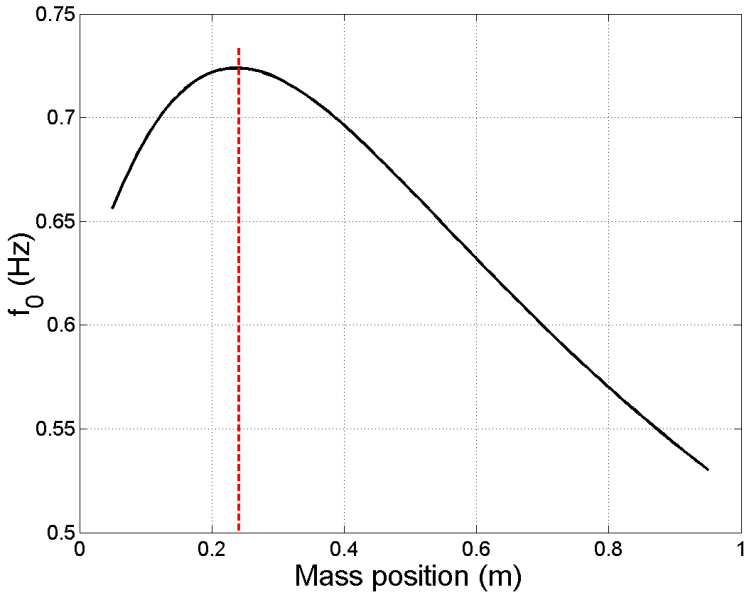


Figure 10.12. Representation of the frequency f_0 as a function of $\bar{s}$. The function has a maximum at about $\bar{s} = 0.237$ m.

10.2. Motion of the pendulum

The swinging motion has been analysed by applying the NSI and the ST-SSI methods separately. This is due to the main drawback affecting them when studying such a time-variant *and* nonlinear system: both methods cannot distinguish the time-variant contribution from the nonlinear one, since the effect of both contributions consists of swinging frequency variations.
As a consequence, the study of the swinging motion has been divided into two parts. In the first one, dedicated to the fixed mass case, the NSI method is used to identify the nonlinear contribution and the "underlying" linear frequency: to validate the method, the results are compared with the known actual values. Then, the system parameters are updated to build a new NSI-based model. In the second part, which focuses on the swings in presence of a moving mass, the ST-SSI method (introduced in Section 6.4) is applied: as a reference, the baseline frequencies described above are computed by using the previously updated parameters, so that all information given in Table 10.2 is no more needed (except for the value m_m of the travelling mass).

10.2.1. Fixed mass

In this section, the NSI method is applied in order to identify the system parameters as only depending upon the nonlinear effects due to the large oscillations. To this aim, the travelling mass has been fixed along the beam in two different positions, namely $\bar{s} = 0.91$ m and $\bar{s} = 0.5$ m.
Relying on the values resulting from both configurations, a parameter updating can be done and adopted for validating the identified nonlinear models.

NSI method

With the mass fixed on the beam, the starting point is (10.5). Moreover, in order to apply the NSI method, it is useful to consider the Taylor expansion of the sine (for $\theta \cong 0$):

$$\sin\theta \cong \theta - \frac{1}{6}\theta^3 + \ldots + \frac{(-1)^n}{(2n+1)!}\theta^{2n+1} + o(\theta^{2n+2}) \tag{10.14}$$

By truncating the expansion in (10.14) to $n = 2$ (for this value the level of accuracy is excellent, when $|\theta| < \pi/2$), (10.5) can be written as:

$$\bar{I}_{tot}\ddot{\theta}(t) + c_v\dot{\theta}(t) + \overline{P}_{tot}\left(\theta(t) + \beta_1\theta(t)^3 + \beta_2\theta(t)^5\right) = 0 \tag{10.15}$$

where $\beta_j = \dfrac{(-1)^j}{(2j+1)!}$ for $j = 1,2$ are the coefficients of the nonlinear terms in the Taylor expansion.

In this way, the linear part of the equation and the nonlinear feedback force can be separated as seen in Section 6.1.1:

$$\bar{I}_{tot}\ddot{\theta}(t) + c_v\dot{\theta}(t) + \overline{P}_{tot}\theta(t) = -\beta_1\overline{P}_{tot}\theta(t)^3 - \beta_2\overline{P}_{tot}\theta(t)^5 \tag{10.16}$$

From (10.16), the discrete state-space model defined in Section 6.1.2 can be identified by means of subspace methods, by only using the system output vector $y = \theta(t)$ measured by the rotary sensor described in Section 10.1.1 and the input (feedback forces) $u = \left[-\theta(t)^3 \quad -\theta(t)^5\right]^T$. The natural frequency of the "underlying" linear system (i.e. linear part of the equation of motion) can be extracted by calculating the eigenvalues of the identified matrix A: in the case of a pendulum, the linear frequency sought for is equal to the frequency f_0 of "small" swings.

The identification of the nonlinear coefficients should be carried on as follows, by exploiting the method used in Section 9.1.2 and the particular form of the nonlinear coefficients defined through (10.14) and (10.15). In fact, they are defined as $\mu_j = \beta_j\overline{P}_{tot}$ for $j = 1,2$, so they are both dependent on $\overline{P}_{tot}$. Equation (9.29) is exploited, which in this case turns into $H_E(\omega) = \left[H\mu_1 \quad H\mu_2\right]$, where $H(\omega)$ is the FRF of the "underlying" linear system.

In particular, when $\omega = 0$ then $H(0) = \overline{P}_{tot}^{-1}$ and an estimate of the coefficients defining the Taylor expansion of the sine in (10.14) can be obtained as

$$H_E(\omega = 0) = \left[\beta_1 \quad \beta_2\right]. \tag{10.17}$$

The NSI method as described above has been repeatedly applied to several time records where the nonlinear contribution was more important. N overlapping time windows of 30 seconds each, covering a range of decreasing amplitudes from 70 to 45 degrees, have been selected: in particular, $N = 20$ for the case $\bar{s} = 0.91$ m and $N = 11$ for $\bar{s} = 0.5$ m, since for the latter the range of amplitudes is swept faster. The results are presented hereafter.

The identified swinging frequencies of the "underlying" linear system with fixed mass in $\bar{s} = 0.91$ m and $\bar{s} = 0.5$ m are shown in Fig. 10.13 and Fig. 10.14, respectively. For each window, an estimate of the frequency is obtained and a comparison with the expected value f_0, computed from the nominal values in Table 10.2, is also given. At this stage, the objective is not to attain a perfect identification, since the knowledge of the expected value is inaccurate due to errors in any of the mass, length or position of the components. Here, it is only verified that the identified values are placed along a constant line, thus validating the removal of the nonlinear contribution carried out by the NSI method. In next section a parameter updating procedure based on these identified values will be performed in order to remove the bias observed especially in Fig. 10.14.

To reconstruct the nonlinear terms, the Taylor expansion in (10.14) of the sine is used, with the estimated values β_1 and β_2 obtained through (10.17). Figs. 10.15 and 10.16 show a comparison between the actual value of $\sin\theta$ and the estimated Taylor expansion, for the first window (i.e. largest amplitudes). Note that similar results, in terms of accuracy, are obtained for each of the windows used. In Fig. 10.15 the case $\bar{s} = 0.91$ m is presented: good agreement can be seen, with an error of 2% in correspondence with the maximum value of θ. In Fig. 10.16 the case $\bar{s} = 0.5$ m is presented: excellent agreement can be seen, with an error of 1%.

These results about the nonlinearity further demonstrate the effectiveness of the performed identification procedures, even if the estimated frequencies are not in perfect agreement with the expected values.

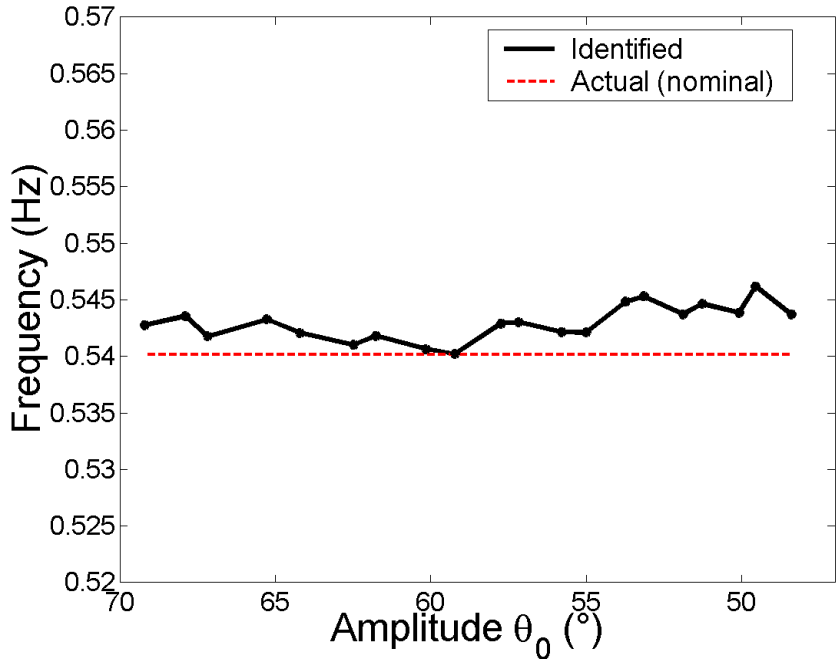


Figure 10.13. Estimates of the frequency of oscillations, for the case $\bar{s} = 0.91$ m. The "expected" value f_0 is represented with a dashed line.

