Amplitude and phase demodulation of multi-carrier signals: Application to gear vibration signals

Elisa Hubert

N°d'ordre NNT : 2019LYSES015

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Jean Monnet

Ecole Doctorale 488
Sciences Ingénierie Santé

Spécialité Traitement du signal :

Soutenue publiquement le 28/06/2019, par :
Elisa Hubert

Démodulation d'amplitude et de phase de signaux multi-porteuses : Application aux signaux vibratoire d'engrenage

---

Devant le jury composé de :

Antoni Jérôme, Professeur des Universités – Université de Lyon — Président
Martin Nadine, Directeur de recherche – CNRS — Rapporteure
Abed-Meraim Karim, Professeur des Universités – Université d'Orléans — Rapporteur
Quadrat Alban, Directeur de recherche – Inria — Examinateur
Randall Bob, Professeur Emérite – University of New South Wales — Examinateur
Renaux Alexandre, Maître de conférence – Université Paris Sud — Examinateur
El Badaoui Mohamed, Professeur des Universités – Université de Lyon, SafranTech — Directeur de thèse
Barrau Axel, Ingénieur de recherche, Chercheur associé – SafranTech, Mines ParisTech — Co-directeur de thèse

Amplitude and phase demodulation of multi-carrier signals: Application to gear vibration signals

Elisa Hubert

October 10, 2019

Résumé Cette thèse contient les résultats de travaux de recherche menés à Safran-Tech et au Laboratoire d'Analyse des Signaux et des Processus Industriels (LASPI) de l'Université de Lyon. Le sujet traité porte sur la surveillance vibratoire des transmissions de puissance aéronautique et plus particulièrement des engrenages.

Traditionnellement, les vibrations sont étudiées par analyse spectrale au moyen d'une représentation du spectre de Fourier. Basé sur ces observations, les vibrations des engrenages ont été représentées par un modèle empirique multiplicatif : d'une part le signal d'engrènement, haute fréquence, et de l'autre les signaux de rotations des roues, basses fréquences. En effet, les vibrations d'engrenage présentent un spectre de raies ayant des caractéristiques similaires à celles de certains signaux de communication comme une porteuse modulée en amplitude.

Dans le but de faire de la détection précoce de défauts, il est intéressant de pouvoir séparer les signaux basses fréquences du reste du signal car ils sont plus souvent porteurs de l'information de défaut.

Partant de ce modèle et de cette constatation, ces travaux étudient la réponse à deux questions:

1. A quel point le signal vibratoire produit lors de la rotation d'un engrenage peut-il être expliqué par la représentation sous la forme d'un produit?

2. Considérant un signal, est-il possible de le reconstruire en estimant ses composantes? Et la solution est-elle unique?

Pour répondre à ces questions, le modèle a été représenté sous la forme d'un problème d'optimisation. D'autre part, un nouvel outil a été défini pour représenter le spectre discret d'un signal vibratoire d'engrenage sous la forme d'une matrice de coefficients de Fourier. Ces travaux ont montré une équivalence entre le produit matriciel de deux vecteurs et la multiplication de deux signaux temporels, et permis de faire le lien entre la séparation du produit de deux signaux (démodulation) et les opérateurs de rang faible.

Cette nouvelle approche de séparation et d'estimation des signaux vibratoire d'engrenage a montré des performances théorique idéales et a permis de détecter de manière précoce les défauts de denture de signaux d'engrenage réels.

Abstract This thesis contains the results of the research studies performed with SafranTech and the Laboratoire d'Analyse des Signaux et des Processus Industriels (LASPI) of the University of Lyon. The main subject focus on vibratory surveillance of aeronautic power transmission systems and more specifically gearboxes.

Usually, vibrations are investigated with spectral analysis by means of the common representation of the Fourier spectrum. Based on these observations,gearbox vibrations were represented by an empirical product model: on one hand the meshing signal, with high frequency, and on the other hand the gears rotation signals, with low frequencies. Indeed, gearbox vibrations develop a line spectrum having similar characteristics with some communication signals, as a carrier signal modulated in amplitude.

For the purpose of incipient fault detection, it is interesting to be able to separate low frequency signals as they usually convey more fault information.

Based on these model and observation, this research work investigate the answer to the two following questions:

1. To which point the vibration signals produced by gears rotation can be explained by the representation as a product ?

2. Given a signal, is it possible to rebuild it by estimating the two component? Is the solution unique?

In order to answer those questions, the given model was formulated as an optimization problem. Then a new tool was defined to represent the discrete spectrum of gearbox vibration signal as a matrix containing the Fourier coefficients. This work has proven an equivalence between the two representations of the matrix product of two vectors and the temporal multiplication of two signals. Furthermore, it allowed us to link the remote fields of signal demodulation and low rank approximation.

This new separation and estimation approach for gearbox vibration signals has shown theoretical interesting performances, close to the ideal and allowed us to perform efficient incipant fault detection on real gearbox vibration dataset.

Acknowledgements

This PhD work has been mostly conducted in France, so I switch to french in the present section.

C'est arrivé! Ces trois années de thèse sont passées à une vitesse folle et n'auraient pas eu le même charme sans la présence de beaucoup. Malgré quelques hauts et bas, surtout à la fin, j'ai passé des années incroyablement riches qui m'ont fait grandir. Et il ne faut pas se le cacher, je me suis vraiment bien amusée!!

Tout d'abord, je tiens à remercier les membres du jury : Nadine Martin et Karim Abed Meraim pour avoir accepté d'être les rapporteurs de mes travaux de thèse et pour leur lecture attentive et les remarques constructives qu'ils ont apporté ; Jérôme Antoni pour avoir accepté de présider mon jury ainsi que Alban Quadrat, Bob Randall et Alexandre Renaux pour l'intérêt qu'ils ont porté à mon travail et leur présence au sein de mon jury.

Ces travaux n'auraient pas pu avoir lieu si Mohamed El Badaoui, mon directeur de thèse, ne m'avait pas accordé sa confiance. Merci de m'avoir laissé autant de liberté et d'autonomie pour explorer les pistes qui nous intéressaient le plus. Un grand merci à Axel Barrau, qui a co-dirigé mes travaux alors que ce n'était pas initialement prévu. Merci pour ces heures de discussions et d'explications et pour ton implication au quotidien. J'ai énormément appris à vos côté et pu gagner en maturité et grâce à votre soutien sans faille, j'ai pu prendre le meilleur des départs dans le monde de la recherche. Je suis très heureuse et fière d'avoir partagé cette expérience extraordinaire qu'est la thèse avec vous.

J'ai eu la chance de pouvoir faire de nombreuses rencontres scientifiques qui m'ont permis d'approcher plusieurs domaines et de monter des collaborations afin de faire des passerelles dans des domaines variés. Merci à Alexandre Renaux pour les après-midis au L2S à m'expliquer la Borne de Cramèr Rao et d'avoir pris le temps de se pencher sur ce sujet inhabituel du traitement de signal qui est l'application aux signaux mécaniques. Merci aussi à Yacine Bouzidi, Roudy Dagher et Alban Quadrat d'avoir pris autant de temps m'aider à résoudre mes problèmes de polynôme. Grâce à vous, le calcul formel et l'algèbre homologique me font moins peur et j'en sais beaucoup plus sur les propriétés de notre fameuse matrice ! Finally, thank you so much Bob Randall, Pietro Borghesani and Wade Smith for inviting me at the University of New South Wales in Sydney. Thanks to you I have eventually learn how mechanical systems produce vibrations and how to exploit their properties for surveillance algorithms. Thanks you for the funny and joyful conviviality Tuesday pizza/beer!

J'ai beaucoup de personnes à remercier, car j'ai été vraiment bien entourée, alors j'espère n'oublier personne. Tout d'abord à SafranTech, merci à Dohy Hong de m'avoir accueillie dans son équipe et à tous les collègues du pôle TSI pour les discussions passionnées et les pauses café animées. Merci aux autres doctorants de Safrantech (Michel, Mina, Paul et EdouardâĂę) d'avoir été là pour mettre de l'ambiance dans l'open space

au quotidien. Merci à l'équipe jeux de plateau, Axel, Séb, Héléna, David, Michel, Paul, Mina, Yosra, pour ces nombreuses soirées à découvrir des jeux avec vous, de Small-world à Room 25 et tant d'autres ! Merci à Luca pour ces pauses café qui font du bien ! Un grand merci à toute l'équipe de Peyresq, qui m'a accompagnée dans mes débuts en conférence et avec qui j'ai pu découvrir d'autres domaines du traitement du signal et partager des supers moments de convivialité, entre plage, rando et tournois de pétanque (Alex, Guillaume, Arnaud, Gilles, Lucien, Joana, Pascal, Eric, Jean-Phi, Fred et tous les autres...).

Mille mercis à mes colocataires, Amad, Kamel et plus récemment JB et Julien, qui m'ont soutenue quand j'en avais besoin. Pour les soirées discussion où on se change les idées ou les petits repas qui remontent le moral !

Un merci infini à Valentine qui est toujours là pour moi et sans qui j'aurai pu avoir envie d'abandonner.

Pour terminer, mes remerciements sont à mes parents et Hélène, ma sÅŞur. Vous n'avez pas toujours compris mes explications au combien peu claires sur le parallèle entre la médecine humaine et l'analyse vibratoire des engrenages mais vous avez toujours été présents pour moi. Mes retours en Normandie ont toujours été synonyme de bol d'air frais et iodé qui non seulement change de la pollution parisienne mais permet de se ressourcer et de recharger les batteries. Merci n'est pas un mot assez fort pour exprimer ma gratitude mais c'est le seul que j'ai à disposition, alors Merci ! Pour m'avoir laissée gérer mon parcours d'étude et pour vos valeurs et votre confiance. Vous êtes la meilleure famille que je puisse rêver avoir.

Contents

General Introduction

Power transmissions such as gears are very common in the industry, as they are one of the most elementary component of mechanical systems. Using power transmission systems offers many advantages such as their robustness, correct power ratio and also their reliability. However, despite those interesting characteristics, when used in hostile environment, gears may become critical. Indeed in the aeronautic industry, aircraft and helicopter engines evolve at both very high rotation speed and heavy load. Those operating conditions may boost the development of wear damage in all mechanical systems but in gearbox it may also lead to more critical failures such as pitting or cracks. Obviously, a damage lately detected in the gearbox may lead to catastrophic failure.

This is why surveillance of power transmission systems is still an actual research topic, even if it has been long studied. Actual maintenance system, called *scheduled maintenance*, plans maintenance operation at specific times. For example, in the aeronautic industry, aircraft engine are checked and dismounted about every thousands hour of flight. This way of working has a cost and may occasionally create more troubles than before the maintenance operation. Indeed, to dismount an engine is not a minor procedure and even if the system was in a good health state, problems are often created during the process. This is why predictive maintenance is of major interest in an economical point of view. It has to be noted that health monitoring allows enabling safety of aircraft as it gives a day-to-day overview of the system's state.

Traditionally, surveillance of mechanical systems is done with vibration analysis, as vibrations generated by rotating machines are usually regarded as a meaningful signature of their health state, instantaneously expressing any change in the structure or operating regime of the system. One of the common operation used for rotating machines' monitoring and diagnostics purposes is a spectrum analysis.

This work is based on those two traditional techniques but a different approach is presented with a new point of view. The thesis has been structured in three parts. The first part recalls the basics about gearbox functioning and monitoring based on techniques extracted from the state of the art. Several models of gearbox are also presented. It ends with a second state of the art explaining different techniques on demodulation followed by the new optimal demodulation approach proposed in this thesis.

The second part details the proposed demodulation technique with the introduction of a new tool that is a matrix representation of spectrum, used for several cases of demodulation case, i.e. amplitude demodulation, both phase and amplitude demodulation and a study of the planetary gearbox case.

The last part is dedicated to some applications of the proposed approach for gearbox signal analysis. The proposed demodulation method is first applied to fault detection and compared to existing fault indicators. The second point concerns signal model testing: indeed multi-carrier demodulation allows to separate in an optimal way signal meshing

as product. Those studies have been done with both simulated and experimental data.

Notations

The notation convention adopted in this manuscript is as follows :

Linear Algebra

a, A :	Scalar quantity
$\boldsymbol{a}$:	Vector quantity
$\boldsymbol{A}$:	Matrix quantity
$(\cdot)^T$:	Matrix transpose operator
$(\cdot)^*$:	Matrix conjugate operator
$(\cdot)^H$:	Matrix conjugate transpose operator
$(\cdot)^\dagger$:	Matrix pseudo-inverse operator
$Tr(\boldsymbol{A})$:	Trace of the matrix $\boldsymbol{A}$
$A_{i,j}$:	i^{th} row and j^{th} column element of the matrix $\boldsymbol{A}$
$\boldsymbol{A}_{:,j}$:	j^{th} column of matrix $\boldsymbol{A}$
$a[i]$:	i^{th} element of vector $\boldsymbol{a}$
$\succcurlyeq$:	$\boldsymbol{A} \succcurlyeq \boldsymbol{B}$ is defined for two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in the sense that $\boldsymbol{A} - \boldsymbol{B}$ is a positive semidefintite matrix
$\|\|\cdot\|\|$:	Modulus if applied to a complex number 2-norm if applied to a T_{tot}-periodic signal i.e. $\|\|x\|\|^2 = \int_0^{T_{tot}} x(t)^2 dt$ if $x(t)$ is a continuous variable $\|\|x\|\|^2 = \sum_0^{T_{tot}} x(t)^2$ if $x(t)$ is a discrete variable
$\|\|\boldsymbol{A}\|\|^2_{\text{Fro}}$:	Frobenius norm of matrix $\boldsymbol{A}$
$(\cdot) * (\cdot)$:	Convolution product defined such as $s_1 * s_2[i] = \sum_{j=-\infty}^{+\infty} s_1[i-j]s_2[j]$
$\otimes$:	Kronecker product
$vec(\boldsymbol{A})$:	Function that turns a matrix $\boldsymbol{A}$ into a vector $\boldsymbol{a}$ containing all the columns put end to end

Signal definitions

$s(t)$:	Temporal periodic signal
H (and K):	Number of harmonics of signal $s_i(t)$, $i = c, m$
C (and M):	Amplitude of the temporal signal s_c (and s_m)
C^R (and C^I):	Real part of C (and Imaginary part of C)
f_i, T_i :	Frequency and period of $s_i(t)$
I_c (or I_m):	Discrete set $[\![-C, C]\!]$ (or $[\![-M, M]\!]$)
T_{tot} :	Lowest Common Multiple (LCM) of the periods T_c and T_m
k_i :	Integer number defined by the factorization $T_{tot} = k_i T_i$
$\mathcal{D}(C, T_c)$:	Set of the T_c-periodic functions whose first C harmonics at most are non-zero (idem. $\mathcal{D}(M, T_m)$)
$\hat{s}$:	Estimated signal of $s(t)$
$\tilde{s}$:	Spectrum of $s(t)$
$\tilde{s}[k]$:	k-th harmonic of the spectrum of $s(t)$ regarded as a T_{tot}-periodic signal, index $k \in \mathbb{Z}$

Part I

State of the art and problem statement

Chapter 1

State of the art

In the aeronautic industry, mechanical systems permanently operate under severe conditions characterized, in the case of aircraft engines, by extreme rotation speed, extreme load and extreme temperature. During the take-off phase of a flight for example, the engine rotation speed outreaches 10000tr/min while handled load is about 30kNm. In this context, power transmission technologies are naturally prone to wear and failure and their monitoring becomes pivotal to the general reliability of a propulsion system. In addition to the issues raised by the involved power levels, the seriousness of a damage lately discovered on an aircraft engine is obviously incomparable to the seriousness of a similar fault occurring on a ground vehicle.

The most hardly exposed mechanical pieces in the process of power transmission are gears: they are mobile, undergo both high stress (in absolute value) and high stress variations, technicians access them with difficulty on ground and the tooth profile has to stay almost perfect for thousands of cycles in order to keep the contact with the opposite gear smooth. For these reasons, the issue of monitoring them indirectly through vibration analysis is gaining growing attention from industry, but also from the academic world. In this chapter we review the main categories of gearbox faults and the detection methods proposed in the existing literature.

1.1 Gearbox surveillance

Among the multiple kinds of power transmission systems, the present work focuses on gearboxes, sometimes simply called *transmissions*. Understanding their behavior from a mechanical point of view is a first step to building efficient fault detection tools but in practice, an approach fully based on physical modeling of the kinematics of a gear is not really tractable due to the complexity of the system and to modeling some phenomena (such as fluid flows) being extremely difficult. In this section dedicated to vibration-based fault detection we give the basics of gear dynamics allowing qualitative understanding of the structure (the spectral content in particular) of vibration signals measured on gearboxes.

Remark 1. *A vibration can be defined as a mechanical oscillation motion in solid bodies.*

The main source of these vibrations is machine rotation, of which they give an image conveying a very rich information: any change within the gearbox's situation or operating regime can instantaneously read in the vibration signal [77, 42]. Thus, since the

very beginning of power transmission monitoring, vibration analysis has been the first, and remains the most used technique [55], mainly due to its implementation simplicity and reliability.

Gearbox vibration can be coarsely understood as an oscillation created by the meshing of the teeth, but a finer qualitative description of the signal can be given even without elaborated physical modeling. This will be the topic of sections 1.1.1 and 1.1.2, while section 1.1.3 will list the main faults possibly affecting both spur gears and planetary gears, and describe their transcription in terms of vibration signal.

1.1.1 General information on gearbox

A basic gearbox is a device made up of at least two gears, the driving and the driven one, the former being usually called pinion for a reducer, illustrated in Figure1.1. The transmitted mechanical power is mainly a function of the rotation speed of the shaft and the torque applied to it. The most important value characterizing a gear is its *transmission ratio*, defined as the ratio between the in-shaft and out-shaft angular speeds. It only depends on the number of teeth of the gears in contact:

$$r = \frac{W_1}{W_2},$$

where W_1 is the teeth number of the pinion and W_2 the teeth number of the driven gear. As explained above in the introductory paragraph of Section 1.1, the gear vibration energy is mainly due to the gears meshing, which occurs at a frequency f_m called *meshing frequency* and related to the shaft rotation frequencies by the following equality:

$$f_m = W_1 f_1 = W_2 f_2,$$

where W_1, W_2 are the teeth numbers of gears $1, 2$ and f_1, f_2 are their rotation frequencies.

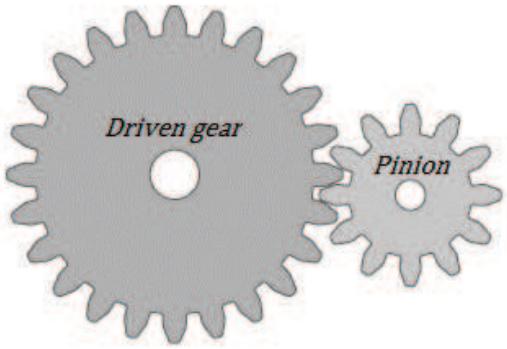
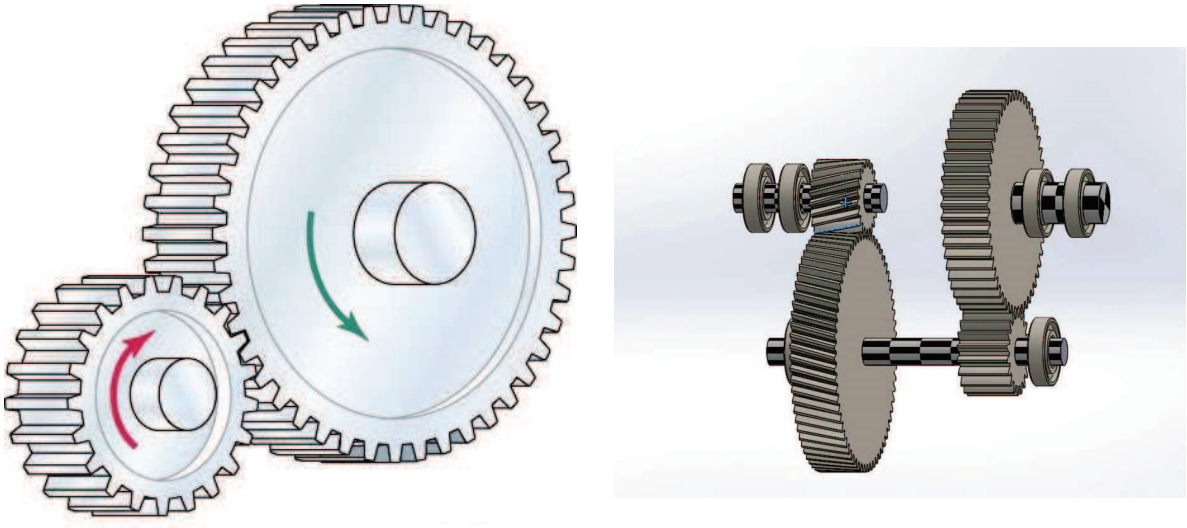


Figure 1.1: Drawing of an elementary gearbox with the name of its components.

Remark 2. *While the most elementary gearbox is made of two gears, as in Figure 1.2(a), several couples of gears can be used in more elaborated systems called* gear trains *or, sometimes,* transmissions*. A classical gear train is made of a series of gears, carried by different shafts and meshing each other, or of several couples of gears forming a* multistage gear *as on Fig. 1.2(b). In the first case the meshing frequency is shared by all the gears of the train:*

$$f_m = W_1 f_1 = W_2 f_2 = W_3 f_3 = \dots.$$



(a) (b)

Figure 1.2: Several geared transmissions: (a) Elementary gear train with one couple of gears, (b) Two-stage gear train with two couples of gears.

In the second case, two gears located at both ends of the same shaft share the same rotation frequency but the corresponding meshing frequencies can be different depending on the teeth number of each shaft.

Although classical gears and gear trains just described are the most common, some more elaborated transmissions have been developed for specific applications, in particular those where bulk is a major concern. This is the case in aircraft engines, which explains the interest raised recently by the planetary or epicyclic gearing, represented in Figure 1.3 scheme. Here, the gear system consists of an outer gear ring meshing with one or more planet gears, themselves revolving around a central gear called *sun gear*. The planet axes are linked together with a carrier plate, rotating around the sun gear too. Depending of the mounting system, the train will be called *epicyclic gearing* if the ring gear is stationary or fixed, or planetary gearing if the carrier plate is stationary or fixed. It can be noticed that all gear axes are coaxial, enabling a more compact design as well as a higher transmission ratio. As in the simpler case of a classical gear, all meshings have the same frequency f_m and it is possible to compute the theoretical rotation frequency of each element of the planetary gearbox as a function of the in-shaft, or out-shaft, rotation frequency:

$$\begin{cases} f_m = W_r f_c, \\ f_p = \left(\frac{W_r}{W_p} - 1\right) f_c, \\ f_s = \left(1 + \frac{W_r}{W_s}\right) f_c, \end{cases}$$

where W_r, W_p and W_s are the teeth number of the ring gear, the planet gear and the sun gear respectively, and f_m is the meshing frequency, f_c, f_p and f_s are the rotation frequencies of the carrier plate, the planet gears and the sun gear. The specificity of epicyclic gear makes traditional surveillance techniques inappropriate. Indeed, vibrations are generated at the same time from many meshing points, i.e. each planet gear meshes with both the ring gear and the sun gear all at once. A more extensive investigation of the particular gear system is done hereafter in Chapter 7.

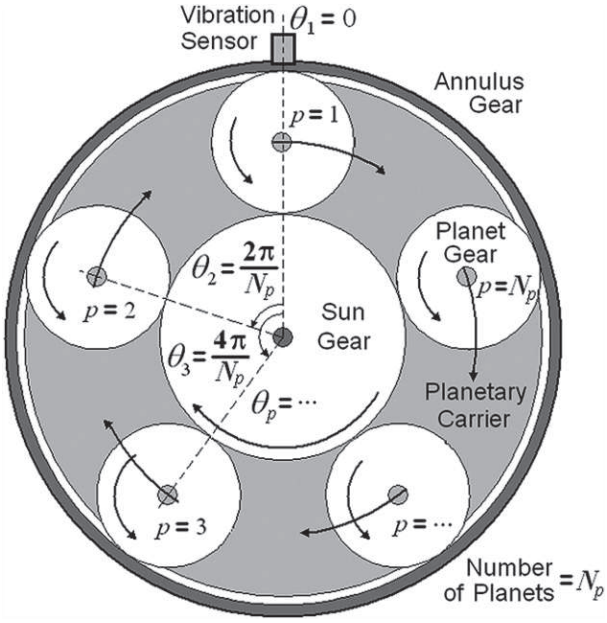


Figure 1.3: Drawing of an epicyclic gearbox with five satellites.

In addition to the general shape of a gearbox (Number of gears, of teeth, classical or planetary, etc.), the second issue to be addressed when designing a mechanical transmission system is the teeth shape. It usually belongs to one of the three following categories:

Spur gear It is the most common shape and allows a correct meshing at moderate speeds but tends to be very noisy at high speeds, Figure 1.4(a).

Helicoidal gear The teeth shape is not parallel to the rotation axis but has an angle which enables a softer and smoother meshing. Meshing is also more progressive, which diminishes the vibrations produced and thus, makes the gear more silent. In return, an axial effort is created that only depends on the inclination angle of the teeth. It has to be be counterbalanced by appropriate bearings, Figure 1.4(b).

Double helicoidal gear It is made with two helicoidal gears mirrored and joined together in a V shape, overcoming the problem of axial thrust of the previous gear, Figure 1.4(c).

1.1.2 Gearbox vibration signal

The issue of modeling the vibrations produced by a gearbox will be discussed in Chapter 2. For now, let us give some qualitative understanding of the general profile of the spectrum obtained when studying such signal. These intuitions are at the ground of most fault detection methods proposed in the literature and reviewed below in Section 1.2.

An ideal gearbox (i.e. where all teeth of each gear are strictly identical and the two gears are perfectly aligned) rotating at a constant frequency is a periodic system, getting back to the exact same state at each meshing. Thus, the only vibration produced

(a) (b) (c)

Figure 1.4: Several types of gears: (a) Spur gear, (b) Helicoidal gear, (c) Double helicoidal gear.

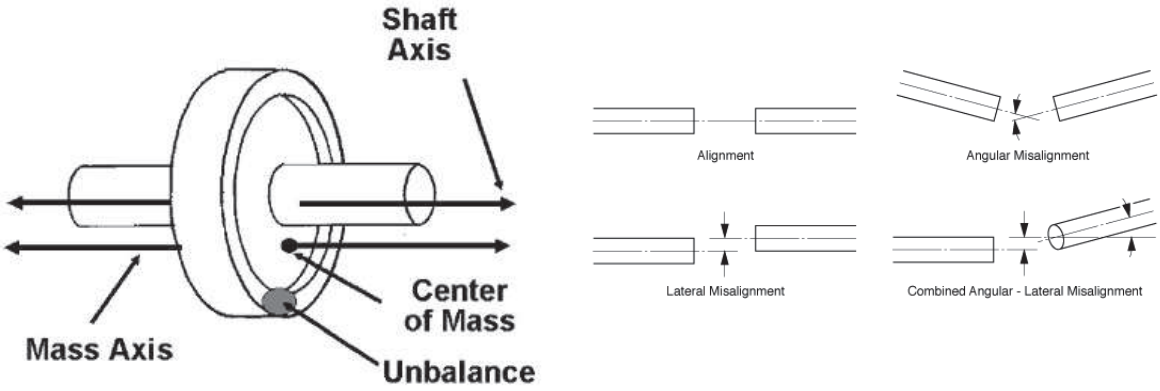
would come from the contact between the couples of gears: it would be a periodic signal repeated at the meshing frequency. In practice, no such a gearbox exists and vibrations are much richer and more complex. The main sources of discrepancy with respect to this ideal situation are the following:

- Manufacturing errors, which result in small variations in the teeth stiffness, induce an amplitude modulation pattern of the gear meshing vibration, changing the envelope shape and the global energy of the signal.

- Some errors regarding the alignment between the input shaft and the output shaft are almost impossible to avoid, even for the simplest gearboxes. Unbalance and misalignment are the two common shaft defects. Unbalance is an eccentric distribution of rotor mass, as in Figure 1.5(a) that produces additional force at the rotation frequency. Another common shaft failure is shaft misalignment, which occurs when the shaft of the pinion and the driven shaft are not coaxial, as illustrated in Figure1.5(b).

- The transfer function between the meshing point and the sensor changes the profile of the spectrum. This function can be decomposed into two parts. First, as any mechanical system, a gearbox is a body subject to the laws of fundamental dynamics, which means that the gearbox has its proper static transfer function, made of resonances and eigenmodes, which are likely to interfere with the signal within the frequency range of interest. Second part of that transfer function is the transfer path between the gearbox location and the sensor. Transfer path analysis gave rise to an important amount of literature [26], but it will not be detailed in the present work.

- Regarding the measurement of those vibration, the sensor is usually mounted as close to gearbox as possible, i.e. on the casing. But in a real system, the sensor can possibly be much further from the vibration source, which manifest into an



(a) (b)

Figure 1.5: Two major shaft defects: (a) Unbalance, (b) Misalignments.

increase of the transfer path between them, even the creation a combination of several paths.

To sum up, even a healthy gearbox already has a rich vibration content. But we will see in Section 1.1.3 that gearbox faults usually have a characteristic effect on the global vibration pattern.

Remark 3. *In vibration analysis of rotating machines, signals generated can be split into two parts:*

- *a deterministic part resulting from the system's kinematic with periodicities mainly related to the rotation speeds,*
- *a random part due to the non-periodic variations of the machine's operating, to the measurement noise...*

1.1.3 Typical faults of transmission systems

In standard operating conditions, direct metallic contact between two gears teeth are as limited as possible using a lubricant, which avoids rapid wear of the teeth surface condition. It is when the gear is poorly lubricated, as during the starting and the ending periods or in case of a faulty lubrication, that teeth surface damage (the most common kind of gear fault) tends to appear. These faults are numerous and varied in nature but mainly belong to two categories: on one hand the faults impacting all teeth of the gear (wear, pitting and spalling in the list below) and on the other hand the ones localized on a specific tooth (crack and galling in the list below). Let us shortly review the main categories of gear faults:

Wear is the progressive abrasion of the gear surface. This is a normal unavoidable phenomenon resulting in a slow and regular loss of the teeth thickness. In unusual faulty situations, wear becomes much faster. This is usually due to an oil problem and comes with the presence of debris.

Pitting takes the form of many very small holes on the surface of the gear. It is a usually explained by an extremely localized corrosion phenomenon caused by an auto

catalytic process. A small surface damage or a change in the chemical composition of the oil film protection can initiate the pitting process.

Spalling also results in the appearance of holes, but fewer and deeper than those characteristic of pitting. It is described in [82] as a macro-scaled contact damage caused by fatigue crack propagation.

Crack refers to a fracture of a material under a stress action. It is generated by the development of a discontinuity initially present at the material surface and usually appears at the bottom of the tooth.

Galling finally, is one of the most dangerous faults for the system. It is caused by adhesion between sliding surfaces, with a transfer of matter from one side to the other. It usually occurs when the oil film protection gets destroyed, due for example to a too high temperature.

As the faults nature and origin can be pretty different, their transcriptions into the vibration signal are consequently highly specific. Indeed generalized fault such as pitting is more likely to modify the envelope of the vibration signal whereas cracks induce a phase modulation of the meshing frequency. This is why for decades, several approaches have been proposed using a wide variety of signal processing techniques for fault detection. In the next section, several methods are introduced based on diverse techniques such as model-based and data-driven methods.

1.2 Fault detection methods

Fault detection methods used in vibration analysis can be classified following two criteria: model-based versus data-driven methods or stationary versus non-stationary methods.

Model-based methods for fault detection assume a model for the process and rely on parameter estimation or state observers. The model is developed *a priori* based on knowledge of the systems' physics. It can take the form of a mathematical function connecting the inputs to the outputs of the system or be more qualitative.

Data-driven methods for fault detection do not make physical assumptions and rely on the data history process only. Some features characterizing faulty behavior are extracted from the data using statistical tools.

Stationary and non-stationary methods refer to the system's operating conditions. Stationary regime means periodic behaviors and makes spectral analysis very efficient. But in practice, the system can undergo fast transient changes in terms of load, speed rotation, or both. More problematic, the load being usually higher during transition phases, some defaults become apparent only when the regime is non-stationary. This observation arouse intense reflection on the development of mathematical tools for non-stationary analysis.

In remaining of this section describe the main fault detection tools, splitting them into stationary (Sect. 1.2.1) and non-stationary (Sect. 1.2.2) methods.

1.2.1 Stationary methods

From a general point of view, a process is told stationary if its characteristics are time-invariant or, at least, periodic. In a stochastic framework, it would mean that its probability distribution is unchanged by a time shift. In particular, statistical moments such as mean and variance are time independents. In the case of gearbox vibration analysis, the regime is considered stationary if speed and torque are constant. Let us review the main fault detection methods used in this situation.

Statistical indicators

The most widespread techniques, data-driven, were established since the 70's and consist in monitoring some statistical features of the signal conveying information regarding the health state of the system [57, 81]. The formation of a fault in a gearbox leads to a change in the vibration signal in terms of energy or envelope shape. These modifications can be detected using statistical indicators computed from the amplitude and/or phase of the time signal, which trigger an alarm when exceeding a given threshold. The most usual indicators are presented below for a sampled time signal x, with N samples and a sample index k.

Root Mean Square (RMS) is defined as the square root of the arithmetic mean of the squares of the signal values and is given by:

$$RMS_x = \sqrt{\frac{1}{N}\left[\sum_{k=1}^{N} x_k^2\right]}.$$

The RMS indicator is relevant for detecting an energy dissipation in the global vibration signal.

Crest Factor (CF) is defined as the ratio of the maximum peak value of the signal to the RMS value:

$$CF = \frac{|x_{peak}|}{RMS_x},$$

where x_{peak} is the maximum peak value of the vibration signal. The crest factor is designed to increase in presence of high amplitude peak, caused by local tooth damage for example. For planetary gearbox diagnosis, a modified version of the Crest Factor (MCF) has been proposed, locally computed on a tooth-wide signal portion and thus returning a time process instead of a single value [20].

Kurtosis is the fourth normalized moment of the signal:

$$Kurtosis = \frac{N\sum_{k=1}^{N}(x_k - \bar{x})^4}{\left[\sum_{k=1}^{N}(x_k - \bar{x})^2\right]^2}.$$

It provides a measure of the impulsive nature of the signal. For a healthy gear we consider that the noise distribution follows a Gaussian distribution, thus the kurtosis equals 3, whereas for a faulty signal it will increase significantly [9].

Many other statistical indicators have been developed and tuned to be highly responsive to some particular faults. Among other we can cite FM0 and FM4 proposed by Stewart in [81], NA4 in [99], M6A [53] or NB4 [100].

Spectral analysis

Spectral analysis uses a representation of the signal in the frequency domain. This change from time domain to frequency domain is done breaking down the signal energy into frequency bands, usually using the widespread Fourier Transform for signals with finite energy:

$$X(f) = TF(x(t)) = \int_{-\infty}^{+\infty} x(t)e^{-i2\pi ft}dt.$$

In the case of a digital signal, i.e. a finite sequence $x_0, \cdots, x_{N-1}$ of uniformly-spaced samples of a continuous function, the computation of its frequency counterpart $X_0, \cdots, X_{N-1}$ is done with the Discrete Fourier Transform (DFT):

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{i2\pi kn}{N}}.$$

Spectral analysis is complementary to time domain analysis. Some hardly detectable phenomena in one of the domains are often clearly visible in the other one. Moreover, many operations are easier in the Fourier domain such as filtering and denoising.

Spectral analysis is an accurate method in machine surveillance as the relation between spectrum and gearbox kinematic is clearly established. Another point of interest is that it makes separation and identification of the vibratory sources possible, regarding the different element characteristics and their rotation speed [11].

Cepstral analysis

Although the presence of a periodic phenomenon at a given frequency can be directly read on the spectrum of the signal, the sharpness of such a phenomenon is rather related to the number and energy of its harmonics, i.e. to a kind of periodicity in the spectrum. In order to identify these periodicities, an operator has been created in 1963 and called *cepstrum* [7]. It is defined as the inverse Fourier transform of the spectrum logarithm:

$$\mathcal{C}_x(t) = TF^{-1}(\log[\tilde{x}(f)]),$$

where $\mathcal{C}_x$ is the cepstrum of signal $x(t)$ and $\tilde{x}(f)$ its spectrum. The cepstrum argument t is called *quefrency* although it has the dimensions of a time variable. Also, cepstrum has the major property of turning a convolution product into an addition, which can allow the separation of the vibration source from the transfer function.

Cepstral analysis can give impressive results when applied to machine surveillance. Indeed, a failure located on a tooth creates a periodic chock and thus a Dirac comb on the vibration spectrum, and finally a single peak on the cepstrum. This means the whole spectral representation of a default can be tracked through the evolution of a single parameter. Randall was one of the first to study the application of cepstrum to gearbox diagnosis [75]. For a healthy gear there are two peaks on the cepstrum, one for each gear quefrency, with the same amplitude. It was shown in [22] that when a periodic fault appears, the peak corresponding to the faulty gear increases while the other one decreases proportionally, making the problem clearly visible on the cepstrum.

Based on the power cepstrum of the vibration signal, an indicator was proposed in [4]. The faults can be detected observing the position of dominant negative rhamonic

response in the cepstrum. In [24] using the previous indicator, the authors developed a technique to differentially diagnose two types of localized gear tooth faults: a spall and a crack in the gear tooth fillet region.

Minimum Entropy Deconvolution (MED)

The minimum entropy deconvolution technique is a system identification method originally developed to aid extraction of reflectivity information in seismic data. It was first presented in [92] by Wiggins. MED is a method that allows to recover the output signal through a optimum set of filters bases on the maximum value of the kurtosis. As kurtosis is usually a good indicator for the detection of pulses in a signal, MED was proved efficient for the deconvolution of impulsive sources in a mixture of signals. MED [34] and related improved methods [23, 54] was proved useful to enhance detection of gear tooth fault.

Auto-Regressive models

An auto-regressive (AR) model is simply a linear regression of the current value of the series against one or more prior values of the series. The AR model of an order p is defined as:

$$x(n) = \sum_{k=1}^{p} a(k)x(n-k) + w(n),$$

where $a(k)$ are the AR coefficients and $w(n)$ is a Gaussian noise. The advantage of using AR model is that its parameters can be determined by solving a linear set of equations. AR modeling method is proven appropriate for the estimation of power spectra with sharp peaks, which is precisely the case of gear meshing vibration spectrum. This technique was used in [88, 23] to detect tooth cracks in gears.

Spectral Kurtosis (SK)

The spectral kurtosis (SK) is a statistical tool which can indicate the presence of series of impulsion and their locations in the frequency domain. In [1], a formalization of the SK by means of the Wold-Cramér decomposition of conditionally non-stationary processes is proposed. The SK was proved to be able to detect transients in the presence of a strong background stationary noise. This property is used in vibration-based condition monitoring of rotating machines [2]. The key idea is to use the high sensitivity of the SK for detecting and characterizing incipient faults that produce impulsive signals. The concept of kurtogram is also introduced, which displays the SK as a function of frequency and of spectral resolution. SK has also been used to detect tooth cracks in planetary gears [5].

Amplitude/Phase Demodulation (AM/FM)

One current technique used in gear fault diagnosis is amplitude and phase demodulation. Indeed, vibration signals of gearbox can be modeled as a carrier having the meshing frequency modulated by two signals having the rotation frequencies of the two shafts. These methods will be developed with more details in Chapter 3.

1.2.2 Non-stationary methods

In non-stationary operating conditions, the vibrations generated by a mechanical system cannot be analyzed with the spectral tools described in the above section. The non-stationarity concept concerns random signals as well as signals with frequency content and/or statistical properties changing over time. This is the case with aircraft engine gearbox' vibrations during take-off, where some important changes appear both regarding rotation speed and torque. To describe and visualize those signals, evolutionary analysis tools such as time-frequency and cyclostationary analysis have to be used.

Spectrogram

The very first technique that has been used to study transient signals is the spectrogram. The spectrogram is the result of the spectrum calculation over a band windowed signal [90, 89]. It is a two-dimensional graphic that represents the spectral energy content variations over time. The energy is usually given by the squared magnitude of the short-time Fourier transform (STFT). To study a signal frequency properties over time, the signal is first multiplied by a time window:

$$x_\tau(t) = x(\tau)h(t-\tau).$$

The STFT about is the Fourier transform of x_τ :

$$\tilde{x}_\tau(\omega) = \int_{-\infty}^{+\infty} x_\tau(t)e^{-2i\pi\omega\tau}d\tau.$$

Thus the energy density spectrum about time is computed as:

$$P(t,\omega) = |\tilde{x}_\tau(\omega)|^2.$$

Spectrogram is used in gear fault detection as it provides a simple representation of transient signals, such as ramp-up. It is also possible to derive time-dependent parameters from it, such as instantaneous energy, mean and median frequencies, and the bandwidth or standard deviation of the mean frequency. The major flaw of spectrogram is that it tends to smooth the characteristics of the signal and can miss some short-time phenomena. Spectrogram was used in [96] in a comparative study in order to perform fault detection on gears.

Wavelet

Unlike the Fourier transform, where the signal is broken down on a basis of sine functions, wavelet analysis uses a class of real and complex bases of non-stationary functions named wavelet. Those are chosen to best fit the signal [59, 32, 51].

The wavelet approach is essentially an adjustable window Fourier spectral analysis with the following general definition:

$$T_y(s,\tau) = \frac{1}{\sqrt{s}}\int_{-\infty}^{+\infty} x(t)\psi\left(\frac{t-\tau}{s}\right)dt.$$

where $\psi(\cdot)$ is the basic wavelet function that satisfies certain very general conditions, s is the dilation factor and τ is the translation of the origin. Although time and frequency do

not appear explicitly in the transformed result, the variable $\frac{\cdot}{s}$ gives the frequency scale and τ the temporal location of an event.

Wavelet analysis can be assimilated to a series of correlation between the signal and the wavelet in time domain, which brings out all moments the signal locally looks like the wavelet shape. Wavelet analysis was successfully used in non-stationary vibration signal processing and fault diagnosis [87, 12, 49]. The major problem with this time-scale representation lies in the choice of the wavelet that best highlights the sought information in the signal.

Wigner-Ville distribution

Wigner-Ville distribution (WVD) is another kind of time-frequency analysis [16]. It is sometimes also referred to as the Heisenberg wavelet. By definition, it is the Fourier transform of the central covariance function. For any time function $x(t)$, the central variance can be defined as

$$c(\tau, t) = x\left(t + \frac{\tau}{2}\right) x\left(t - \frac{\tau}{2}\right)^* .$$

Then the WVD is set as:

$$W_x(t, f) = \int_{-\infty}^{+\infty} c(\tau, t) e^{-2i\pi f \tau} d\tau.$$

It can be noticed that even for times that are far from t, the window weight is the same as if they were near t, which makes the WVD highly non local. Furthermore, as Wigner-Ville distribution is not a linear transform, a cross term appears when the source is the sum of two signals. This interference may be useful for identification of multicomponent signals but it makes in general the interpretation of the distribution harder. This is why a compromise has to be done between the precision of the spectral content and the importance of the interferences due to the cross term. The application of this method to gear faults in particular began with the works [27] by Forrester. He applied the Wigner-Ville distribution to averaged gear vibration signals and showed that different faults such as a tooth crack and pitting could be detected in the WVD plot. McFadden and Wang applied the usual WVD and a weighted version of it to gear failure analysis in order to improve the detection capabilities of the method [90, 89].

Cyclostationarity

A *cyclostationary process* is a specific case of non-stationary random signal, the statistical properties of which are time varying but periodic [8]. This kind of signals can be obtained for example when a periodic signal undergoes a random uncorrelated disturbance with periodic amplitude, as it is the case for rotating machines and more specifically for gearboxes. A random process can be called n^{th} *order cyclostationary* if all its statistical moments until the n^{th} order are periodically time-varying. In the case of second order cyclostationarity, it means that the average m_x and autocorrelation functions r_{xx} of a random process $x(t)$ verify: $m_x(t + T_0) = m_x(t)$ and $r_{xx}(t + T_0, \tau) = r_{xx}(t, \tau)$, where T_0 is called the *cyclic period*.

Remark 4. *The Wiener–Khintchine theorem states that the autocorrelation function of a wide-sense-stationary random process has a spectral decomposition given by the power*

spectrum of that process, which allows us to link the autocorrelation function with Fourier series

$$R_{xx}(t,\tau) = E\left[x\left(t+\frac{\tau}{2}\right)x\left(t-\frac{\tau}{2}\right)\right] = \sum_{\alpha\in A} R_{xx}^{\alpha}(\tau)e^{2i\pi\alpha t)}.$$

The set of all the cyclic frequencies α, defined as $\alpha = \frac{1}{T_0}$, is called the *cyclic spectrum* of the signal. This variable is used to define the *spectral correlation*:

$$C_s(f,\alpha) = E\left[S\left(f+\frac{\alpha}{2}\right)S^*\left(f-\frac{\alpha}{2}\right)\right].$$

Spectral correlation can be seen as a measurement of the correlation degree between all the signal frequencies components. Cyclostationarity analysis can be employed to inspect the non-linearity induced by a breathing crack during fatigue damage. It has been shown that the magnitude of the cyclic frequency increases with crack depth, and it can be used as an indicator for fatigue damage [10]. The cyclostationary approach is very well suited to machine diagnostics. Indeed, a fault occurring in a rotating component will produce a repetitive release of vibration energy, which creates a signal with a strong cyclostationary behavior. The presence of a fault, such as tooth spalling or tooth crack, usually affects the vibration signals by periodically imposing random modulations, which is a good example of cyclostationary behavior. It was shown that the appearance of a default increases the second order cyclostationary component level [43].

Time Synchronous Average (TSA)

The time synchronous average (TSA) is a method of background noise reduction, used for periodic signals [56] generated by rotating element of mechanical systems. Usually the synchronous average is calculated using a trigger signal convoluted with the time signal. In the frequency domain, it is equivalent to multiply the Fourier transform of the signal with a comb filter [11]:

$$C(f) = \frac{1}{N}\frac{\sin(\pi NTf)}{\sin(\pi Tf)}.$$

The TSA is particularly well suited for gearbox analysis, where it allows the vibration signature of the gear under analysis to be separated from other gears and noise sources in the gearbox that are not synchronous with that gear. In order to consider both gears, TSA can thus be done on the lower common multiple (LCM) of the two gears rotation periods. In order to detect gearbox failure, analysis can be performed on both the synchronous average of the gear at the frequency of the default and on the residual signal. Indeed, periodic faults synchronous with the gear frequencies are enhanced in the TSA signals and thus become more easily detectable. However, removing the TSA from the global vibration signal, allows getting rid of deterministic component of the gear signal, and helps emphasizing asynchronous perturbations that are yet representative of faults.

TSA can also be used as a tool for extracting gear mesh vibrations from composite vibration signals, for example in a planetary gearbox. After performing a TSA, the resulting vibration signal corresponds to one complete revolution of the gear under consideration. Thus changes in the vibration waveform due to damage on individual teeth can be identified [20, 19]. Synchronous average has been used to detect fatigue cracks on the carrier of a planetary gearbox. The method is based on detecting changes in the modulation pattern of the fundamental gear mesh vibration created by the crack [6].

Kalman filter

In statistics and control theory, *Kalman filtering* is an technique that estimates variables state from a series of measurements observed over time that may be incomplete or noisy.

Within the category of stochastic approaches, a general fault detection and diagnosis procedure was first expressed in [58]. The author proposes to use residuals of a Kalman filters to perform diagnosis of dynamic systems. The faults are diagnosed using statistical indicators reflecting whiteness, mean and covariance of the residuals. Further research has led to using modified Kalman filter techniques, such as *extended Kalman filters* (EKFs) [80] or *unscented Kalman filters* (UKFs).

Artificial intelligence methods

In [38], *artificial intelligence* (AI) has been recently defined by Kaplan and Haenlein as a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation. Historically, the starting point of this topic is the 50's with Alan Turing's research work, who wonders if machines can "think" [84].

Some have considered fault diagnosis as pattern recognition problem, and AI was proved to be a powerful pattern recognition tool [50]. Due to the complexity of mechanical response signals, it is impossible to directly extract and recognize fault patterns. This is why the process is cut into two parts: a data extraction and preprocessing step followed by fault recognition. Here we will briefly present some of the most used methods for fault diagnosis, and more specifically classifiers and statistical learning methods. Extensive literature exists on the topic, but it is not the central interest of the present thesis.

k-nearest neighbor *k-nearest neighbor* (k-NN) is the simplest nonparametric decisions procedure that assigns to unclassified observation the label of the nearest sample [17]. It is used for both classification and regression. There are three basic elements in the k-NN algorithm: the number of measured instances k, the distance metric and the decision rule for classification. k-NN was proved to be an efficient way to perform multi-fault diagnosis in gearboxes [45] but also to detect cracks in [86].

Bayesian classifiers *Naive Bayes classifiers* are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features [63]. Naive Bayes first learns the joint probability distribution of the input and output by the conditional probability distribution based on the conditional independence assumption. Then, based on the learned model, the output label with the biggest posterior probability, for a given input, can be calculated via Bayes' Theorem. Naive Bayes classifiers was used for fault detection for induction machines [79], induction motors [66] as well as incipient bearing failures [64].

Support Vector Machine *Support vector machine* (SVM) are supervised learning models used for classification and regression analysis. To do so, the SVM builds a hyperplan to discriminate the set in a non-linear space. Some researchers have used SVM-based methods in order to classify several health state conditions of gearboxes [41]. Cracks

were also detected in the carrier plate of planetary gearbox with SVM combined with least square [40].

Neural Network *The neural network* (NN) is a framework for many different machine learning algorithms to work together and process complex data inputs. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. In [71], the authors used a hierarchical neural network to perform classification for bearing fault detection. An NN-based procedure was used for fault detection and identification of gearboxes using a vector extracted from standard deviation of wavelet packet coefficients of vibration signals [73].

Chapter 2

Mechanical modeling of gearboxes

This thesis is mainly focused on empirical modeling of vibrations produced by gear rotation. As gearboxes are highly complex mechanical systems, it is important to have a some understanding of its functioning from a mechanical point of view before proposing simplified models. Gearbox mechanics has been a major research field in the last decades, mostly driven by the need for models helping designing industrial products. Several approaches exist, each of them with its own specificity and benefits. Gear modeling can be split into two main categories: empirical signal modeling and dynamic modeling. The former can be used as a first approximation of the vibration signal, especially when the problem does not require a deep analysis or contains too many unpredictable factors, as is for example the case of operating system monitoring. The latter is based on a more fundamental analysis of the gear mechanism, modeling all physical interactions using either Newton's laws of motion, or equilibrium reasoning and quasi-static approximations. In the present Chapter, three different approaches are briefly described in order to help the reader locating the work presented in the remaining of the manuscript in the big picture of the gear vibration theory.

2.1 Empirical signal modeling

A first route to gearbox vibration modeling is looking for a general shape retaining the main features of the vibration signals observed in practice, both on time and frequency representations.

To that end, qualitative understanding of the phenomena generating the vibration is necessary. Considering a pair of gears, the main source of vibration comes from the meshing force applied at the contact point between the teeth of the pinion and driven gear. The time profile of this force can vary from one tooth to the other, and what is observed in practice is that this variation takes the shape of two amplitude modulations, each of them having the frequency of one of the two wheels.

Remark 5. *One important thing to note is that in the perfect case of an ideal couple of gears, i.e. where the stiffness is the same for all teeth, vibrations would be a pure periodic signal at the tooth-meshing rate.*

According to this observation, a simple modulation model was proposed as an approximation of the vibration stemming from the gear rotation [13].

$$s(t) = s_{carrier}(t) \times (1 + s_{gear1}(t) + s_{gear2}(t)) \,, \tag{2.1}$$

where $s_{carrier}(t)$ is the high frequency component representing the gear mesh and $s_{gear1}(t), s_{gear2}(t)$ are the two modulations coming from the pinion and driven gear rotations. $s(t)$ represents here the vibration at the meshing point between the gears.

In reality, the accelerometer used to measure the vibrations is located on the casing of the gear, which can be more or less distant from the meshing point. That is why a transfer function is usually also added to the vibration model along with white Gaussian noise in order to have a more realistic representation of the measured signal

$$x(t) = s(t) * h(t) + w(t), \tag{2.2}$$

where $x(t)$ is the representation of the measured vibration at the accelerometer point, $h(t)$ is the transfer function which can be seen as the path taken by the mechanical wave from the meshing point through the casing, and $w(t)$ the additive white Gaussian noise.

But in a real vibration signal, even in the case of a healthy gear, many additional components also contribute besides the gear meshing. Those have various origins, such as tooth deflection under load or geometrical errors in the tooth profile. Machining errors is one of the source of tooth profile differences that may produce either random or periodic variations. The latter are sometimes called ghost components, and correspond to the integer number of teeth of the gear. Therefore, they appear at the rotation frequency of the gear. Moreover, operating parameters such as load and rotation speed have been assumed so far to be constant over time and teeth spacing for each gear was supposed to be identical. Fluctuations of some of the above parameters will generate additional modulations. Sensitivity of vibration to tooth loading is more likely to generate an amplitude modulation pattern while variation of the rotation speed or tooth spacing are more susceptible to engender a frequency modulation of the signal generated by the gear mesh.

Remark 6. *As in the vibration the sidebands generated by both types of modulation have the same frequency components, in the empirical model 2.1, the frequency modulation phenomenon is neglected in order to simplify the problem.*

For amplitude and phase modulation issue, a different model has been used in order to represent the vibration signal of the gear. Based on an extension of the previous amplitude modulation model to include the phase modulation Eq.2.3 gives the usual representation of both phase and amplitude modulated signal,

$$s(t) = s_{carrier}(t + s_{\Phi}(t)) \times (1 + s_{gear1}(t) + s_{gear2}(t)), \tag{2.3}$$

where all variables are the same as in Eq.2.1 and $s_{\Phi}(t)$ is the periodic phase modulation of the carrier signal.

Remark 7. *It has to be mentioned that the formation of faults obviously produce major changes on the gear and thus in its vibrations and therefore in the signal model itself. Some faults may alter mainly the amplitude of the vibration without changing its spectral content but it is also known that for example, crack formation generates some frequency modulation sidebands such as the stiffness of the tooth is modified.*

2.2 Lumped parameter modeling

An approach often used for mechanical systems modeling, more elaborated than the empirical model described above, is lumped parameter modeling. This representation

is based on a simplified description of the systems behavior. The mechanical structure under study is decomposed into elementary components, i.e. solid parts with given mass and inertia moments, whose interactions are described by a system of springs and dampers. In other words, several assumptions are made to simplify the complex nature of gears, namely, all objects are regarded as perfectly rigid bodies while all interactions between those rigid bodies are split into spring and damper actions.

The major interest of this representation is to reduce the number of degree of freedom (dof) before solving the equations describing the system.

In the case of gears, setting the main body of the gear as a rigid body means concentrating all the deformations in the teeth part.

In the literature, many studies have used lumped-parameters models in order to enhance understanding of gear mechanisms [65, 85]. As an example, simple lumped-parameter representations can not only provide insight into the mechanisms that generate forces and moments but also help for elucidating the non-linearity of the gear mesh and explain how tilting/twisting moments impact gear vibration.

In [68], the authors modeled the contact of a spur gear pair in order to investigate its non-linear dynamic response. When interesting in studying the gear mesh, a simple representation such as illustrated in Figure 2.1, considering a single degree of freedom model is enough.

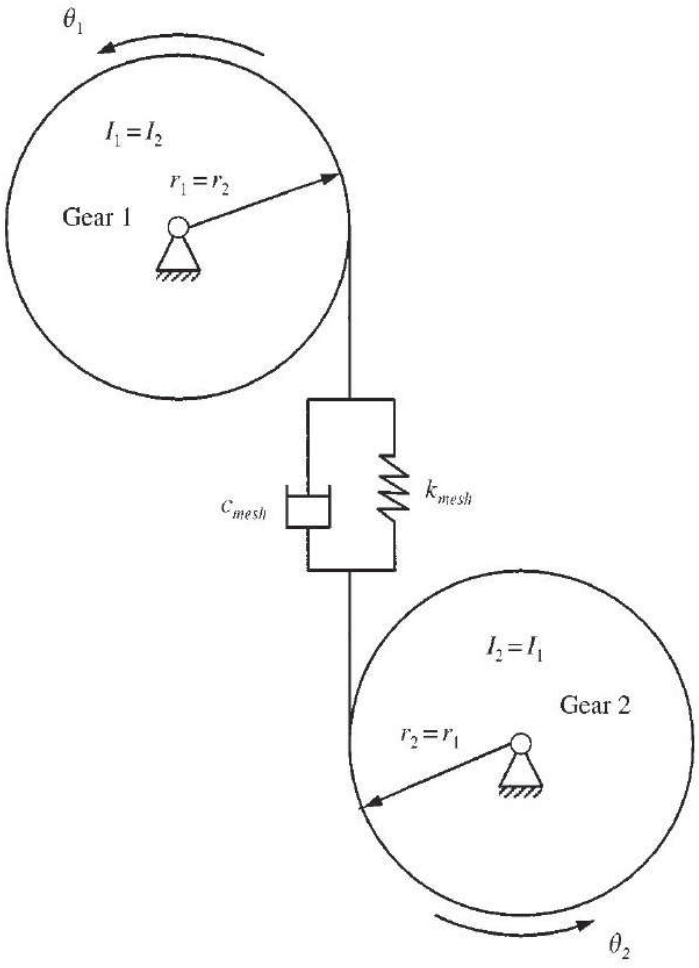


Figure 2.1: Two-gear system modeled with a single degree of freedom.

The unique contact modeling permits dynamic response analyses (as opposed to the

much simpler static and natural frequency/vibration mode analyses) with a reasonable number of degrees of freedom.

In order to improve the accuracy of the model, some more complex modeling has to be considered with a bigger number of degree-of-freedom. In [67], the authors proposed a 6-dof non-linear model. The model includes four inertia, prime mover, pinion and gear. The torsional compliance of shafts and the transverse compliance of bearings combined with those of shafts are included in the model. The 6-dof lumped-parameter model is illustrated in Figure 2.2.

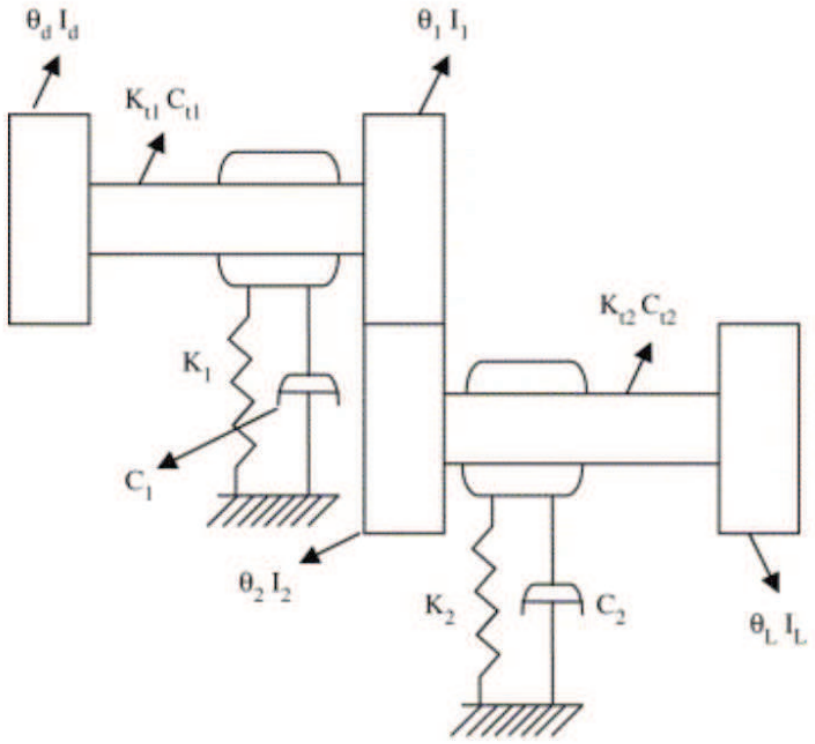


Figure 2.2: 6-degree-of-freedom nonlinear model of the fixed-shaft gearbox.

With this model, the response, including modulations due to transverse and torsional vibration stemming from bearing and shaft compliance, can be calculated.

Then, in order to improve modeling, several upgrades of this type of model have been proposed. In [44], the authors have combined the 6-dof transverse-torsional dynamic model with an elastohydrodynamic lubrication model of the spur gear. This tribodynamics model allowed to capture all transient effects associated with the gear kinematics but also to measure tooth surface roughness profiles. Some studies proposed to add failures into models in order to better understand the way it affects the dynamic response of a spur gear. In [14], the authors studied crack propagation in the tooth by modifying a 6-dof model in order to consider crack mechanical characteristics.

2.3 Finite element modeling

The last approach to mechanical system modeling is the *Finite Element Method* (FEM). This technique is widely employed in the industry for dynamic modeling of gears since it allows describing space- and time-dependent problems. Overall, FEM provides an

approximate solution of a partial differential equation (PDE) defined on a compact domain for specific boundary conditions, describing the physical behavior of a system under study. This numerical technique is performed in three steps. The discretization, first, consists in dividing the whole system into simpler and smaller pieces called finite elements, each of them described a few parameters. Then, an equation on these parameters is written, making the discretized system a good approximation of the real, continuous-time/continuous space system. Finally, the equation is solved by numerical means. This method also needs a proof of the convergence toward the system solution of the approximted solution. Looking deeper into the discretization step, we see it involves defining a mesh on the whole space, the cells of which are the finite elements. Usually, these elements are triangles or quadrangles, but more complex shapes can be chosen.

Remark 8. *It has to be noticed that the more vertices the shape contains, the higher is the dimension of the obtained equation: a compromise has to be done between the precision/detail of the mesh and the computational cost of the algorithm.*

A way to dramatically reduce the complexity of such a model is noticing that drawing a regular mesh is not mandatory. Indeed, in some cases, only a specific zone of the whole system is of interest: when studying crack formation on gears for example, it is more interesting to have a precise representation of the system in the fillet part of the tooth than on the entire gear, as shown on Figure 2.3.

Once a mesh is defined, continuous functions describing the dynamics of the system (stress or deformations for example) are approximated by a combination of basis functions attached to the nodes: in practice, a system of piecewise linear functions is generally chosen but it is also common to find piecewise polynomial functions. The PDEs can then be substituted with a system of ordinary differential equations (or classical algebraic equations in the quasi-static case). Finally, a solver is chosen among several numerical algorithms which usually belong to one of the two main categories: iterative and direct solvers. The final solution is eventually extracted during a post-processing phase, where the best representation of the solution is selected depending on the application. For example, in Figure 2.4, the constraints generated by the contact between the two gears are represented in a 2D plot.

Remark 9. *There can be a last step where approximation errors are computed and compared to a threshold of acceptability. If the latter is exceeded, one goes back to the discretization step and starts over with a more precise mesh.*

In the specific case of gears, the most elaborated models of the literature are tridimensional and describe several physical quantities such as elasticity and stiffness along with numerous sources of excitation such as the gearmesh, mounting faults and manufacturing errors.

In spite of the rich description of gears they provide, these models also have strong limitations. They are limited to quasi-static regime due to the very high dimension of the vector describing the state of the system: solving ODE's of this size is in general out of reach. Indeed, the mesh has to be precise enough to capture small deflections near the teeth. Note that under this quasi-static assumption, all phenomena related to inertia are discarded. Another restriction is that the limit conditions of the equations should depend on all other mechanical components interfering with the gear, such as bearings, shafts and casing, which is impossible in practice. More generally, the mechanical behavior of a system is highly dependent on its (unknown) environment.

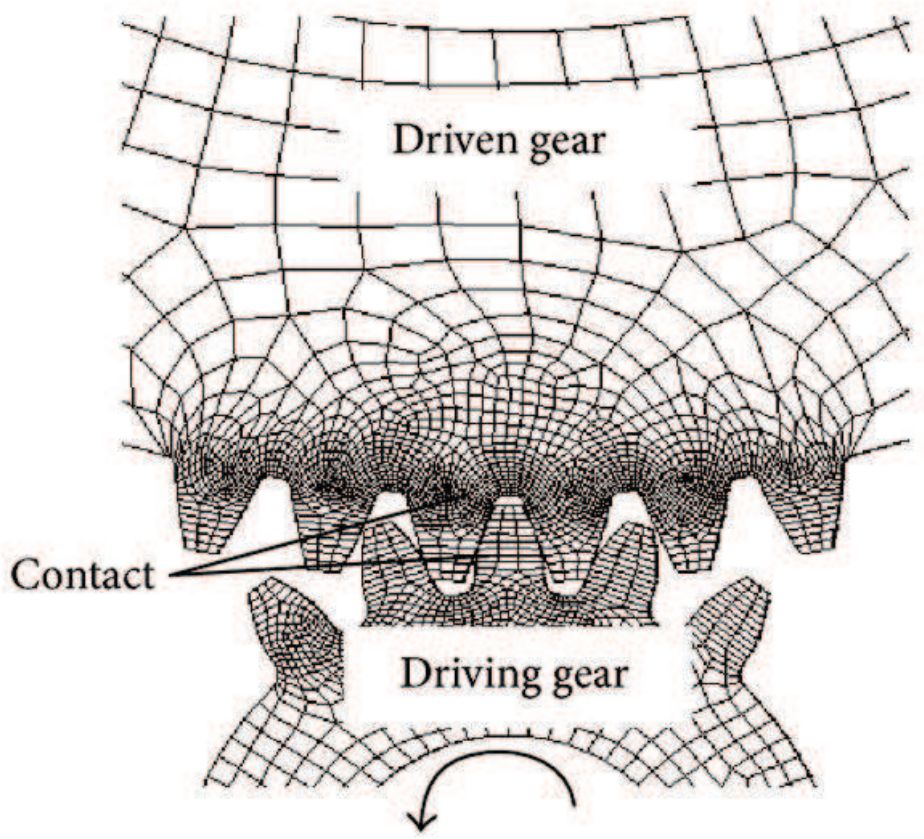


Figure 2.3: Example of a gear discretized with quadrangles elements. Here the meshing is getting denser from the gear centre over the teeth.

2.4 Conclusion

This chapter was dedicated to describing the most common modeling techniques applicable to gear dynamics. Each of them comes with its own benefits and drawbacks, which can be summed up as follows:

Empirical modeling allows direct description of the signal actually observed, retaining its main features while sparing a lot of computation time. It is the most natural approach to fault detection: on the one hand if for instance a crack happens to create a kind of modulation, then demodulation should provide a good indicator and should be tried. On the other hand, mere observation can give a false idea of the true nature of a phenomenon. Presence of sidebands around the meshing frequency, for instance, evokes a modulation but can result from a amplitude or phase modulation, a combination of both or even a different operation. This issue will be discussed in Chapter 3.

Lumped parameter modeling allows covering dynamical behavior and in particular inertia effects. As opposed to empirical modeling, it allows testing the relation between a physical property (say for instance stiffness variations) and a given characteristic of the observed signal. Yet, it relies on strong assumptions (infinitely

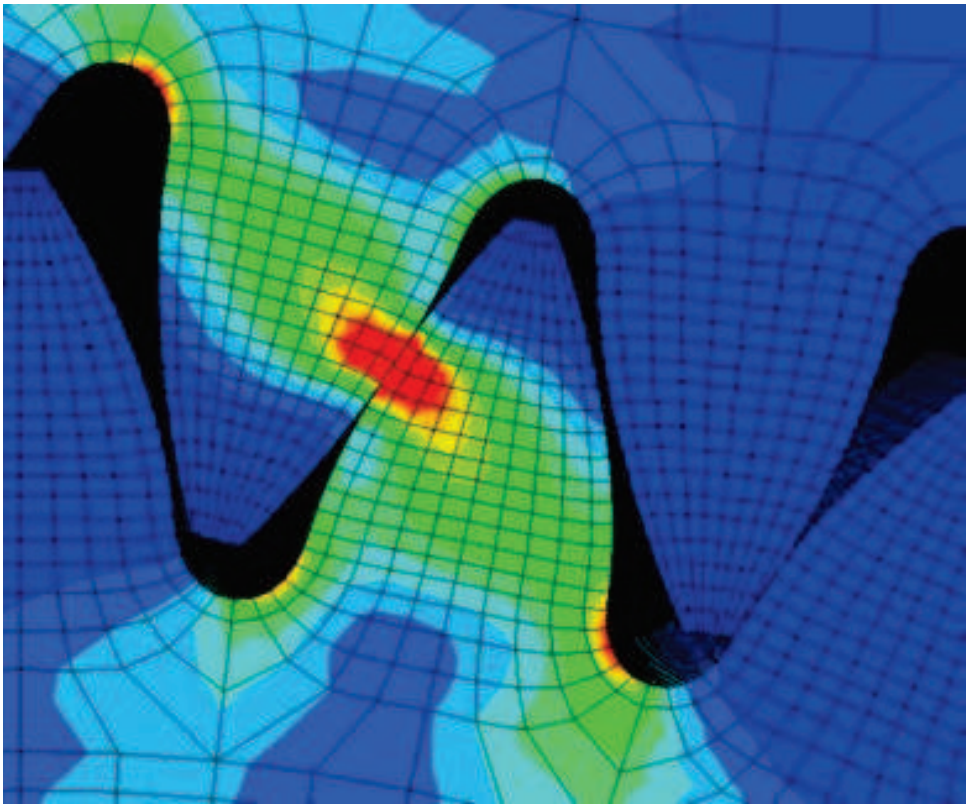


Figure 2.4: Numerical simulation using Finite Element Method for a couple of gears. Here is shown the force applied at the contact point between the pinion and the driven gear along with the propagation of that force toward the center of the gears.

rigid pieces) and requires a first additional modeling step relating the lumped parameters to the gear geometry and used materials. Moreover, it does not help estimating internal constraints.

Finite Element Modeling is an attempt to take into account the continuous nature of the pieces at play. This results in a state representation having a great number of degrees of freedom. It follows a high-dimensional differential equation which cannot always be solved without a quasi-static assumption. In this case, all inertia-related effects covered by lumped parameter modeling are lost. On the other hand, parameterization is less problematic as all required parameters have physical meaning. To retain the benefits of both approaches some hybrid models was proposed where finite elements are used only near the contact surfaces.

The remainder of this thesis will be dedicated to determining if empirical models developed until now for gearbox monitoring do provide a sufficient representation of the signals measured in practice.

Chapter 3

Our new approach to demodulation

3.1 Some reminders about modulation

Modulation is the process of applying a low-frequency perturbation to a high-frequency wave called "carrier", usually with the aim of encoding information. This perturbation can be of several kinds, but the two most commonly encountered classes are amplitude modulation (AM) and angle modulation. In the field of telecommunications this perturbation is deliberate: it is used to transmit a message. In the case of mechanical signals the perturbation reflects imperfect operating conditions and thus, can contain information regarding the health condition of the system. Let us describe in more details these two families of modulated signals (amplitude and angle modulated signals).

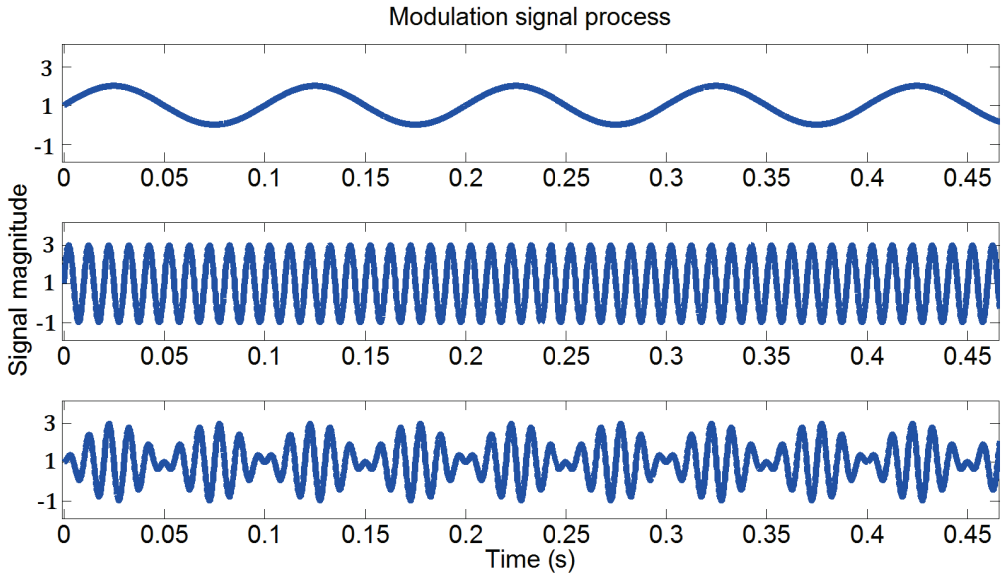


Figure 3.1: From top to bottom: Low frequency signal usually constituting the useful information, high frequency signal used to "carry" the information and amplitude modulated signal computed as the product of the low and high frequency signals.

Amplitude modulation of a signal $s_c(t)$ by a signal $s_m(t)$ returns a signal $s(t)$ defined as the product below:

$$s(t) = s_c(t)s_m(t). \tag{3.1}$$

In more visual terms, this operation can be thought of as a modifying of the envelope of the carrier signal $s_c(t)$ to give it the same shape as the modulating signal $s_m(t)$ as shown in Figure 3.1.

In the Fourier domain, a product signal such as Eqn.(3.1) becomes a circular convolution of the spectra $\tilde{s}_c$ and $\tilde{s}_m$ of the carrier and modulating signals: $\tilde{s} = \tilde{s}_c * \tilde{s}_m$. The result is represented in Figure 3.2: the spectrum of the modulation is repeated at all multiples of the carrier frequency and multiplied by the corresponding harmonic.

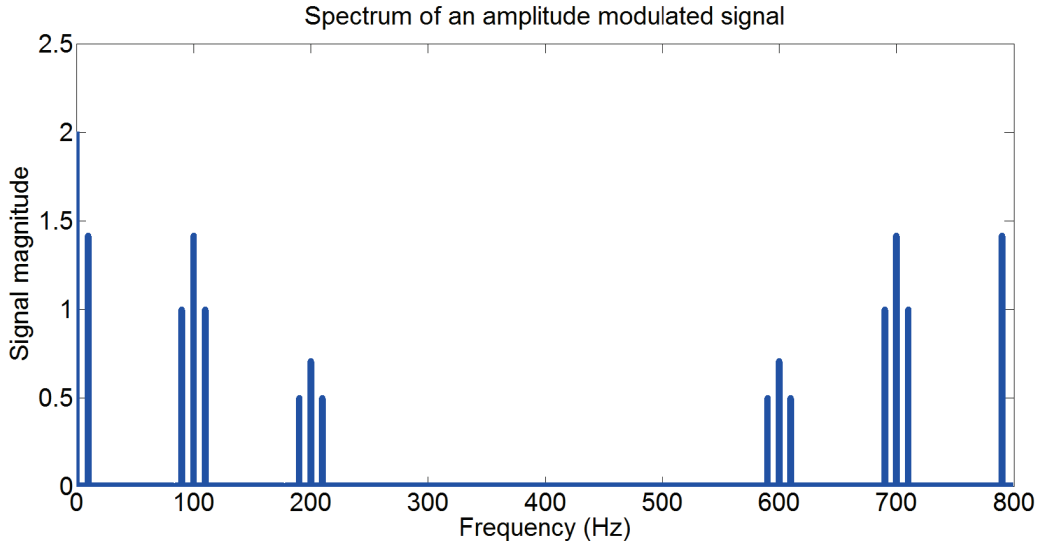


Figure 3.2: Spectrum of an amplitude modulated signal where the carrier signal has two harmonics.

That modulation pattern repeated about frequencies f_c, $2f_c$, etc. is proportional to the spectrum of $s_m(t)$. This characteristic gives an immediate way to access its spectrum and thus its temporal representation through the demodulation process developed hereafter in Section 3.2.

In the case of angle modulation, the information is encoded as a perturbation of the phase or the frequency of the carrier signal, which are referred to as phase modulation (PM) and frequency modulation (FM). In FM the variations of the carrier frequency are controlled by both the frequency and the amplitude of the modulating wave. In PM, the instantaneous phase deviation $\Phi(t)$ of the carrier is controlled by the modulating waveform, such that the frequency remains constant. Generally speaking, the perturbation used is a real periodic signal named s_Φ. The output $s(t)$ of the phase or frequency modulation of a "carrier" signal s_c by a signal s_Φ is then defined as below:

$$s(t) = s_c\left(t + s_\Phi(t)\right). \tag{3.2}$$

The profile of the resulting signal is illustrated on Figure 3.3 in the case of frequency modulation: we see that the shape of the signal taken on one period is unchanged but a time-varying horizontal scaling is applied.

In the Fourier domain the effect of an angle modulation is more complicated than for amplitude modulation, but a formula can be derived in the simplified case of a sine wave modulated by a second sine wave:

$$s(t) = \cos(\omega_c t + \alpha \sin(\omega_m t)), \tag{3.3}$$

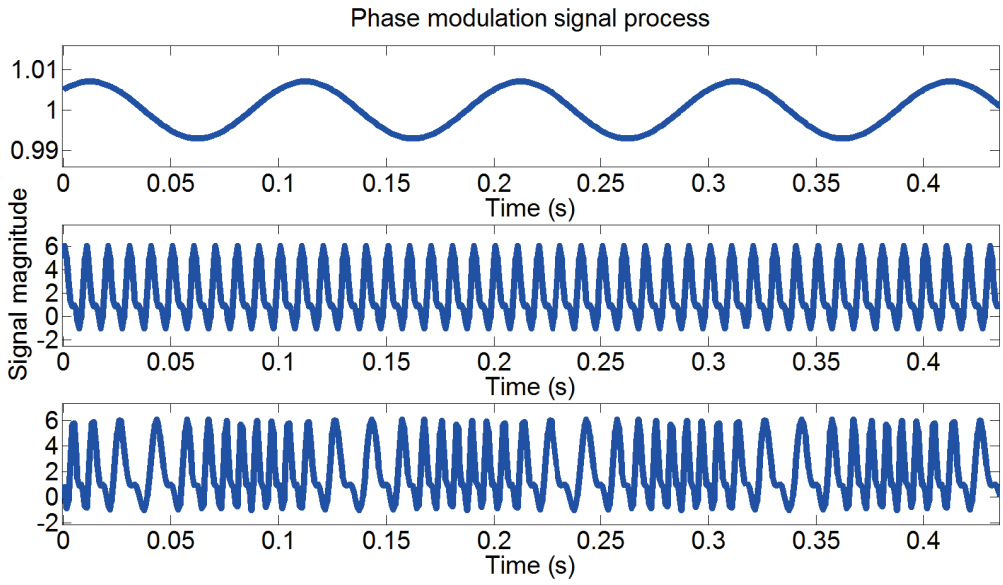


Figure 3.3: From top to bottom: low-frequency signal sometimes referred to as the "message", high-frequency carrier signal and phase modulated signal constructed from Eqn. (3.2).

where ω_c and ω_m are the pulsations of the carrier and the modulating signal respectively and the factor α is called the modulation index. The spectrum of an angle modulated signal is given by the amplitudes of the Bessel functions J_k according to the signal's modulation index α, where k represents the number of sidebands. The impact of the modulation index α on the specific case of phase modulated signal is illustrated by Figure 3.4.