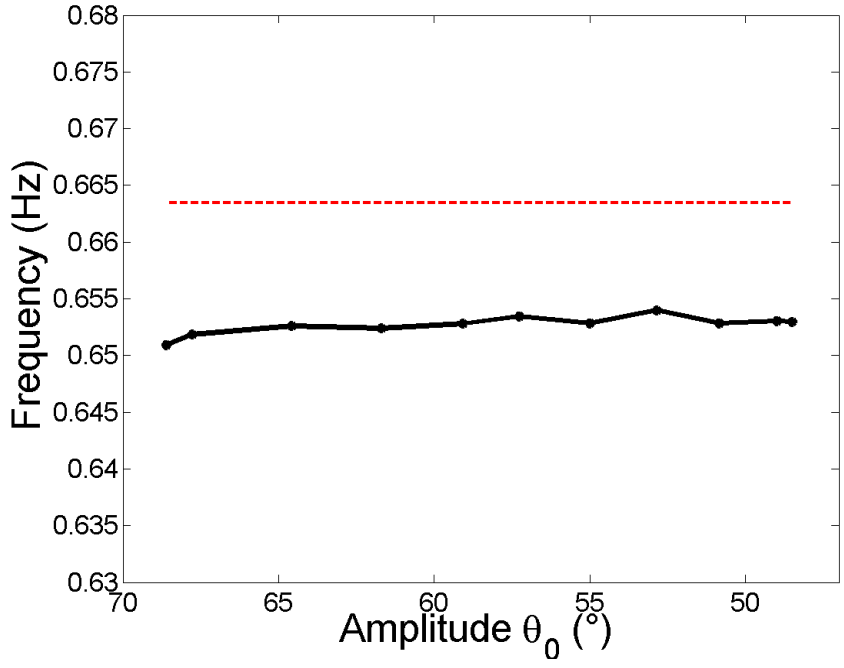


Figure 10.14. Estimates of the frequency of oscillations, for the case $\bar{s} = 0.5$ m. The "expected" value f_0 is represented with a dashed line.

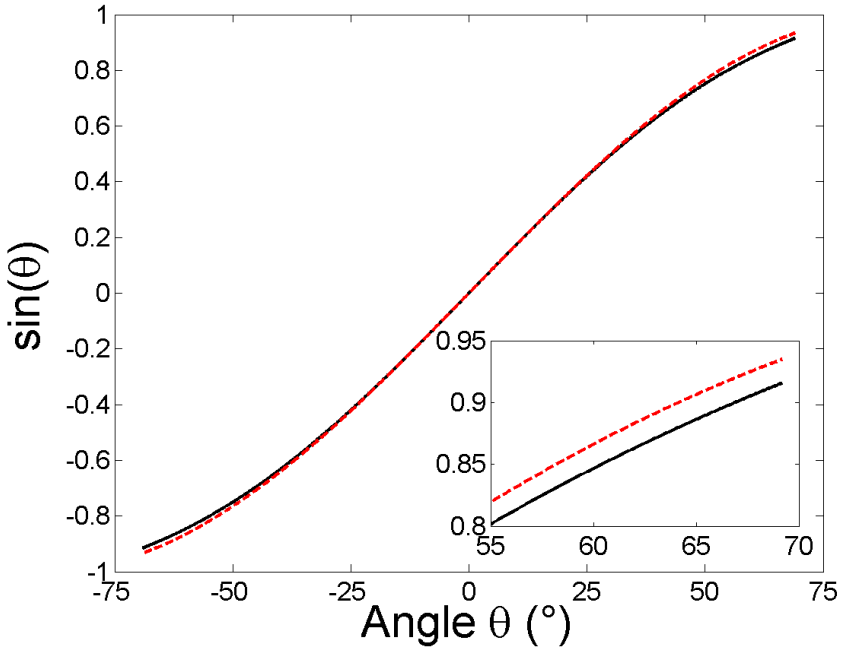


Figure 10.15. Estimates of the sinusoidal term, for the case $\bar{s} = 0.91$ m. The actual value $\sin\theta$ is represented with a dashed line. A magnification for large amplitudes is also shown.

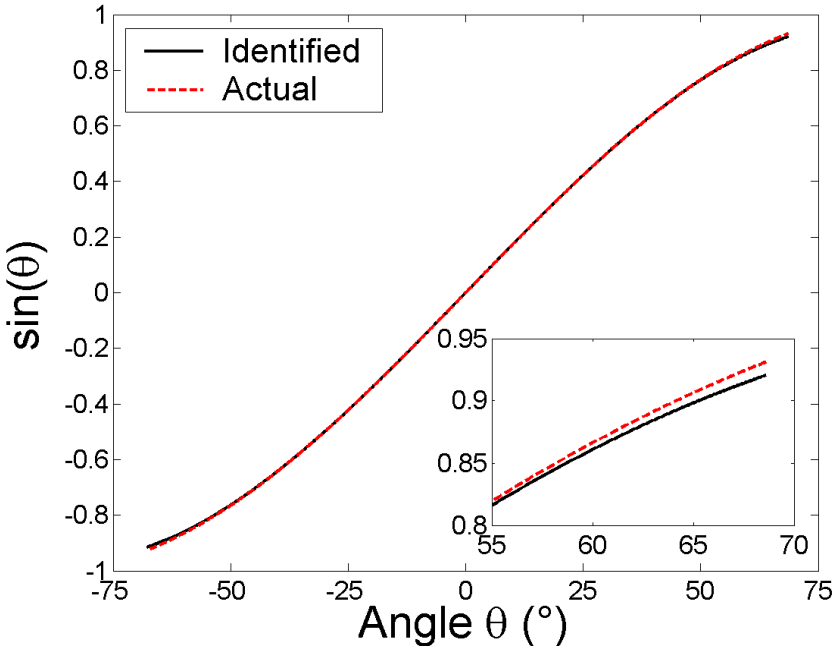


Figure 10.16. Estimates of the sinusoidal term, for the case $\bar{s} = 0.5$ m. The actual value $\sin\theta$ is represented with a dashed line. A magnification for large amplitudes is also shown.

Parameter updating and validation

As explained above, the bias seen in Fig. 10.14 may be caused by inaccuracies in the characteristics of the components or by the presence of instrumentation (accelerometers, cables). An updating procedure is then performed, in order to build a new model based on identified results.

In more detail, consider Eq. (10.13), which defines the swinging frequency for small amplitudes, and assume that the values of I and P have to be updated to fit the identified model. In the identification step, the NSI estimates $\hat{f}_{0,j} = \hat{f}_0(\bar{s}_j)$ for two fixed mass positions $\bar{s}_1 = 0.91$ m and $\bar{s}_2 = 0.5$ m have already been computed: in order to have a single value for each position, the mean value over the N estimates (Figs. 10.13 and 10.14) is computed. By assuming that the moving mass $m_m = 0.5025$ kg is also known, the following system of equations can be obtained from (10.13), in the new unknowns I_{up} and P_{up}:

$$\begin{bmatrix} -(2\pi\hat{f}_{0,1})^2 & 1 \\ -(2\pi\hat{f}_{0,2})^2 & 1 \end{bmatrix} \begin{Bmatrix} I_{up} \\ P_{up} \end{Bmatrix} = \begin{Bmatrix} (2\pi\hat{f}_{0,1})^2 m_m \bar{s}_1{}^2 - g m_m \bar{s}_1 \\ (2\pi\hat{f}_{0,2})^2 m_m \bar{s}_2{}^2 - g m_m \bar{s}_2 \end{Bmatrix}. \tag{10.18}$$

A comparison between the nominal and the updated values is given in Table 10.3: as expected, the updated quantities are higher because of the influence of instrumentation.