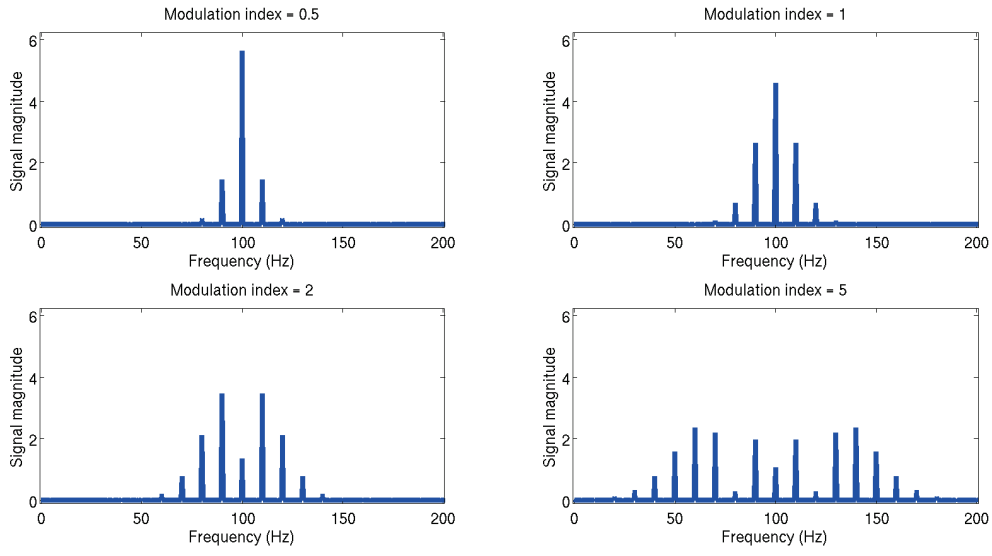


Figure 3.4: Spectrum of a phase modulated signal for four different values of the modulation index α: $\alpha = 0.5$, $\alpha = 1$, $\alpha = 2$ and $\alpha = 5$.

Contrary to amplitude modulation, where the number of sidebands created around each harmonic of the carrier is the number of harmonics of the modulating signal, angle

modulation always generates an infinite number of sidebands. Thus, it can be shown that 98% of the signal's energy is included in a frequency bandwidth around the carrier frequency, called Carson bandwith, computed as $B_C = 2(f_m + \Delta f)$. In the phase modulation example above, it can be noticed that for $\alpha < 0.5$, i.e. $B_C \approx 2\Delta f$, the spectrum of the angle modulated signal is similar to the amplitude modulation one. The spectrum's energy is mostly concentrated on the peak coming from the carrier signal and the modulations present around are quickly negligible. However, when the modulation index increases, the corresponding spectrum has less and less negligible modulation harmonics as the energy of the carrier frequency is distributed over the harmonics of the modulating signal.

Remark 10. *Amplitude modulation changes the global signal energy, whereas angle modulation only changes the distribution of this energy over the spectrum.*

Finally, a signal can be both phase and amplitude modulated at the same time. The process is defined as a combination of the two previous models (3.1) and (3.2):

$$s(t) = s_c\left(t + s_\Phi(t)\right) \times s_m(t). \tag{3.4}$$

There are many situations where both amplitude and phase modulations are present, and where instantaneous amplitude and phase are commonly used for the detection of broken rotor bars, eccentricity fault and bearing defects.

Although initially motivated by telecommunications, the theory of modulation finds applications to other domains of signal processing and, in particular, to the analysis of mechanical signals. Indeed, we showed in the previous Chapter that vibration signals generated by rotating machinery have a spectrum recalling the modulation patterns just described. As a consequence, modeling the impact of a fault as a modulation (either of the phase or the amplitude) of a reference periodic signal has become a standard idea leveraged by numerous works [52]. While this reference signal characterizing a healthy gearbox has the same frequency as the meshing, the modulation assumed to reflect a fault has the frequency at which the default goes through the meshing (i.e. usually the rotation frequency of the damaged gear). For this reason, in order to improve fault detection, we are interested in recovering a modulated signal as precisely as possible. This issue, called demodulation, is the topic of the remaining of the present chapter.

3.2 Classical demodulation

Demodulation was first defined as recovering a low-frequency signal and a high-frequency signal out of their product, in the case of amplitude demodulation, and has been then extended to include angle modulation.

It has been used in telecommunications for several decades and the first implementations were entirely analogical. The ideas they were based on were later transposed to the digital digital world and improved. In the present thesis, they will be referred to as "classical demodulation" and described in the sections 3.2.1 and 3.2.2 below.

3.2.1 Monocomponent signals

When considering the simplest case of amplitude modulated signal, the amplitude modulation of the carrier is composed of a single sine wave. The classical solution to that

demodulation problem is illustrated on Figure 3.5 and detailed hereafter. One carrier harmonic is first selected and the rest of the spectrum is filtered out. The remaining spectrum is then shifted to zero, which makes it proportional to the spectrum of the modulation signal. Naturally, its time counterpart is then proportional to the time-varying modulating signal.

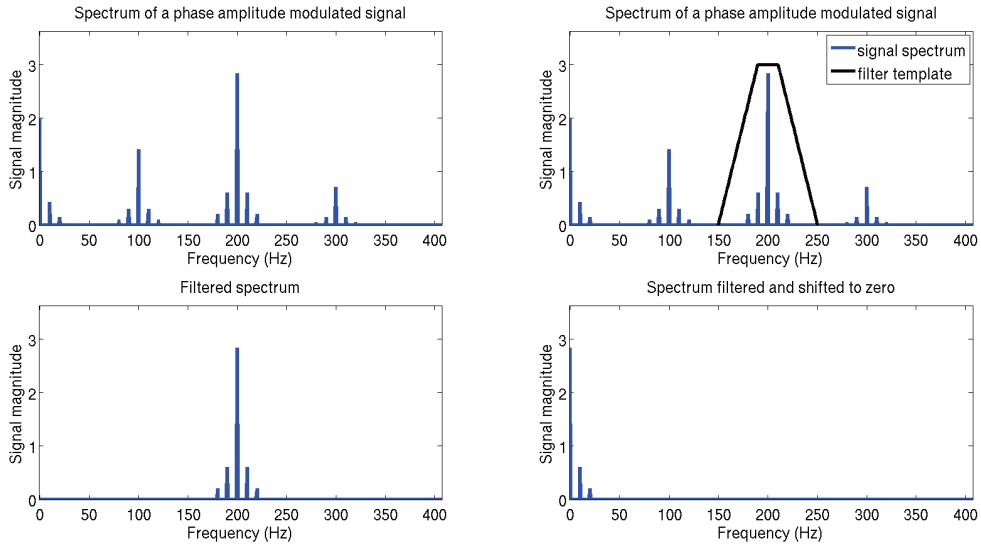


Figure 3.5: Classical demodulation steps: the frequency spectrum of the amplitude modulated signal (top left) is filtered around the most energetic carrier harmonic (top right) and the rest of the signal is set to zeros (bottom left). The remaining of the spectrum is then shifted to zero (bottom right).

Remark 11. *In practice, the harmonic with the highest energy is chosen in order to maximize the signal-to-noise ratio and thus to have a better estimation. But this procedure is problematic for some recent applications of signal demodulation.*

Then for the combined amplitude and phase modulation case, the usual demodulation method traditionally consists in using *Hilbert transform* to estimate the envelope of the modulated signal [61, 78]. For an arbitrary signal, the Hilbert transform is defined as:

$$\mathcal{H}(s(t)) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{s(\tau)}{t-\tau} d\tau. \tag{3.5}$$

From (3.5), the analytical signal of $s(t)$ can be defined as:

$$z(t) = s(t) + j\mathcal{H}(s(t)) = a(t)e^{j\Phi(t)}, \tag{3.6}$$

where, the modulus $a(t)$ and phase derivative $\dot{\Phi}(t)$ can serve as generally approximate estimates for the envelope and instantaneous frequency of $s(t)$. Indeed, Hilbert transform allows a complex demodulation analysis that suits a signal made of a single modulated sine wave. Theoretically, it can be used for any kind of modulated signal, but in practice it makes sense to use it only for narrowband signal.

Instead of using Hilbert transform, demodulation can also be performed using other techniques such as energy separation algorithm. It usually is computed by means of nonlinear diferential operators [70]. For example the Teager-Kaiser energy operator,

derived in [37], can be used to extract the amplitude envelope and the instantaneous frequency of modulated signals. The Teager-Kaiser Energy Operator of a continuous signal is defined as:

$$\Psi(s(t)) = \left(\frac{ds(t)}{dt}\right)^2 - s(t)\frac{ds(t)^2}{dt^2}, \tag{3.7}$$

while its discrete version is defined as:

$$\Psi(s(n)) = s^2(n) - s(n-1)s(n+1). \tag{3.8}$$

In order to improve the energy separation algorithm, an iterative generalized demodulation method was developed in [25].

When considering phase modulated signal, in order to estimate the modulations, the main purpose is to make the signal equivalent to an amplitude modulated signal. Thus, similar demodulation techniques can be used to retrieve the modulations.

In the classical approach of demodulation problems, the carrier energy is usually concentrated on one harmonic , i.g. in telecommunication. Considering this point, the demodulation process is the strictly opposite operation of modulation and thus recovers exactly the message signal. But it can be easily noticed that if several harmonics of the carrier are visible, all of them but one are filtered out during the process although they also contain information and could contribute to noise cancellation.

3.2.2 Multicomponent signals

Even if the use mono-component allows many applications, it is sometimes too simplistic for more elaborated signals, which is why some methods have been developed in order to use all the information contained in those signals. Multicomponent signal analysis have been obviously used in telecommunication [29, 21] but also in many other fields such as biomedical engineering [94], speech processing [31, 3] and also in mechanics [55].

Demodulation performed on multicomponent signals is usually made of two steps: the signal is first decomposed into mono-component signals and then those are demodulated individually.

The most classical multicomponent demodulation method selects one carrier harmonic and the modulations around, which involves that the rest of the spectrum is filtered out, as illustrated in Figure 3.6. In practice, the harmonic with the most energy is chosen in order to maximize the signal-to-noise ratio and thus to have a better estimation. The remaining spectrum is shifted to zero, which makes the resulting spectrum proportional to the one of signal $s_p(t)$. As a consequence, its time counterpart is proportional to $s_p(t)$.

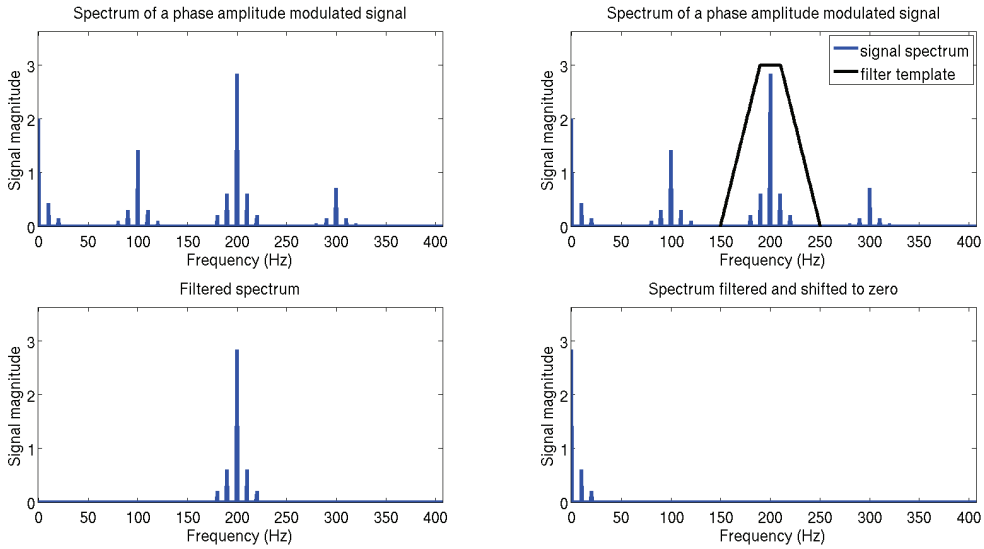


Figure 3.6: Classical demodulation steps: the frequency spectrum of the amplitude modulated signal (top left) is filtered around the most energetic carrier harmonic (top right) and the rest of the signal is set to zeros (bottom left). The whole is then shifted to zero (bottom right).

Some other methods of multicomponent AM-FM demodulation have been proposed, mainly in the communication community. In [78] the signal separation is made with a periodicity-based algebraic algorithm and the demodulation is then performed on the mono-components so obtained with energy-based algorithm (PASED). In order to compute the matrix algebraic separation, the signal has to be composed of an additive mixture of narrowband periodic signals. The problem was solved using some linear algebra techniques elaborated in [62, 78]. To perform the signal separation, some others have prefered to use several kind of filters such as bandpass filters or comb-filters [28]. Another method named *Hilbert-Huang transform* and introduced in [35] analizes the signal by taking it apart into intrinsic mode functions (IMF) using the empirical mode decomposition. A Hilbert transform is then performed on the obtained IMFs. A more recent approach was introduced in [30] where an iterative Hilbert transform is computed on the low-pass filtered amplitude envelope of the signal. The instantaneous frequencies and amplitudes are extracted from the resulting components. This algorithm has the main advantage to have a low computational complexity, which makes it suitable for online use. Regarding the former three methods, even if they are very effective in terms of demodulation accuracy, they all present an important lack of theoretical support and background understanding. For non-stationary analysis, wavelet based methods have the advantage to allow detection on signal whose time-frequency distributions are curved paths [97], but the interpretation is much more difficult since image processing techniques have to be used as a final step of the analysis.

3.3 Limits of the usual approach

In the previous section we have seen that in the literature there is a large choice of demodulation methods with application to many signal processing domains. But even though the use of demodulation algorithms is widely spread, there are conditions on

the signal that have to be respected to properly perform it. Those conditions can be set regarding the signal's frequency content as well as its time properties. Those signal restrictions may make the demodulation techniques listed above not properly applicable.

A first common requirement is to have a narrowband signal. Indeed many techniques lose accuracy when used on wideband signals, even worse, they can theoretically be not applicable. The main cause which turns signal to be wideband is the noise. This leads to some considerations on the background noise: thanks to its statistical properties and appropriate representation of experimental noise, background noise is usually considered additive white Gaussian. Usually while the noise level stays under a threshold, the signal can be considered narrowband as it does not spread too much in the spectrum. But when the noise level increases, the signal's frequency content expands and thus does not allow the narrowband hypothesis to be fulfilled.

In our application, i.e. gearbox monitoring, there are many systems that contribute to total vibration (bearings, rotors...) and thus they amplify the noise level. This one can easily reach (and even exceed) the level of the signal of interest. As the main consequence, gearbox vibration spectrum content is very rich as illustrated in Figure 3.8.

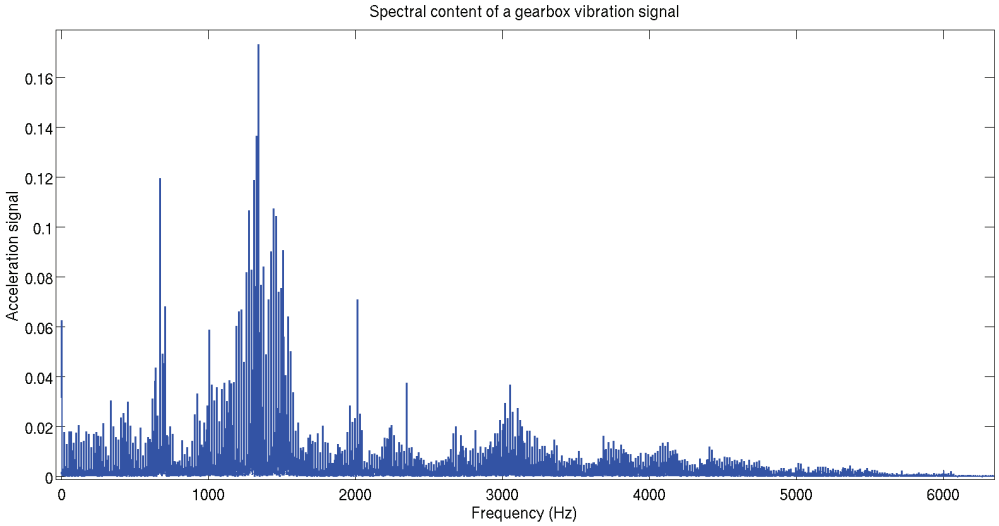


Figure 3.7: Example of vibration acquisition on a gearing system: it is difficult to identify the gearbox component as they are shrouded in noise.

Another limitation of demodulation techniques for the specific application of gearbox monitoring is the filtering step. By removing a large part of the available information, demodulation techniques turn to be sub-optimal. Indeed as explained in Chapter 2, vibration signals are periodic with several harmonics for both low and high frequency signals, as illustrated in Figure 3.8. Therefore by filtering the signal, the demodulation algorithms are losing a significant part of the available information, even in the case of multi-component techniques and last but not least limitation, they does not take advantage of the specific structure of the gearbox vibration.

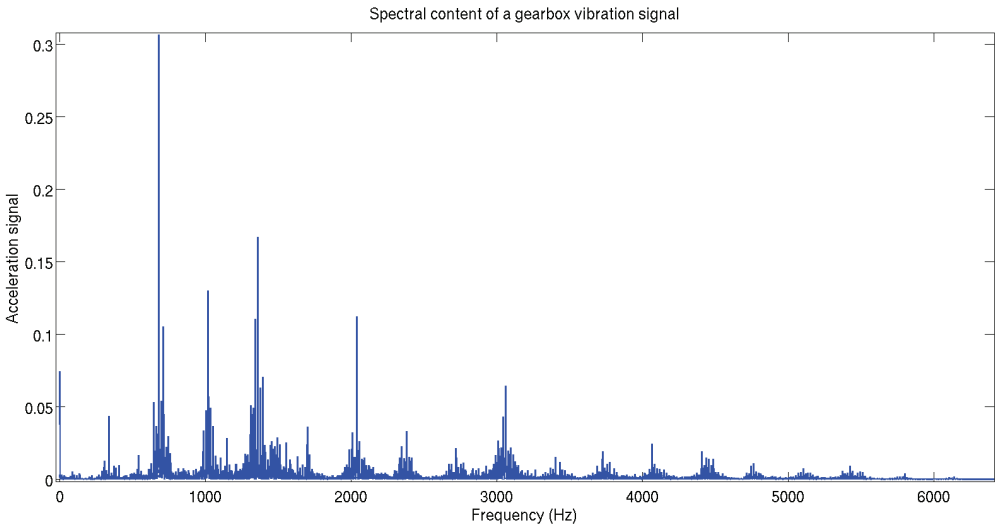


Figure 3.8: Gearbox vibration spectrum: a multicomponent signal with a particular structure.

Remark 12. *It is not enough to just rely only on the observation of a signal's spectrum to claim that this signal stems from a modulation process. Indeed even if a modulated signal will always have the characteristic spectral pattern made of a carrier lines with sidebands, the reciprocal is not necessarily true. A signal composed of several purely additive periodic signals is an example of signal that may have a spectrum similar to that of a modulation but which is not one.*

In order to free ourselves from the limitations listed above, we propose to rewrite the demodulation issue as an optimization problem. The main interest is that the modulated signal can be considered in its entirety, so that no informations are lost.

3.4 Demodulation as an optimization problem

In this section we present a first idea that can be summed up as follows: casting the demodulation problem into an optimization framework leads to more accurate results. This alternative approach recasts the demodulation problem as the problem of finding both carrier and modulation signals that best fit the data.

3.4.1 Signal framework

In the following sections, a discrete framework will be used, since it better suits the demodulation methods that will be presented thereafter than the continuous time one.

From a mathematical point of view, for a given sequence of sampling times $(t_n)_{n\geqslant 0}$, the signal $s(t_n)$ output obtained by a modulation process is the product of a high-frequency carrier $s_c(t_n)$ with a low-frequency modulation $s_m(t_n)$:

$$s(t_n) = s_c(t_n)s_m(t_n). \tag{3.9}$$

In the Fourier domain, the Discrete Fourier Transform (DFT) of the obtained signal $s(t_n)$ (denoted here with a tilde sign $\tilde{s}$) is a circular convolution of the carrier and modulation

DFT: $\tilde{s} = \tilde{s}_c * \tilde{s}_m$. The discrete spectrum $\tilde{s}_m$ of the modulation is repeated at all multiples of the carrier frequency and multiplied by the corresponding harmonic of $\tilde{s}_c$. In the case where the pattern centered on two consecutive harmonic do not overlap, the spectrum of the modulation is directly accessible from the spectrum of $\tilde{s}$. This is the condition making demodulation possible:

Hypothesis 1. *A periodic carrier of frequency f_c and a modulation of* maximum *frequency F_m can be recovered from their product under the following hypothesis:*

$$2F_m < f_c. \tag{3.10}$$

In discrete time an additional assumption is actually needed:

Hypothesis 2. *The signal duration has to be a multiple of the carrier period. Otherwise, the carrier won't have a pure line spectrum.*

Under Hypothesis 1, there is at most one carrier/modulation couple reversing the modulation operation, up to a multiplicative factor. Finding this couple will be referred to as Problem 1:

Problem 1 (Exact demodulation)**.** *Given a discrete signal $(s(t_n))_{n\in[\![1,N]\!]}$ (with $N \in \mathbb{N}$) of sampling period T_s and duration $T_{tot} = N \cdot T_s$, given a frequency $f_c = k_c \cdot f_{tot}$ (with $f_{tot} = 1/T_{tot}$ and $k_c \in \mathbb{N}$), find a carrier/modulation couple $(s_c(t_n), s_m(t_n))$ verifying for any $n \in [\![1, N]\!]$:*

$$s(t_n) = s_c(t_n)s_m(t_n), \tag{3.11}$$

with the temporal signal $s_c(t_n)$ of frequency f_c, and $(s_c(t_n), s_m(t_n))$ verifying Hypothesis 1.

3.4.2 Proposed optimization framework

Casting the problem of demodulation into an optimization means replacing Problem 1 with the following optimization problem:

Problem 2 (Optimal demodulation)**.** *Given a discrete signal $(s(t_n))_{n\in[\![1,N]\!]}$ (with $N \in \mathbb{N}$) of sampling period T_s and duration $T_{tot} = N \cdot T_s$, given a frequency $f_c = k_c \cdot f_{tot}$ (with $f_{tot} = 1/T_{tot}$ and $k_c \in \mathbb{N}$), find a carrier/modulation couple $(s_c(t_n), s_m(t_n))$ minimizing the following cost function:*

$$C(s_c, s_m) = \sum_{n=1}^{N} |s_c(t_n)s_m(t_n) - s(t_n)|^2$$

over all carrier/modulation couples verifying Hypothesis 1, with f_c the frequency of the temoral signal $s_c(t_n)$.

A first remark is that Problem 2 is non-quadratic, actually it is even not necessarily convex.

Remark 13. *The cost function $C(s_c, s_m)$ is the l_2-norm of the difference $s_c(t_n)s_m(t_n) - s(t_n)$ but other choices can be done. But we will see later striking properties making the l_2-norm particularly suited to numerical implementation.*

3.5 Conclusion

Demodulation problems are widespread and can be found under many types and multiple situations. Using demodulation tools is more recent in the field of mechanical signal analysis than in telecommunications. This is why existing techniques are mostly adapted to the specificity of telecommunication signals. We saw that vibration signals stemming from the rotation of mechanical systems are extremely rich in term of frequency components and often highly noisy. Furthermore, still by comparison with the telecommunication domain, the modulation pattern is not just present about one carrier, but repeated about several harmonics of that carrier, which makes most of the available demodulation techniques less adapted. This is the main reason why we proposed to recast the demodulation problem into a new optimization framework.

Part II

New tools for optimization-based demodulation

Chapter 4

Matrix representation of modulated spectra

The optimization approach for demodulation problems introduced in the previous Chapter is set for any kind of modulation problems, as long as the signal to be demodulated can be represented as a product of two signals. This general set-up, set as Problem 2, is defined in the temporal domain, but an equivalence can be done into the spectral domain. In this Chapter, we will introduce an original representation of the Fourier spectrum which proves to be more relevant in order to solve the optimization problem. We will also establish the striking properties of that representation.

4.1 Modulation in the discrete Fourier domain

First, let remind some technicalities regarding the Discrete Fourier transform (DFT) at play. The considered signal is measured at discrete times $t_n = nT_s$, with T_s the sampling period. Some technical assumptions are also made to ensure product signals of the form $s_c(t_n)s_m(t_n)$ have a modulation spectrum such as those of Figure 4.1, for example avoiding the sides effects. Furthermore, in order to respect Hypothesis 2, the acquisition duration T is an integer multiple of the carrier period T_c, which is itself an integer multiple of the sampling period T_s:

$$T = k_c T_c \qquad \text{and} \qquad T_c = N_c T_s, \qquad \text{with } k_c, N_c \in \mathbb{N}.$$

Dividing by T_s we obtain that $s_c(t_n)$ has period N_c as a discrete signal and that the relation $N = k_c N_c$ is verified, where N is the total length of the signal. In the Fourier domain, this implies that the modulation pattern has a discrete length of k_c and is repeated N_c times (with different scale factors) over the total length N of the spectrum. These notations are illustrated by Figure 4.1. Note these assumptions are not restrictive, but re-sampling the signal and discarding a possibly remaining fraction of the period T_c, is usually necessary to fulfill them.

Remark 14. *When talking about the DFT of a signal, we use the usual convention that the mean value of the signal ("frequency zero") is the first element of the DFT.*

But the following abuse of notations will be used to greatly alleviate computations:

Remark 15. *Indices of the DFT $\tilde{s}$ of a signal $s(t_n)$ of length N have to be understood modulo N. So, $\tilde{s}[0], \tilde{s}[-1], \tilde{s}[-2]$ mean respectively $\tilde{s}[N]$, $\tilde{s}[N-1]$, $\tilde{s}[N-2]$. This allows*

representing DFTs with harmonic zero in the middle as done on Figure 4.1, where the DFT coefficients $\tilde{s}[k]$ are represented not for k ranging not from 1 to N but from $N_l = -\lfloor N/2 \rfloor$ to $N_r = +\lfloor (N-1)/2 \rfloor$ (with the modulo N convention for negative indices). For a given integer N this sequence of shifted indices will be denoted by:

$$I_N = [\![-\lfloor N/2 \rfloor, +\lfloor (N-1)/2 \rfloor]\!].$$

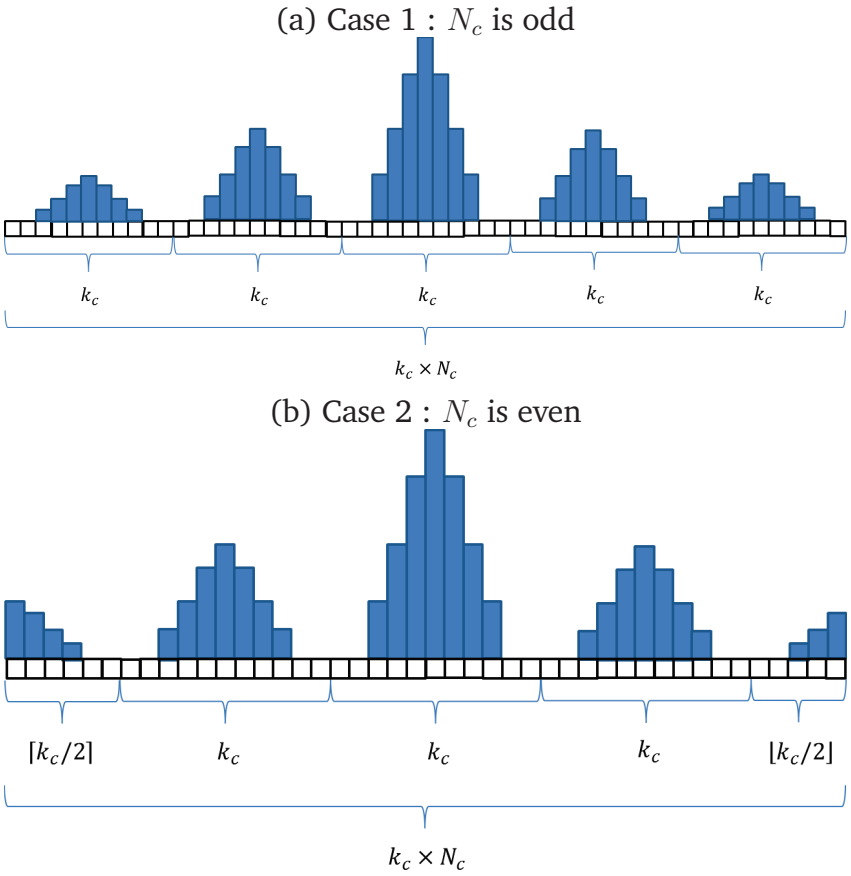


Figure 4.1: DFT of a discrete modulated signal, with indices ranging in I_N (see Remark 15 for the meaning of I_N). Note that when N_c is even, one modulation pattern is cut in the middle.

After these clarifications, let us transpose Problems 1 and 2 to the Fourier domain:

Proposition 1 (Exact demodulation in the Fourier domain)**.** *In the Fourier domain, Problem 1 becomes finding the DFT coefficients $\tilde{s}_c[k_c i + 1]$ for $i \in [\![0, N_c[\![$ and $\tilde{s}_m[j+1]$ for $j \in I_{k_c}$ verifying for all couple (i, j):*

$$\tilde{s}[ik_c + j + 1] = \tilde{s}_c[k_c i + 1]\tilde{s}_m[j + 1].$$

Proof. A DFT applied to the equality appearing in Problem 1 gives the equality $\tilde{s}[k] = \tilde{s}_c * \tilde{s}_m[k]$ for all $k \in [\![1, N]\!]$ (the product becomes a convolution product). But Hypothesis 1 ensures there is no overlap in the convolution product $\tilde{s}_c * \tilde{s}_m$ so we have the equality $(\tilde{s}_c * \tilde{s}_m)[ik_c + j + 1] = \tilde{s}_c[k_c i + 1]\tilde{s}_m[j + 1]$ for $i \in [\![0, N_c[\![$ and $j \in I_{k_c}$. As the expression $ik_c + j + 1$ covers, modulo N, the whole interval $[\![1, N]\!]$, Proposition 1 is verified. □

Proposition 2 (Optimal demodulation in the Fourier domain)**.** *In the discrete Fourier domain, Problem 2 becomes finding the DFT coefficients* $\tilde{s}_c[k_c i + 1]$ *for* $i \in [\![0, N_c[\![$ *and* $\tilde{s}_m[j+1]$ *for* $j \in I_{k_c}$ *that minimize the following cost function:*

$$\tilde{C}(\tilde{s}_c, \tilde{s}_m) = \sum_{i=0}^{N_c-1} \sum_{j=I_{k_c}} \left| \tilde{s}_c[k_c i + 1] \tilde{s}_m[j+1] - \tilde{s}[i k_c + j + 1] \right|^2 .$$

Proof. The cost function C of Problem 2 is the l_2-norm of the difference $s(t_n) - s_m(t_n)s_c(t_n)$. Using the Plancherel theorem, it is equal to $\frac{1}{N} ||\tilde{s}_c * \tilde{s}_m - \tilde{s}||_{l_2}^2 = \frac{1}{N} \sum_{k=1}^{N} \left| (\tilde{s}_c * \tilde{s}_m)[k] - \tilde{s}[k] \right|^2$. As in the proof of Proposition 1, Hypothesis 1 ensures there is no overlap in the convolution product $\tilde{s}_c * \tilde{s}_m$ so we have the equality $(\tilde{s}_c * \tilde{s}_m)[i k_c + j + 1] = \tilde{s}_c[k_c i + 1]\tilde{s}_m[j+1]$ for $i \in [\![0, N_c[\![$ and $j \in I_{k_c}$. As the expression $i k_c + j + 1$ covers, modulo N, the whole interval $[\![1, N]\!]$, we have $C(s_c, s_m) = \frac{1}{N} \sum_{i=0}^{N_c-1} \sum_{j \in I_{k_c}} \left| \tilde{s}_c[k_c i + 1]\tilde{s}_m[j+1] - \tilde{s}[i k_c + j + 1] \right|^2$. As removing the factor $\frac{1}{N}$ does not change the optimum, we end up with the cost function of Proposition 2. □

4.2 Matrix representation of spectrum construction

We propose and discuss a new tool for demodulation problems that we call *Matrix representation of a spectrum*.

Let $s(t_n)_n \in [1, N]$ be a discretized time signal and $\tilde{s}$ its DFT. Each Fourier coefficient of the spectrum can be decomposed into the product of one element of the carrier signal and one element of the modulation signal as represented in Figure 4.2.

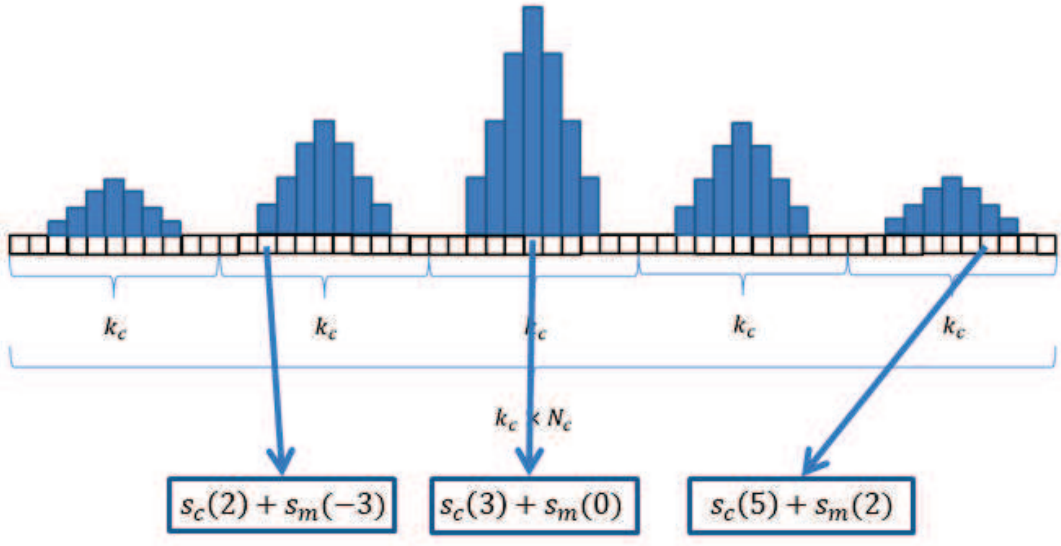


Figure 4.2: Discrete spectrum represented as the product of two vectors.

In a similar way, Figure 4.3 represents the product of a column vector with a line vector, building a matrix. It is easy and fast to notice that both notations are identical which allows un to define an equivalence of both representation, mapped in Figure 4.4. Based on the above mentionned observation, the matrix representation we propose consists in cutting $\tilde{s}$ into buckets centered on each harmonic of the carrier, then stacking its values vertically as illustrated by Figure 4.5.

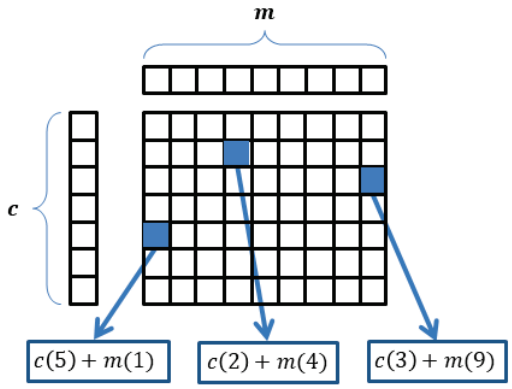


Figure 4.3: Matrix built as the product of a column vector with a line vector.

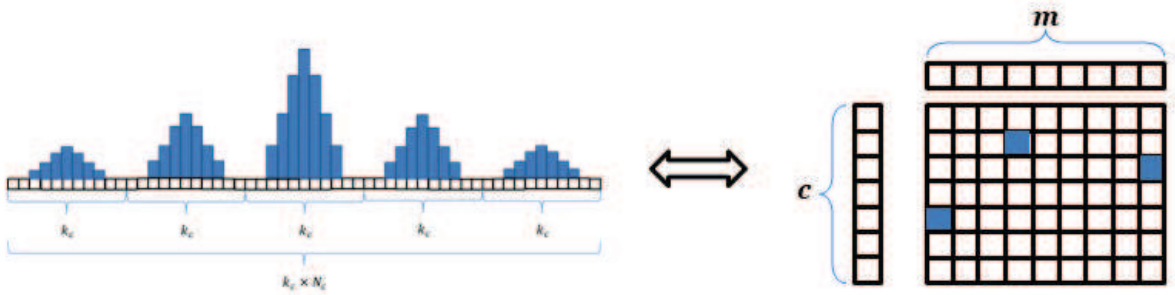


Figure 4.4: Equivalence between the discrete spectrum of a temporal product signal and the matrix product of two vectors.

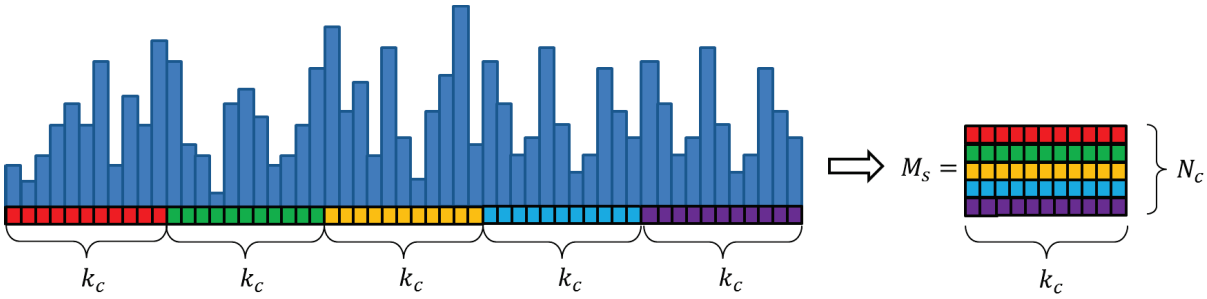


Figure 4.5: Matrix spectrum construction from the line spectrum vector.

Let us write down a mathematical definition of this matrix $\mathbf{M_s}$:

Definition 1 (Matrix representation of a spectrum)**.** *Let $(s(t_n))_{n\in[\![1,N]\!]}$ be a discrete time signal of length N, $\tilde{s}[\cdot]$ its DFT and k_c an integer dividing N. We call "matrix representation of $\tilde{s}$ for k_c periods" the matrix $\mathbf{M_s}$ defined as:*

$$[M_s]_{i+1,:} = \tilde{s}\left[I_{k_c} + k_c i\right], \tag{4.1}$$

where $I_{k_c} + k_c i$ denotes the sequence I_{k_c} (see Remark 15) with $k_c i$ added to each element. This definition is illustrated by Figure 4.5.

4.3 On the properties of the matrix representation of a spectrum

The matrix representation of a spectrum was proposed because it shows to have interesting properties for the resolution of the demodulation issue previously exposed. More specifically, we will see that it builds a bridge between the theories of signal modulation and low-rank operators.