As a further control, the contributions I_{acc} and P_{acc} due to the accelerometers can be evaluated by considering the information given in Table 10.1:

$$I_{acc} = m_{tri} s_{tri}{}^2 + m_{mono} \sum_{j=1}^{4} s_{mono,j}{}^2 = 0.0295 \text{ kg m2},$$

$$P_{acc} = g\left(m_{tri} s_{tri} + m_{mono} \sum_{j=1}^{4} s_{mono,j} \right) = 0.3327 \text{ kg m2 s-2}.$$

These quantities are in good agreement with ΔI and ΔP shown in Table 10.3, this highlighting the contribution of the accelerometers in the dynamics of the pendulum.

In order to validate the updated model, a first step consists of computing the new values of the “expected” oscillation frequency for both the cases $\bar{s} = 0.91$ m and $\bar{s} = 0.5$ m. It is called updated frequency and it is defined through (10.13) as

$$f_{0,up}(\bar{s}) = \frac{1}{2\pi}\sqrt{\frac{\left(P_{up} + m_m g\bar{s}\right)}{\left(I_{up} + m_m \bar{s}^2\right)}}\,. \tag{10.19}$$

The updated frequency and the NSI estimates (those obtained in the previous subsection) are shown in Fig. 10.17 and Fig. 10.18: observe that now, in comparison with Fig. 10.14, the NSI estimates are much closer to the actual updated value, confirming the accuracy of the identified model.

In this way, by parameter identification (for the nonlinear terms) and updating (for the linear terms) a general model representing the experimental pendulum is defined:

$$\left(I_{up} + m_m \bar{s}^2\right)\ddot{\theta}(t) + c_v\dot{\theta}(t) + \left(P_{up} + m_m g\bar{s}\right)\left(\theta + \beta_1\theta^3 + \beta_2\theta^5\right) = 0\,, \tag{10.20}$$

As a concluding verification, the model (10.20) can be used to perform numerical simulations, starting from the same initial conditions, and compare the measured and the obtained "estimated" time history of $\theta(t)$. This is shown in Fig. 10.19 for the case $\bar{s} = 0.91$ m: an excellent level of agreement is observable. Note that the same accuracy can be obtained for all the time windows considered and also for the other case, $\bar{s} = 0.50$ m.

At this point, all information given in Table 10.2 is no more needed (except for the value m_m of the travelling mass): the updated parameters I_{up} and P_{up} can be used to compute the baseline frequencies of Eqs. (10.12) and (10.13), as a reference for the application of ST-SSI.

Table 10.3. Comparison between nominal and updated I and P.

	I (kg m^2)	P (kg m^2 s^{-2})
Nominal	0.0973	1.4294
Updated	0.1292	1.8380
Δ=Updated-Nominal	$\Delta I = 0.0319$	$\Delta P = 0.4086$

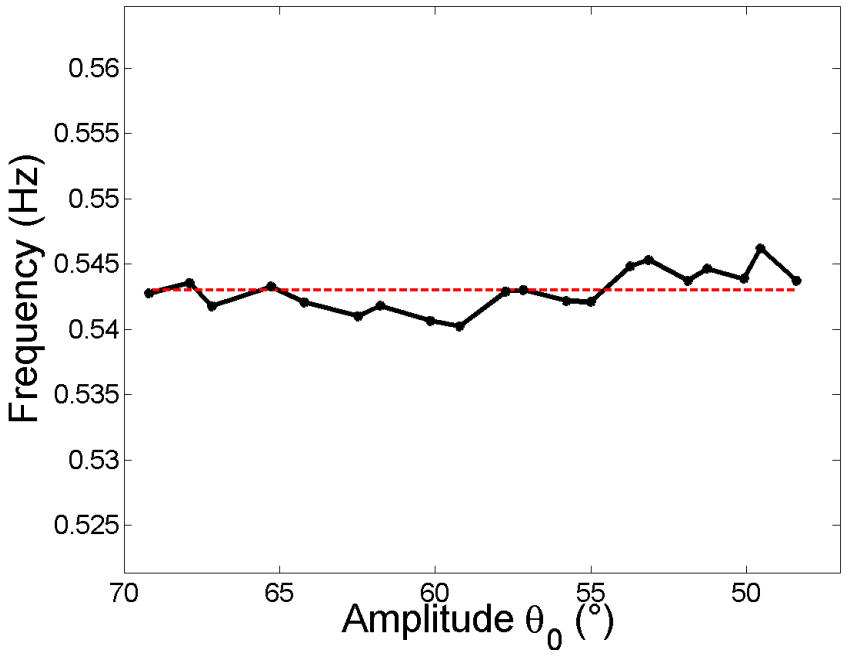


Figure 10.17. Estimates of the swinging frequency, for the case $\bar{s} = 0.91$ m. The updated value $f_{0,up}$ is represented with a dashed line.

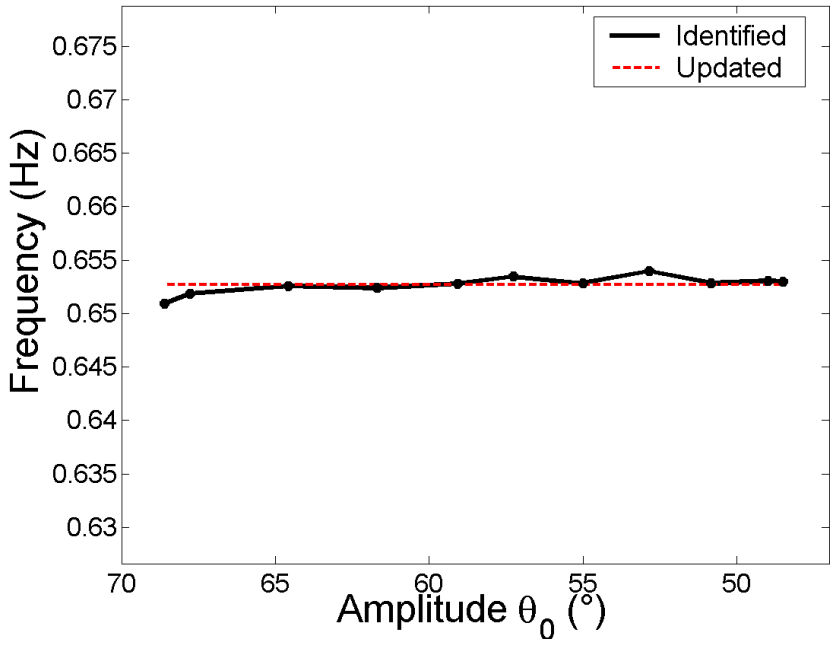


Figure 10.18. Estimates of the swinging frequency, for the case $\bar{s} = 0.5$ m. The updated value $f_{0,up}$ is represented with a dashed line.

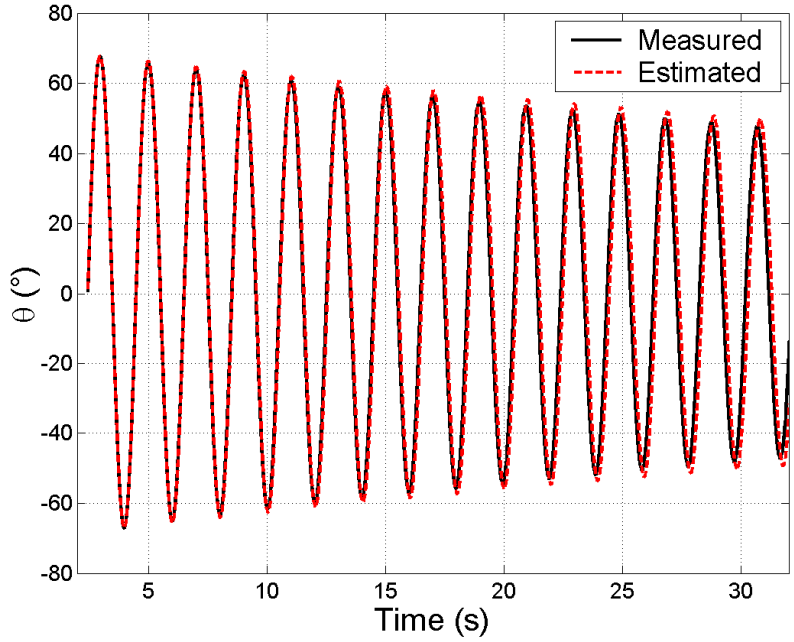


Figure 10.19. Comparison between the measured time history of $\theta(t)$ and the estimate obtained with the identified and updated parameters. The case is $\bar{s} = 0.91$ m.

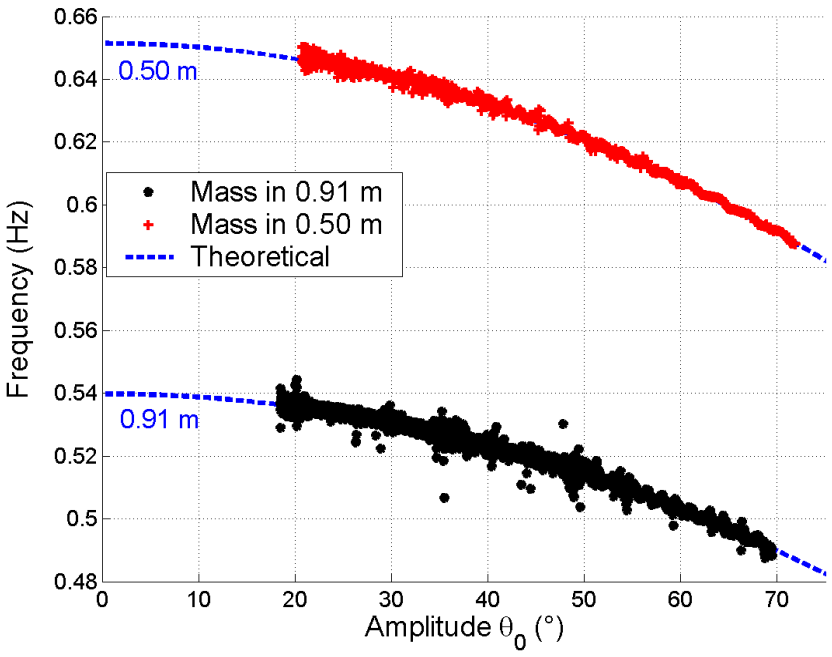


Figure 10.20. Comparison of two cases with fixed mass on the pendulum (in $\bar{s} = 0.91$ m and $\bar{s} = 0.5$ m) with the theoretical frequencies calculated on the same mass position.

Estimation of the frequencies with ST-SSI

To identify the swinging frequencies the ST-SSI method is finally applied to the angular position signal. In this case, the identified values can be depicted as a function of the amplitude θ_0 and compared with the baseline frequencies described in Section 10.1.3.

The ST-SSI identification is applied on signals decimated with a factor of 12. 60 samples are considered in each window, corresponding to a time duration of 2.81 s, with an overlap of 59 samples. With these parameters, each window included from 1.5 to 2 periods of the first natural frequency.

The results are shown in Fig. 10.20, where the theoretical curves have been calculated for two different positions of the mass ($\bar{s} = 0.91$ m and $\bar{s} = 0.5$ m), for different angular amplitudes. The identified and theoretical frequencies are very close, confirming that the adopted model is able to predict the evolution of the frequencies, in the case of swings with a fixed mass on the pendulum.

10.2.2. Moving mass

Let us consider two cases in which the load is travelling on the pendulum, this producing both nonlinear and time-varying effects:

- case M1: the mass is moving upward
- case M2: the mass is moving firstly upward and then downward

For both cases, the time histories of the angle and the load position are shown in Fig. 10.21, together with the frequencies identified by means of the ST-SSI method and compared to those obtained by applying Eqs. (10.12) and (10.13) with the updated parameters I_{up} and P_{up}.

In order to have a final visualization, Fig. 10.22 shows the identified frequencies and compares them with the baseline frequencies $\bar{f}$ introduced in Section 10.1.3. Ten different mass positions on the pendulum (from 0.05 m to 0.95 m) have been depicted to verify the correspondence among frequency, angle amplitude and mass position (in a certain time instant) for the theoretical and the identified models. Some mass positions measured by the linear potentiometer have been marked on the graph by means of diamonds, both for the case M1 and the case M2. The identified values are very close to the predicted frequencies, for both the experiments.

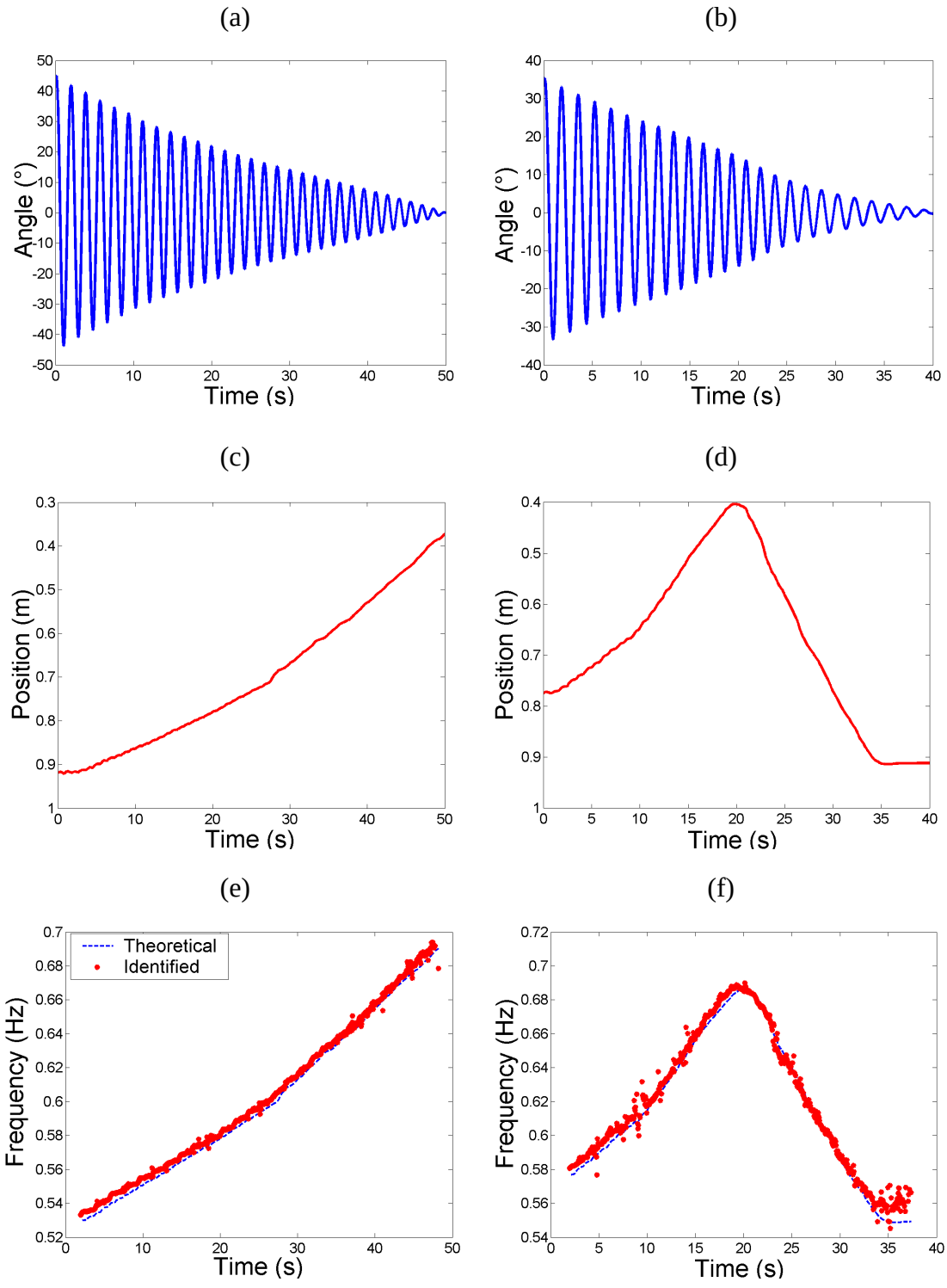


Figure 10.21. Nonlinear oscillations for cases M1 (a-c-e) and M2 (b-d-f): time evolution of the oscillation angle (a-b), of the mass position (c-d) and frequencies calculated with ST-SSI, compared with the theoretical ones (e-f).