4.3.1 Link with low-rank operators

In the case where $s(t_n)$ stems from a modulation of a carrier $s_c(t_n)$, as illustrated by Figure 4.1, we see that each line of the matrix $\mathbf{M_s}$ is a representation of the same pattern up to a multiplicative scale factor. This precisely means that $\mathbf{M_s}$ has rank one, which is where the interest of this reshaping lies:

Proposition 3. *Problem 1 can be solved for signal $s(t_n)$ and carrier period $N_c = N/k_c$ if and only if the matrix representation $\mathbf{M_s}$ for k_c periods has rank one.*

For any couple of spectra $(\tilde{s}_c, \tilde{s}_m)$ of size N we define two "reduced" spectra ($\tilde{s}_{rc}$ and $\tilde{s}_{rm}$) of sizes N_c and k_c respectively as $\tilde{s}_{rm}[j+1] = \tilde{s}_m[k_c j + 1]$ for $j \in [\![0, N_c - 1]\!]$ and $\tilde{s}_{rc}[i] = \tilde{s}_c[i]$ for $i \in [\![1, k_c]\!]$ (Note that $\tilde{s}_{rc}$ is simply obtained keeping only the N_c first elements of $\tilde{s}_c$).

Proof. For any couple of "reduced" spectra $(\tilde{s}_{rc}, \tilde{s}_{rm})$, we have the following equivalences:

Problem 1 is solved by (s_c, s_m)		
$\Leftrightarrow \forall n,$	$s(t_n) = s_c(t_n)s_m(t_n)$	
$\Leftrightarrow \forall i \in [\![0, N_c[\![, \forall j \in I_{k_c}$	$\tilde{s}[k_c i + j + 1] = \tilde{s}_c[k_c i + 1]\tilde{s}_m[j+1]$	(see Prop. 1)
$\Leftrightarrow \forall i \in [\![0, N_c[\![, \forall j \in I_{k_c}$	$[M_s]_{i+1, j-\lfloor k_c/2 \rfloor + 1} = \tilde{s}_c[k_c i + 1]\tilde{s}_m[j+1]$	(see Def. 2)
$\Leftrightarrow$	$\mathbf{M_s} = \tilde{s}_{rc}\tilde{s}_{rm}^T$	.

But when the couple $(\tilde{s}_{rc}, \tilde{s}_{rm})$ describes $\mathbb{R}^{k_c} \times \mathbb{R}^{N_c}$, the matrix $\tilde{\boldsymbol{s}}_{\boldsymbol{c}}\tilde{\boldsymbol{s}}_{\boldsymbol{m}}{}^T$ exactly describes the set of all matrices of size $k_c \times N_c$ and rank one. So we obtain the following equivalence:

$$\text{Problem 1 can be solved} \Leftrightarrow rank(\mathbf{M_s}) = 1.$$

□

4.3.2 The case of periodic modulations

As it usually occurs for mechanical systems, the signal $s_m(t_n)$ stemming from of its rotation is periodic. This implies that the modulation pattern of Figure 4.1 itself is a line spectrum, meaning additional holes are present (see Figure 4.6). Note this happens only if the acquisition duration T is chosen to be a multiple of T_c the carrier period and T_m the period of the modulation: $T = k_m T_m = k_c T_c$ with $k_m, k_c \in \mathbb{N}$ (otherwise the expected clean peaks spread around the expected position of each harmonic). In this case the values of $\tilde{s}[\cdot]$ at the position of the holes play no role in the optimization Problem 2 and may be ignored. Concretely, this means that a reduced form of the matrix representation can be used, which is illustrated by Figure 4.6.

Remark 16. *The reduced matrix representation of a spectrum will always be used in the further algorithms, as it allows to reduce the computation load.*

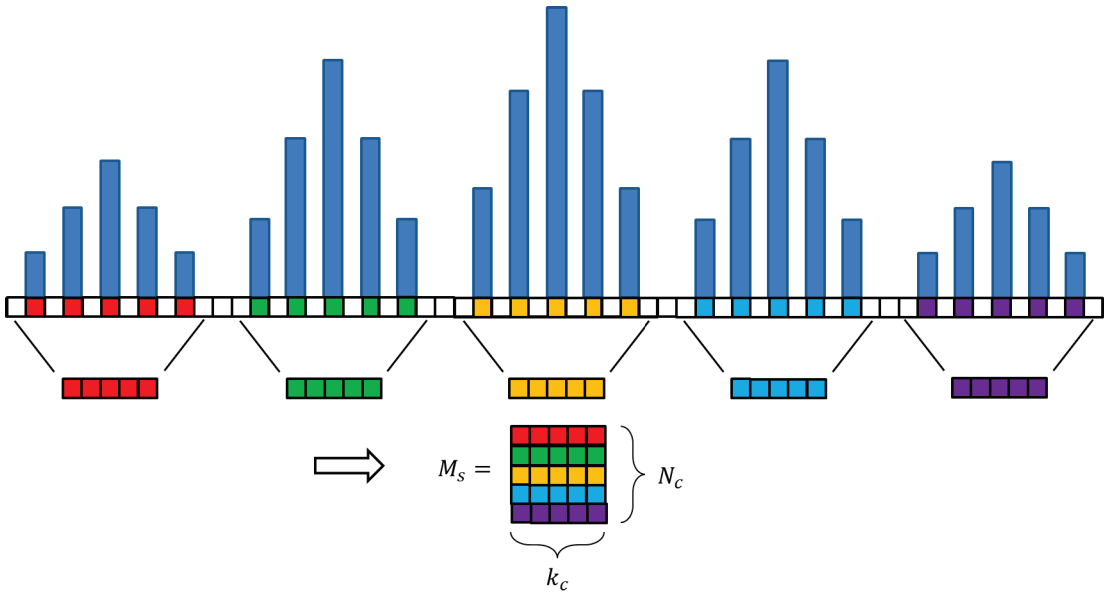


Figure 4.6: Periodic modulation and reduced matrix representation

Its formal definition is as follows:

Definition 2 (Reduced matrix representation of a spectrum)**.** *Let $k_c, k_m \in \mathbb{N}$ and $(s(t_n))_{n\in[\![1,N]\!]}$ be a discrete time signal with N a multiple of k_c and of k_m, and $\tilde{s}[\cdot]$ its DFT. We call "matrix representation of $\tilde{s}$ for k_c carrier periods and k_m modulation periods" the matrix $\mathbf{M_s}$ defined as:*

$$[M_s]_{i+1,:} = \tilde{s}[I_{k_c;k_m} + k_c i + 1], \tag{4.2}$$

where $I_{k_c;k_m} = I_{k_c}[1, 1+k_m, 1+2k_m, ...]$ is the set I_{k_c} defined in Remark 15 where one over k_c indices is kept. This definition is illustrated by Figure 4.6.

This situation is specific to the application to mechanical systems, as in telecommunications the transmitted signal has no reason to be periodic.

4.3.3 Complements on centro-symmetric matrices

The results of this section regard the specific structure that we call "centro-symmetric" of the matrix representation of a spectrum. They will only be used in Chapter 7 and do not have to be read immediately but since they are not related to a specific application, we chose to present them here. The general idea we develop is that the spectrum of a real signal being conjugate-symmetrical its matrix representation has a similar property. First, let us give a few definitions.

Definition 3. *For a given integer $N \in \mathbb{N}$, we call "symmetry operator", denoted by $\sigma_N : [1, N] \to [1, N]$, the operator switching the indices of "symmetrical" harmonics of a discrete spectrum. With the convention that the zero harmonic (the average value of the signal) is in first position (the usual FFT convention) we have $\sigma_N(1) = 1$ and $\sigma_N(1+i) = N+1-i$ for $i \in [0, N-1]$. If the zero harmonic is at position $I = \lceil (N+1)/2 \rceil$ as it can also be the case we have $\sigma_N(I) = I$ and $\sigma_N(I+i) = I-i$ if $I+i \in [1, N]$ and $I-i \in [1, N]$. Note that in any case, σ_N is a bijection.*

This operator allows writing down the symmetry property verified by the matrix representation of a specrum regardless of the DFT index convention used.

Definition 4 (Centro-symmetric matrix)**.** *A complex matrix $M \in \mathcal{M}_{N,P}(\mathbb{C})$ is told centro-symmetric if it verifies the following property for any $i \in [1, N]$ and $j \in [1, P]$:*

$$M_{\sigma_N(i),\sigma_P(j)} = \bar{M}_{i,j}$$

where $\bar{x}$ denotes the conjugate of complex number $x \in \mathbb{C}$. The definition naturally expands to a vector $x \in \mathbb{R}^N$ as:

$$x_{\sigma_N(i)} = \bar{x}_i$$

The space of centro-symmetric matrices (resp. vectors) will be denoted by $SC(N, P)$ (resp. $SC(N)$).

It can be easily checked that matrix representations of spectra are centro-symmetric matrices, which is why we dedicate a section to giving their main properties. First, the "size" of this space is given by Proposition 4 below:

Proposition 4. *The set $SC(N)$ endowed with vector addition and multiplication by a real factor is a real N-dimensional vector space.*

Proof. First, note this result is not as obvious as it looks like as we consider dimension of a set of complex vectors as a real vector space. Let V be $\mathbb{C}^N$ seen as a $\mathbb{R}$-vector space. $\mathbb{R}^N$ is then a vector subspace of V. The DFT operator is linear and invertible on $\mathbb{C}^N$, and in particular on V. Thus the image of $\mathbb{R}^N$ through this operator is also a vector subspace of V, and it has dimension N over $\mathbb{R}$. □

The following result allows us to use centro-symmetric matrices as linear operators on the $\mathbb{R}$-vector space $SC(N)$:

Proposition 5. *The product of two centro-symmetric matrices is also a centro-symmetric matrix. In particular, the operation $x \mapsto Mx$ defines a linear operator on $SC(N)$ if M is centro-symmetric.*

Proof. Let A, B be two centro-symmetric matrices of respective dimensions $N \times P$ and $P \times Q$. We have:

$$\begin{aligned}
(\bar{AB})_{ij} &= \sum_{k=1}^{P} \bar{A}_{i,k}\bar{B}_{k,j} \\
&= \sum_{k=1}^{P} A_{\sigma_N(i),\sigma_P(k)} B_{\sigma_P(k),\sigma_Q(j)} \\
&= \sum_{l=1}^{P} A_{\sigma_N(i),l} B_{l,\sigma_Q(j)} \qquad \text{(where we introduced } l = \sigma_P(k)\text{)} \\
&= (AB)_{\sigma_N(i),\sigma_Q(j)}
\end{aligned}$$

□

These basic properties will be used later to distinguish between the $\mathbb{C}$-rank and the $\mathbb{R}$-rank of a centro-symmetric matrix.

4.4 Conclusion

The matrix representation of a spectrum is a new tool introduced in this work to facilitate signal demodulation when it is set as an optimization problem. We have seen that it is possible to link the optimization issue with the low-rank operator theory. Thanks to that correspondance, the resolution turns to become equivalent to a search for a rank-one matrix, for which an abundant litterature exists.

Chapter 5

Amplitude demodulation

The matrix representation of a spectrum, introduced in Chapter 4, was shown to have extremely convenient properties regarding demodulation. They will be leveraged in the present chapter to solve the optimal amplitude demodulation problem proposed in Chapter 3.

But as previously mentioned, the properties of the matrix spectrum representation require a non-overlapping condition (Hypothesis 1): the spectral support of the amplitude modulation should be finite and smaller than half the carrier frequency. Since the case where this hypothesis is not verified can occur in practice, it is also studied but we will see that the problem becomes much more complicated when the matrix spectrum representation we introduced cannot be used.

In Section 5.1 a statistical interpretation of the optimization framework proposed in Chapter 3 is given. In Section 5.2 a new demodulation algorithm is derived for the case where the matrix spectrum representation can be used, based on the relation established in the previous chapter between modulated signals and rank-one operators. In Section 5.3 the case where the matrix spectrum representation cannot be used is addressed using a more classical optimization algorithm. Simulations results are presented for all proposed algorithms.

5.1 Statistical model formulation

Let us cast optimal demodulation into the framework of parameter estimation. Problem 2 can be seen as the maximum-likelihood estimator of the couple (s_c, s_m) if we assume the following statistical model for the measured signal:

$$s(t_n) = s_c(t_n) \times s_m(t_n) + w(t_n), \tag{5.1}$$

where the sequence $w(t_n)$ is a white Gaussian noise verifying $w(t_n) \sim \mathcal{N}(0, \sigma^2)$. As both signals $s_c(t_n)$ and $s_m(t_n)$ are considered periodic, let us define their frequencies f_c and f_m and set a maximum number of harmonics H and K. Each signal is defined in the time domain by the following formulae:

$$\begin{cases} s_c(t_n) = \frac{1}{N} \sum_{h=-H}^{H} C_h e^{2i\pi h f_c t_n}, \\ s_m(t_n) = \frac{1}{N} \sum_{k=-K}^{K} M_k e^{2i\pi k f_m t_n}. \end{cases} \tag{5.2}$$

All C_h and M_k are unknown complex values which verify the following constraints. The carrier is set to have zero-mean $C_0 = 0$ and an arbitrary constant mean is set for the

modulation such as $M_0 = 1$. Also, in order to ensure $s_c(t_n)$ and $s_m(t_n)$ are real signals we set the additional constraint $C_{-h} = C_h^*$ and $M_{-k} = M_k^*$.

In order to adopt a real-valued vector parameterization we introduce the sine/cosine Fourier Coefficients:

$$C_h^R = 2\Re\left(\tilde{s}_c[hk_c]\right), \qquad C^I = 2\Im\left(\tilde{s}_c[hk_c]\right), \qquad for\, h \in [1, H]$$
$$M_k^R = 2\Re\left(\tilde{s}_m[k]\right), \qquad M_k^I = 2\Im\left(\tilde{s}_m[k]\right), \qquad for\, k \in [1, K]$$

We obtain a real-valued vector of parameters $\boldsymbol{\theta}$ defined as:

$$\boldsymbol{\theta} = \left[C_1^R, \cdots, C_H^R, C_1^I, \cdots, C_H^I, M_1^R, \cdots, M_K^R, M_1^I, \cdots, M_K^I\right]^T. \tag{5.3}$$

This statistical interpretation and the real-valued parameterization will be used later to define the Cramér - Rao Lower Bound of the demodulation problem while the parameterization of Chapter 4, based on DFT coefficients will be preferred when using the matrix representation as in Secttion 5.2 below.

5.2 Amplitude demodulation without overlapping

We are first interested in the case where there is no overlapping between the modulation sidebands. This precisely means that the maximum number of harmonics for the carrier and modulation are $H = \lfloor N_c/2 \rfloor$ and $K = \lfloor k_c/2 \rfloor$ respectively. The set of all couples (s_c, s_m) verifying Hypothesis 1 can be parameterized by the complex-valued DFT coefficients of $\tilde{s}_m$ and $\tilde{s}_c$, as defined previously in Section 4.1

5.2.1 Amplitude demodulation with the matrix representation of a spectrum

In the previous Chapter, we have seen that Proposition 3 allows us to replace a search for a modulated signal fitting at best the measurements as in Problem 2, with a search for a rank-one matrix fitting at best the matrix $\mathbf{M_s}$ of Definition 2. This idea is formalized by Proposition 6 below:

Proposition 6. *Let $\mathcal{M}_1$ be the set of rank-one $k_c \times N_c$ matrices and $\mathbf{M_s}$ the matrix representation of the spectrum of the measured signal $s(t_n)$ as defined in Chapter 4. Consider the following cost function defined for matrices $\mathbf{M} \in \mathcal{M}_1$:*

$$C_{mat}(\mathbf{M}) = ||\mathbf{M} - \mathbf{M_s}||_{Fro}^2.$$

Then we have the following equivalence:

$$\textit{The matrix } \tilde{s}_{rc}\tilde{s}_{rm}^T \textit{ minimizes } C_{mat} \textit{ over } \mathcal{M}_1 \Leftrightarrow (\tilde{s}_c, \tilde{s}_m) \textit{ minimizes } \tilde{C},$$

with indices $\cdot_{rc}$ and $\cdot_{rm}$ defined as in Proposition 3, $\tilde{C}$ defined as in Proposition 2 and $||\cdot||_{Fro}$ the matrix Frœbenius norm.

This result means that finding the optimal demodulation of a signal in the sense of Problem 2 is equivalent to approximating the matrix $\mathbf{M_s}$ by a rank-one matrix. The literature for this second problem is extensive, which gives a general approach to optimal demodulation:

Algorithm 1. *The optimization Problem 2 can be optimally solved in the Fourier domain through the following steps:*

1. *Building the matrix* $\mathbf{M_s}$*,*
2. *Applying a rank-one approximation method to* $\mathbf{M_s}$*,*
3. *Decomposing the obtained matrix* $\hat{\mathbf{M}}$ *as* $\hat{\mathbf{M}} = \tilde{s}_{rc}\tilde{s}_{rm}^T$*,*
4. *Returning the corresponding spectra* $\tilde{s}_c$ *and* $\tilde{s}_m$ *(see Proposition 3) or their time counterparts.*

Note that Step 3 is easy as $\hat{\mathbf{M}}$ *has rank one. One could for instance take* $\tilde{s}_{rc}$ *as the first column of* $\hat{\mathbf{M}}$ *and* $\tilde{s}_{rm}$ *as its first line, divided by* $\hat{\mathbf{M}}_{1,1}$*. Several algorithms can be used at Step 2. For the simulations of the next section, the SVD-factorization* $\mathbf{M_s} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ *was performed, with* $\boldsymbol{U} \in \mathbb{C}^{(N_c+1)\times(N_c+1)}$*,* $\boldsymbol{V} \in \mathbb{C}^{(k_c+1)\times(k_c+1)}$ *and* $\boldsymbol{D} \in \mathbb{R}^{(N_c+1)\times(k_c+1)}$ *a diagonal matrix. Then, the estimated reduced spectra* $\tilde{s}_{rc}, \tilde{s}_{rm}$ *are defined as:*

$$\begin{aligned} \tilde{s}_{rc} &= \boldsymbol{U}_{:,1}, \\ \tilde{s}_{rm} &= D_{1,1}\boldsymbol{V}_{:,1}. \end{aligned} \tag{5.4}$$

Remark 17. *Performing the optimization using Algorithm 1 requires one step since the problem has a closed-form solution.*

The observation we have in hand is the discrete signal $s(t_n)$, i.e. a vector of dimension N, related to the vector of parameters $\boldsymbol{\theta}$ to be estimated through the relations (5.1) and (5.2). Algorithm 1 provides an estimator of $\boldsymbol{\theta}$, whose precision we want to numerically compare with the theoretical limit given by the CRLB to be computed in the Section 5.2.2 below.

5.2.2 Performance Estimation

The precision of an unbiased estimator $\hat{\boldsymbol{\theta}}$, usually measured by its variance $VAR_{\hat{\theta}}(\boldsymbol{\theta}) = \mathbb{E}\left((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right)$, has a theoretical limit called *Cramér-Rao bound* [18]:

$$\boldsymbol{VAR}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) \succcurlyeq \boldsymbol{F}(\boldsymbol{\theta})^{-1},$$

where $\boldsymbol{F}(\boldsymbol{\theta})$ is the *Fisher Information Matrix* (FIM) given by: $\boldsymbol{F}(\boldsymbol{\theta}) \triangleq -\mathbb{E}\left(\frac{\partial^2 \ln p(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)$, with $p(\boldsymbol{y};\boldsymbol{\theta})$ probability density function of the measurement $\boldsymbol{y}$ for the parameter vector $\boldsymbol{\theta}$.

In order to simplify the computation of the Cramér-Rao Lower Bound, we rewrite Problem 5.1 with sine/cosine Fourier Coefficients. Thus Equation 5.2 becomes:

$$\begin{cases} s_c(t_n) = \frac{1}{N}\sum_{h=0}^{H} C_h^R \cos(2i\pi h f_c t_n) + \sum_{h=1}^{H} C_h^I sin(2i\pi h f_c t_n), \\ s_m(t_n) = \frac{1}{N}\sum_{k=0}^{K} M_k^R \cos(2i\pi k f_m t_n) + \sum_{k=0}^{K} M_k^I sin(2i\pi k f_m t_n). \end{cases} \tag{5.5}$$

Proposition 7. *The FIM for the parameter estimation problem* (5.1) *reads:*

$$\boldsymbol{F} = \frac{1}{\sigma^2}\sum_{t_n=1}^{N} \begin{bmatrix} s_m(t_n)\boldsymbol{\gamma_n} \\ s_c(t_n)\boldsymbol{\mu_n} \end{bmatrix} \begin{bmatrix} s_m(t_n)\boldsymbol{\gamma_n} \\ s_c(t_n)\boldsymbol{\mu_n} \end{bmatrix}^T, \tag{5.6}$$

with $\boldsymbol{\gamma_n}$ *and* $\boldsymbol{\mu_n}$ *the vectors defined as* $\boldsymbol{\gamma_n} = \left[\cos(2\pi\frac{k_c n}{N})\dots\cos(2\pi\frac{Hk_c n}{N})\right]^T$ *and* $\boldsymbol{\mu_n} = \left[\cos(2\pi\frac{n}{N})\dots\cos(2\pi\frac{Kn}{N})\right]^T$.

Proof. In the model Eqation (5.1) the observation is an i.i.d. sequence of Gaussian variables with means parameterized by $\boldsymbol{\theta}$. Consequently, the computation of the FIM can be carried out using the so-called Slepian-Bang formula (see [Kay] p.47) which takes, in the case where the noise variance σ^2 is known, the simplified form $\boldsymbol{F} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\left[\boldsymbol{\nabla_\theta s}\right]\left[\boldsymbol{\nabla_\theta s}\right]^T$, with $\boldsymbol{\nabla_\theta s}$ the gradient of $s(t_n)$ w.r.t. $\boldsymbol{\theta}$. Let us compute separately the gradient of $s(t_n)$ with respect to the carrier part $\boldsymbol{\theta_c} = \left[C_0^R,\dots,C_H^R,C_1^I,\dots C_H^I\right]^T$ of $\boldsymbol{\theta}$ and its modulation part $\boldsymbol{\theta_m} = \left[M_0^R,\dots,M_K^R,M_1^I,\dots M_K^I\right]^T$. Formula (5.5) can be rewritten as $s_c(t_n) = \boldsymbol{\gamma_n}^T\boldsymbol{\theta_c}$ and $s_m(t_n) = \boldsymbol{\mu_n}^T\boldsymbol{\theta_m} + 1$, which allows rewriting (5.1) as $s(t_n) = \boldsymbol{\theta_c}^T\boldsymbol{\gamma_n}(\boldsymbol{\mu_n}^T\boldsymbol{\theta_m} + 1) + w_n$. The gradients of $s(t_n)$ with respect to $\boldsymbol{\theta_c}$ and $\boldsymbol{\theta_m}$ are thus $\boldsymbol{\nabla_{\theta_c} s} = \boldsymbol{\gamma_n}(\boldsymbol{\mu_n}^T\boldsymbol{\theta_m} + 1)$ and $\boldsymbol{\nabla_{\theta_m} s} = \boldsymbol{\mu_n}\boldsymbol{\gamma_n}^T\boldsymbol{\theta_c}$. Using $s_c(t_n) = \boldsymbol{\gamma_n}^T\boldsymbol{\theta_c}$ and $s_m(t_n) = \boldsymbol{\mu_n}^T\boldsymbol{\theta_m}$, they simplify as $\boldsymbol{\nabla_{\theta_c} s} = s_m(t_n)\boldsymbol{\gamma_n}^T$ and $\boldsymbol{\nabla_{\theta_m} s} = s_c(t_n)\boldsymbol{\mu_n}^T$. Finally, stacking $\boldsymbol{\nabla_{\theta_c} s}$ and $\boldsymbol{\nabla_{\theta_m} s}$ to build $\boldsymbol{\nabla_\theta s}$ and injecting the result into the Slepian-Bang formula $\boldsymbol{F} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\left[\boldsymbol{\nabla_\theta s}\right]\left[\boldsymbol{\nabla_\theta s}\right]^T$ we obtain Equation (5.6). □

Remark 18. *The variance* σ^2 *of the noise could also be considered as an unknown parameter to be estimated and added to the vector* $\boldsymbol{\theta}$ *(say at the last position). The FIM can still be computed, using full Slepian Bang formula (including its variance term):*

$$F_{i,j} = \frac{1}{\sigma^2}\sum_{t=1}^{L}\left(\frac{\partial s(t)}{\partial\theta[i]}\frac{\partial s(t)}{\partial\theta[j]}\right) + \frac{L}{2\sigma^4}\frac{\partial\sigma^2}{\partial\theta[i]}\frac{\partial\sigma^2}{\partial\theta[j]}. \tag{5.7}$$

We see that the right-hand term will be zero, except if i *and* j *are both the last index of* $\boldsymbol{\theta}$*, corresponding to the variance, while the left-hand term will be zero if* i *or* j *is this last index. This gives:*

$$\boldsymbol{F_{tot}}(\boldsymbol{\theta}) = \begin{bmatrix}\boldsymbol{F} & \mathbf{0}^{N\times 1}\\ \mathbf{0}^{1\times N} & \frac{1}{2\sigma^4}\end{bmatrix} \tag{5.8}$$

with $\boldsymbol{F}$ *the FIM of Proposition 7.*

5.2.3 Simulations

The performance of the proposed amplitude estimator is assessed in several situations. We are interested in fine in gearbox monitoring, so as to be close to evaluate it in relevant conditions, i.e. in the operating conditions we will have with real dataset in Chapter 8, the two parameters of influence studied are the length of the signal (i.e. the number of samples) and the signal to noise ratio (SNR). For each case, the estimator range of validity is first assessed with a bias study and then, the Mean Square Error (MSE) is compared to the Cramér-Rao Lower Bound computed in the previous subsection in order to conclude on its asymptotic statistical efficiency.

A general set up was defined for the simulations. A synthetic vibration signal is generated according to Equation (5.1) and Equation (5.5) and $N_{MC} = 10000$ Monte-Carlo draws are run out to obtain the variance of the estimated parameter $\hat{\boldsymbol{\theta}}$. The signal settings used in the simulations are a sampling frequency $f_s = 10kHz$, signal frequencies $f_c = 500Hz$, $f_m = 20Hz$, number of harmonics $H = 9$ and $K = 5$. Consequently, one has 27 parameters to estimate. The amplitudes coefficients $\mathbf{c^R}$, $\mathbf{c^I}$, $\mathbf{m^R}$ and $\mathbf{m^I}$ are randomly set with real positive values. It has to be noted that the sampling frequency was

chosen high enough to ensure the Nyquist-Shannon sampling theorem and the numbers of harmonics H and K are chosen to guarantee the Hypothesis 3.10, i.e. to avoid aliasing in the signal spectrum.

The two evaluation criteria are defined by:

$$bias = \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \right), \tag{5.9}$$

where the considered bias, i.e. $bias_{\mathbf{c}^\mathbf{R}}$ (resp. $bias_{\mathbf{c}^\mathbf{I}}, bias_{\mathbf{m}^\mathbf{R}}, bias_{\mathbf{m}^\mathbf{I}}$) is the mean of the $\mathbf{c}^\mathbf{R}$ vector bias (resp. $\mathbf{c}^\mathbf{I}, \mathbf{m}^\mathbf{R}, \mathbf{m}^\mathbf{I}$), and:

$$MSE = \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^2. \tag{5.10}$$

The MSE can be calculated for each vector parameter separately, i.e. $MSE_{\mathbf{c}^\mathbf{R}}$ (resp. $MSE_{\mathbf{c}^\mathbf{I}}$, $MSE_{\mathbf{m}^\mathbf{R}}$ and $MSE_{\mathbf{m}^\mathbf{I}}$) from Equation (5.10) by averaging the correponding part of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ as defined in Equation(5.3).

Influence of the signal-to-noise ratio

Here, the impact of the noise level on the estimator performances is evaluated. The length of the signal is set to $N = 10000$ and the signal-to-noise ratios are spanning the range from $SNR = -10dB$ to $SNR = 20dB$.

The first step of the estimator assessment is to verify that it is unbiased with respect to the SNR. Figure 5.1 displays the bias for all the parameter vectors for several SNRs.

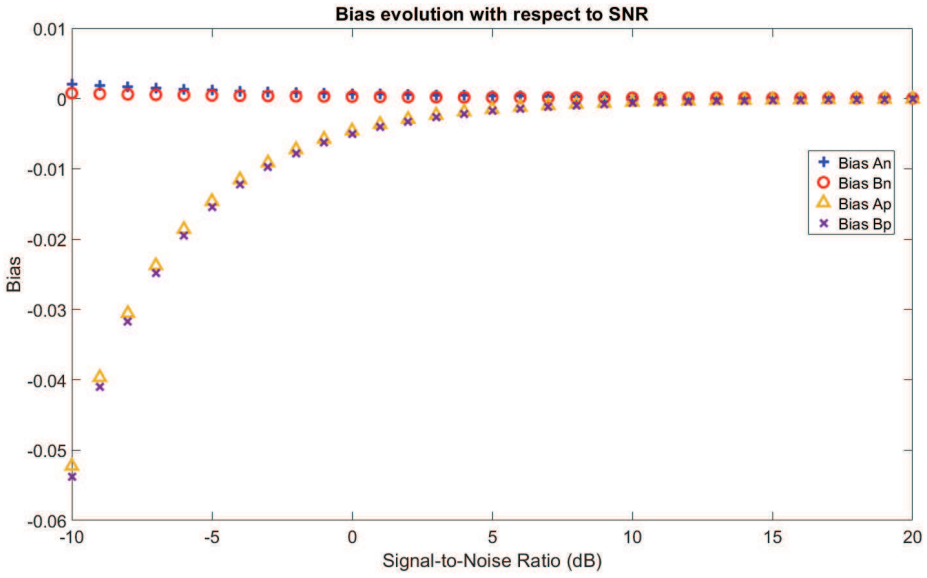


Figure 5.1: Overlay of the evolution of the bias for all four parameter vectors with respect to the Signal-to-Noise Ratio.

It can be seen that for $\mathbf{c}^\mathbf{R}$ and $\mathbf{c}^\mathbf{I}$, the bias, that was normalized to the magnitude of the estimated vector, is really close to zero even for low SNR, and for $\mathbf{m}^\mathbf{R}$ and $\mathbf{m}^\mathbf{I}$ the bias tends to slowly move away from zero for low SNR.

The second step consists in visualizing the impact of the noise level on the reconstruction error, computed from the MSE. Figure 5.2 presents the performances of the amplitude estimator for the previously defined range of SNR.

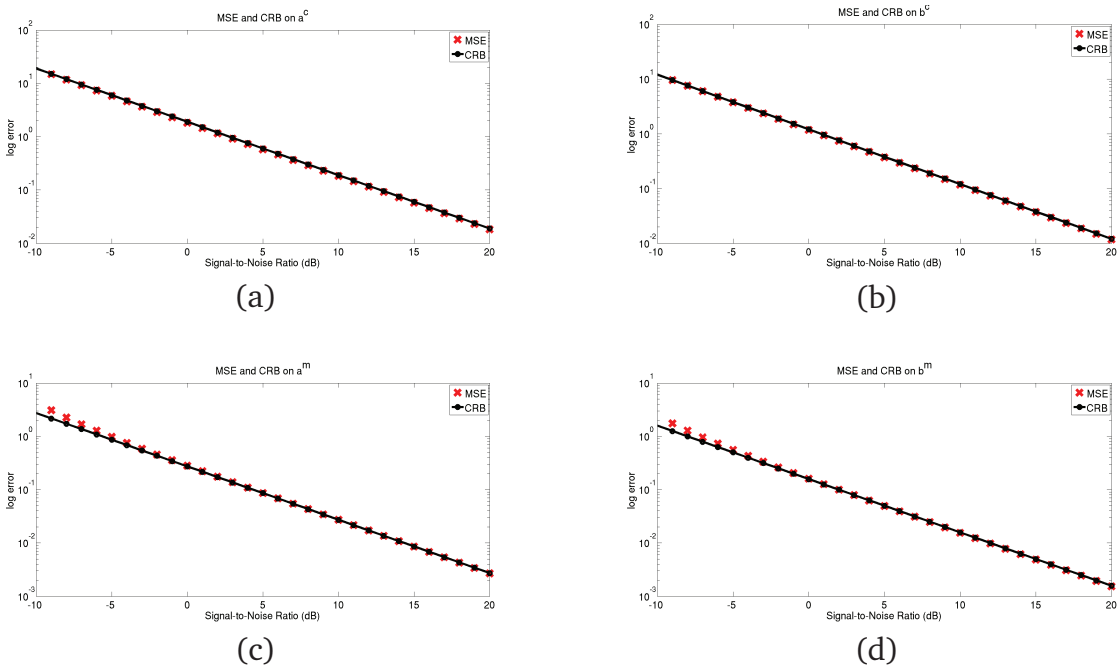


Figure 5.2: Graphical representations of the MSE and the CRB in a given range of signal-to-noise ratios for the four vector parameters (a) $\mathbf{c^R}$, (b) $\mathbf{c^I}$, (c) $\mathbf{m^R}$ and (d) $\mathbf{m^I}$.

The MSE is plotted along with its corresponding CRB. It can be seen that for all the parameters the mean square errors computed on the estimated amplitudes achieve the bound. For the parameter vectors $\mathbf{m^R}$ and $\mathbf{m^I}$, for low SNR, the estimator is moving away from the bound, which corresponds to the emergence of the bias. This can be explained with the loss of resolution for very low frequency signals. This results will be verified in the study of the signal's length influence.

Influence of the length of the signal

The second criterion for the estimator assessment leads to the analysis of the signal's length effect on the performances of the amplitude estimation. For this simulation, it is evaluated for a given signal-to-noise ratio of $SNR = 0dB$ and with the signal length ranging from $N = 1000$ to $N = 50000$.

As before, the estimator's bias is first evaluated. In Figure 5.3, the bias evolution was plotted for each parameter. It can be seen that in average, the bias for all parameter vectors oscillates around zero. For signals with very few samples, it can be noticed that the bias tends to increase.

The reconstruction error is then computed to complete the performance study. Figure 5.4 displays the MSE versus the data-length at $SNR = 0dB$. Again it is possible to note that the proposed amplitude estimator achieves the CRB for all the parameters and whatever the signal length, except for really short signals ($N < 2000$) where the resolution of the temporal signal is not sufficient.

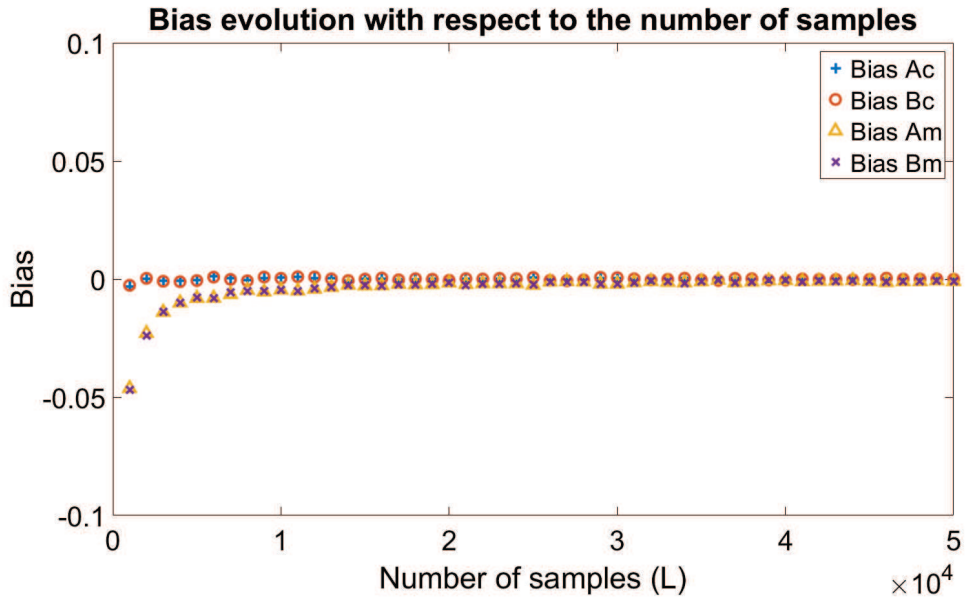
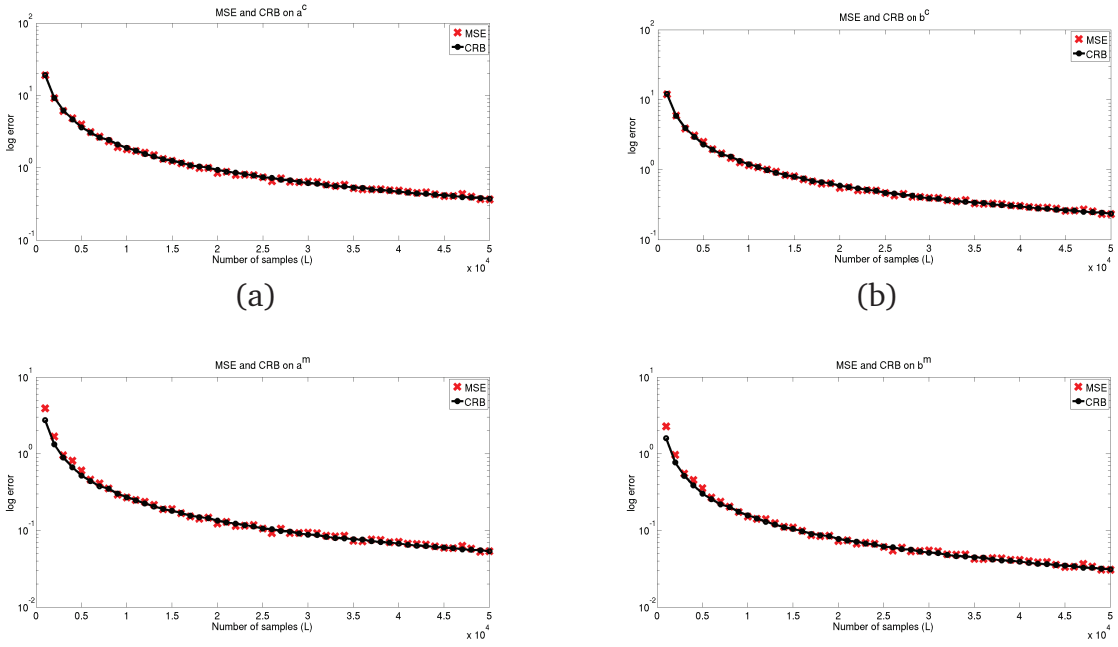


Figure 5.3: Overlay of the evolution of the bias for all four parameter vectors with respect to the number of samples.



(a) (b) (c) (d)

Figure 5.4: Graphical representations of the MSE and the CRB for several length of the signal and for the four vector parameters (a) $\mathbf{c}^{\mathbf{R}}$, (b) $\mathbf{c}^{\mathbf{I}}$, (c) $\mathbf{m}^{\mathbf{R}}$ and (d) $\mathbf{m}^{\mathbf{I}}$.