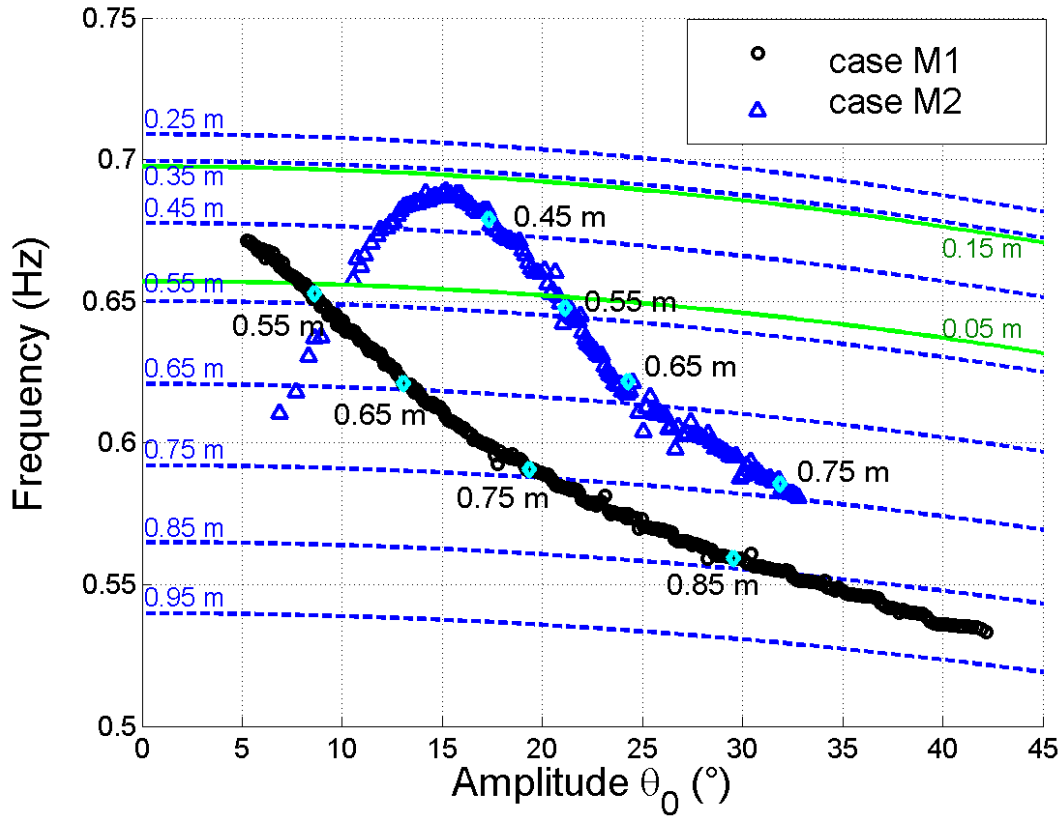


Figure 10.22. Nonlinear oscillations: comparison among the frequencies of the experimental cases M1/M2 and the baseline frequencies built for ten different load positions. The diamonds on the experimental curves highlight specific mass positions, measured by the linear potentiometer.

Conclusions

In this research work two main topics were investigated: damage detection and dynamic identification. For this reason, two main parts can be recognised.

In the first part, the well-known Principal Component Analysis (PCA) method for reducing a complex data set to a lower dimension was presented. PCA can successfully be adopted in diagnostics, since it provides useful tools for eliminating the influence of the operational and environmental conditions on a system. Changes in these conditions for structures or machines are known to have considerable effects on system features. In fact, if the whole range of conditions is not span by the data used to characterise the normal operating condition, incorrect diagnostics will occur.

After a structural example was used to show how the PCA-based diagnostics works, a bearing diagnostic application was considered. The bearing test rig was conceived to carry out an exhaustive experimental campaign in controlled laboratory conditions. Some different damaged bearings were available and the huge amount of performed tests was useful for giving an overview on how the PCA-based method for damage detection can be applied on a complicated real-life machine.

Moving to the second part of the work, in many cases damage causes a structure that initially behaves in a predominantly linear manner to exhibit nonlinear response: the application of nonlinear system identification methods to the feature-extraction process can also be used as a direct diagnosis of damage. For these reasons, a detailed study of the subspace-based identification methods was given: in particular, the Nonlinear Subspace Identification (NSI) method was considered.

Particular attention was given to the problem concerning computational memory limitations, which affect the classical data-driven subspace method. Due to a data matrix which can be too large to be stored nor factorised, the method undergoes

severe limitations in its applicability, in particular as regards large MDOF systems or systems having many nonlinear terms. In order to find a way for overcoming these problems, two techniques were introduced and demonstrated on numerical and experimental applications. Moreover, a modal counterpart of the NSI method was developed, to extend the current method to be applied also on realistic large engineering structures.

In a conclusive application, two of the main sources of non-stationary dynamics, namely the time-variability and the presence of nonlinearity, were analysed through the analytical and experimental study of a time-varying inertia pendulum. The pendulum undergoes large swinging amplitudes, so that its equation of motion is definitely nonlinear, and hence becomes a nonlinear time-varying system. A general model representing the experimental pendulum was correctly defined, by means of parameter identification (for the nonlinear terms) and updating (for the linear terms).

Main contributions

The main achievements of this research work are summarised in the following.

Diagnostics

- The PCA method was presented, with a deep mathematical insight. An existing PCA-based technique for damage detection was deeply investigated through a structural numerical application. Then, the more challenging issue of applying the PCA-based method for bearing diagnostics was addressed, by considering an experimental bearing test rig.
- The motivation for exploiting PCA on the experimental test rig was illustrated: the influence of the operational and environmental conditions on the extracetd features was studied. Then, the obtained results were described by investigating the main objectives of diagnostics: damage detection and false-positive verification.

  The assumption of linearity can be verified, but this assumption may be too strong when trying to apply PCA to different subsets of data: as a consequence the PCA-based detection method was not robust when features are identified in a limited range of operational and environmental variations.

In conclusion, the only way for preventing missed or false detections to occur consisted in using a full range data set, including all values of the conditions that can occur in practice.

- After a correct damage detection, two other diagnostics issues were addressed. A simple damage localisation technique, which is proper for bearing diagnostics, was introduced and applied to the experimental test rig. In this way, each sensor (or group of channels) can be evaluated in order to determine which is the most sensitive to damage: this is expected to be strictly related to its distance from the damaged bearing. Moreover, an attempt of applying the PCA-based method for damage extent evaluation was also proposed.

Identification

- The data-driven subspace identification (DDSI) method was enforced by the development of a new algorithm to compute the QR factorisation in a *Matlab*® environment, for overcoming the memory limitation problems in those cases in which the data matrix is too large to be stored nor factorised. This new algorithm, which exploited some useful properties of the Householder transformations, allowed the nonlinear DDSI method to reach more accurate results in the parameter estimation.
- A multivariate formulation in the time domain for modal parameter identification using covariances was developed, with the aim of proposing a complete input-output covariance-driven subspace identification (CDSI) method. It can be applied in the same way as its well-established DDSI counterpart with similar results: for both methods, the quality of results was excellent and they can be both used in practical situations, depending on the size of the data sets that have to be managed. The CDSI method, in fact, was not suffering from the memory limitation problems, such as in the DDSI method.
- In order to extend the NSI method to be applied also on realistic large engineering structures, a model reduction is needed and can be performed by selecting a set of modes that span the dominant dynamics. To this purpose, a modal counterpart of the NSI method was developed, together with ideas for handling a large complex nonlinear system as simple modal single degree of freedom systems.

  As one of the main advantages of the present modal approach, it was not required to perform a mode by mode excitation, so multi-exciters were not

necessary. By dealing with separate modal single degree of freedom systems, a single point excitation on the structure was sufficient to obtain several modal forces, with also a little gain in terms of testing time. However, a free response analysis can also be useful in order to perform a characterisation of modal nonlinearities, in particular for large structures, when forced tests are often uneasy.

Future works

In addition to the main conclusions and results obtained throughout this research work, both the topics of PCA-based bearing diagnostics and subspace identification have shown that some drawbacks or incorrect results can be further studied and fixed. In particular, the following issues should be considered for future developments.

Diagnostics

- Unfortunately, some of the tests performed on the bearing test rig were affected by the problem of having a few number of data. In particular, data of Test #1 (involving all the damaged types of bearing) have not been acquired in an optimum way for carrying out a PCA-based detection, since this was not the main objective at that time. This is the reason why damage extent evaluation had to be considered only as an “attempt”: with a few number of reference data, the influence of all the conditions cannot be completely eliminated by the PCA-based method. A more suited analysis for this issue should be carried out by performing a more accurate test for each of the seven types of bearing.
- As demonstrated by the experimental application involving a bearing test rig, PCA is limited by its linearity and may sometimes be too simple for dealing with real-world data especially when the relations among the features are nonlinear. The proposed solution, which consists in using full range data of reference and monitored system, is effective but only when the nonlinearity is weak.

  In future works, data collected on the test rig can be analysed by means of nonlinear generalisations of PCA. This should allow the detection method to

be robust even when features are identified in a limited range of operational and environmental variations.

Identification

- The presented CDSI method was only demonstrated on linear identification procedures, although on large systems with many degrees of freedom. Some computational difficulties leading to wrong results were encountered when trying to apply the method on nonlinear systems, in the same way as the DDSI method was demonstrated to be capable of. The reasons of this failure are currently under investigation: the procedure for estimating the state-space matrices B and D should probably be revisited.
- The study of a particular nonlinear time-varying system, such as the time-varying inertia pendulum of last chapter, revealed a drawback affecting the NSI method. In fact, the method was not capable of distinguishing the time-variant contribution from the nonlinear one, since the effect of both contributions was the same: it consisted of swinging frequency variations. Further investigation on the method should be carried out, so that the two contributions can correctly be separated and quantified.

Bibliography

[1] A. K.S. Jardine, D. Lin, D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance", *Mechanical Systems and Signal Processing*, vol. 20, 2006, pp. 1483-1510.

[2] K. F. Martin, "A review by discussion of condition monitoring and fault-diagnosis in machine-tools", *International Journal of Machine Tools and Manufacture*, vol. 34, 1994, pp. 527-551.

[3] N. Sawalhi, R. B. Randall, "Vibration response of spalled rolling element bearings: Observations, simulations and signal processing techniques to track the spall size", *Mechanical Systems and Signal Processing*, vol. 25, 2011, pp. 846-870.

[4] K. Worden, J.M. Dulieu-Barton, "An overview of intelligent fault detection in systems and structures", *International Journal of Structural Health Monitoring*, vol. 3, n. 1, 2004, pp. 85-98.

[5] N.V. Kirianaki, S.Y. Yurish, N.O. Shpak, V.P. Deynega, *Data Acquisition and Signal Processing for Smart Sensors*, Wiley, Chichester, West Sussex, England, 2002.

[6] R. Xu, C. Kwan, "Robust isolation of sensor failures", *Asian Journal of Control*, vol. 5, 2003, pp. 12–23.

[7] R. B. Randall, J. Antoni, “Rolling element bearing diagnostics – A tutorial”, *Mechanical Systems and Signal Processing*, vol. 25, 2011, pp. 485-520.

[8] H.L. Balderston, “The detection of incipient failure in bearings”, *Material Evaluation*, vol. 27, 1969, pp. 121–128.

[9] S. Braun, “The extraction of periodic waveforms by time domain averaging”, *Acoustica*, vol. 23, n. 2, 1975, pp. 69–77.

[10] S. Braun, B. Datner, “Analysis of roller/ball bearings”, *Journal of Design*, vol. 101, n. 1, 1979, pp. 118–128.

[11] K. Worden, W.J. Staszewski, J.J. Hensman, “Natural computing for mechanical systems research: a tutorial overview”, *Mechanical Systems and Signal Processing*, vol. 25, n. 1, 2011, pp. 4-111.

[12] A. Widodo, E. Y. Kim, J.-D. Son, B.-S. Yang, A. C.C. Tan, D.-S. Gu, B.-K. Choi, J. Mathew, “Fault diagnosis of low speed bearing based on relevance vector machine and support vector machine”, *Expert Systems with Applications*, vol. 36, 2009, pp. 7252-7261.

[13] Q. He, F. Kong, R. Yan, “Subspace-based gearbox condition monitoring by kernel principal component analysis”, *Mechanical Systems and Signal Processing*, vol. 21, 2007, pp. 1755-1772.

[14] C. J. Li, S. Y. Li, “Acoustic emission analysis for bearing condition monitoring”, *Wear*, vol. 185, 1995, pp. 67-74.

[15] D. Ho, R.B. Randall, “Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals”, *Mechanical Systems and Signal Processing*, vol. 14, n. 5, 2000, pp. 763–788.

[16] P.D. McFadden, J.D. Smith, “Model for the vibration produced by a single point defect in a rolling element bearing”, *Journal of Sound and Vibration*, vol. 96, n. 1, 1984, pp. 69-82.

[17] C.S. Sunnersjo, “Varying compliance vibrations of rolling bearings”, *Journal of Sound and Vibration*, vol. 58, n. 3, 1978, pp. 363-373.

[18] C.S. Sunnersjo, “Rolling bearing vibrations - geometrical imperfections and wear”, *Journal of Sound and Vibration*, vol. 98, n. 4, 1985, pp. 455-474.

[19] A. Choudhury, N. Tandon, “A theoretical model to predict vibration response of rolling bearings to distributed defects under radial load”, *Journal of Vibration and Acoustics*, vol. 120, n. 1, 1998, pp. 214-220.

[20] N. Tandon, A. Choudhury, “A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings”, *Tribology International*, vol. 32, 1999, pp. 469-480.

[21] N. Sawalhi, R.B. Randall, H. Endo, “The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis”, *Mechanical Systems and Signal Processing*, vol. 21, n. 6, 2007, pp. 2616-2633.

[22] R.B. Randall, J. Antoni, S. Chobsaard, “The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals”, *Mechanical Systems and Signal Processing*, vol. 15, n. 5, 2001, pp. 945-962.

[23] J. Antoni, R.B. Randall, “Differential diagnosis of gear and bearing faults”, *Journal of Vibration and Acoustics*, vol. 124, 2002, pp. 165-171.

[24] M.S. Kay, S.L. Marple, ”Spectrum analysis – a modern perspective”, *Proceedings of the IEEE*, vol. 69, n. 11, 1981, pp. 1380-1419.

[25] B. Widrow, S. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985, pp. 349–351.

[26] D. Ho, R.B. Randall, “Effects of time delay, order of fir filter and convergence factor on self adaptive noise cancellation”, *ICSV5*, Adelaide, 1997.

[27] J. Antoni, R.B. Randall, “Unsupervised noise cancellation for vibration signals: part I – evaluation of adaptive algorithms”, *Mechanical Systems and Signal Processing*, vol. 18, 2004, pp. 89-101.

[28] J. Antoni, R.B. Randall, “Unsupervised noise cancellation for vibration signals: part II – a novel frequency-domain algorithm, *Mechanical Systems and Signal Processing*, vol. 18, 2004, pp. 103-117.

[29] P.D. McFadden, “A revised model for the extraction of periodic waveforms by time domain averaging”, *Mechanical Systems and Signal Processing*, vol. 1, n. 1, 1987, pp. 83-95.

[30] S. Braun, “The synchronous (time domain) average revisited”, *Mechanical Systems and Signal Processing*, vol. 25, n. 4, 2011, pp. 1087-1102.

[31] R.A. Wiggins, “Minimum Entropy Deconvolution”, *Geoexploration*, vol. 16, 1978, pp. 21-35.

[32] H. Endo, R.B. Randall, “Application of a minimum entropy deconvolution filter to enhance autoregressive model based gear tooth fault detection technique”, *Mechanical Systems and Signal Processing*, vol. 21, n. 2, 2007, pp. 906-919.

[33] D.E. Newland, “Wavelet analysis of vibration signals”, in: M. Crocker (Ed.), *Handbook of Noise and Vibration Control*, Wiley, 2007, (Chapter 49).

[34] Z.K. Peng, F.L. Chu, “Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography”, *Mechanical Systems and Signal Processing*, vol. 18, 2004, pp. 199-221.

[35] F. Bonnardot, R.B. Randall, J. Antoni, “Enhanced unsupervised noise cancellation using angular resampling for planetary bearing fault diagnosis”, *International Journal of Acoustics and Vibration*, vol. 9, n. 2, 2004.

[36] J. Antoni, “The spectral kurtosis: a useful tool for characterising nonstationary signals”, *Mechanical Systems and Signal Processing*, vol. 20, n. 2, 2006, pp. 282–307.

[37] J. Antoni, R.B. Randall, “The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines”, *Mechanical Systems and Signal Processing*, vol. 20, n. 2, 2006, pp. 308–331.

[38] J. Antoni, “Fast computation of the kurtogram for the detection of transient faults”, *Mechanical Systems and Signal Processing*, vol. 21, n. 1, 2006, pp. 108–124.

[39] W. Bartelmus, R. Zimroz, “Vibration condition monitoring of planetary gearbox under varying external load”, *Mechanical Systems and Signal Processing*, vol. 23, n. 1, pp. 246-257.

[40] W. Bartelmus, R. Zimroz, “A new feature for monitoring the condition of gearboxes in nonstationary operation conditions”, *Mechanical Systems and Signal Processing*, vol. 23, n. 5, pp. 1528-1534.

[41] J. Tuma, “Gearbox noise and vibration prediction and control”, *International Journal of Acoustics and Vibration*, vol. 14, n. 2, 2009, pp. 1-11.

[42] P.D. McFadden, J.D. Smith, “An explanation for the asymmetry of the modulation sidebands about the tooth meshing frequency in epicyclic gear vibration”, *Proceedings of the Institution of Mechanical Engineers, Part C: Mechanical Engineering Science*, vol. 199, n. 1, 1985, pp. 65-70.

[43] P.D. McFadden, “Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing vibration”, *Journal of Vibration Acoustics Stress and Reliability in Design*, vol. 108, 1986, pp. 165-170.

[44] R.B. Randall, “State of the art in monitoring rotating machinery – Part 1”, *Sound and Vibration*, vol. March, 2004, pp. 14-21.

[45] R.B. Randall, “A New Method of Modeling Gear Faults”, *Journal of Mechanical Design*, vol. 104, 1982, pp 259-267.

[46] J. Antoni, R.B. Randall, “Differential Diagnosis of Gear and Bearing Faults”, *Journal of Vibration and Acoustics*, vol. 124, 2002, pp. 165-171.