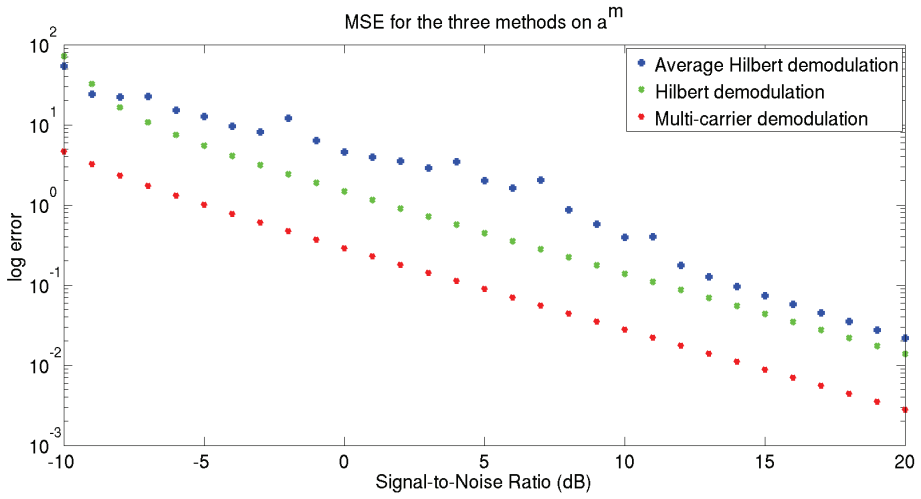
Those two studies might lead to the conclusion that first, by experiment, all vector parameters are asymptotically unbiased accordingly to both the SNR and the number of samples, and second that it is also asymptotically statistically efficient as the CRB is achieved in a large majority of cases.

5.2.4 Comparison with classical demodulation methods

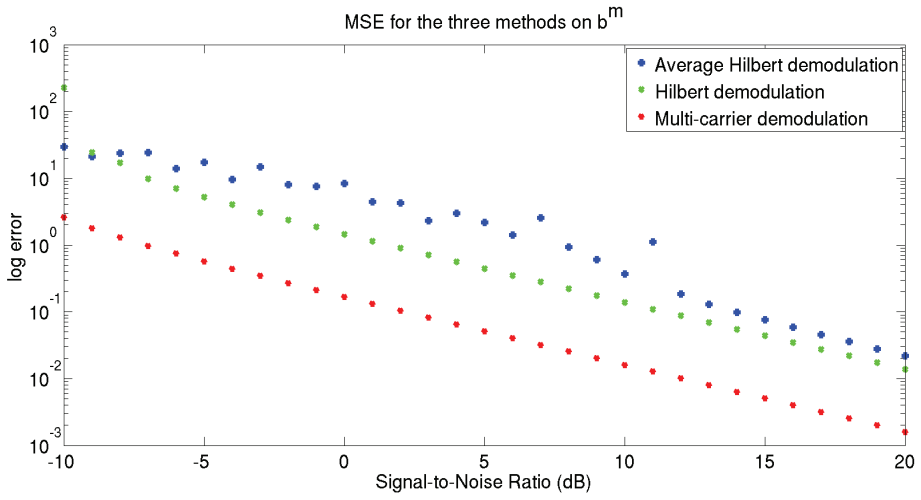
In this subsection the capacity of the presented method is compared to the widespread demodulation algorithm to estimate the modulations of an amplitude modulated signal.

In order to have a fair comparison, the proposed optimal demodulation algorithm is confronted to the classical demodulation and to an averaged demodulation. This last one is the classical demodulation introduced in Section 3.1, computed on every carrier and then averaged to estimate the modulations.

In Figure 5.5 a Monte-Carlo simulation is done for 10000 draws in order to compare the performance of the presented optimal demodulation algorithm (red triangles) with the one of the single classical demodulation (blue squares) and the averaged demodulation (green diamonds) on the estimation of the modulation parameters $\mathbf{m^R}$ and $\mathbf{m^I}$.



(a)



(b)

Figure 5.5: Overlay of the MSE for the same modulation estimation with the classical demodulation and the averaged demodulation methods and the proposed (optimal) multi-carrier demodulation method for both parameters (a) $\mathbf{m^R}$ and (b) $\mathbf{m^I}$.

It can be seen that the MSE on the estimation of the modulation amplitudes with the multi-carrier demodulation is always lower than the one performed on the amplitudes

estimated with both the classical and averaged demodulation methods until a signal-to-noise ratio of $-10dB$. The average Hilbert demodulation is worse than the single one as the average is done with demodulation for all carrier harmonics, unlike Hilbert demodulation which has been done on the carrier harmonics with the best signal-to-noise ratio. The amplitudes of the modulation signal are estimated with more accuracy with the proposed multi-carrier demodulation method than with traditional algorithms. It has to be noticed that the Hilbert demodulation does not allow the estimation of a carrier with several harmonics. To do so, a second distinct step has to be done to estimate the high frequency (carrier) signal with a synchronous average algorithm for example, whereas it is estimated at the same time with the multi-carrier demodulation.

5.3 Amplitude demodulation with overlapping

This study has been conducted collectively with the condition monitoring research team of the School of Mechanical and Manufacturing Engineering of the University of New South Wales of Sydney.

In the previous case we considered that the spectral support of the signal under study was bounded, ensuring that any overlap between the modulations was avoided. However, it is also interesting to consider the case where modulations do overlap, i.e. the upper limit of the spectral support of the modulation signal is bigger than half the carrier frequency, in other words Hypothesis 1 set in Chapter 3 is not fulfilled.

Based on the model 5.1 defined in section 5.1, we can compute the signals DFT over a finite length of N samples and at a sampling rate F_s. Here we specify that the carrier frequency f_c is an integer multiple of the modulation's fundamental frequency $f_c = Gf_m$. The components corresponding to the carrier harmonics and sidebands will be identified by $\tilde{s}(t_n) = \tilde{s}(nf_m)$ and can thus be expressed according to the model as

$$\tilde{s}(t_n) = \sum_{h=-H}^{H} C_h M_{n-hG} + Y_n, \tag{5.11}$$

with $Y_n \sim \mathcal{CN}(0, \frac{\sigma^2}{N})$ and $\mathcal{CN}$ representing a circular symmetric complex normal distribution. The quantities $\tilde{s}(t_n)$ will be therefore stochastic with the following distribution:

$$\tilde{s}(t_n) \sim \mathcal{CN}\left(\mu_n(\boldsymbol{\theta}), \frac{\sigma^2}{N}\right). \tag{5.12}$$

with

$$\mu_n(\boldsymbol{\theta}) = \sum_{h=-H}^{H} C_h M_{n-hG}.$$

5.3.1 Maximum Likelihood Estimator

Assuming H and K known, the estimation of the parameter vector $\boldsymbol{\theta}$, i.e. the carrier and modulation amplitudes coefficients, can be formulated as a maximum-likelihood problem. The negative-log-likelihood function of a set of $\tilde{s}(t_n)$ with $n \in \mathcal{W}$ is:

$$\bar{\Lambda} = L\, log\left(\frac{\pi\sigma^2}{N}\right) + N \sum_{n\in\mathcal{W}} \frac{|\tilde{s}(t_n) - \mu_n(\boldsymbol{\theta})|^2}{\sigma^2},$$

where L is the size of the set $\mathcal{W}$ of selected frequency points. The problem of minimizing $\bar{\Lambda}$ against μ_n is independent on the variance σ^2, so it can be rewritten as the following least-squares problem:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{argmin} \sum_n |\tilde{s}(t_n) - \mu_n(\boldsymbol{\theta})|^2 \\ &= \underset{\boldsymbol{\theta}}{argmin} \left\{ \sum_n \left[\tilde{s}_n^R - \mu_n^R(\boldsymbol{\theta})\right]^2 + \sum_n \left[\tilde{s}_n^I - \mu_n^I(\boldsymbol{\theta})\right]^2 \right\}.\end{aligned} \tag{5.13}$$

In order to be able to solve this non-linear problem, we propose to use the well-known Gauss-Newton descent algorithm on the residuals $\mathbf{r} = \begin{bmatrix} \mathbf{r^R} \\ \mathbf{r^I} \end{bmatrix}$, where

$$\begin{aligned} r_n^R &= \tilde{s}_n^R - \mu_n^R(\boldsymbol{\theta}) \\ r_n^I &= \tilde{s}_n^I - \mu_n^I(\boldsymbol{\theta}), \end{aligned} \tag{5.14}$$

and

$$\begin{aligned} \mu_n^R(\boldsymbol{\theta}) &= \sum_{h=-H}^{H} \Re\left\{C_h M_{n-hG}\right\} = \sum_{h=-H}^{H} \left\{C_h^R M_{n-hG}^R - C_h^I M_{n-hG}^I\right\} \\ \mu_n^I(\boldsymbol{\theta}) &= \sum_{h=-H}^{H} \Im\left\{C_h M_{n-hG}\right\} = \sum_{h=-H}^{H} \left\{C_h^I M_{n-hG}^R + C_h^R M_{n-hG}^I\right\}. \end{aligned} \tag{5.15}$$

Remark 19. *The following notation has been chosen to represent real and imaginary parts of the considered object :* $(\cdot)^R$ *and* $(\cdot)^I$.

Using the symmetry of the spectrum of both the carrier and the modulation, we can rewrite expressions in 5.15 as

$$\begin{aligned} \mu_n^R(\boldsymbol{\theta}) &= \sum_{h=1}^{H} \left\{C_h^R M_{|n-hG|}^R + C_h^R M_{n+hG}^R - C_h^I M_{|n-hG|}^I sgn(n-hG) + C_h^I M_{n+hG}^I\right\} \\ \mu_n^I(\boldsymbol{\theta}) &= \sum_{h=1}^{H} \left\{C_h^I M_{|n-hG|}^R - C_h^I M_{n+hG}^R + C_h^R M_{|n-hG|}^I sgn(n-hG) + C_h^R M_{n+hG}^I\right\}. \end{aligned} \tag{5.16}$$

The Jacobian of the residual $\boldsymbol{r}$ with respect to the unknown parameters $\boldsymbol{\theta}$ is defined as

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J^{R,C^R}} & \boldsymbol{J^{R,C^I}} & \boldsymbol{J^{R,M^R}} & \boldsymbol{J^{R,M^I}} \\ \boldsymbol{J^{I,C^R}} & \boldsymbol{J^{I,C^I}} & \boldsymbol{J^{I,M^R}} & \boldsymbol{J^{I,M^I}} \end{bmatrix} \tag{5.17}$$

which all sub-matrices are composed by the following elements:

$$
\begin{array}{rcl}
j_{n,h}^{R,C^R} & = & -\left\{M_{|n-hG|}^R + M_{n+hG}^R\right\} \\
j_{n,h}^{R,C^I} & = & -\left\{-M_{|n-hG|}^I sgn(n-hG) + M_{n+hG}^I\right\} \\
j_{n,k}^{R,M^R} & = & -\sum_{h=1}^{H}\left\{C_h^R \delta_{|n-hG|,k} + C_h^R \delta_{n+hG,k}\right\} \\
j_{n,k}^{R,M^I} & = & -\sum_{h=1}^{H}\left\{-C_h^I sgn(n-hG)\delta_{|n-hG|,k} + C_h^I \delta_{n+hG,k}\right\} \\
j_{n,h}^{I,C^R} & = & -\left\{M_{|n-hG|}^I sgn(n-hG) + M_{n+hG}^I\right\} \\
j_{n,h}^{I,C^I} & = & -\left\{M_{|n-hG|}^R - M_{n+hG}^R\right\} \\
j_{n,k}^{I,M^R} & = & -\sum_{h=1}^{H}\left\{C_h^I \delta_{|n-hG|,k} - C_h^I \delta_{n+hG,k}\right\} \\
j_{n,k}^{I,M^I} & = & -\sum_{h=1}^{H}\left\{C_h^R sgn(Mn-hG)\delta_{|n-hG|,k} + C_h^R \delta_{n+hG,k}\right\}.
\end{array}
$$

The Gauss-Newton method can be then implemented by iteratively refining the estimation of the parameters $\boldsymbol{\theta}$ as

$$
\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} - \left\{\left(\boldsymbol{\theta}^{(s)}\right)\right\}^{\dagger} \boldsymbol{r}\left(\boldsymbol{\theta}^{(s)}\right),
$$

where the operator $(\cdot)^{\dagger}$ represents the pseudo-inverse operator.

5.3.2 Confidence interval

Confidence intervals are a range of potentials values of the unknown parameter and defined as the relation between all the parameters, which is represented by the covariance matrix. In order to compute that covariance matrix, we use its standard approximation by the Hessian matrix, made possible by the formulation of our problem with the Gauss-newton algorithm. In other words, the covariance matrix $\boldsymbol{\Sigma}$ is defined as $\boldsymbol{\Sigma} = \boldsymbol{H}^{-1}$ where $\boldsymbol{H}$ is the Hessian matrix. The Hessian matrix is the second derivative of the residual with respect to each parameter in the full parameter vector $\boldsymbol{\beta} = [\boldsymbol{\theta}\sigma]$:

$$
\boldsymbol{H} = \frac{\partial^2 \Lambda}{\partial \boldsymbol{\beta}^2}, \tag{5.18}
$$

which can be detailed as:

$$
\boldsymbol{H} = \begin{bmatrix}
\boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{C^R}} & \boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{C^I}} & \boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{M^R}} & \boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{M^I}} & \boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{\sigma}} \\
\boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{C^R}} & \boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{C^I}} & \boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{M^R}} & \boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{M^I}} & \boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{\sigma}} \\
\boldsymbol{H}^{\boldsymbol{M^R},\boldsymbol{C^R}} & \boldsymbol{H}^{\boldsymbol{M^R},\boldsymbol{C^I}} & \boldsymbol{H}^{\boldsymbol{M^R},\boldsymbol{M^R}} & \boldsymbol{H}^{\boldsymbol{M^R},\boldsymbol{M^I}} & \boldsymbol{H}^{\boldsymbol{M^R},\boldsymbol{\sigma}} \\
\boldsymbol{H}^{\boldsymbol{M^I},\boldsymbol{C^R}} & \boldsymbol{H}^{\boldsymbol{M^I},\boldsymbol{C^I}} & \boldsymbol{H}^{\boldsymbol{M^I},\boldsymbol{M^R}} & \boldsymbol{H}^{\boldsymbol{M^I},\boldsymbol{M^I}} & \boldsymbol{H}^{\boldsymbol{M^I},\boldsymbol{\sigma}} \\
\boldsymbol{H}^{\boldsymbol{\sigma},\boldsymbol{C^R}} & \boldsymbol{H}^{\boldsymbol{\sigma},\boldsymbol{C^I}} & \boldsymbol{H}^{\boldsymbol{\sigma},\boldsymbol{M^R}} & \boldsymbol{H}^{\boldsymbol{\sigma},\boldsymbol{M^I}} & \boldsymbol{H}^{\boldsymbol{\sigma},\boldsymbol{\sigma}}
\end{bmatrix}
$$

Remark 20. *It is possible to notice that $\boldsymbol{H}$ is symmetric, i.e. $\boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{C^I}} = \boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{C^R}}$ for each couple of parameters.*

Based on Remark 20, it is not necessary to calculate separately all the sub-matrices.

The essentials sub-matrices are defined as below:

$$
\begin{aligned}
\boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{C^R}} &= \tfrac{2H}{\sigma^2}\left(1+2\sum_{k=1}^{K}\left(M_k^R+M_k^I\right)\right)\\
\boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{C^I}} &= \mathbf{0}^{H\times H}\\
\boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{M^R}} &= \tfrac{4K}{\sigma^2}\sum_{h=1}^{H}C_h^R\\
\boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{M^I}} &= \tfrac{4K}{\sigma^2}\sum_{h=1}^{H}C_h^R\\
\boldsymbol{H}^{\boldsymbol{C^R},\boldsymbol{\sigma}} &= -\tfrac{4}{\sigma^3}\Big(\sum_{h=1}^{H}\left(-\tilde{s}_{h,0}^R+C_h^R\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,k}^R+C_h^RM_k^R-C_h^IM_k^I\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,k}^I+C_h^RM_k^I+C_h^IM_k^R\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,-k}^R+C_h^RM_k^R+C_h^IM_k^I\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(\tilde{s}_{h,-k}^I+C_h^RM_k^I-C_h^IM_k^R\right)\Big)\\
\boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{C^I}} &= \tfrac{2H}{\sigma^2}\left(1+2\sum_{k=1}^{K}\left(M_k^R+M_k^I\right)\right)\\
\boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{M^R}} &= \tfrac{4K}{\sigma^2}\sum_{h=1}^{H}C_h^I\\
\boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{M^I}} &= \tfrac{4K}{\sigma^2}\sum_{h=1}^{H}C_h^I\\
\boldsymbol{H}^{\boldsymbol{C^I},\boldsymbol{\sigma}} &= -\tfrac{4}{\sigma^3}\Big(\sum_{h=1}^{H}\left(-\tilde{s}_{h,0}^I+C_h^I\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(\tilde{s}_{h,k}^R-C_h^RM_k^R+C_h^IM_k^I\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,k}^I+C_h^RM_k^I+C_h^IM_k^R\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,-k}^R+C_h^RM_k^R+C_h^IM_k^I\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,-k}^I-C_h^RM_k^I+C_h^IM_k^R\right)\Big)\\
\boldsymbol{H}^{\boldsymbol{M^R},\boldsymbol{M^R}} &= \tfrac{4K}{\sigma^2}\sum_{h=1}^{H}\left(C_h^R+C_h^I\right)\\
\boldsymbol{H}^{\boldsymbol{M^R},\boldsymbol{M^I}} &= \mathbf{0}^{K\times K}\\
\boldsymbol{H}^{\boldsymbol{M^R},\boldsymbol{\sigma}} &= -\tfrac{4}{\sigma^3}\Big(\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,k}^R+C_h^RM_k^R-C_h^IM_k^I\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,k}^I+C_h^RM_k^I+C_h^IM_k^R\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,-k}^R+C_h^RM_k^R+C_h^IM_k^I\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,-k}^I-C_h^RM_k^I+C_h^IM_k^R\right)\Big)\\
\boldsymbol{H}^{\boldsymbol{M^I},\boldsymbol{M^I}} &= \tfrac{4K}{\sigma^2}\sum_{h=1}^{H}\left(C_h^R+C_h^I\right)\\
\boldsymbol{H}^{\boldsymbol{M^I},\boldsymbol{\sigma}} &= -\tfrac{4}{\sigma^3}\Big(\sum_{h=1}^{H}\sum_{k=1}^{K}\left(\tilde{s}_{h,k}^R-C_h^RM_k^R+C_h^IM_k^I\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,k}^I+C_h^RM_k^I+C_h^IM_k^R\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(-\tilde{s}_{h,-k}^R+C_h^RM_k^R+C_h^IM_k^I\right)\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(\tilde{s}_{h,-k}^I+C_h^RM_k^I-C_h^IM_k^R\right)\Big)\\
\boldsymbol{H}^{\boldsymbol{\sigma},\boldsymbol{\sigma}} &= \tfrac{2}{\sigma^2}\Bigg(-\tfrac{2}{3}+\sum_{h=1}^{H}\tfrac{\left(\tilde{s}_{h,0}^R-C_h^R\right)^2}{\sigma^2}\\
&\quad+\sum_{h=1}^{H}\tfrac{\left(\tilde{s}_{h,0}^I-C_h^I\right)^2}{\sigma^2}\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\tfrac{\left(\tilde{s}_{h,k}^R-C_h^RM_k^R+C_h^IM_k^I\right)^2}{\sigma^2}\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\tfrac{\left(\tilde{s}_{h,k}^I-C_h^RM_k^I-C_h^IM_k^R\right)^2}{\sigma^2}\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\tfrac{\left(\tilde{s}_{h,-k}^R-C_h^RM_k^R-C_h^IM_k^I\right)^2}{\sigma^2}\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\tfrac{\left(\tilde{s}_{h,-k}^I+C_h^RM_k^I-C_h^IM_k^R\right)^2}{\sigma^2}\Bigg)
\end{aligned}
$$

5.3.3 Model selection

So far we have assumed that the number of harmonics is known for both carrier and modulation signals, which is usually not the case in practice. Fitting a model with an approximate H is not so critical, as long as higher harmonics are simply negligible in amplitude, but the selection of the correct K is indeed an important task, especially in case of overlapping modulation sidebands, i.e. $\left(K > \frac{G}{2}\right)$ where significantly biased results could be obtained with the wrong assumption on the number of actual modulation harmonics. In order to automatically select the best number of modulation harmonics, the procedure described in the previous section is repeated with a fixed number H of carrier harmonics (for instance the maximum $H = \lfloor \frac{max(n)}{G} \rfloor$) and a series of possible $K^{(q)}, q = 1, \cdots, Q$. The log-likelihood of each model is computed as

$$\Lambda^{(q)} = -L\, log\left(\frac{\pi \sigma^{2\,(q)}}{N}\right) + N \sum_{n \in \mathcal{W}} \frac{|\tilde{s}(t_n) - \mu_n\left(\boldsymbol{\theta}^{(q)}\right)|^2}{\sigma^2}^{(q)} \quad . \tag{5.19}$$

The optimal model can be selected using the maximum likelihood ratio and Wilks' theorem. In statistics, a likelihood test ratio is a statistical test used for comparing the quality of fitting for two statistical models. The test expresses which models of two is a better representation of the data, usually choosing between an alternative model against the null model. When it is the logarithm of the likelihood test that is computed, the Wilks' theorem provides an asymptotic distribution of the ratio statistic. This result has been demonstrated in [93] and states that when a variate is distributed in large samples, then the distribution of $-2\, log\Lambda$ is similar to a χ^2 distribution.

With this in mind, it is possible to compute the following log-likelihood differences

$$\Delta\Lambda^{q,q'} = \Lambda^{(q)} - \Lambda^{(q')}, \tag{5.20}$$

with $q = 1, \cdots, Q$ and $q' = 1, \cdots, q$. Then, Wilks' theorem allows to calculate the p-values. In statistical hypothesis testing p-values are the probality under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value [91]. p-values for the previous log-likelihood differences in 5.20 can be calculated as follow:

$$p^{(q,q')} = \chi^2_{K^{(q)} - K^{(q')}}\left(2\Delta\Lambda^{q,q'}\right), \tag{5.21}$$

where χ^2 represents the chi-squared cumulative distribution function .

The selection of the model that best fit the data can be done by setting an arbitrary threshold α for the p-values and choosing q as

$$q_{opt} = max\left(q | \min_{q'} p^{q,q'} > \alpha\right). \tag{5.22}$$

Remark 21. *The optimal model will be the most complex one that shows the higher p-value against all models along with the lower number of parameters.*

5.3.4 Numerical example

A signal is generated following the model of Eq.5.1 with $N = 100,000$, $F_s = 10kHz$, a fundamental frequency $f_m = 10Hz$, and a gearmesh order $G = 10$. The components of

the signal are shown in Figure 5.6 and include a 20-harmonic carrier and a 15-harmonic modulating signal (thus with different carrier harmonics presenting significantly overlapping sidbands). The resulting signal-to-noise ratio has been evaluated at $6,5dB$.

The actual number of harmonics for carrier and modulation signals are respectively $H = 20$ and $K = 15$.

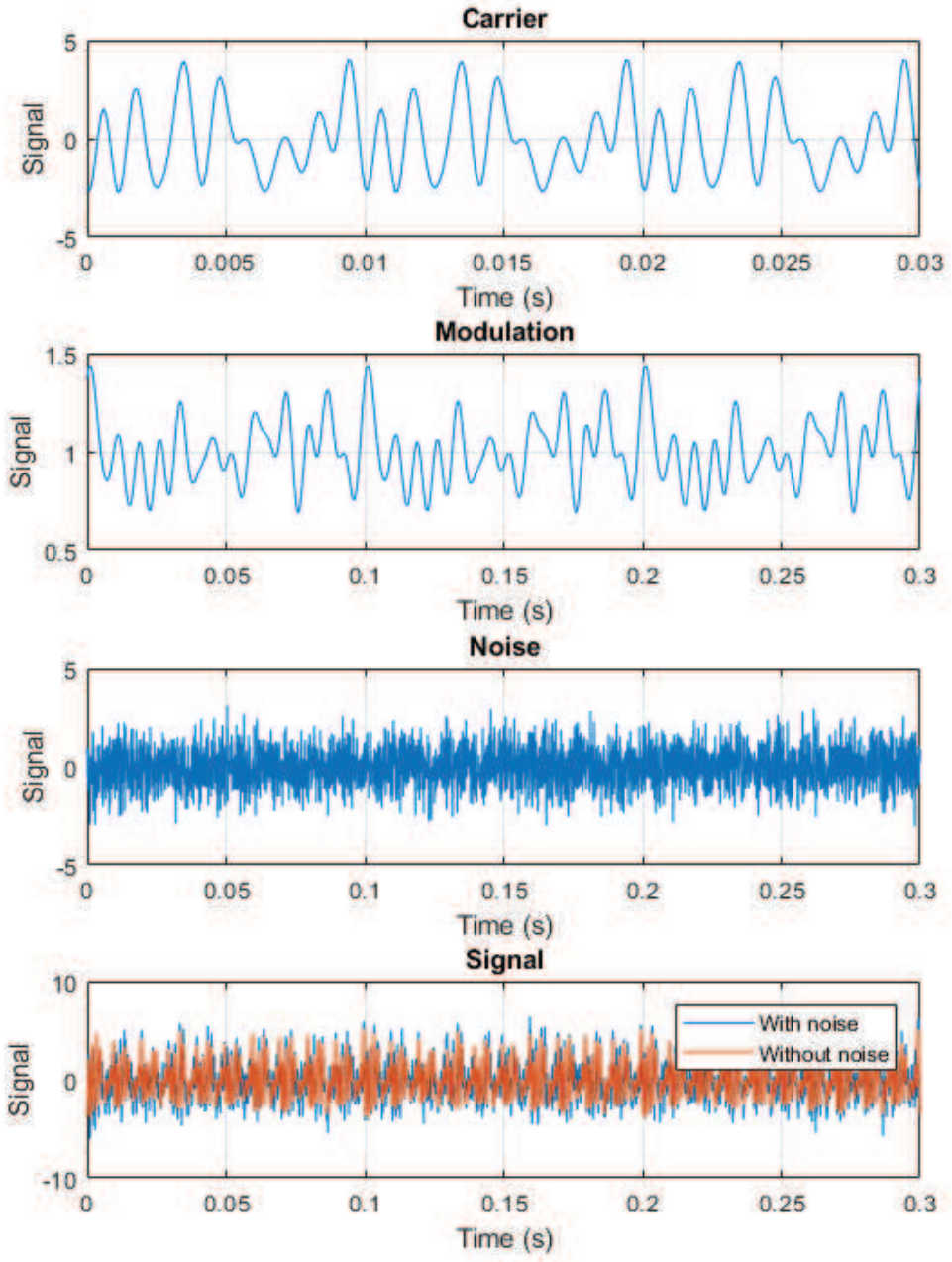


Figure 5.6: Simulation of the temporal signal numerically generated. From top to bottom: Carrier signal, Modulation signal, Noise signal and Overlay of noisy signal set according to Eq.5.1 and the modulated signal only .

The procedure described in the previous sections is applied using all available har-

monics of the fundamental frequency $\tilde{s}_n$ with $n = 1, \cdots, \frac{F_s}{2f_m})$. The maximum number of possible carrier harmonics is set for the model parameters as $\hat{H} = \lfloor \frac{max(n)}{G} \rfloor$ and a search for the optimal value of K is done over the interval $K^{(q)} = [1, \cdots, 30]$, i.e. with the maximum extent of the sidebands set at 3 times the carrier. The likelihood is calculated for each of the 30 optimized models and a p-value matrix comparing each pair is shown in Figure 5.7 (left). This allows the calculation of the combined p-value (minimum value across teach row) and the selection of model 15, which is the most complex model with combined p-value above the predetermined threshold of $\alpha = 0.999$. It has to be highlighted that, despite α being arbitrarily chosen, in this case all combined p-values below 15 have a value of 1, thus leading to selection of model 15 for any α.

Remark 22. *This is obviously theoretically impossible (chi-squared distribution are not right-bounded), but it shows that the confidence is superior to the numerical truncation error of a MATLAB* chi2cdf *function.*

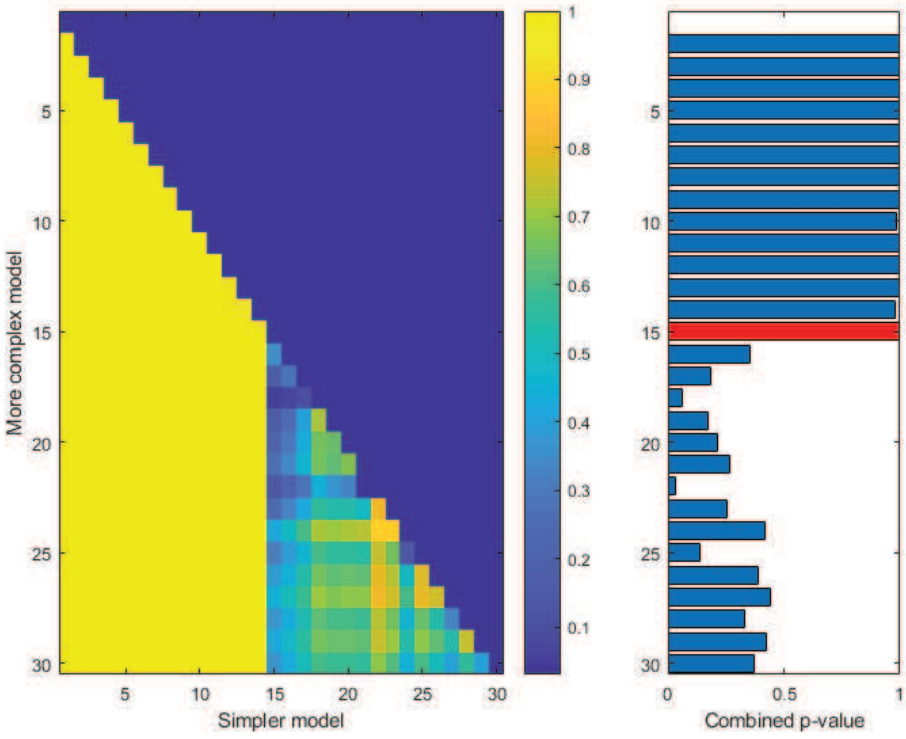


Figure 5.7: Model selection matrix (left) and combined p-value (right). In red the combined p-value of the selected model $(\alpha = 0.999)$.

We evaluate the fitting of the selected model for every amplitude coefficient, i.e. for both carrier and modulation and for real and imaginary parts, as illustrated in Figure 5.8. A temporal overlay of the generated theoretical signal without noise with the estimated amplitude modulated signal is shown in Figure 5.9.

Three periods of the modulation signal and its estimations are shown in Figure 5.9. It is clear how the estimation results in an order-of-magnitude gain in demodulation accuracy.

The MLE method applied to overlapping signal, without previous knowledge of the number of modulation harmonics has been compared to the classical demodulation

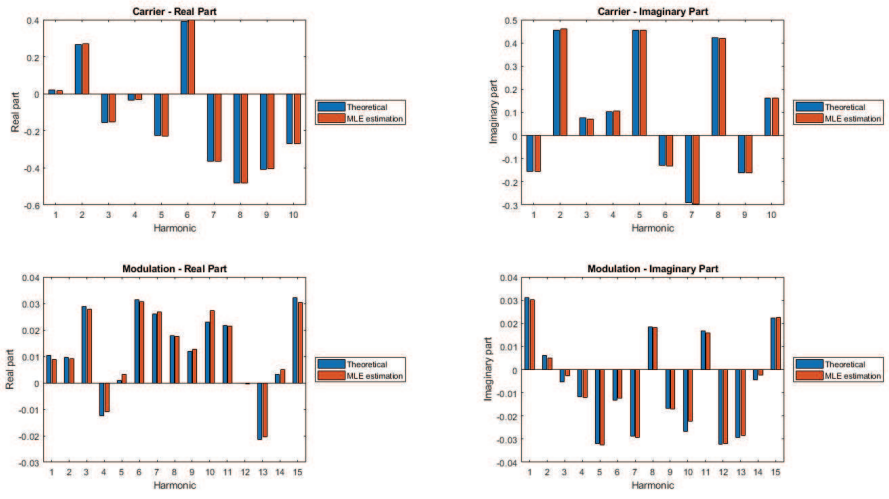


Figure 5.8: Fitting of both the real (left) and imaginary (right) parts of the signal for the carrier (top) and the modulation (bottom) signals.

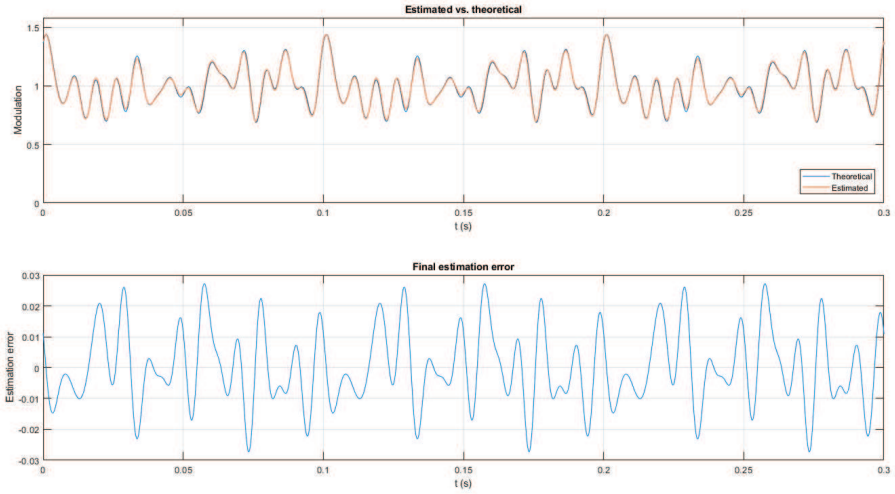


Figure 5.9: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

method. The estimated amplitude coefficients, represented in Figure 5.10, are clearly badly estimated when using the usual demodulation technique as it is not adapted to overlapping sidebands. Figure 5.11 presents the temporal reconstruction of the estimated signals with both methods, and in that case there is no possible comparison in the results as the classical demodulation algorithm is completely wrong in its estimation meanwhile the MLE-based estimation is more than correct.

5.4 Conclusion

The amplitude demodulation formulated as an optimization problem has been studied in this chapter, considering two situations. First the case where we consider that the sidebands of the modulation do not overlap, which allows us to use the multi-carrier

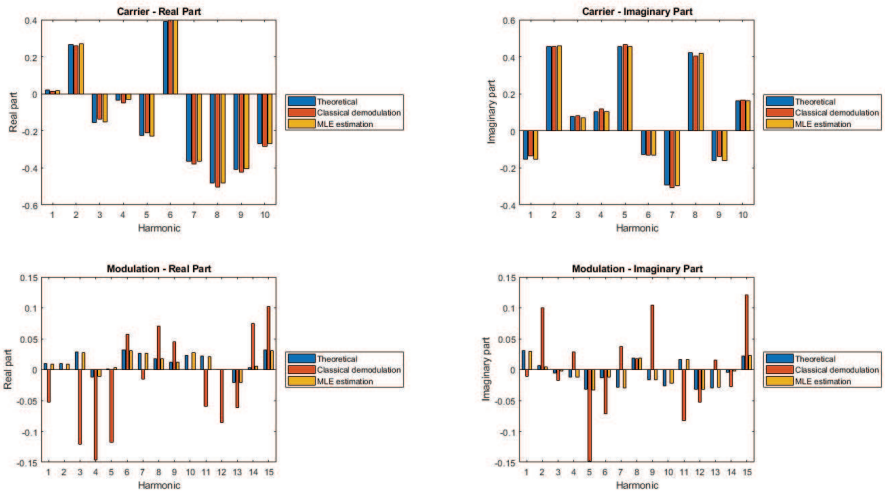


Figure 5.10: Results of the fitting given by the two estimation methods for both the real (left) and imaginary (right) parts of the signal for the carrier (top) and the modulation (bottom) signals.

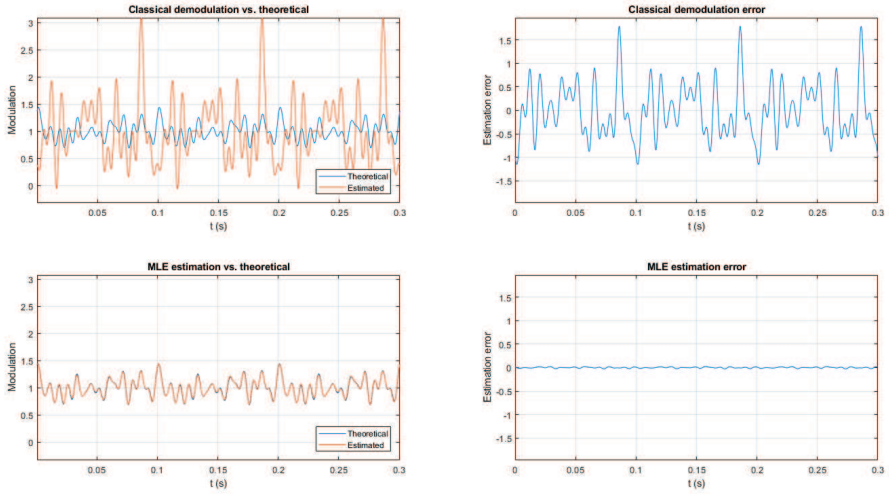


Figure 5.11: Overlay of the estimated signal given by the classical demodulation (top) and MLE-based demodulation (bottom) with the theoretical one (left) and estimation error representation on the temporal representation (right).

demodulation algorithm based on the matrix representation of a spectrum tool defined in Chapter 4. Then in the second case, we considered that sidebands of the modulation can overlap, which implies that the optimization problem has to be tracked with a descent algorithm. Here we used a MLE-based demodulation in a Gauss-Newton descent algorithm. The proposed algorithm has shown interesting properties such as a good estimation of the model number of parameters and also the right amplitude coefficients.

In both cases, the algorithms have been compared to the classical demodulation method which is always the worse estimator.

Chapter 6

Phase and amplitude demodulation

In Chapter 5, we were only focused on the amplitude demodulation problem . But generally speaking, in signal modulation studies, phase modulation is also very common and usually much more difficult to detect and analyze. Indeed, one of phase modulation characteristics is that it generates an infinite spectrum support but with little energy and fast decrease.

For example, little variation in the rotational speed of a rotating machine affects its vibration and is reflected in the measured signal by a phase modulation. For that reason, extensive research has been conducted on combined phase and amplitude demodulation processes. Here we are interested in both the formulation of phase and amplitude demodulation as an optimization problem and the analysis of the properties of that system.

6.1 Problem statement

6.1.1 Model formulation

There are major differences between amplitude modulation and both phase and amplitude modulation signals, but from a statistical point of view their expression are very similar. Based on the previous amplitude modulation statistical model, it is possible to extend it by adding a phase modulation of the carrier. This new formulation is expressed below in Problem 3.

Problem 3 (Exact phase demodulation)**.** *Given a discrete signal* $(s(t_n))_{n\in[\![1,N]\!]}$ *(with* $N \in \mathbb{N}$*) of sampling period* T_s *and duration* $T_{tot} = N \cdot T_s$*, given a frequency* $f_c = k_c \cdot f_{tot}$ *(with* $f_{tot} = 1/T_{tot}$ *and* $k_c \in \mathbb{N}$*), find* $(s_c(t_n), s_m(t_n), s_\Phi(t_n))$ *verifying for any* $n \in [\![1, N]\!]$*:*

$$s(t_n) = s_c\left(t_n + s_\Phi(t_n)\right) s_m(t_n), \tag{6.1}$$

where s_c *and* s_m *are the same carrier and amplitude modulation as before, and* s_Φ *is the new element representing the phase modulation.*

As previously exposed, those signals are periodic and defined with complex coeffi-

cients as below:

$$\begin{cases} s_c(t_n) = \dfrac{1}{N} \sum\limits_{h=-H}^{H} C_h e^{2i\pi h f_c t_n}, \\ s_m(t_n) = \dfrac{1}{N} \sum\limits_{k=-K}^{K} M_k e^{2i\pi k f_m t_n} \\ s_\Phi(t_n) = \dfrac{1}{N} \sum\limits_{j=-\infty}^{\infty} P_j e^{2i\pi j f_\Phi t_n}. \end{cases} \tag{6.2}$$

A study of this general system is difficult, therefore we reckon an additional hypothesis.