[47] K. Worden, C.R. Farrar, “An introduction to structural health monitoring”, *Philosophical Transactions of the Royal Society*, vol. 365, 2007, pp. 303-315.

[48] K. Worden, C.R. Farrar, J. Haywood, M. Todd, “A review of nonlinear dynamics applications to structural health monitoring”, *Structural Control and Health Monitoring*, vol. 15, 2008, pp. 540-567.

[49] H. Sohn, C.R. Farrar, F.M. Hemez, J.J. Czarnecki, D.D. Shunk, D.W. Stinemates, B.R. Nadler, “A review of structural health monitoring literature: 1996-2001”, *Los Alamos National Laboratory Report*, 2004.

[50] L. Ljung, *System Identification: Theory for the User*, Second edition, Prentice-Hall, Upper Saddle River, NJ, USA, 1999.

[51] P. Van Overschee, B. De Moor, *Subspace identification for linear systems: theory, implementation, applications*, Kluwer Academic Publishers, Boston / London / Dordecht, 1996.

[52] C.R. Farrar, P.J. Cornwell, S.W. Doebling, M.B. Prime, “Structural health monitoring studies of the alamosa canyon and I-40 brdges”, *Los Alamos National Laboratory Report*, 2000.

[53] A.-M. Yan, G. Kerschen, P. De Boe, J.-C. Golinval, “Structural damage diagnosis under varying environmental conditions – Part I: A linear analysis”, *Mechanical Systems and Signal Processing*, vol. 19, 2005, pp. 847-864.

[54] B. Peeters, G. De Roeck, “Reference-based stochastic subspace identification for output-only modal analysis”, *Mechanical Systems and Signal Processing*, vol. 13, 1999, pp. 855-878.

[55] E. Reynders, G. De Roeck, “Reference-based combined deterministic-stochastic subspace identification for experimental and operational modal analysis”, *Mechanical Systems and Signal Processing*, vol. 22, 2008, pp. 617-637.

[56] B. Peeters, G. De Roeck, “Stochastic system identification for operational modal analysis: a review”, *Journal of Dynamic Systems, Measurement, and Control*, vol. 123, 2001, pp. 659-667.

[57] M. Viberg, “Subspace-based methods for the identification of linear time-invariant systems”, *Automatica*, vol. 31, 1995, pp. 1835-1853.

[58] P. Van Overschee, B. De Moor, “N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems”, *Automatica*, vol. 30, 1994, pp. 75-93.

[59] E. Reynders, R. Pintelon, G. De Roeck, “Uncertainty bounds on modal parameters obtained from stochastic subspace identification”, *Mechanical Systems and Signal Processing*, vol. 22, 2008, pp. 948-969.

[60] A. Guyader, L. Mevel, “Covariance driven subspace methods: input/output vs output-only”, *Proceedings of the 21st International Modal Analysis Conference*, Kissimmee, 2003.

[61] L. Mevel, A. Benveniste, M. Basseville, M. Goursat, B. Peeters, H. Van der Auweraer, A. Vecchio, “Input/output versus output-only data processing for structural identification – Application to in-flight data analysis”, *Journal of Sound and Vibration*, vol. 295, 2006, pp. 531-552.

[62] G. Kerschen, K. Worden, A. F. Vakakis, J.-C. Golinval, “Past, present and future of nonlinear system identification in structural dynamics”, *Mechanical Systems and Signal Processing*, vol. 20, 2006, pp. 505-592.

[63] S. F. Masri, T. K. Caughey, “A nonparametric identification technique for nonlinear dynamic problems”, *Journal of Applied Mechanics*, vol. 46, 1979, pp. 433-447.

[64] K. Worden, D. Hickey, M. Haroon, D. E. Adams, “Nonlinear system identification of automotive dampers: a time and frequency-domain analysis”, *Mechanical Systems and Signal Processing*, vol. 23, 2009, pp. 104-126.

[65] K. S. Mohammad, K. Worden, G. R. Tomlinson, “Direct parameter estimation for linear and nonlinear structures”, *Journal of Sound and Vibration*, vol. 152, 1991, pp. 471-499.

[66] C. M. Richards, R. Singh, “Identification of multi-degree-of-freedom non-linear systems under random excitations by the reverse-path spectral method”, *Journal of Sound and Vibration*, vol. 213, 1998, pp. 673-708.

[67] G. Kerschen, V. Lenaerts, S. Marchesiello, A. Fasana, “A frequency domain versus a time domain identification technique for nonlinear parameters applied to wire rope isolators”, *Journal of Dynamic Systems Measurement and Control-Transactions of the ASME*, vol. 123, n. 4, 2001, pp. 645-650,

[68] D. E. Adams, R. J. Allemang, “A frequency domain method for estimating the parameters of a non-linear structural dynamic model through feedback”, *Mechanical Systems and Signal Processing*, vol. 14, n. 4, 2000, pp. 637-656.

[69] A. Fasana, L. Garibaldi, S. Marchesiello, “Performances analysis of frequency domain nonlinear identification techniques”, in: *Proceedings of ISMA*, 2004, pp. 2115-2128.

[70] S. Marchesiello, L. Garibaldi, “A time domain approach for identifying nonlinear vibrating structures by subspace methods”, *Mechanical Systems and Signal Processing*, vol. 22, 2008, pp. 81-101.

[71] C.M. Bishop, “Novelty detection and neural network validation”, *IEEE Proceedings – Vision and Image Signal Processing*, vol. 141, 1994, pp. 217-222.

[72] K. Worden, “Structural fault detection using a novelty measure”, *Journal of Sound and Vibration*, vol. 201, 1997, pp. 85-101.

[73] A. Widodo, B.-S. Yang, T. Han, “Combination of independent component analysis and support vector machines for intelligent faults diagnosis of induction motors”, *Expert Systems with Applications*, vol. 32, n. 2, 2007, pp. 299-312.

[74] A. Widodo, B.-S. Yang, “Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motor”, *Expert Systems with Applications*, vol. 33, n. 1, 2007, pp. 241-250.

[75] V.N. Vapnik, *The nature of statistical learning theory*, Springer, New York, 1999.

[76] M.E. Tipping, A. Smola, “Sparse Bayesian learning and the Relevance Vector Machine”, *Journal of Machine Learning Research*, vol. 1, 2001, pp. 211-244.

[77] M. Basseville, M. Abdelghani, A. Benveniste, “Subspace-based fault detection algorithms for vibration monitoring”, *Automatica*, vol. 36, 2000, pp. 101-109.

[78] A.-M. Yan, J.-C. Golinval, “Null subspace-based damage detection of structures using vibration measurements”, *Mechanical Systems and Signal Processing*, vol. 20, n. 3, 2006, pp. 611-626.

[79] L. Mevel, M. Basseville, M. Goursat, A. Hassim, “A subspace detection approach to damage localization”, *Proceedings of the 21st International Modal Analysis Conference*, Kissimmee, 2003.

[80] M. Basseville, L. Mevel, M. Goursat, “Statistical model-based damage detection and localization: subspace-based residuals and damage-to-noise sensitivity ratios”, *Journal of Sound and Vibration*, vol. 275, n. 3-5, 2004, pp. 769-794.

[81] K.M. Holford, “Acoustic emission in structural health monitoring”, *Key Engineering Materials*, vol. 413-414, 2009, pp. 15-28.

[82] D. Mba, Raj B.K.N. Rao, “Development of acoustic emission technology for condition monitoring and diagnosis of rotating machines: bearings, pumps, gearboxes, engines, and rotating structures”, *The Shock and Vibration Digest*, vol. 38, n. 1, 2006, pp. 3-16.

[83] J. Shlens, “A tutorial on Principal Component Analysis”, available online: http://www.cs.cmu.edu/~elaw/papers/pca.pdf, 2005.

[84] H. Anton, C. Rorres, *Elementary linear algebra*, 10th edition, Wiley, 2011.

[85] G. H. Golub, C. Van Loan, *Matrix computations*, John Hopkins Univ. Press, Baltimore, 1983.

[86] C. L. Lawson, R. J. Hanson, *Solving least squares problems*, Prentice-Hall, Englewood Cliffs, N.J., 1974.

[87] B. Schölkopf, A. J. Smola, K.-R. Müller, “Kernel principal component analysis”, in: *Advances in Kernel Methods – Support Vector Learning*, MIT Press, Cambridge MA, 1999, pp. 327-352.

[88] T.-W. Lee, M. Girolami, A. J. Bell, T. J. Sejnowski, “A unifying information-theoretic framework for independent component analysis”,

Computers & Mathematics with Applications, vol. 39, n. 11, 2000, pp. 1-21.