Hypothesis 3. *In this study we consider that the phase fluctuation is small, i.e.* $s_\Phi(t_n) << 2\pi$*, which means that the delay induced by the phase modulation does not cause any lag bigger than a short percentage of the carrier period.*

Under the constraint given by Hypothesis 3, we can use a first-order approximation w.r.t. s_Φ using Taylor's series, that allows us to rewrite the phase modulation of the carrier as follow:

$$s_c\left(t_n + s_\Phi(t_n)\right) \simeq s_c(t_n) + {s_c}'(t_n) s_\Phi(t_n),$$

where ${s_c}'(t_n)$ is the temporal derivative of the carrier signal $s_c(t_n)$. It is possible then to reformulate Eq. 6.1 of Problem 3 into

$$s(t_n) = s_c(t_n) s_m(t_n) + {s_c}'(t_n) s_\Phi(t_n) s_m(t_n). \tag{6.3}$$

Setting $s_{c1} = s_c$, $s_{c2} = {s_c}'$, $s_{m1} = s_m$ and $s_{m2}(t_n) = s_\Phi(t_n) s_m(t_n)$, Eq. 6.3 yields to:

$$s(t_n) = s_{c1}(t_n) s_{m1}(t_n) + s_{c2}(t_n) s_{m2}(t_n). \tag{6.4}$$

Remark 23. *It can be noticed that Eq. 6.4 can be seen as an extension of the amplitude modulation formulation of Eq. 3.11 introduced in Chapter 3. This observation has lead us to set a more general class of signals that are of the form:*

$$s(t_n) = \sum_{i=1}^{q} s_{ci}(t_n)\, s_{mi}(t_n).$$

Remark 24. *Some conditions can be set in order to have* s_{m2} *and* s_{m1} *having zeros at the same locations to be able to recover* $s_\Phi(t_n)$*.*

Remark 25. *In the remain of the study, the frequency of the phase modulation signal is the same as the one of the amplitude modulation, i.e.* $f_\Phi = f_m$*.*

Optimal demodulation In Chapter 3, we have expressed amplitude demodulation as an optimization problem. This concept can be extended to the amplitude and phase modulation issue expressed in Problem3. The same analogy that has been done in amplitude demodulation for Problems 1 and 2, can be repeated for Problem 3.

Problem 4 (Optimal phase demodulation)**.** *Given a discrete signal* $(s(t_n))_{n\in[\![1,N]\!]}$ *(with* $N \in \mathbb{N}$*) of sampling period* T_s *and duration* $T_{tot} = N \cdot T_s$*, given a frequency* $f_c = k_c \cdot f_{tot}$ *(with* $f_{tot} = 1/T_{tot}$ *and* $k_c \in \mathbb{N}$*), estimate all signals* $(s_{c1}, s_{c2}, s_{m1}, s_{m2})$ *minimizing the following cost function:*

$$C(s_c, s_m) = \sum_{n=1}^{N} \left| s_{c1}(t_n)\, s_{m1}(t_n) + s_{c2}(t_n)\, s_{m2}(t_n) - s(t_n) \right|^2$$

with f_c *the frequency of* $s_c(t_n)$*.*

6.1.2 Matrix formulation

In practice, transposing Problems 3 and 4 in the Fourier domain and using the matrix formulation introduced in Chapter 4 allows a very handy reformulation of the problems.

Problems 3 and 4 can be both reformulated within a matrix formulation.

Problem 5 (Exact matrix formulation)**.** *If $\mathbf{c_1}$ (resp., $\mathbf{c_2}, \mathbf{m_1}, \mathbf{m_2}$) denotes the vector containing the Fourier coefficients of $s_{c1}(t_n)$ (resp., $s_{c2}(t_n), s_{m1}(t_n), s_{m2}(t_n)$), and ^{H} the Hermitian transpose operator, then Problem 3 is equivalent to finding the vectors $\mathbf{c_1}, \mathbf{c_2}$ and the vectors $\mathbf{m_1}, \mathbf{m_2}$ satisfying:*

$$\mathbf{M_s} = \mathbf{c_1}\,\mathbf{m_1}^H + \mathbf{c_2}\,\mathbf{m_2}^H.$$

Problem 6 (Optimal matrix formulation)**.** *If $\mathbf{c_1}$ (resp., $\mathbf{c_2}, \mathbf{m_1}, \mathbf{m_2}$) denotes the vector containing the Fourier coefficients of $s_{c1}(t_n)$ (resp., $s_{c2}(t_n), s_{m1}(t_n), s_{m2}(t_n)$), and ^{H} the Hermitian transpose operator, then Problem 4 is equivalent to finding the vectors $\mathbf{c_1}, \mathbf{c_2}$ and the vectors $\mathbf{m_1}, \mathbf{m_2}$ satisfying:*

$$C(\mathbf{c_1}, \mathbf{c_2}, \mathbf{m_1}, \mathbf{m_2}) = \left|\left|\mathbf{c_1}\mathbf{m_1}^H + \mathbf{c_2}\mathbf{m_2}^H - \mathbf{M_s}\right|\right|_{Fro}^2.$$

Remark 26. *All vectors $\mathbf{c_1}, \mathbf{c_2}, \mathbf{m_1}, \mathbf{m_2}$ are vectors containing the complex Fourier coefficients of each signal spectrum, i.e. the carrier, the amplitude modulation and the product of phase and amplitude modulation respectively.*

We recall that $\mathbf{c_2} = \mathbf{c_1}' = \mathbf{c}'$ is the derivative of $\mathbf{c}$ in the frequency domain. This derivative operation can be expressed as $\mathbf{c}' = \boldsymbol{D}\,\mathbf{c}$, where $\boldsymbol{D}$ is a diagonal matrix that allows the computation of the derivative of the carrier spectrum in the frequency domain, i.e.

$$\boldsymbol{D} = i\,2\,\pi\,f_c\,diag(-H \cdots H).$$

Thus, we obtain the following general matrix formulation of the phase and amplitude demodulation problem:

$$\mathbf{M_s} = \mathbf{c}\mathbf{m_1}^H + \boldsymbol{D}\mathbf{c}\mathbf{m_2}^H.$$

Remark 27. *Intrinsically, the matrix $\mathbf{M_s}$ has a very specific structure, that is complex centrosymmetric, as it is built with all the complex coefficients of the signal's spectrum. In the case where $\mathbf{M} \in \mathbb{C}^{3\times3}$ the matrix can basically be represented as*

$$\mathbf{M_s} = \begin{bmatrix} a & \bar{b} & \bar{d} \\ b & c & \bar{e} \\ d & e & \bar{a} \end{bmatrix}.$$

This property will be of interest in order to study the indeterminations of the optimization problem 4.

6.2 The exact problem

To address the matrix decomposition problem obtained in the previous section the author contacted the Ouragan Inria research team, who proposed a reconstruction algorithm for the exact case. Section 6.2.1 below is the justification of the algorithm provided by the team, and cannot be considered a part of this PhD work as the author has no contribution to it. It is given only for the sake of self-sufficiency of the manuscript. The reader interested only in the final solution can go straight to Section 6.2.2. The results presented in Sect. 6.3 stem from discussions with the same team, but should be considered part of the PhD work.

6.2.1 Solution to the exact problem

Motivated by Problem 5, we can now state the main problem as: **Main problem** Given $M \in \mathbb{K}^{n\times m}$ and $D \in \mathbb{K}^{n\times n}$, determine – if they exist – $u \in \mathbb{K}^{n\times 1}$ and v_1, $\mathbf{m_2} \in \mathbb{K}^{1\times m}$ satisfying:

$$M = u\, v_1 + D\, u\, \mathbf{m_2}. \tag{6.5}$$

We shall name this problem the *rank 2 decomposition problem*.

Problem 5 corresponds to the case $\mathbb{K} = \mathbb{C}$, $q = 2$, i.e., the phase and amplitude modulation problem, $u = \mathbf{c}$, $D\, u = \boldsymbol{D}\, \mathbf{c}$, $v_1 = {\mathbf{m_1}}^H$, and $\mathbf{m_2} = {\mathbf{m_2}}^H$.

Annexe 10.2 recalls principles of homological and linear algebra for solving inhomogeneous linear systems.

Let us note:

$$A(u) := (u \quad D\, u) \in \mathbb{K}^{n\times 2},\ v := ({v_1}^T \quad {\mathbf{m_2}}^T)^T \in \mathbb{K}^{2\times m}.$$

Then, (6.5) can be rewritten as:

$$A(u)\, v = M. \tag{6.6}$$

Problem defined in (6.6) is bilinear in u and $\mathbf{m}$.

Remark 28. *If $M = 0$, then $u = 0$ or $v_1 = \mathbf{m_2} = 0$ solves the problem. Hence, in what follows, we suppose that $M \neq 0$.*

Remark 29. *The $\mathbb{K}$-vector space $\mathrm{im}_{\mathbb{K}}(A(u).)$ is generated by the two vectors u and $D\, u$, which shows that:*

$$\mathrm{rank}_{\mathbb{K}}(A(u)) := \dim_{\mathbb{K}}(\mathrm{im}_{\mathbb{K}}(A(u).)) \leq 2.$$

A necessary condition for the solvability of (6.5) is then:

$$\mathrm{rank}_{\mathbb{K}}(M) \leq 2. \tag{6.7}$$

Remark 30. *If (6.6) is solvable with a non full row rank matrix $\mathbf{m}$, then there exists $\alpha := (\alpha_1 \quad \alpha_2) \in \mathbb{K}^{1\times 2}$ such that $\alpha\, v = \alpha_1\, v_1 + \alpha_2\, \mathbf{m_2} = 0$, which yields:*

$$\begin{cases} M = \left(\left(D - \alpha_2\, \alpha_1^{-1}\, I_n\right)\, u\right)\, \mathbf{m_2}, & \text{if } \alpha_1 \neq 0, \\ M = \left(\left(I_n - \alpha_1\, \alpha_2^{-1}\, D\right)\, u\right)\, v_1, & \text{if } \alpha_2 \neq 0. \end{cases}$$

Hence, $M \neq 0$ must satisfy $\mathrm{rank}_{\mathbb{K}}(M) = 1$. A necessary condition for the solvability of (6.6) for a matrix M satisfying $\mathrm{rank}_{\mathbb{K}}(M) = 2$ is then that $\mathbf{m}$ has full row rank.

In what follows, we shall consider the case of a full row rank matrix $\mathbf{m}$. By Remark 30, this case includes the case of a matrix M satisfying $\mathrm{rank}_{\mathbb{K}}(M) = 2$ and the case $\mathrm{rank}_{\mathbb{K}}(M) = 1$ by first considering $D = 0$, solving (6.5) with a full row rank matrix $\mathbf{m}$ to get – if it exists – $M = u\, v_1$, and finally considering the solution:

$$M = u\, v_1 + D\, u\, 0.$$

Hence, the rank 1 decomposition problem will be considered as a particular case of the general case with $D = 0$.

Necessary condition for the solvability of Problem (6.5)

Let us solve (6.6). Let $L \in \mathbb{K}^{p\times n}$ be a matrix whose rows generate a basis of the $\mathbb{K}$-vector space $\ker_{\mathbb{K}}(.M)$, i.e., L is a full row rank matrix satisfying:

$$\ker_{\mathbb{K}}(.M) = \mathrm{im}_{\mathbb{K}}(.L) = \mathbb{K}^{1\times p}\, L.$$

Let us explicitly characterize p in terms of the rank of M:

$$\begin{aligned} p &= \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(.M)) = n - \dim_{\mathbb{K}}(\mathrm{im}_{\mathbb{K}}(.M)) \\ &= n - \dim_{\mathbb{K}}(\mathrm{im}_{\mathbb{K}}(M.)) = n - \mathrm{rank}_{\mathbb{K}}(M). \end{aligned} \tag{6.8}$$

Now, (6.6) yields:

$$L\, A(u)\, v = L\, M = 0. \tag{6.9}$$

Since $\mathbf{m}$ has full row rank, we get $L\, A(u) = 0$, i.e., u must satisfy the following $\mathbb{K}$-linear system:

$$\left\{ \begin{array}{l} L\, u = 0, \\ L\, D\, u = 0. \end{array} \right. \tag{6.10}$$

Remark 31. *Since the p rows of L are $\mathbb{K}$-linearly independent, the dimension of the $\mathbb{K}$-vector solution space of $L\, u = 0$, i.e., $\ker_{\mathbb{K}}(L.)$, is $n - p = \mathrm{rank}_{\mathbb{K}}(M)$ by (6.8). The dimension of the solution space of (6.10) is then at most $\mathrm{rank}_{\mathbb{K}}(M)$ (exactly $\mathrm{rank}_{\mathbb{K}}(M)$ if, e.g., $D = I_n$ or $L\, D = 0$).*

Let us now derive an equivalent characterization of (6.10). By definition of L, we have the following exact sequence:

$$0 \longrightarrow \mathbb{K}^{1\times p} \xrightarrow{.L} \mathbb{K}^{1\times n} \xrightarrow{.M} \mathbb{K}^{1\times m}.$$

Applying the exact functor $\mathrm{hom}_{\mathbb{K}}(\,\cdot\,, \mathbb{K})$ to it, we obtain the following dual exact sequence of $\mathbb{K}$-vector spaces:

$$0 \longleftarrow \mathbb{K}^{p\times 1} \xleftarrow{L.} \mathbb{K}^{n\times 1} \xleftarrow{M.} \mathbb{K}^{m\times 1}.$$

Using $\ker_{\mathbb{K}}(L.) = \mathrm{im}_{\mathbb{K}}(M.)$, we get $L\, u = 0$ is equivalent to the existence of $w \in \mathbb{K}^{m\times 1}$ such that $u = M\, w$. Thus, the second equation of (6.10) is equivalent to $L\, (D\, M\, w) = 0$, which in turn is equivalent to the existence of $w' \in \mathbb{K}^{m\times 1}$ such that $D\, M\, w = M\, w'$, which can be rewritten as:

$$(M \quad -D\, M) \begin{pmatrix} w' \\ w \end{pmatrix} = 0.$$

Using (6.7) and the upper bound of *Sylvester's inequality*

$$\begin{aligned} &\mathrm{rank}_{\mathbb{K}}(D) + \mathrm{rank}_{\mathbb{K}}(M) - n \\ &\leq \mathrm{rank}_{\mathbb{K}}(D\, M) \leq \min\{\mathrm{rank}_{\mathbb{K}}(D), \mathrm{rank}_{\mathbb{K}}(M)\}, \end{aligned}$$

then (6.10) is equivalent to:

$$\mathrm{rank}_{\mathbb{K}}(M \quad -D\, M) \leq 3. \tag{6.11}$$

Lemma 1. *With the above notations, a necessary condition on u for the existence of a solution of Problem (6.5) is (6.10) with $p \geq n - 2$, or equivalently (6.7) and (6.11).*

Necessary and sufficient conditions for the solvability of Problem (6.5)

Let u be a non-trivial solution solution of (6.10). We can now form the matrix $A(u) = (u \quad D\,u)$ and we are then led to the study of the linear inhomogeneous system $A(u)\,v = M$. Let $L' \in \mathbb{K}^{p'\times n}$ be a full row rank matrix whose rows form a basis of $\ker_{\mathbb{K}}(.A(u))$, i.e., $\ker_{\mathbb{K}}(.A(u)) = \mathrm{im}_{\mathbb{K}}(.L')$, and $p' = \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(.A(u)))$. By Theorem 2, there exists $v \in \mathbb{K}^{2\times m}$ which satisfies $A(u)\,v = M$ iff the following compatibility condition holds:

$$L'\,M = 0. \tag{6.12}$$

Notice that the compatibility condition (6.12) depends on u, and thus we seek for u – if it exists – in the solution space of (6.10) so that (6.12) holds.

Let us reinterpret (6.12) and (6.9) in a more intrinsic mathematical setting. By definition of L, we have the following exact sequence of $\mathbb{K}$-vector spaces:

$$0 \longrightarrow \mathbb{K}^{1\times p} \xrightarrow{.L} \mathbb{K}^{1\times n} \xrightarrow{.M} \mathbb{K}^{1\times m}\,.$$

Then, $L\,A(u) = 0$ iff u satisfies (6.10). If so, then we get the following complex of $\mathbb{K}$-vector spaces:

$$0 \longrightarrow \mathbb{K}^{1\times p} \xrightarrow{.L} \mathbb{K}^{1\times n} \xrightarrow{.A(u)} \mathbb{K}^{1\times 2}\,.$$

The defect of exactness of this complex at $\mathbb{K}^{1\times n}$ is then:

$$H(\mathbb{K}^{1\times n}, u) := \ker_{\mathbb{K}}(.A(u))/\mathrm{im}_{\mathbb{K}}(.L).$$

Now, by definition of L', we have the following exact sequence of $\mathbb{K}$-vector spaces:

$$0 \longrightarrow \mathbb{K}^{1\times p'} \xrightarrow{.L'} \mathbb{K}^{1\times n} \xrightarrow{.A(u)} \mathbb{K}^{1\times 2}\,. \tag{6.13}$$

Hence, we obtain:

$$\begin{aligned} H(\mathbb{K}^{1\times n}, u) &= \ker_{\mathbb{K}}(.A(u))/\ker_{\mathbb{K}}(.M) \\ &= \mathrm{im}_{\mathbb{K}}(.L')/\mathrm{im}_{\mathbb{K}}(.L). \end{aligned}$$

Since $\mathrm{im}_{\mathbb{K}}(.L) \subseteq \mathrm{im}_{\mathbb{K}}(.L')$, $L \in \mathrm{im}_{\mathbb{K}}(.L')$, and thus there exists $L'' \in \mathbb{K}^{p\times p'}$ such that $L = L''\,L'$. Since L' has full row rank, we obtain the following isomorphism:

$$H(\mathbb{K}^{1\times n}, u) = \mathrm{im}_{\mathbb{K}}(.L')/\mathrm{im}_{\mathbb{K}}(.L) \cong \mathbb{K}^{1\times p'}/\mathrm{im}_{\mathbb{K}}(.L'').$$

Hence, $H(\mathbb{K}^{1\times n}, u) = 0$ iff $\mathrm{im}_{\mathbb{K}}(.L'') = \mathbb{K}^{1\times p'}$, i.e., iff there exists $X \in \mathbb{K}^{p'\times p}$ such that $X\,L'' = I_{p'}$, i.e., iff L'' admits a left inverse, which is also equivalent to the injectivity of the $\mathbb{K}$-linear map $L''.: \mathbb{K}^{p'\times 1} \longrightarrow \mathbb{K}^{p\times 1}$.

Applying the exact functor $\mathrm{hom}_{\mathbb{K}}(\,\cdot\,, \mathbb{K})$ to the exact sequence (6.13), we obtain the exact sequence:

$$0 \longleftarrow \mathbb{K}^{p\times 1} \xleftarrow{L'.} \mathbb{K}^{n\times 1} \xleftarrow{A(u).} \mathbb{K}^{2\times 1}.$$

Then, applying the exact functor $\cdot \otimes_{\mathbb{K}} \mathbb{K}^{1\times m}$ to the last exact sequence, we get the following exact sequence:

$$0 \longleftarrow \mathbb{K}^{p\times m} \xleftarrow{L'.} \mathbb{K}^{n\times m} \xleftarrow{A(u).} \mathbb{K}^{2\times m}.$$

Hence, M belongs to $\mathrm{im}_{\mathbb{K}^{1\times m}}(A(u).) = A(u)\,\mathbb{K}^{2\times m}$, i.e., there exists $v \in \mathbb{K}^{2\times m}$ such that $M = A(u)\,v$, iff:

$$L'\,M = 0.$$

By definition of L, we have $L\,M = 0$, i.e., $L''\,(L'\,M) = 0$. Therefore, $L\,M = 0$ yields $L'\,M = 0$ iff $L''.$ is injective, i.e., iff L'' admits a left inverse, i.e., iff

$$H(\mathbb{K}^{1\times n}, u) \cong \mathbb{K}^{1\times p'}/\mathrm{im}_{\mathbb{K}}(.L'') = 0,$$

i.e., iff:

$$\ker_{\mathbb{K}}(.A(u)) = \ker_{\mathbb{K}}(.M). \tag{6.14}$$

Problem (6.5) is reduced to finding $0 \neq u$ satisfying (6.10) such that the $\mathbb{K}$-vector space $H(\mathbb{K}^{1\times n}, u)$ is trivial, i.e.:

$$\dim_{\mathbb{K}}(H(\mathbb{K}^{1\times n}, u)) = 0.$$

Remark 32. *Let us give a direct interpretation of (6.14). Let us first suppose that (6.5) or equivalently that (6.6) is solvable and let us compare the $\mathbb{K}$-vector spaces:*

$$\ker_{\mathbb{K}}(.A(u)) = \{\lambda \in \mathbb{K}^{1\times n} \mid \lambda\,A(u) = 0\},$$
$$\ker_{\mathbb{K}}(.M) = \{\lambda \in \mathbb{K}^{1\times n} \mid \lambda\,M = 0\}.$$

We clearly have $\ker_{\mathbb{K}}(.A(u)) \subseteq \ker_{\mathbb{K}}(.M)$. Now, if we consider $\lambda \in \ker_{\mathbb{K}}(.M)$, then $\lambda\,A(u)\,v = \lambda\,M = 0$, which yields $\lambda\,A(u) = 0$ since $\mathbf{m}$ is full rank rank. Thus, if a solution exists for Problem (6.5), then (6.14) holds.

Conversely, if u is a non-trivial solution of (6.10) such that (6.14), then $\mathbb{K}^{1\times p'}\,L' = \mathbb{K}^{1\times p}\,L$, which shows that $p' = p$ and there exist $U, V \in \mathbb{K}^{p\times p}$ such that $L' = U\,L$ and $L = V\,L'$, which yields $(U\,V - I_p)\,L' = 0$ and $(V\,U - I_p)\,L = 0$, i.e., $U\,V = I_p$ and $V\,U = I_p$ since both L and L' have full row rank. The compatibility condition $L'\,M = 0$ is thus equivalent to $L\,M = 0$, which is satisfied by definition of L. Then, Theorem 2 shows that there exists $v \in \mathbb{K}^{2\times m}$ such that $A(u)\,v = M$, which solves (6.5).

Let us state the first main result of this section.

Theorem 1. *With the above notations, Problem (6.5) is solvable iff there exists $0 \neq u \in \mathbb{K}^{n\times 1}$ satisfying*

$$\left\{\begin{array}{l} L\,u = 0, \\ L\,D\,u = 0, \end{array}\right.$$

and such that $H(\mathbb{K}^{1\times n}, u) = \ker_{\mathbb{K}}(.A(u))/\ker_{\mathbb{K}}(.M) = 0$.

Let us study the $\mathbb{K}$-vector space $H(\mathbb{K}^{1\times n}, u)$. The Euler-Poincaré characteristic of the short exact sequence

$$0 \longrightarrow \ker_{\mathbb{K}}(.M) \xrightarrow{i} \ker_{\mathbb{K}}(.A(u)) \xrightarrow{\pi} H(\mathbb{K}^{1\times n}, u) \longrightarrow 0,$$

yields:

$$\begin{array}{rcl} \dim_{\mathbb{K}}(H(\mathbb{K}^{1\times n}, u)) & = & \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(.A(u)) - \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(.M)) \\ & = & p' - p. \end{array} \tag{6.15}$$

Let us now characterize p'. Considering the following two short exact sequences of $\mathbb{K}$-vector spaces

$$0 \longrightarrow \ker_{\mathbb{K}}(.A(u)) \longrightarrow \mathbb{K}^{1\times n} \longrightarrow \mathrm{im}_{\mathbb{K}}(.A(u)) \longrightarrow 0,$$

$$0 \longleftarrow \mathrm{im}_{\mathbb{K}}(A(u).) \longleftarrow \mathbb{K}^{2\times 1} \longleftarrow \ker_{\mathbb{K}}(A(u).) \longleftarrow 0,$$

the Euler-Poincaré characteristic then yields:

$$\left\{ \begin{array}{ll} \dim_{\mathbb{K}}(\mathrm{im}_{\mathbb{K}}(.A(u))) & = n - \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(.A(u))) = n - p', \\ \dim_{\mathbb{K}}(\mathrm{im}_{\mathbb{K}}(A(u).)) & = 2 - \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(A(u).)). \end{array} \right.$$

Since we have

$$\mathrm{rank}_{\mathbb{K}}(A(u)) = \dim_{\mathbb{K}}(\mathrm{im}_{\mathbb{K}}(.A(u))) = \dim_{\mathbb{K}}(\mathrm{im}_{\mathbb{K}}(A(u).)),$$

we obtain:

$$p' = n - 2 + \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(A(u).)).$$

By (6.8), we have $p = n - \mathrm{rank}_{\mathbb{K}}(M)$. Hence, we obtain:

$$\begin{aligned} H(\mathbb{K}^{1\times n}, u) = 0 &\Leftrightarrow p' = p \\ &\Leftrightarrow \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(A(u).)) = 2 - \mathrm{rank}_{\mathbb{K}}(M). \end{aligned}$$

Corollary 1. *With the above notations, Problem (6.5) is solvable iff there exists* $0 \neq u \in \mathbb{K}^{n\times 1}$ *satisfying*

$$\left\{ \begin{array}{l} L\,u = 0, \\ L\,D\,u = 0, \end{array} \right.$$

and one of the following two equivalent conditions holds:

1. $p' = p$*, i.e.,* $\dim_{\mathbb{K}}(\ker_{\mathbb{K}}(.A(u))) = \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(.M))$*,*
2. $\dim_{\mathbb{K}}(\ker_{\mathbb{K}}(A(u).)) = 2 - \mathrm{rank}_{\mathbb{K}}(M)$.

Let us now study the $\mathbb{K}$-vector space:

$$\ker_{\mathbb{K}}(A(u).) = \{w \in \mathbb{K}^{2\times 1} \mid A(u)\,w = 0\}.$$

If $w = (w_1 \quad w_2) \in \ker_{\mathbb{K}}(A(u).)$, i.e., $u\,w_1 + D\,u\,w_2 = 0$, then, using $u \neq 0$, we have $w = 0$ if $w_2 = 0$, or $D\,u = -w_1\,w_2^{-1}\,u$ if $w_2 \neq 0$, i.e., u is an eigenvector of D with the eigenvalue $-w_1\,w_2^{-1} \in \mathbb{K}$. Hence, if u is not an eigenvector of D, then $\ker_{\mathbb{K}}(A(u).) = 0$. We then get the following exact sequence

$$0 \longleftarrow \mathbb{K}^{p'\times 1} \xleftarrow{L'.} \mathbb{K}^{n\times 1} \xleftarrow{A(u).} \mathbb{K}^{2\times 1} \longleftarrow 0,$$

and we find again that $p' = n - 2$.

Now, if u is an eigenvalue of D with eigenvalue $\lambda \in \mathbb{K}$, then $\ker_{\mathbb{K}}(A(u).) = \mathrm{im}_{\mathbb{K}}(K.)$, where $K = (-\lambda \quad 1)^T$, is a $\mathbb{K}$-vector space of dimension 1. We get the exact sequence

$$0 \longleftarrow \mathbb{K}^{p'\times 1} \xleftarrow{L'.} \mathbb{K}^{n\times 1} \xleftarrow{A(u).} \mathbb{K}^{2\times 1} \xleftarrow{K.} \mathbb{K} \longleftarrow 0,$$

and we find again that $p' = n - 2 + 1 = n - 1$.

Corollary 2. *With the above notations, Problem (6.5) is solvable iff there exists* $0 \neq u \in \mathbb{K}^{n\times 1}$ *satisfying*

$$\left\{ \begin{array}{l} L\,u = 0, \\ L\,D\,u = 0, \end{array} \right.$$

and such that:

1. *If* $\mathrm{rank}_{\mathbb{K}}(M) = 2$*, then* u *is not an eigenvector of* D *for an eigenvalue* $\lambda \in \mathbb{K}$*. Then, there exists a unique* $v \in \mathbb{K}^{2\times m}$ *satisfying* $A(u)\,v = M$ *defined by*

$$v = E\,M,$$

where $E \in \mathbb{K}^{2\times n}$ *denotes a left inverse of* $A(u)$*.*

2. *If* $\mathrm{rank}_{\mathbb{K}}(M) = 1$*, then* u *is an eigenvector of* D *with an eigenvalue* $\lambda \in \mathbb{K}$*. Then, all the solutions* $v \in \mathbb{K}^{2\times m}$ *satisfying* $A(u)\,v = M$ *are defined by*

$$\forall\, z \in \mathbb{K}^{1\times m}, \quad v = E\,M + K\,z,$$

where $K = (-\lambda \quad 1)^T$ *and* $E \in \mathbb{K}^{2\times n}$ *denotes a generalized inverse of* $A(u)$*.*

6.2.2 Resolution method

In the above section we have proved that there exist necessary and sufficient conditions that allow us to retrieve each couples of carrier and modulation signals. Here we propose an algorithm to compute the solution when it exists.

Algorithm 2. *Given two known matrices* $\mathbf{M} \in \mathbb{C}^{n\times p}$ *and* $\boldsymbol{D} \in \mathbb{C}^{n\times n}$*, and three unknown vectors* $\mathbf{c} \in \mathbb{C}^{n\times 1}$*,* $\mathbf{m_1} \in \mathbb{C}^{p\times 1}$ *and* $\mathbf{m_2} \in \mathbb{C}^{p\times 1}$ *where* n *and* p *are the dimensions of vectors* $\mathbf{c}$ *and* $\mathbf{m}$ *respectively. The matrix equation* $\mathbf{c}{\mathbf{m_1}}^H + \boldsymbol{D}\mathbf{c}{\mathbf{m_2}}^H = \mathbf{M}$ *is feasible if and only if*

- $rank(\mathbf{M}) \leq 2$
- $rank\left([\boldsymbol{D}\mathbf{M}, -\mathbf{M}]\right) \leq 3.$

Under the above conditions, the solution(s) can be found as follows

1. *Perform the SVD of* $\mathbf{M}$

$$\mathbf{M} = \boldsymbol{U}\boldsymbol{V}\boldsymbol{W}^H = \boldsymbol{X}\boldsymbol{Y}^H.$$

2. *Select a transformation matrix* $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$*, such that*

$$vec(\boldsymbol{A}) \in ker\left([\boldsymbol{D}\boldsymbol{X}, -\boldsymbol{X}]\right).$$

3. *The target factorization reveals the solution vectors:*

$$\begin{aligned}\boldsymbol{X}\boldsymbol{A} &= [\mathbf{c}, \boldsymbol{D}\mathbf{c}]\\ \boldsymbol{Y}\boldsymbol{A}^{-H} &= [\mathbf{m_1}, \mathbf{m_2}]\,.\end{aligned}$$

6.3 The optimal problem

If a noisy phase and amplitude modulated signal is studied, there is no way to find a closed-form solution to the optimization problem. In the best case-scenario, some elements can be studied in order to characterized the space where the solutions are, such as the conditions of existence of the solution, the shape of the solution space, the number of minimums, the existence of a global minimum ...

In order to solve this optimization problem, descent techniques have to be implemented. One need to first compute the gradient of the scalar function $C(\mathbf{c}, \mathbf{m_1}, \mathbf{m_2})$ with respect to the real and imaginary part of each variables

For more generality, let consider a more general formulation of the rank-2 problem

$$arg \min_{\mathbf{c_1},\mathbf{c_2},\mathbf{m_1},\mathbf{m_2}} \mathcal{C}(\mathbf{c_1}, \mathbf{c_2}, \mathbf{m_1}, \mathbf{m_2}) = \left|\left|\mathbf{c_1}{\mathbf{m_1}}^H + \mathbf{c_2}{\mathbf{m_2}}^H - \mathbf{M_s}\right|\right|_{Fro}^2,$$

where in the specific case of phase and amplitude demodulation, the parameters are set as $\mathbf{c_1} = \mathbf{c}$ and $\mathbf{c_2} = \boldsymbol{D}\mathbf{c}$.

6.3.1 Gradient computation

In the general formulation defined above, the implementation of the gradient of $\mathcal{C}$ can be reduced to the computation of an equivalent problem. By splitting $\mathcal{C}$ into two sub-problems w.r.t. the variables with the same index, i.e. 1 or 2, we obtain a rank-1 formula which is much simpler. As variables $\mathbf{c_2}$ and $\mathbf{m_2}$ can be considered constant against $\mathbf{c_1}$ and $\mathbf{m_1}$, the two sub-problems can then be rewritten as

$$arg\min_{\mathbf{c_1},\mathbf{m_1}} \mathcal{C}_1(\mathbf{c_1}, \mathbf{m_1}) = \left|\left|\mathbf{c_1}{\mathbf{m_1}}^H - \mathbf{M_1}\right|\right|_{Fro}^2,$$

where $\mathbf{M_1} = \mathbf{c_2}{\mathbf{m_2}}^H - \mathbf{M_s}$ and

$$arg\min_{\mathbf{c_2},\mathbf{m_2}} \mathcal{C}_2(\mathbf{c_2}, \mathbf{m_2}) = \left|\left|\mathbf{c_2}{\mathbf{m_2}}^H - \mathbf{M_2}\right|\right|_{Fro}^2,$$

where $\mathbf{M_2} = \mathbf{c_1}{\mathbf{m_1}}^H - \mathbf{M_s}$. This transformation is allowed as long as each variable is independent from the others.

The main idea at this stage is to study the gradient of the global problem by starting with its rank-1 version. Once both rank-1 gradient are computed (i.e. $\nabla\mathcal{C}_1$ and $\nabla\mathcal{C}_2$), they can then be stacked into a single vector expressing the global gradient.

$$\nabla\mathcal{C} = \begin{bmatrix} \nabla\mathcal{C}_1 \\ \nabla\mathcal{C}_2 \end{bmatrix}.$$

At this stage, we have computed a global gradient for a general rank-2 problem with independent variables. But in our case, we have a link between two variables. Which is why we define a transformation function $\mathcal{H}$ which allows to turn our 3-variables problem into a 4-variables problem.

$$\mathcal{H}\begin{pmatrix} \mathbf{c} \\ \mathbf{m_1} \\ \mathbf{m_2} \end{pmatrix} \rightarrow \begin{pmatrix} \mathbf{c} \\ \mathbf{m_1} \\ \boldsymbol{D}\mathbf{c} \\ \mathbf{m_2} \end{pmatrix}.$$

The transformation matrix $\mathcal{H}$ is built according to the following scheme

$$\mathcal{H} = \begin{bmatrix} \mathbf{Id}^{2(n+p)\times 2(n+p)} & \mathbf{0}^{2(n+p)\times 2p} \\ \mathbf{D_d} & \mathbf{0}^{2(n+p)\times 4p} \\ \mathbf{0}^{2p\times 2(n+p)} & \mathbf{Id}^{2p\times 2p} \end{bmatrix},$$

where matrix $\mathbf{D_d}$ is defined as $\mathbf{D_d} = \begin{bmatrix} \Re(\boldsymbol{D}) & -\Im(\boldsymbol{D}) \\ \Im(\boldsymbol{D}) & \Re(\boldsymbol{D}) \end{bmatrix}$.

It now possible to compute the gradient of the 3-variables problem using the chain rule on the new cost function $\mathcal{L}$.

$$\mathcal{L} = \mathcal{C} \circ \mathcal{H} \iff \mathcal{L} \begin{pmatrix} \mathbf{c} \\ \mathbf{m_1} \\ \mathbf{m_2} \end{pmatrix} = \mathcal{C} \left(\mathcal{H} \begin{pmatrix} \mathbf{c} \\ \mathbf{m_1} \\ \mathbf{m_2} \end{pmatrix} \right).$$

$$\nabla \mathcal{L} = \mathcal{H}^H \nabla \mathcal{C}. \tag{6.16}$$

Now, we will expand the complete detail of the rank-1 gradient computation. The rank-1 problem is written as :

$$\left|\left| M - \mathbf{c}\mathbf{m}^H \right|\right|$$

. All variables are complex, which brings us to the following notations: $M = M_R + jM_I$ (resp.$\mathbf{c}$ and $\mathbf{m}$).

The problem is developed with the complex notation below:

$$\begin{aligned} \left|\left| \mathbf{M} - u\mathbf{m}^H \right|\right| &= Tr\left(\left(\mathbf{M} - \mathbf{c}\mathbf{m}^H \right)^H \left(\mathbf{M} - \mathbf{c}\mathbf{m}^H \right) \right) \\ &= Tr\left(\left(\mathbf{M}^H - \mathbf{m}\mathbf{c}^H \right) \left(\mathbf{M} - \mathbf{c}\mathbf{m}^H \right) \right) \\ &= Tr\left(\mathbf{M}^H M \right) - Tr\left(\mathbf{M}^H \mathbf{c}\mathbf{m}^H \right) - Tr\left(\mathbf{m}\mathbf{c}^H M \right) + Tr\left(\mathbf{m}\mathbf{c}^H \mathbf{c}\mathbf{m}^H \right). \end{aligned}$$

As we are interested in estimating the gradient of the previous optimization problem, the terms which are independent from the variables $\mathbf{c}$ and $\mathbf{m}$ can be ignored.

$$\begin{aligned} -Tr\left(\mathbf{M}^H \mathbf{c}\mathbf{m}^H \right) &= -Tr\left(\left(\mathbf{M}_R^H - iM_I^H \right) \left(\mathbf{c}_R + i\mathbf{c}_I \right) \left(\mathbf{m}_R^H - i\mathbf{m}_I^H \right) \right) \\ &= -Tr\left(\mathbf{M}_R^H \mathbf{c}_R \mathbf{m}_R^H \right) + Tr\left(iM_R^H \mathbf{c}_R \mathbf{m}_I^H \right) - Tr\left(iM_R^H \mathbf{c}_I \mathbf{m}_R^H \right) - Tr\left(\mathbf{M}_R^H \mathbf{c}_I \mathbf{m}_I^H \right) \\ &\quad + Tr\left(iM_I^H \mathbf{c}_R \mathbf{m}_R^H \right) + Tr\left(\mathbf{M}_I^H \mathbf{c}_R \mathbf{m}_I^H \right) - Tr\left(\mathbf{M}_I^H \mathbf{c}_I \mathbf{m}_R^H \right) + Tr\left(iM_I^H \mathbf{c}_I \mathbf{m}_I^H \right). \end{aligned}$$

$$\begin{aligned} -Tr\left(\mathbf{m}\mathbf{c}^H M \right) &= -Tr\left(\left(\mathbf{m}_R + i\mathbf{m}_I \right) \left(\mathbf{c}_R^H - i\mathbf{c}_I^H \right) \left(\mathbf{M}_R + iM_I \right) \right) \\ &= -Tr\left(\mathbf{m}_R \mathbf{c}_R^H M_R \right) - Tr\left(i\mathbf{m}_I \mathbf{c}_R^H M_R \right) + Tr\left(i\mathbf{m}_R \mathbf{c}_I^H M_R \right) - Tr\left(\mathbf{m}_I \mathbf{c}_I^H M_R \right) \\ &\quad - Tr\left(i\mathbf{m}_R \mathbf{c}_R^H M_I \right) + Tr\left(\mathbf{m}_I \mathbf{c}_R^H M_I \right) - Tr\left(\mathbf{m}_R \mathbf{c}_I^H M_I \right) - Tr\left(i\mathbf{m}_I \mathbf{c}_I^H M_I \right). \end{aligned}$$

$$\begin{aligned} -Tr\left(\mathbf{M}^H \mathbf{c}\mathbf{m}^H \right) - Tr\left(\mathbf{m}\mathbf{c}^H M \right) & \\ &= -2Tr\left(\mathbf{M}_R^H \mathbf{c}_R \mathbf{m}_R^H \right) - 2Tr\left(\mathbf{M}_R^H \mathbf{c}_I \mathbf{m}_I^H \right) \\ &\quad + 2Tr\left(\mathbf{M}_I^H \mathbf{c}_R \mathbf{m}_I^H \right) - 2Tr\left(\mathbf{M}_I^H \mathbf{c}_I \mathbf{m}_R^H \right). \end{aligned}$$

$$Tr\left(\mathbf{m}\mathbf{c}^H \mathbf{c}\mathbf{m}^H \right) = ||\mathbf{c}_R||^2 ||\mathbf{m}_R||^2 + ||\mathbf{c}_R||^2 ||\mathbf{m}_I||^2 + ||\mathbf{c}_I||^2 ||\mathbf{m}_R||^2 + ||\mathbf{c}_I||^2 ||\mathbf{m}_I||^2.$$

The new optimization problem can be expressed as:

$$\begin{aligned} \mathcal{C}_R : \left|\left| \mathbf{M} - \mathbf{c}\mathbf{m}^H \right|\right|_{Fro}^2 = &- 2\mathbf{m}_R^H M_R^H \mathbf{c}_R - 2\mathbf{m}_I^H M_R^H \mathbf{c}_I + 2\mathbf{m}_I^H M_I^H \mathbf{c}_R - 2\mathbf{m}_R^H M_I^H \mathbf{c}_I \\ &+ ||\mathbf{c}_R||^2 ||\mathbf{m}_R||^2 + ||\mathbf{c}_R||^2 ||\mathbf{m}_I||^+ ||\mathbf{c}_I||^2 ||\mathbf{m}_R||^2 + ||\mathbf{c}_I||^2 ||\mathbf{m}_I||^2 \end{aligned}$$

The next step is to calculate the partial derivatives w.r.t each variable of the problem, i.e. $\mathbf{c}_R$, $\mathbf{c}_I$, $\mathbf{m}_R$ and $\mathbf{m}_I$.

$$\begin{aligned}
\mathcal{C}_1(\mathbf{c}_R + d\mathbf{c}_R) &= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\mathbf{m}_R^H M_R^H d\mathbf{c}_R + 2\mathbf{m}_I^H M_I^H d\mathbf{c}_R \\
&+ \left(||\mathbf{m}_R||^2 + ||\mathbf{m}_I||^2\right) ||\mathbf{c}_R + d\mathbf{c}_R|| \\
&= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\mathbf{m}_R^H M_R^H d\mathbf{c}_R + 2\mathbf{m}_I^H M_I^H d\mathbf{c}_R \\
&\quad + \left(||\mathbf{m}_R||^2 + ||\mathbf{m}_I||^2\right) \left(||\mathbf{c}_R||^2 + ||d\mathbf{c}_R||^2 + 2\mathbf{c}_R^H d\mathbf{c}_R\right) \\
&= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\left(\mathbf{m}_R^H M_R^H + \mathbf{m}_I^H M_I^H + \left(||\mathbf{m}_R||^2 + ||\mathbf{m}_I||^2\right)\mathbf{c}_R^H\right) d\mathbf{c}_R.
\end{aligned}$$

$$\begin{aligned}
\mathcal{C}_1(\mathbf{c}_I + d\mathbf{c}_I) &= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\mathbf{m}_I^H M_R^H d\mathbf{c}_I - 2\mathbf{m}_R^H M_I^H d\mathbf{c}_I \\
&+ \left(||\mathbf{m}_R||^2 + ||\mathbf{m}_I||^2\right) ||\mathbf{c}_I + d\mathbf{c}_I|| \\
&= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\mathbf{m}_I^H M_R^H d\mathbf{c}_I - 2\mathbf{m}_R^H M_I^H d\mathbf{c}_I \\
&\quad + \left(||\mathbf{m}_R||^2 + ||\mathbf{m}_I||^2\right) \left(||\mathbf{c}_I||^2 + ||d\mathbf{c}_I||^2 + 2\mathbf{c}_I^H d\mathbf{c}_I\right) \\
&= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\left(\mathbf{m}_I^H M_R^H + \mathbf{m}_R^H M_I^H + \left(||\mathbf{m}_R||^2 + ||\mathbf{m}_I||^2\right)\mathbf{c}_I^H\right) d\mathbf{c}_I.
\end{aligned}$$

$$\begin{aligned}
\mathcal{C}_1(\mathbf{m}_R + d\mathbf{m}_R) &= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\mathbf{c}_R^H M_R d\mathbf{m}_R - 2\mathbf{c}_I^H M_I d\mathbf{m}_R \\
&+ \left(||\mathbf{c}_R||^2 + ||\mathbf{c}_I||^2\right) ||\mathbf{m}_R + d\mathbf{m}_R|| \\
&= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\mathbf{c}_R^H M_R d\mathbf{m}_R - 2\mathbf{c}_I^H M_I d\mathbf{m}_R \\
&\quad + \left(||\mathbf{c}_R||^2 + ||\mathbf{c}_I||^2\right) \left(||\mathbf{m}_R||^2 + ||d\mathbf{m}_R||^2 + 2\mathbf{m}_R^H d\mathbf{m}_R\right) \\
&= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\left(\mathbf{c}_R^H M_R + \mathbf{c}_I^H M_I + \left(||\mathbf{c}_R||^2 + ||\mathbf{c}_I||^2\right)\mathbf{m}_R^H\right) d\mathbf{m}_R.
\end{aligned}$$

$$\begin{aligned}
\mathcal{C}_1(\mathbf{m}_I + d\mathbf{m}_I) &= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\mathbf{c}_I^H M_R d\mathbf{m}_I + 2\mathbf{c}_R^H M_I d\mathbf{m}_I \\
&+ \left(||\mathbf{c}_R||^2 + ||\mathbf{c}_I||^2\right) ||\mathbf{m}_I + d\mathbf{m}_I|| \\
&= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\mathbf{c}_I^H M_R d\mathbf{m}_I + 2\mathbf{c}_R^H M_I d\mathbf{m}_I \\
&\quad + \left(||\mathbf{c}_R||^2 + ||\mathbf{c}_I||^2\right) \left(||\mathbf{m}_I||^2 + ||d\mathbf{m}_I||^2 + 2\mathbf{m}_I^H d\mathbf{m}_I\right) \\
&= \left|\left|\mathbf{M} - \mathbf{c}\mathbf{m}^H\right|\right|^2 - 2\left(\mathbf{c}_I^H M_R + \mathbf{c}_R^H M_I + \left(||\mathbf{c}_R||^2 + ||\mathbf{c}_I||^2\right)\mathbf{m}_I^H\right) d\mathbf{m}_I.
\end{aligned}$$

The gradient can though be expressed as:

$$\vec{\nabla}\mathcal{C}_1 = 2 \times \begin{pmatrix}
-\mathbf{m}_R^H M_R^H + \mathbf{m}_I^H M_I^H + \left(||\mathbf{m}_R||^2 + ||\mathbf{m}_I||^2\right)\mathbf{c}_R^H \\
-\mathbf{m}_I^H M_R^H - \mathbf{m}_R^H M_I^H + \left(||\mathbf{m}_R||^2 + ||\mathbf{m}_I||^2\right)\mathbf{c}_I^H \\
-\mathbf{c}_R^H M_R - \mathbf{c}_I^H M_I + \left(||\mathbf{c}_R||^2 + ||\mathbf{c}_I||^2\right)\mathbf{m}_R^H \\
-\mathbf{c}_I^H M_R + \mathbf{c}_R^H M_I + \left(||\mathbf{c}_R||^2 + ||\mathbf{c}_I||^2\right)\mathbf{m}_I^H
\end{pmatrix}. \tag{6.17}$$

6.3.2 Resolution method

Based on the gradient computation as we have defined previously, we have developed several descent algorithms in order to estimate the three unknown vectors. In this work

we will present you two different algorithms. The first algorithm is called Line search: it basically looks for the best descent step solving a polynomial. Here all the vectors are estimated at the same time. The second one is called alternated descent. It is based on a gradient descent too, but this time, the product variables are separated into two parts that are estimated independently and repeatedly.

Line search Algorithm

In order to solve this optimization problem, we compute a gradient descent with a line search assimilated method. The gradient is implemented considering the 3 variables we are interested to estimate $\mathbf{c}$, $\mathbf{m_1}$ and $\mathbf{m_2}$. Then a polynomial of degree 4 is implemented from the cost function. The descent step value is chosen from the minimum root of the polynomial previously computed. The gradient is implemented for each step of the optimization.

Algorithm 3. *Given two known matrices $\mathbf{M} \in \mathbb{C}^{n\times p}$ and $\boldsymbol{D} \in \mathbb{C}^{n\times n}$, and three unknown vectors $\mathbf{c} \in \mathbb{C}^{n\times 1}$, $\mathbf{m_1} \in \mathbb{C}^{p\times 1}$ and $\mathbf{m_2} \in \mathbb{C}^{p\times 1}$ where n and p are the dimensions of vectors $\mathbf{c}$ and $\mathbf{m}$ respectively. The following process is repeated until a satisfying error threshold is reach.*

1. *Compute the gradient w.r.t. the full unknown vector,*
2. *Vectors $\mathbf{c}$, $\mathbf{m_1}$ and $\mathbf{m_2}$ will vary in the direction of the gradient of the cost function,*
3. *Computation of the polynomial coefficients,*
4. *Choosing of the root minimizing the polynomial,*
5. *Incrementation of the estimated vector,*
6. *Update of the reshaped estimated vectors*

We will detail some of the algorithm steps here. Step 2 is called vector variation, and is mere splitting the variation vector obtained with the gradient computed during step 1 Eq. 6.18 into three parts corresponding to the parameters to be estimated.

$$\vec{\nabla}\mathcal{C} = \begin{bmatrix} \vec{\nabla}\mathcal{C}\mathbf{c}_R \\ \vec{\nabla}\mathcal{C}\mathbf{c}_I \\ \vec{\nabla}\mathcal{C}\mathbf{m_1}_R \\ \vec{\nabla}\mathcal{C}\mathbf{m_1}_I \\ \vec{\nabla}\mathcal{C}\mathbf{m_2}_R \\ \vec{\nabla}\mathcal{C}\mathbf{m_2}_I \end{bmatrix} \tag{6.18}$$

$$\begin{aligned} \Delta u &= \vec{\nabla}\mathcal{C}\mathbf{c}_R + i\vec{\nabla}\mathcal{C}\mathbf{c}_I \\ \Delta v_1 &= \vec{\nabla}\mathcal{C}\mathbf{m_1}_R + i\vec{\nabla}\mathcal{C}\mathbf{m_1}_I \\ \Delta \mathbf{m_2} &= \vec{\nabla}\mathcal{C}\mathbf{m_2}_R + i\vec{\nabla}\mathcal{C}\mathbf{m_2}_I. \end{aligned} \tag{6.19}$$

In order to chose the descent step of the optimization, we express the cost function as a 4^{th} degree polynomial depending on a constant parameter α. The polynomial computation of step 3 is given in the Appendix 2.

Two simulations have been run to measure the method performances on the estimation of all parameters vectors. As a first step, the algorithm is studied for a matrix $\mathbf{M}$ built with known vectors $\mathbf{c}$, $\mathbf{m_1}$ and $\mathbf{m_2}$, such as $\mathbf{M} = \mathbf{c}\mathbf{m_1}^H + \boldsymbol{D}\mathbf{c}\mathbf{m_2}^H$. The initialization of the gradient descent is given by the solution with an additive Gaussian noise.

Simulation 1 The initial error on the initialization is set to $1e^{-3}$.

Remark 33. *The initialization error may seem small to the reader when compared to the amplitude modulation, that is true. However, when compared to the phase modulation values, it is very important.*

We first look at the global evolution of the reconstruction error for the global descent as illustrated by Figure 6.1 and the first 1000 descent steps in Figure 6.2.

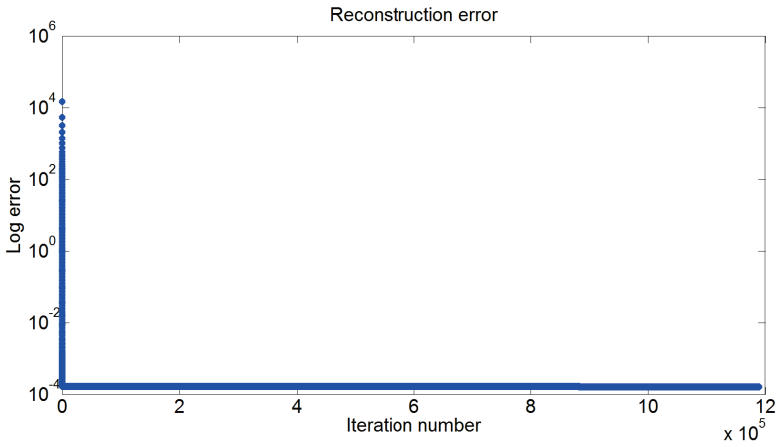


Figure 6.1: Reconstruction error of the estimated vectors

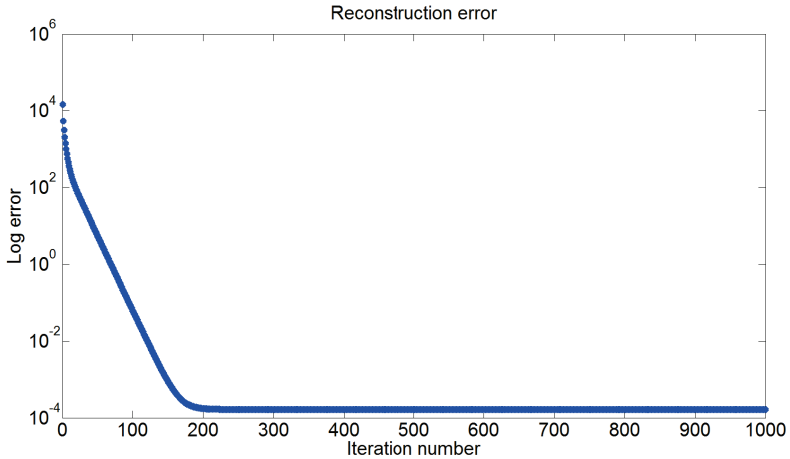


Figure 6.2: Reconstruction error of the estimated vectors zoom on the 1000 first steps

One first thing to note is that the algorithm does not converge as quickly as we thought it would. On the contrary, after reaching a certain error value, it seems not to descent anymore. This is not true, the process continue to go down but at a very low speed.

We also the descent step values, in order to observe the behavior of the algorithm.

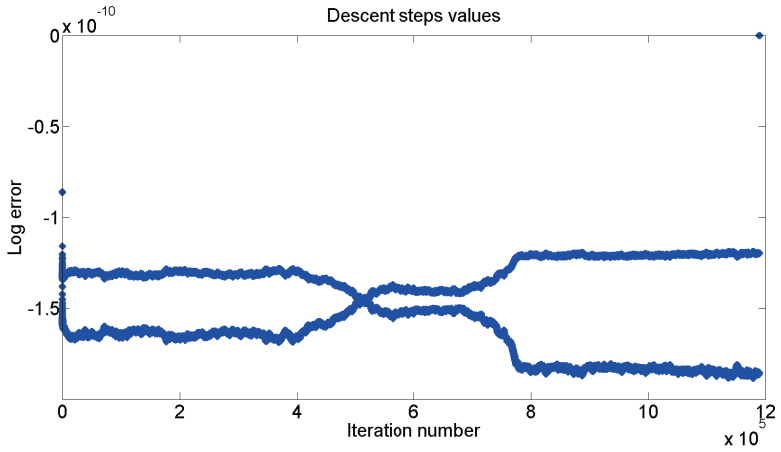


Figure 6.3: Descent steps

We can see the optimal descent step values chosen by the algorithm are very small and draw an almost symmetric pattern. We do not have an explanation for the oscillating behavior.

Now we get interested in the evolution of the values estimated for the vector $\mathbf{c}$. As the final space in which the solution is living is of very high dimensions, we look at the projection on the line of its two first principal components, for the last 500 000 steps of the descent:

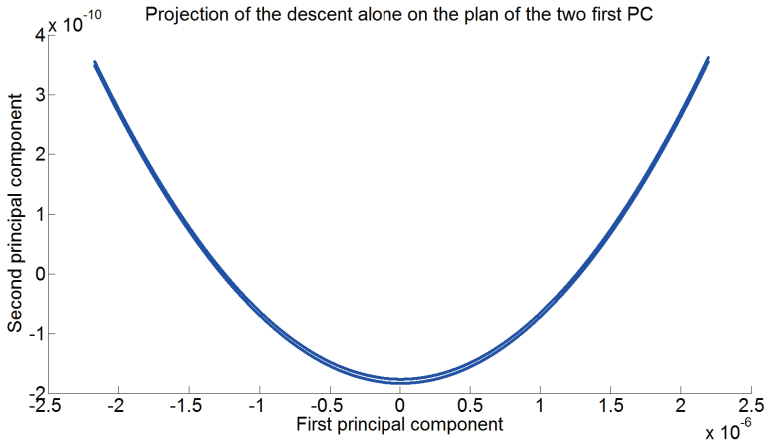


Figure 6.4: Evolution of the vector $\mathbf{c}$ along the plan stem from the two PC

Figure 6.5 represents the evolution of the values estimated for the vector $\mathbf{c}$ projected on the line of its two first principal components, for the last 250 000 steps of the descent:

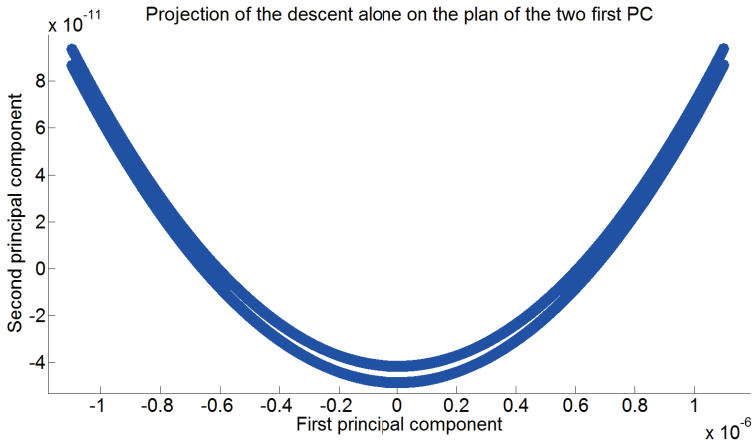


Figure 6.5: Evolution of the vector U

In order to have a better understanding of the algorithm behavior, we want to have a visual representation of the trajectory given by all the obtained values. To do so, we plot the the projection of the estimated value on the three first principal components for $\mathbf{c}$ and $\mathbf{m}$ for the first 250 steps, when the descent is the quickest.

The evolution of the values estimated for the vector $\mathbf{c}$ projected on the space of its three first principal alone is represented in Figure 6.6 while Figure 6.7 adds some values of the cost function chosen randomly.

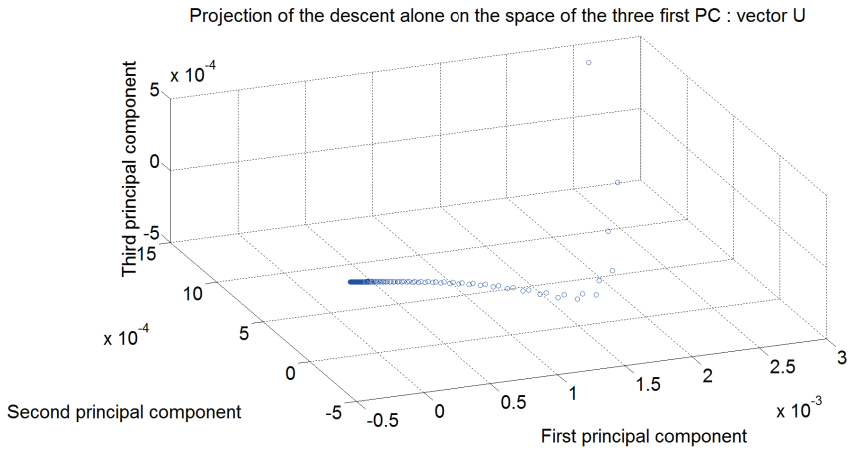


Figure 6.6: Projection of the descent on the space given by the three first principal components

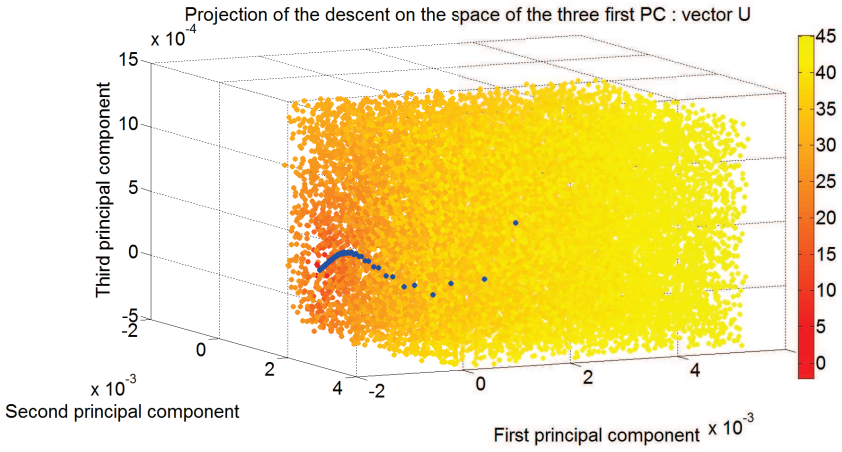


Figure 6.7: Projection of the descent on the projection space with some randomly chosen values of the cost function

The same representation is done for the vector $\mathbf{m}$ in Figures 6.8 and 6.9.

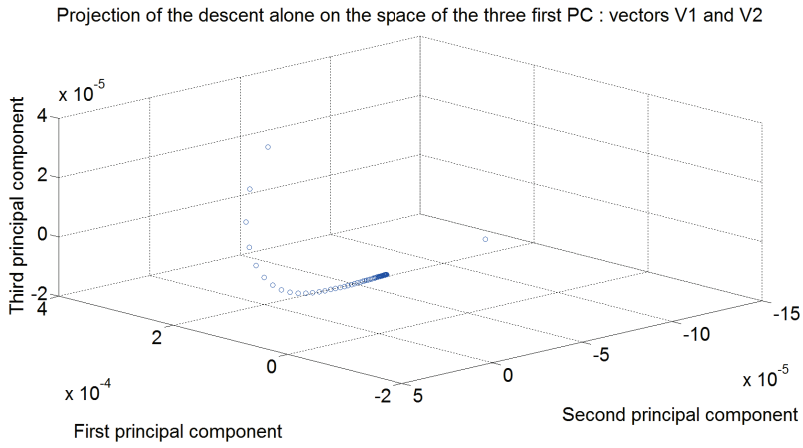


Figure 6.8: Projection of the descent on the space given by the three first principal components

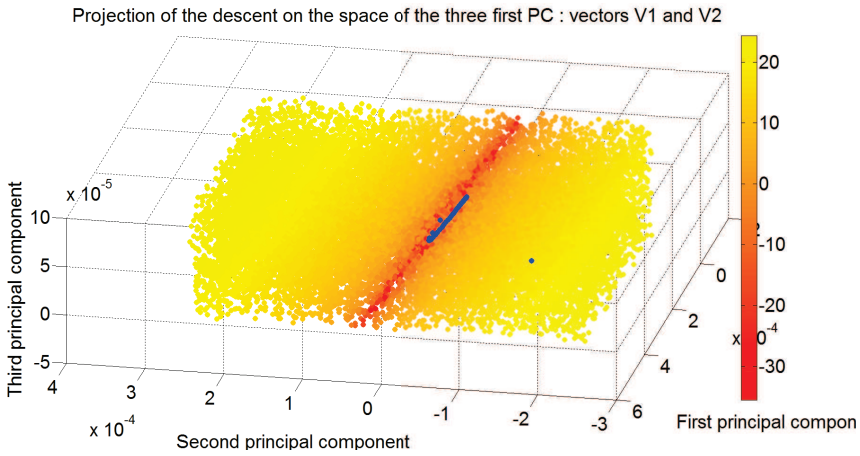


Figure 6.9: Projection of the descent on the projection space with some randomly chosen values of the cost function

For the second vector, we see that the descent values seem to ”live" in a plan more than a 3D volume. A representation in the plan given by the two PC is illustrated by Figure

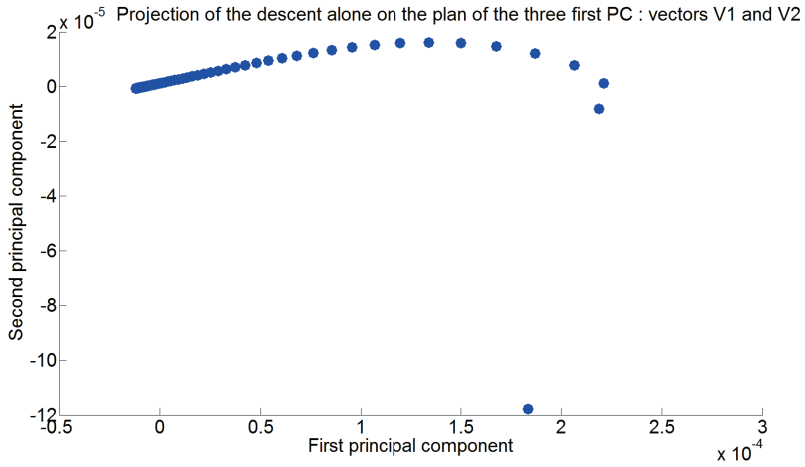


Figure 6.10: Projection of the descent on the plan given by the two first principal components

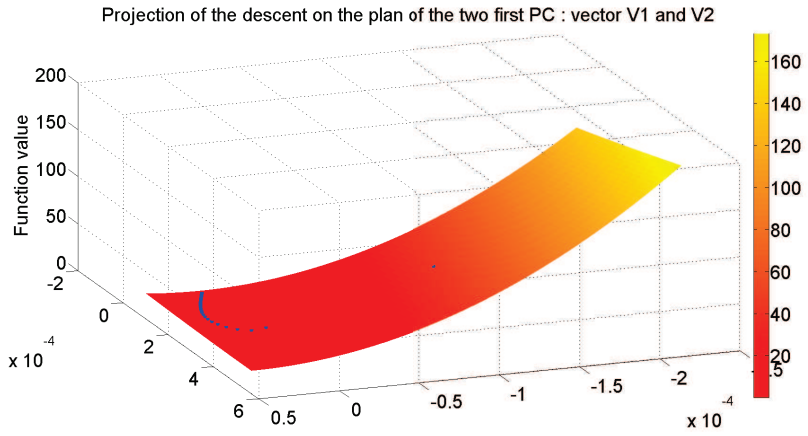


Figure 6.11: Projection of the descent on the projection surface with the values of the cost function

It is clearly visible that the algorithm is going in the direction of the minimum of the space, but when a valley is reach, it slows down a lot.

Simulation 2 For the second simulation, the problem is the same but with a different initialization. The same representations as the ones of simulation 1 have been illustrated. Initial error on the initialization = $1e^{-3}$

Evolution of the reconstruction error :

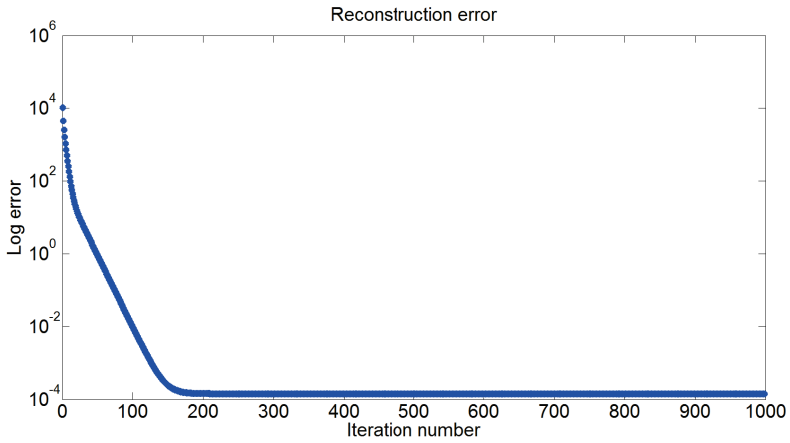


Figure 6.12: Reconstruction error of the estimated vectors

Evolution of the descent step values:

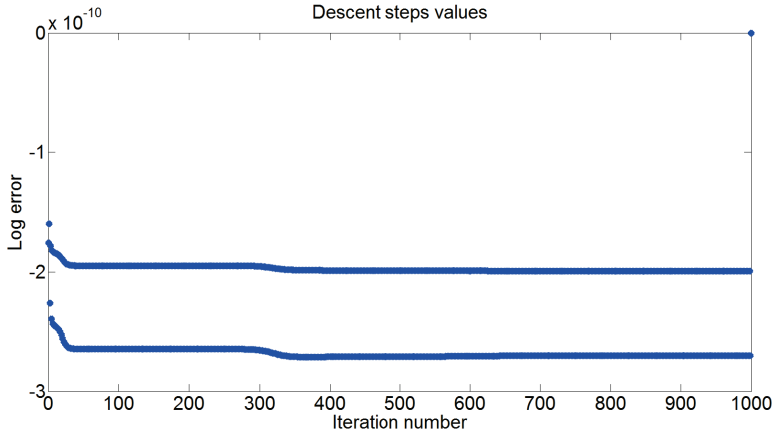


Figure 6.13: Descent steps

Evolution of the values estimated for the vector $\mathbf{c}$ projected on the space of its three first principal components:

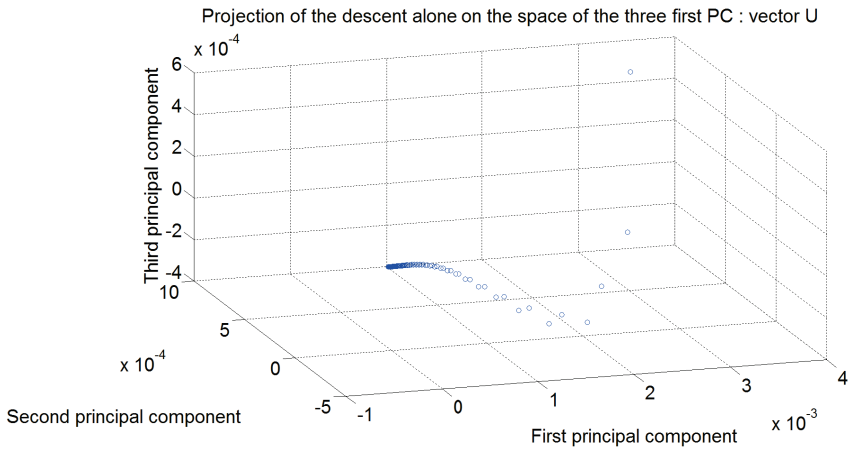


Figure 6.14: Projection of the descent on the space given by the three first principal components

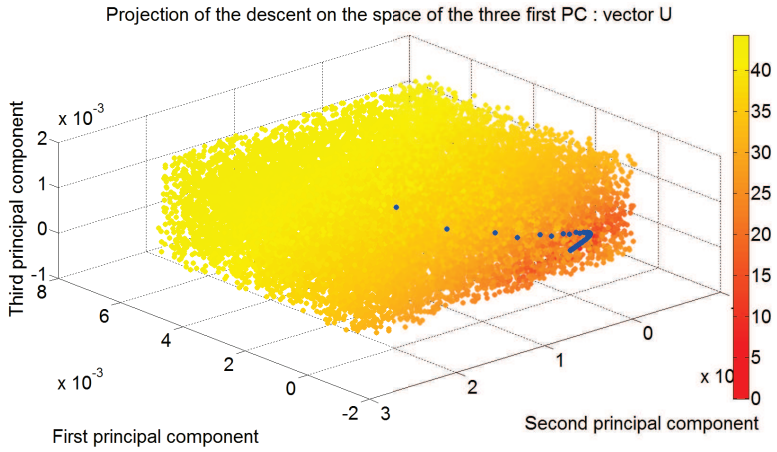


Figure 6.15: Projection of the descent on the projection space with some randomly chosen values of the cost function

Evolution of the values estimated for the vector $\mathbf{m}$ projected on the space of its three first principal components:

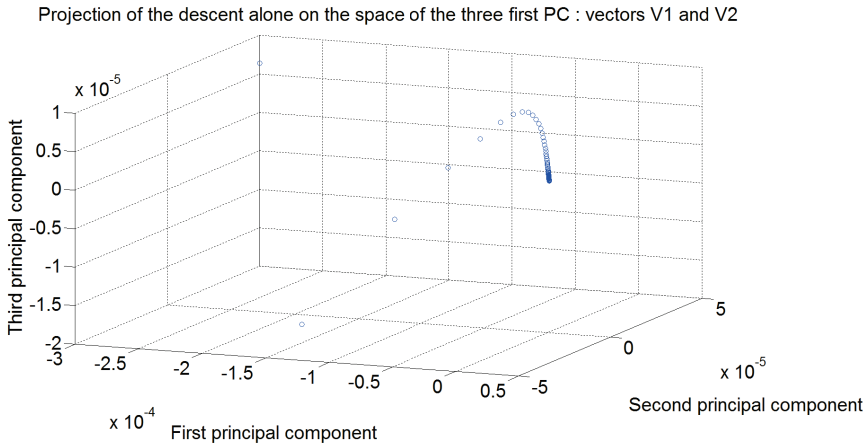


Figure 6.16: Projection of the descent on the space given by the three first principal components

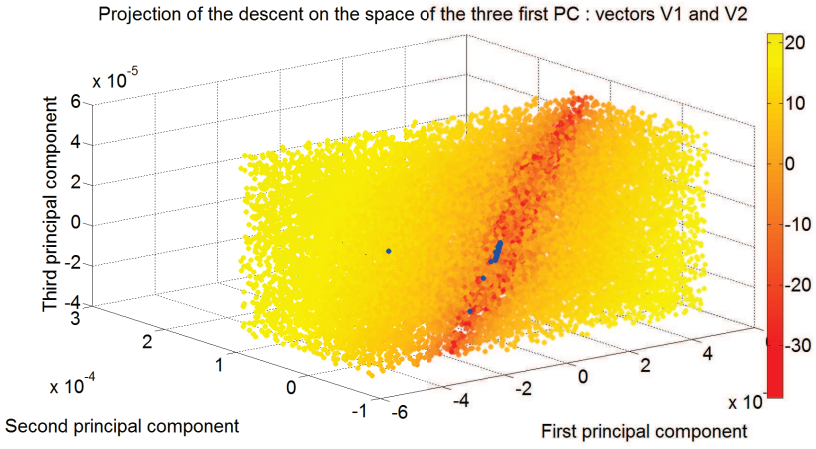


Figure 6.17: Projection of the descent on the projection space with some randomly chosen values of the cost function

Generally speaking, there is no major difference between the results of the two simulations. The obtained parameter vectors, i.e. $\mathbf{c}$ and $\mathbf{m}$ are approximately well estimated. The main concern is that we have not proved that the algorithm always converge.

Alternated descent Algorithm

To solve the optimization problem, in this second algorithm, two steps are considered: one part (vector $\mathbf{c}$) of the product is estimated while the other one is considered constant and then the second part (vectors $\mathbf{m_1}$ and $\mathbf{m_2}$) of the product are estimated while considering the the latter constant.

Algorithm 4. *Given two known matrices $\mathbf{M} \in \mathbb{C}^{n\times p}$ and $\boldsymbol{D} \in \mathbb{C}^{n\times n}$, and three unknown vectors $\mathbf{c} \in \mathbb{C}^{n\times 1}$, $\mathbf{m_1} \in \mathbb{C}^{p\times 1}$ and $\mathbf{m_2} \in \mathbb{C}^{p\times 1}$ where n and p are the dimensions of vectors $\mathbf{c}$ and $\mathbf{m}$ respectively. The following process is repeated until a satisfying error threshold is reach.*

1. *Vector $\mathbf{c}$ estimation with $\mathbf{m_1}$ and $\mathbf{m_2}$ known,*
2. *Update of the reconstruction error,*
3. *Vectors $\mathbf{m_1}$ and $\mathbf{m_2}$ estimation with $\mathbf{c}$ known,*
4. *Update of the reconstruction error,*

and repeat the process until a satisfying reconstruction error is obtained.

We have done the two same simulations (same initialization point and same ground truth vectors) with the alternating descent to compare the results in terms of number of iteration needed to reach the reconstruction error wanted and the speed of the descent.

Simulation 1 Number of iteration needed to reach a relative error of reconstruction of 10^{-5} : 2410.