[89] G. Kerschen, J.-C. Golinval, "Nonlinear generalisation of principal component analysis: from a global to a local approach", *Journal of Sound and Vibration*, vol. 254, n. 5, 2002, pp. 867-876.

[90] 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", *Mechanical Systems and Signal Processing*, vol. 19, 2005, pp. 865-880.

[91] A. Bellino, A. Fasana, L. Garibaldi, S. Marchesiello, "PCA-based detection of damage in time-varying systems", *Mechanical Systems and Signal Processing*, vol. 24, 2010, pp. 2250-2260.

[92] K. Worden, G. Manson, N.R.J. Fieller, "Damage detection using outlier analysis", *Journal of Sound and Vibration*, vol. 229, n. 3, 2000, pp. 647-667.

[93] H. Sohn, C.R. Farrar, "Damage diagnosis using time series analysis of vibration signals", *Smart Materials & Structures*, vol. 10, 2000, pp. 1-6.

[94] M. Pirra, E. Gandino, A. Torri, L. Garibaldi, J. M. Machorro-Lopez, "PCA algorithm for detection, localization and evolution of damages in gearbox bearings ", *Journal of Physics: Conference Series*, vol. 305, 2011.

[95] A. Bellino, J.M. Machorro-López, S. Marchesiello, L. Garibaldi, "Damage detection in structures under variations of temperature and clamping conditions", *VCB2012*, Paris, July 3-5 2012.

[96] S. Marchesiello, L. Garibaldi, "Subspace-based identification of nonlinear structures", *Shock and Vibration*, vol. 14, 2007, pp. 1-10.

[97] S. Marchesiello, L. Garibaldi, “Identification of clearance-type nonlinearities”, *Mechanical Systems and Signal Processing*, vol. 22, 2008, pp. 1133-1145.

[98] G. Kerschen, V. Lenaerts, J.-C. Golinval, “Identification of a continuous structure with a geometrical non-linearity. Part I: Conditioned reverse path method”, *Journal of Sound and Vibration*, vol. 262, 2003, pp. 889-906.

[99] A. Bellino, L. Garibaldi, S. Marchesiello, “Time-Varying Output-Only Identification of a cracked beam”, *Key Engineering Materials*, vol. 413-414, 2009, pp. 643-650.

[100] S. Marchesiello, S. Bedaoui, L. Garibaldi, P. Argoul, “Time-dependent identification of a bridge-like structure with crossing loads”, *Mechanical System and Signal Processing*, vol. 23, 2009, pp. 2019-2028.

[101] K. Liu, “Extension of Modal Analysis to Linear Time-Varying Systems”, *Journal of Sound and Vibration*, vol. 226, 1999, pp. 149-167.

[102] E. Gandino, S. Marchesiello, “Identification of a Duffing oscillator under different types of excitation”, *Mathematical Problems in Engineering*, vol. 2010, 2010.

[103] A. H. Nayfeh, D. T. Mook, *Nonlinear oscillations*, Wiley, New York, 1979.

[104] K. Worden, “On jump frequencies in the response of a Duffing oscillator”, *Journal of Sound and Vibration*, vol. 198, n. 4, 1996, pp. 522-525.

[105] M. I. Friswell, J. E. T. Penny, “The accuracy of jump frequencies in series solutions of the response of a Duffing oscillator”, *Journal of Sound and Vibration*, vol. 169, n. 2, 1994, pp. 261-269.

[106] P. Malaktar, A. H. Nayfeh, “Calculation of the jump frequencies in the response of s.d.o.f. non-linear systems”, *Journal of Sound and Vibration*, vol. 254, n. 5, 2002, pp. 1005-1011.

[107] M. J. Brennan, I. Kovacic, A. Carrella, T. P. Waters, “On the jump-up and jump-down frequencies of the Duffing oscillator”, *Journal of Sound and Vibration*, vol. 318, 2008, pp. 1250-1261.

[108] C. M. Richards, R. Singh, “Feasibility of identifying non-linear vibratory systems consisting of unknown polynomial forms”, *Journal of Sound and Vibration*, vol. 220, n. 3, 1999, pp. 413-450.

[109] L. Mevel, A. Benveniste, M. Basseville, M. Goursat, “Blind subspace-based eigenstructure identification under nonstationary excitation using moving sensors”, *Transactions on Signal Processing*, vol. 50, 2002, pp. 41-48.

[110] A. Benveniste, L. Mevel, “Nonstationary consistency of subspace methods”, *Transactions on automatic control*, vol. 52, 2007, pp. 974-984.

[111] E. Gandino, L. Garibaldi, S. Marchesiello, “Covariance-driven subspace identification: a complete input-output approach”, *Journal of Sound and Vibration*, submitted.

[112] M. Basseville, A. Benveniste, M. Goursat, L. Hermans, L. Mevel, H. Van der Auweraer, “Output-only subspace-based structural identification: from theory to industrial testing practice”, *Journal of Dynamic System Measurement and Control*, vol. 123, 2001, pp. 668-676.

[113] J. S. Bendat, A. G. Piersol, *Engineering applications of correlation and spectral analysis*, Wiley, New York, 1980, re-issued 1993.

[114] P. Stoica, R. L. Moses, *Introduction to Spectral Analysis*, Prentice-Hall, Englewood Cliffs, 1997.

[115] P. R. G. Kurka, H. N. Cambraia, “Application of a multivariable input-output subspace identification technique in structural analysis”, *Journal of Sound and Vibration*, vol. 312, 2008, pp. 461-475.

[116] R.J. Allemang, D.L. Brown, “A complete review of the complex mode indicator function (CMIF) with applications”, *Proceedings of the International Conference on Noise and Vibration Engineering*, Leuven, September 2006, pp. 3209-3246.

[117] M.F. Platten, J.R. Wright, G. Dimitriadis, J.E. Cooper, “Identification of multi-degree of freedom non-linear systems using an extended modal space model”, *Mechanical Systems and Signal Processing*, vol. 23, 2009, pp. 8-29.

[118] L. Bornn, C. R. Farrar, G. Park “Damage detection in initially nonlinear systems”, *International Journal of Engineering Science*, vol. 48, 2010, pp. 909-920.

[119] E. Gandino, L. Garibaldi, S. Marchesiello, “Pescara benchmarks: nonlinear identification”, *Journal of Physics: Conference Series*, vol. 305, 2011.

[120] A. Bellino, L. Garibaldi, S. Marchesiello, “Determination of moving load characteristics by output-only identification over the Pescara beams”, *Journal of Physics: Conference Series*, vol. 305, 2011.

[121] W. Szyszkowski, E. Sharbati, “On the FEM modeling of mechanical systems controlled by relative motion of a member: A pendulum-mass interaction test case”, *Finite Elements in Analysis and Design*, vol. 45, 2009, pp. 730-742.

[122] W. Szyszkowski, D.S.D. Stilling, “On damping properties of a frictionless physical pendulum with a moving mass”, *International Journal of Non-Linear Mechanics*, vol. 40, 2005, pp. 669-681.

[123] W. Szyszkowski, D.S.D. Stilling, “Controlling angular oscillations through mass reconfiguration: a variable length pendulum case”,

International Journal of Non-Linear Mechanics, vol. 37, 2002, pp. 89-99.

[124] A.A. Zevin, L.A. Filonenko, "A qualitative investigation of the oscillations of a pendulum with a periodically varying length and a mathematical model of a swing", *Journal of Applied Mathematics and Mechanics*, vol. 71, 2007, pp. 892-904.

[125] F.M.S. Lima, P. Arun, "An accurate formula for the period of a simple pendulum oscillating beyond the small angle regime", *American Journal of Physics*, vol. 74, 2006, pp. 892-895.

[126] C.G. Carvalhes, P. Suppes, "Approximations for the period of the simple pendulum based on the arithmetic-geometric mean", *American Journal of Physics*, vol. 76, n. 12, 2008, pp. 1150-1154.

[127] A. Beléndez, E. Arribas, M. Ortuño, S. Gallego, A. Márquez, I. Pascual, "Approximate solutions for the nonlinear pendulum equation using a rational harmonic representation", *Computers and Mathematics with Applications*, vol. 64, n. 6, 2012, pp. 1602-1611.

[128] A. Bellino, A. Fasana, E. Gandino, L. Garibaldi, S. Marchesiello, "A time-varying inertia pendulum: analytical modelling and experimental identification", *Mechanical Systems and Signal Processing*, submitted.

[129] T. Yokoyama, "Vibrations of a hanging Timoshenko beam under gravity", *Journal of Sound and Vibration*, vol. 141, n. 2, 1990, pp. 245-258.

[130] M. Akbarzade, Y. Khan, "Dynamical model of large amplitude non-linear oscillations arising in the structural engineering: Analytical solutions", *Mathematical and Computer Modelling*, vol. 55, 2012, pp. 480-489.