Evolution of the reconstruction error:

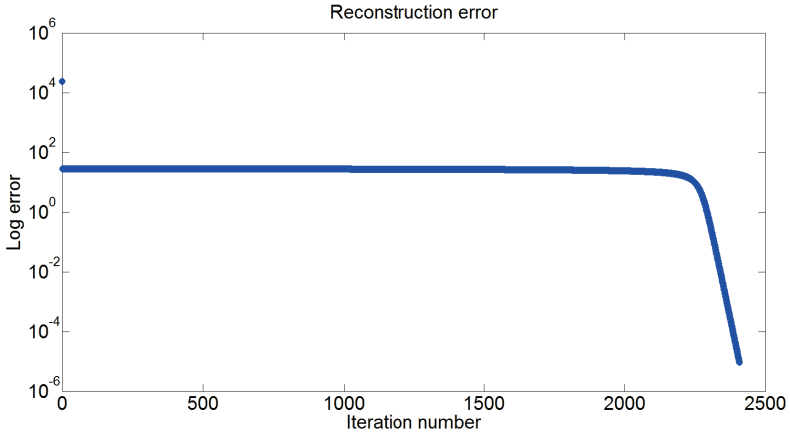


Figure 6.18: Reconstruction error of the estimated vectors

Evolution of the values estimated for the vector $\mathbf{c}$ projected on the space of its three first principal components:

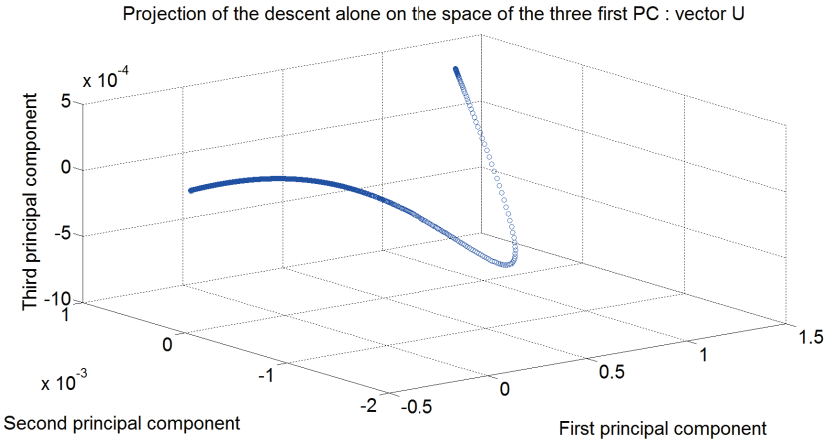


Figure 6.19: Projection of the descent on the space given by the three first principal components

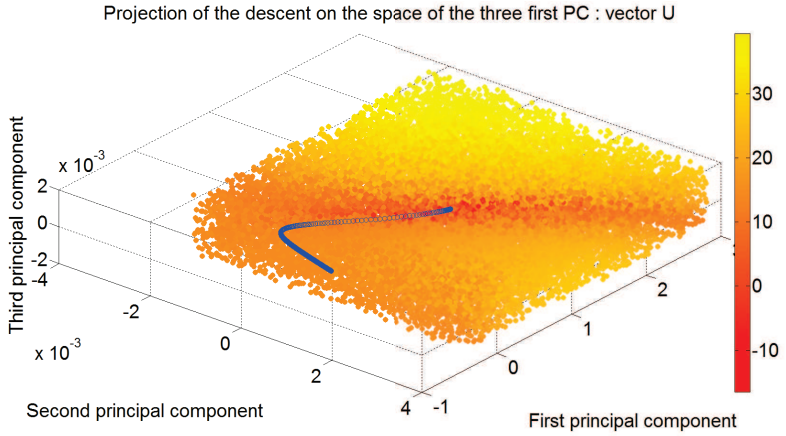


Figure 6.20: Projection of the descent on the projection space with some randomly chosen values of the cost function

Evolution of the values estimated for the vector $\mathbf{m}$ projected on the plan of its two first principal components:

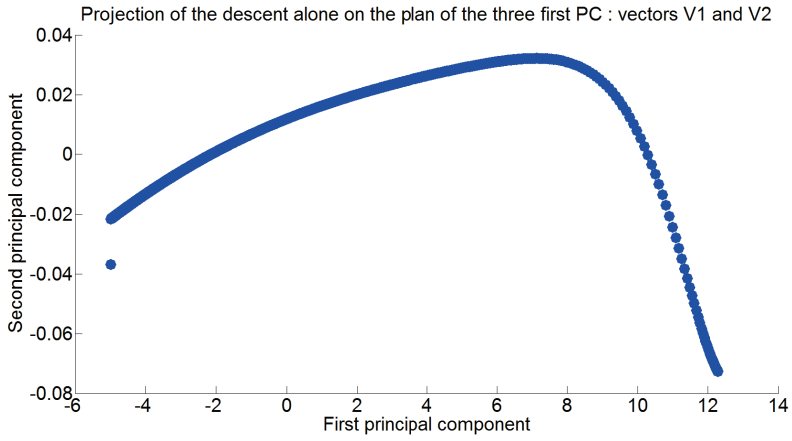


Figure 6.21: Projection of the descent on the plan given by the two first principal components

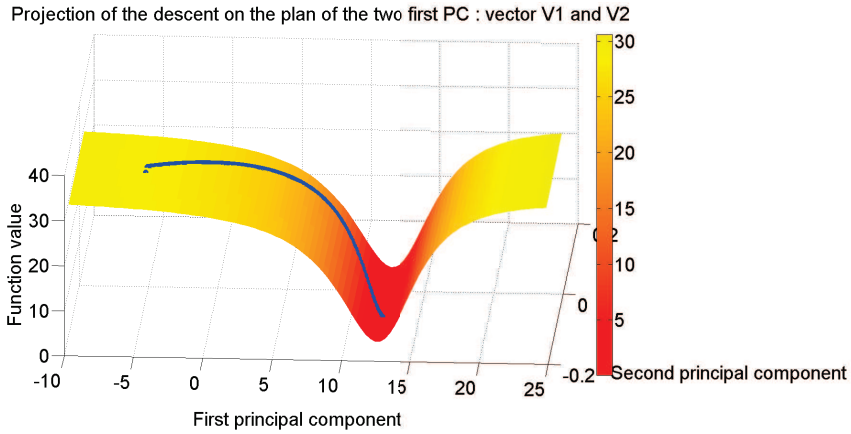


Figure 6.22: Projection of the descent on the projection surface with the values of the cost function

Simulation 2 Number of iteration needed to reach a relative error of reconstruction of 10^{-5} : 217.

Evolution of the reconstruction error:

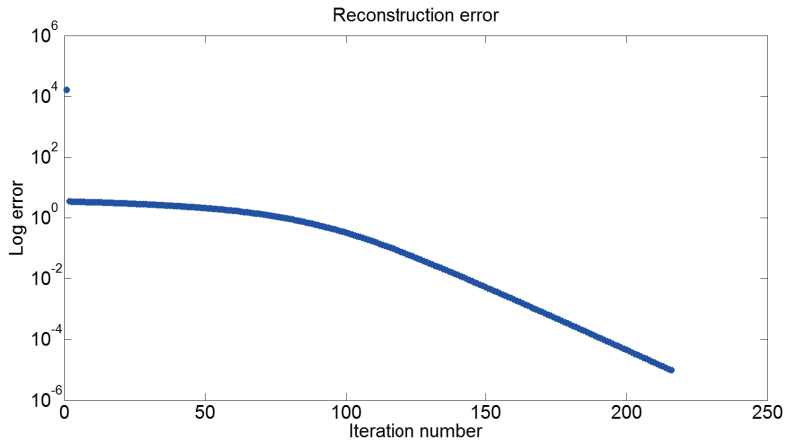


Figure 6.23: Reconstruction error of the estimated vectors

Evolution of the values estimated for the vector $\mathbf{c}$ projected on the space of its three first principal components:

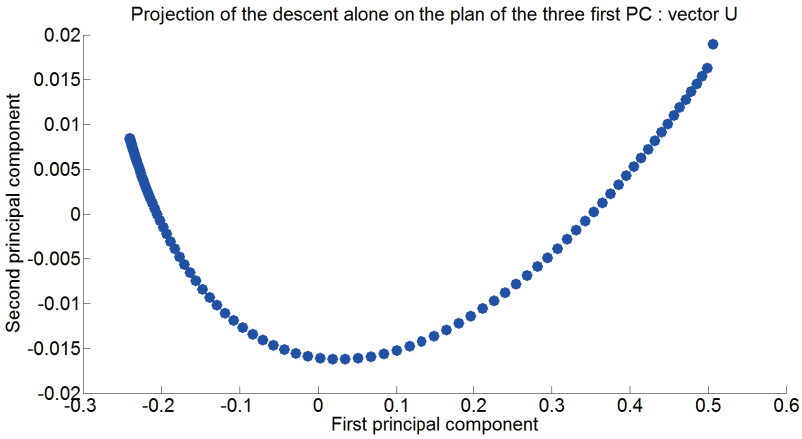


Figure 6.24: Projection of the descent on the plan given by the two first principal components

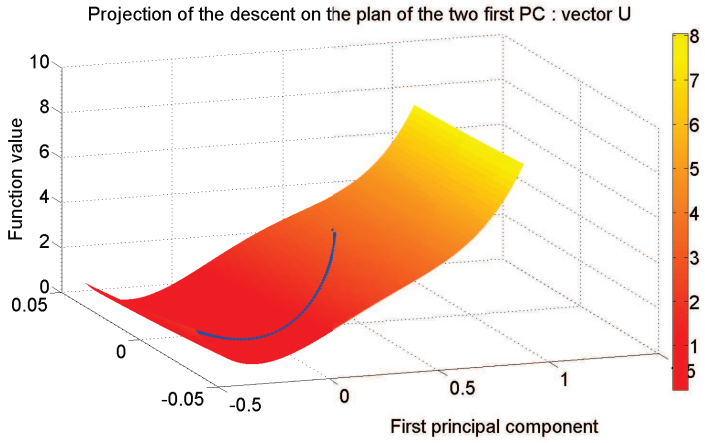


Figure 6.25: Projection of the descent on the projection surface with the values of the cost function

Evolution of the values estimated for the vector $\mathbf{m}$ projected on the plan of its two first principal components:

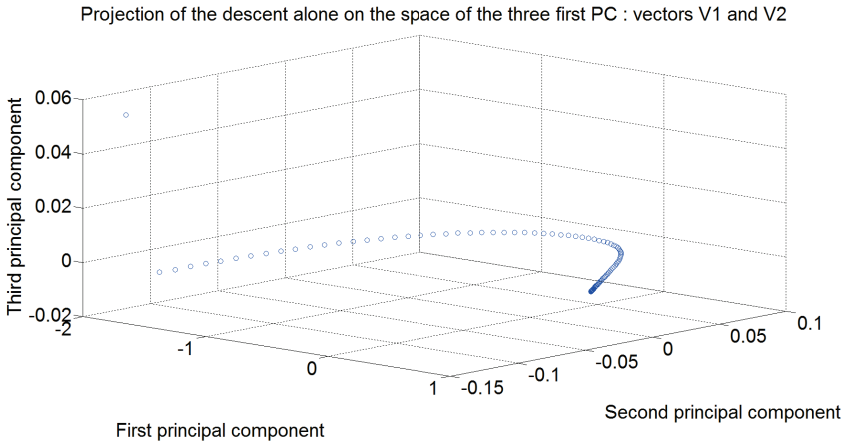


Figure 6.26: Projection of the descent on the space given by the three first principal components

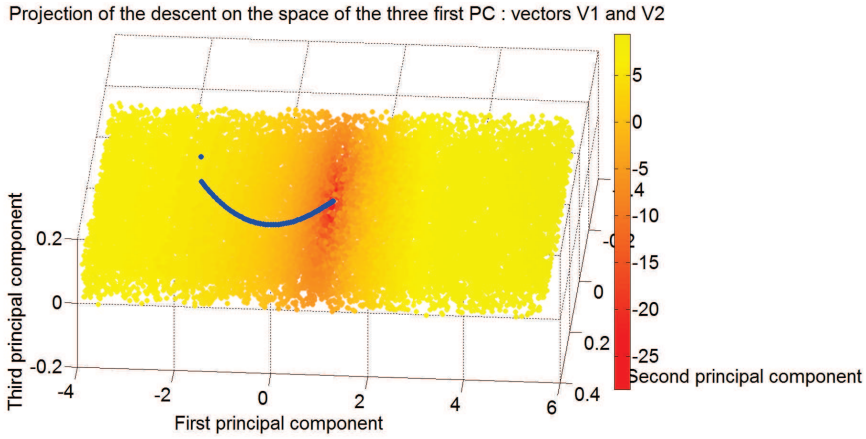


Figure 6.27: Projection of the descent on the projection space with some randomly chosen values of the cost function

The alternated descent algorithm has shown interesting properties regarding the convergence. Indeed, for a given threshold, the estimation error has always reach a smaller value.

6.4 Conclusion

Several points were addressed in this chapter. First of all, the class of both amplitude and phase modulated signals has been cast into the same matrix framework as amplitude modulated signals only. This reformulation allowed us to express phase and amplitude demodulation either as an exact matrix decomposition (noiseless case) or as an optimization problem (noisy case).

For the exact case, we were able to prove that a solution could be computed under some necessary and sufficient conditions. For the optimal case, we proposed two different algorithms based on a gradient computation, namely the Line search algorithm and the alternate descent algorithm.

Based on simulations, it seems that the Line search algorithm is faster at the beginning of the descent but then stay stuck into a minimum and do not converge to the solution. However, the alternated descent algorithm is more efficient, as it converges to the solution.

We have to say that, we do not have a precise idea of the structure of the solution-space. We do not know if there is a global minimum, nor even if there is a finite number of local minimums. Those questions are very interesting and will surely be part of a future work.

Chapter 7

The planetary gearbox case

Previous chapter presented demodulation issues raised by the study of fixed-shaft gearing systems. Now we go to a different configuration briefly described in the general introduction on gearboxes of Chapter 1: epicyclic gearing.

These devices are of major importance to aeronautics as they can be found in helicopters and more recently in some aircraft engines. As their functioning is very different and way more complicated than the one of classical gearing systems, both dedicated surveillance systems and signal processing techniques have to be developed. We will see in this chapter that the matrix representation of a spectrum we introduced previously helps analyzing the models currently used to describe vibration signals produced by this family of gearboxes. More precisely, it allows in some cases distinguishing the respective contribution of each planetary wheel to the total signal measured by a static sensor.

7.1 Introduction

7.1.1 Functioning of planetary gearing systems

The specificity of a planetary (or epicyclic) gear lies in its arrangement, which is going to be briefly recalled here.

Remark 34. *For the anecdote, epicyclic gears have been named after their earliest application, invented by the Greek about* $500BC$*, that is the representation of the planets's movements as epicycles, i.e. circles moving in a circular orbit. This theory has been formalized by Ptolemy in the* Almagest *in* $148AD$ *[83].*

Unlike fixed-shaft gearing systems that usually have parallel axes, except for bevels gears that have an angle between input and output shafts, in the case of planetary gears, planets, sun and ring gears are coaxial. This particular arrangement allows systems with reduced overall dimension, which can result in significant space savings. But the major interest is that planetary gearing systems provide both higher power density than comparable parallel axis gear trains and increased torque capability as the load is shared among the multiple planet gears.

In plain terms and as illustrated in Figure 7.1, planetary gearing systems are composed of a central sun gear around which several planet gears rotate. Those gears are identical and equally spaced around the sun gear. Planet gears are also meshing at the same time with the ring gear and are linked together by a carrier plate. In order to

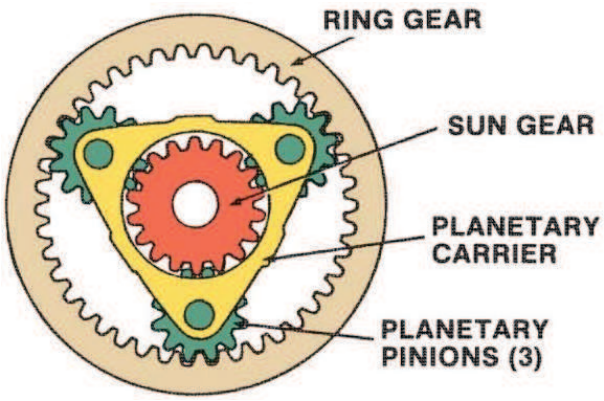


Figure 7.1: Illustration of a planetary/epicyclic gearing system with the name of its components.

be operational, one element has to be fixed. Depending on the chosen element, it is possible to obtain two different mountings:

- *planetary gear* when the carrier plate is fixed,
- *epicyclic gear* when the ring gear is fixed.

7.1.2 Vibration signal of planetary gearing systems

Generally speaking and whatever be the considered type of gearbox, if the tooth profiles of a gear transmission were perfect and the teeth were considered as rigid bodies, then no vibration would be generated during the rotation of the system. But in reality tooth profiles differ a lot from perfection, which generate vibrations.

Moreover we have seen that by construction, each planet gear meshes simultaneously with both the ring and the sun gears, creating two vibration sources per planet gear, and as all of them are meshing at the same time, this kind of gearing system therefore has a multitude of vibration sources emitting all at once.

The vibrations are usually measured in the outer part of the ring gear since it is the closest access to the vibration sources, as represented in Figure 7.2. From the sensor's point of view, the global measured vibration can be seen as a sum of each planet gear contribution. Those are supposed to be approximately identical but delayed due to the spatial shift of planet gear positions.

During the acquisition, the planet gears revolve around the sun gear. This makes the contact points between each planet gear and the ring gear, i.e. the source location of the vibrations, to follow the internal circumference of the ring gear, as illustrated in Figure 7.3.

As the sensor remains steady, the distance between the sensor and the sources of the vibrations is thus variable during the rotation of the system. The variable distance produces a variable transmission path, which turns into an amplitude modulation effect on each planet periodic vibration. Hence, the sensor experiences an increase in the amplitude of the vibration as the i^{th} planet gear approaches the sensor position and a decrease in the amplitude of the vibration, as the planet gear moves away from it. One rotation period of the global vibration signal is illustrated by Figure 7.4. It is possible to see that the global vibration signal is modulated by the passage of each planet gear.

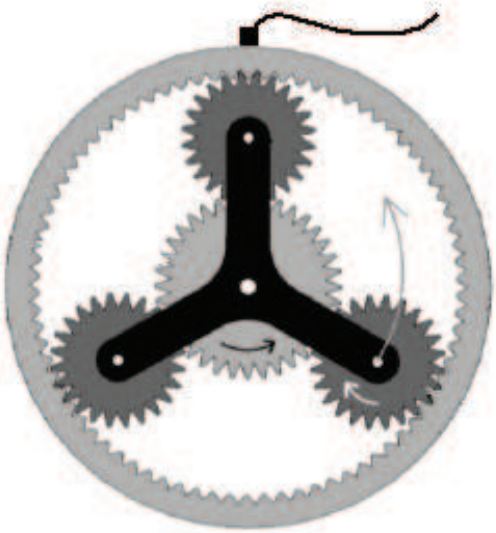


Figure 7.2: Representation of a measurement mounting

The structure of the vibration spectrum of planetary gear has been studied in [60]. The authors have classified vibration spectrum in four groups based on the planet distribution around the sun gear and the phase of gear mesh process.

Remark 35. *Here the phase of the gear mesh process is defined as the ratio of the ring gear's number of tooth and the number of planets. The gear mesh process is stated* in-phase *if the ratio is an integer number and* out-of-phase *if the ratio is not.*

Quickly, the four groups of planetary gear transmissions are separated after those characteristics:

- group A: Planetary gear transmissions with equally-spaced planet gears and in-phase gear mesh processes. The spectrum presents lines at the gear mesh frequency and harmonics. In addition, each of these lines presents a symmetrical distribution of spectral lines.
- group B: Planetary gear transmissions with equally-spaced planet gears and out-of-phase gear mesh processes. The spectrum presents no line at the gear mesh frequency and harmonics. Additionally, an asymmetrical (in magnitude and frequency) distribution of spectral lines is observed around the frequency axis defined by the gear mesh frequency and its harmonics.
- group C: Planetary gear transmissions with unequally-spaced planet gears and in-phase gear mesh processes. The spectrum presents non-zero magnitude lines at the gear mesh frequency and harmonics. A symmetrical distribution of spectral lines is present around the gear mesh frequency and its harmonics.
- group D: Planetary gear transmissions with unequally-spaced planet gears and out-of-phase gear mesh processes. There is no typical vibration spectrum structure. It can be stated that in this group transmissions will present a vibration spectrum with non-zero magnitude lines at the gear mesh frequency and harmonics, each with a magnitude-asymmetrical distribution of spectral lines.

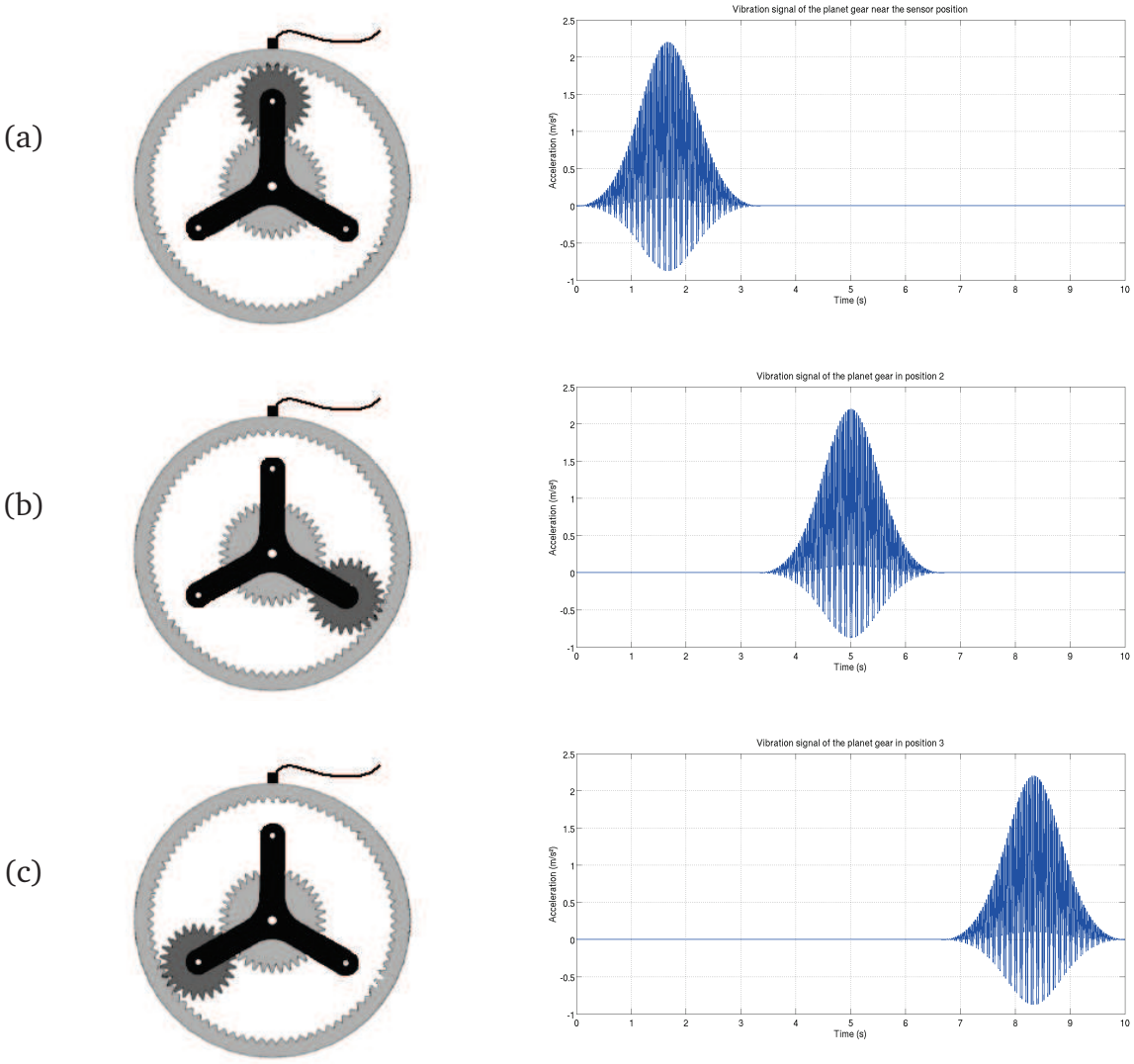


Figure 7.3: Illustration of the rotation of a planet around the sun gear and the vibration generated relatively to the sensor position. (a),(b),(c): Relative position of the planet gear and associated vibration signal in time domain

Faulty gear characteristics

As planetary gearboxes are used in particularly difficult conditions such as high load, high rotation speed or extreme functioning conditions, they are also quickly susceptible to wear and failures. The appearance of faults in the gear changes its kinematic and brings new rotation frequencies, specific to the fault [74]. For an epicyclic gearbox with N satellites, the values of these possible frequencies are:

where f'_c, f'_p and f'_s are the fault frequencies and of the carrier plate, the planetary gears and the sun gear respectively and W_r, W_p, W_s are the number of teeth of the ring gear, planet gear and sun gear respectively.

Basically, epicyclic gear faults are similar to the ones of fixed-shaft gear and so are the fault detection techniques: the main principle remains identical, i.e. use of spectral or cepstral analysis, statistical indicators . . . , but some tricks are found in order to adapt them to the specific characteristics of epicyclic gearing.

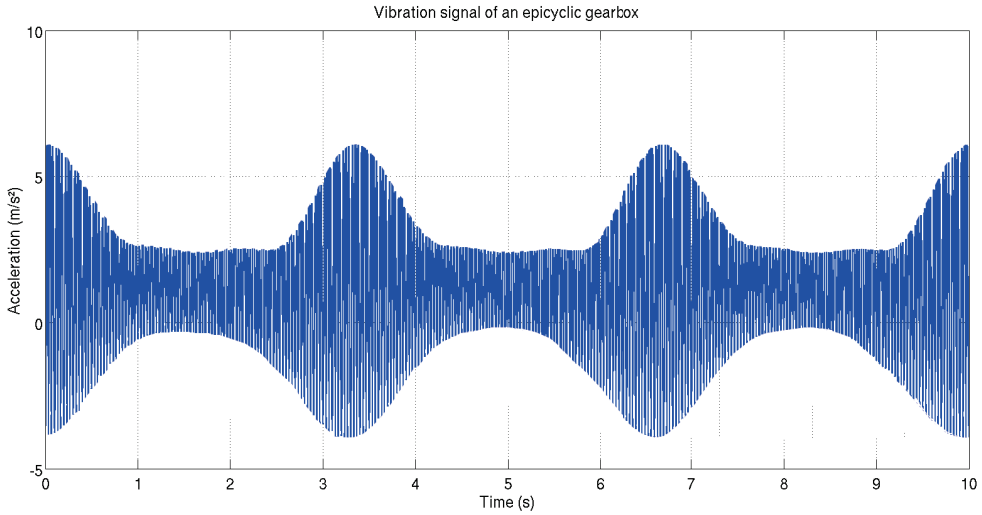


Figure 7.4: Illustration of the global vibration signal generated by a planetary gear made of 3 planet gears.

7.2 Modeling

Recently, planetary gearing systems rose an increasing interest as many industrial systems are equipped with it, such as wind turbines, helicopter engines, bucket wheels excavators... which are critical equipment. Monitoring them properly requires having good understanding of their functioning. That is why research have been made to propose representative modeling of planetary gear behavior. Plenty of models can be found in the recent literature which, as fixed-shaft gearing systems, can be of several types: lumped parameters, finite elements and mathematical modelings. In this section, some references on these representations are briefly given and a new empirical approach is proposed.

7.2.1 Vibration models in the literature

Planetary gear dynamics and vibration has been studied intensively by means of lumped-parameter models. Figure 7.5 shows a typical 2-D lumped parameter modeling of a planetary gear set introduced in [46]. Each gear has three degrees of freedom: angular rotation and transverse motions in the x- and y-directions. Each gear mesh interface and each bearing are modeled as a spring-damper system [33]. Other models have been proposed previously to simulate dynamic vibrations. In [48] the proposed model admits three planar degrees of freedom for each of the sun, ring, carrier and planets. It includes key factors affecting planetary gear vibration such as gyroscopic effects and time-varying stiffness, while in [36] the model includes several manufacturing errors and assembly variations, and can accommodate tooth separations and time-varying gear mesh stiffness.

Finite Element Modeling for spur gear has been extended to planetary gear to investigate its dynamic response [69]. This finite element/contact mechanics approach did not require a highly refined mesh at the contacting tooth surfaces. That model has then been used as a basis for further studies. For example, it has been used for the analysis of quasi-static loads [47] and the root stresses [72] in planetary gears. Furthermore the ef-

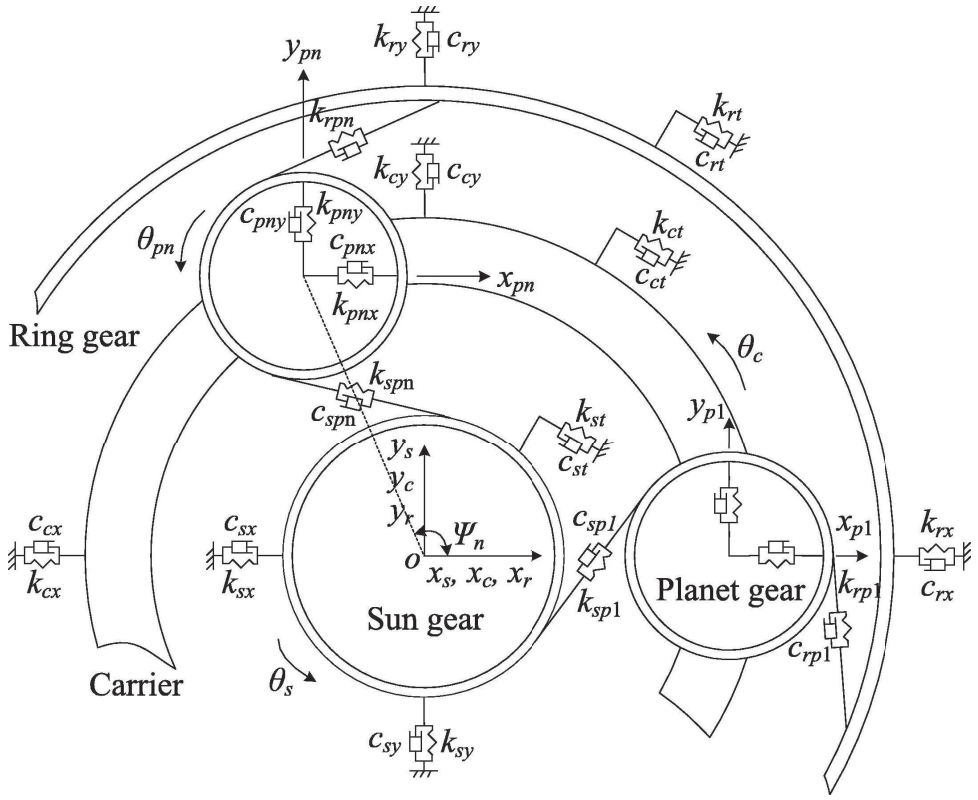


Figure 7.5: Dynamic modeling of a planetary gearbox using a 2-D lumped parameter model.

fect of manufacturing errors [15] and wear [98] on planetary gear dynamics have been both investigated to enhance fatigue mechanisms. Torsional effects have been deeply investigated in [95] with models including gear mesh stiffness evaluation, crack and spalling modeling and gear dynamic characteristics.

7.2.2 New vibration signal modeling

In the case of fixed-shaft gears, existing literature [52] proposes empirical models mostly based on the shape of spectra obtained on real data. Such work has not really be conducted for epicyclic gears, except in [60] where a sum of multiplicative signals is proposed (one for each gear), encoding the increase and decrease of the intensity of the contribution of each gear depending on its (varying) distance to the sensor. This model reads:

$$s(t_n) = \sum_{p=0}^{P-1} s_{mesh,p}(t_n) s_{planet,p}(t_n), \tag{7.1}$$

with $s(t_n)$ the signal measured by the sensor, P the number of planet gears, $s_{mesh,p}$ the signal produced by the meshing of gear p and $s_{planet,p}$ the amplitude modulation due to the distance between the meshing of gear p and the sensor. This model was used in [60] to provide a convincing explanation of a counter-intuitive phenomenon observed on real data : the meshing frequency is absent from the spectrum.

But looking at the peaks in spectra obtain from real data we see that some harmonics of significant intensity are not explained by the model of Eq. (7.1). Thus, we investigated possible improvements ensuring the model predicts all frequencies observed in practice.

Remark 36. *It has to be noted that we are interested in the specific case of epicyclic gear - and not planetary - as the distance between the sensor and meshing point varying over time is pivotal to both the model of Eq.* (7.1) *and the one we propose in the remaining of this section.*

Building on the empirical approach leading to the classical model of fixed-shaft gear vibration we first identify what combinations of the rotation frequencies of interest seem to be present on the spectra obtained from real data then propose the simplest possible model compatible with the presence of these harmonics.

Remark 37. *The task is actually more complicated than in the case of fixed-shaft gears as the superposition of 5 contributions creates artifacts on the spectrum. Thus, the Fourier analysis has to be completed with an observation of the time signal. In particular, it can be observed in the time domain that a modulation pattern is repeated P times per planet carrier period, with P the number of planets as on Figure 7.4. This pattern, easily interpreted as an intensity variation due to the distance between a meshing and the sensor, is visible in the time domain but completely hidden in the spectrum as all planet gears have the same rotation frequency.*

Using a visual identification of the spectrum's lines, we have been able to pinpoint the contribution of each epicyclic gear rotating element:

- lines at the meshing frequency
- modulations of the sun gear around the meshing frequency
- modulations of the planet gears around the meshing frequency
- carrier plate rotation frequency.

An interesting fact noticed looking deeper into the spectrum typical features is that there are no frequency combinations between the sun and planet gears. This observation can be translated mathematically as a pure additive relation between the sun and planet gears lines, while the combinations between the meshing frequency and both the sun and planet gears, imply a nonlinear relations the simplest model of which is a product.

Based on these observations of the spectrum and temporal signal, we propose a model for a planetary gearbox vibration $s(t_n)$ inspired by the empirical model of fixed-shaft gearbox:

$$s(t_n) = \sum_{p=0}^{P-1} s_{mesh,p}(t_n)\left(1 + s_{carrier,p}(t_n)\right)\left(1 + s_{sun,p}(t_n) + s_{planet,p}(t_n)\right), \tag{7.2}$$

where $s_{mesh,p}(t_n)$ represents the p^{th} meshing component, $s_{carrier,p}(t_n)$ the modulation at the carrier plate frequency, $s_{sun_p}(t_n)$ the sun gear rotation contribution to planet p and $s_{planet_p}(t_n)$ the intrinsic contribution of planet gear p. Each involved temporal signal is periodic, with the following periods:

$$\begin{cases} T_m = W_r T_c, \\ T_p = \left(\frac{W_r}{W_p} - 1\right) T_c, \\ T_s = \left(1 + \frac{W_r}{W_s}\right) T_c, \end{cases}$$

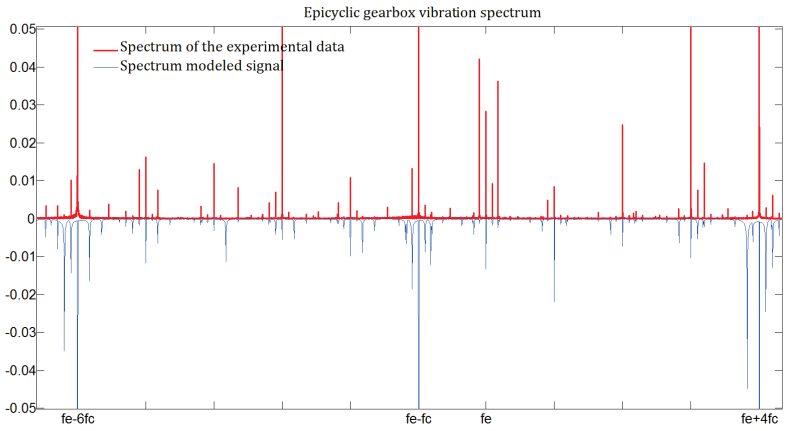


Figure 7.6: Overlay of the vibration spectrum computed from the experimental data and from simulation data using the proposed model.

where T_c is the carrier plate period taken as the reference.

Figure 7.6 illustrates that the proposed model allows to identify every characteristic peak in the vibration spectrum. However the magnitude are not estimated properly.

An immediate application of this model is predicting what peaks should be affected by a fault on the carrier, on the sun gear or on a planet gear. But this work has been jeopardized by the lack of usable real data. Instead, we provide in the next section a theoretical analysis, based on the matrix representation of spectra we introduced in Chapter 4, of the possibility to separate the contributions of the planetary gears.

7.3 Planet separation

When studying fault detection in planetary gearbox, a crucial topic is the localization and identification of the faulty gear. We have seen in the previous section that a planetary gearbox is made of three types of gears, each of them impacting specific frequencies. This observation allows finding out if a detected fault occurred on the sun gear, on the ring gear or on a planet gear. However, when the fault is on the planet gear, the identification of the faulty one becomes tricky, as we have seen that their individual spectral contribution are overlaid. Moreover, classical fault detection methods are designed for a single meshing signal and cannot be transposed if several of them are combined. For instance, the ration between the meshing harmonic and its side bands is usually considered a good indicator of the health condition of a fixed-shaft gear. There are two ways to address this issue. The first one is designing from scratch new indicators for epicyclic gears. The second one, to be discussed below, is isolating the contribution of each gear.

7.3.1 Formalization of the planet separation problem

The possibility to break down the vibration signal produced by a planetary gearbox will be studied in the simplified framework proposed in [60], where the contribution of each gear is simply a periodic signal having the period of the meshing, as in Eq. (7.1). Let us

recall this model already written above in Eq. (7.1):

$$s(t_n) = \sum_{p=0}^{P-1} s_{mesh,p}(t_n) s_{planet,p}(t_n), \tag{7.3}$$

where $s_{mesh,p}(t_n)$ and $s_{planet,p}(t_n)$, which denote respectively the signal produced by the k^{th} gear and the modulation of this signal encoding the energy loss between the meshing of this gear and the sensor, are two periodic signals with respective frequencies f_m and f_c. While the P meshing signals $s_{mesh,p}$ can be different, the amplitude modulations $s_{planet,p}(t_n)$ are related the physical position of the gear only, and thus are related through a mere time shift:

$$s_{planet,p+1}(t_n) = s_{planet,p}(t_n + T_c/P), \tag{7.4}$$

where P is the number of planet gears and T_c the period of the carrier rotation.

A signal of this kind is displayed on Fig. 7.4. To describe the model in a more wordy manner we know that all planet gears contribute to the global vibration signal although there is not direct way to identify individual contributions. From the point of view of the sensor mounted on the ring gear, the measured signal is a weighted sum of the vibrations generated by each gear, the weight assigned to a gear becoming higher when the gear is close. We obtain the exact separation problem 7 below:

Problem 7 (Exact planet separation)**.** *Given a discrete signal $(s(t_n))_{n\in[\![1,N]\!]}$ (with $N \in \mathbb{N}$) of sampling period T_s and duration $T_{tot} = N \cdot T_s$, given an integer number P stating for the number of planet gears, find P carrier signals $s_{c,p}(t_n)$ and a modulation signal $s_m(t_n)$ verifying for any $n \in [\![1, N]\!]$:*

$$s(t_n) = \sum_{p=0}^{P-1} s_{c,p}(t_n) s_m(t_n + p\frac{T_c}{P}) \tag{7.5}$$

This separation problem recalls the first-order phase and amplitude demodulation of Eq. (6.4) (Chapter 6): a signal has to be broken down into a sum of products verifying a given linear condition. The specificity here, beyond the higher number of components involved, is that the linear condition to be respected by the components is a time shift instead of the second component being the time derivative of the first one.

Remark 38. *Planet gears meshing signals $c_p(t_n)$ are supposed to be identical if all planet gears are interchangeable.*

In order to solve problem 7, we introduce a general methodology based on the matrix representation introduced in Chapter 4.

7.3.2 Matrix formulation of the separation problem

As shown in Chapter 4, the matrix representation of a spectrum provides a comfortable framework for analysis of modulated signals. Let us see how it can help solving Problem 7. First, lets us use it to reformulate the problem:

Proposition 8. *Let $s(t_n)$ be a time signal and M_s the matrix representation of its spectrum for carrier frequency f_c and modulation frequency f_m. Let $s_m, s_{c,0}, \ldots, s_{c,P-1}$ candidate modulation and carrier signals for Problem 7 respecting Hypotheses 1 and 2 and*

$\tilde{s}_{rm}, \tilde{s}_{rc_0}, \tilde{s}_{rc_{P-1}}$ their reduced spectra as defined in the proof of Proposition 3. Then Eq. (7.5) *is equivalent to:*

$$M_s = \sum_{p=0}^{P-1} S^p \tilde{s}_{rm} \tilde{s}_{rc_p}^T,$$

with S the diagonal "time shift" matrix defined by $S[k,k] = \exp(2i\pi kN/P)$ and S^p the power p of the matrix, representing p times the shift represented by S, which could also be explicitly written as $S[k,k]^p = \exp(2i\pi kpN/P)$.

This reformulation of the problem answers the question of the uniqueness of the decomposition in Problem 7 as shown by the proposition below:

Proposition 9. *Let $s_m, s_{c,0}, \ldots, s_{c,P-1}$ be modulation and carrier signals verifying Hypotheses 1 and 2. Denoting by $s(t_n) = \sum_{p=0}^{P-1} s_{c,p}(t_n) s_m(t_n + p\frac{T_{tot}}{P})$ their combination as in Problem 7, then the decomposition of $s(t_n)$ is not unique and the space of solutions regarding u is a real vector space of dimension smaller than P.*

Proof. Vectors u, v_p and matrix S being centro-symmetric, all matrices $S^p u v_p^T$ are centro-symmetric operators of rank 1 (as an $\mathbb{R}$-linear operator). As a consequence, M is a centro-symetric operator of (real) rank at most P, i.e., we have $Rank_{\mathbb{R}}(M) \leqslant P$. But the solution regarding u necessarily lies in $Im_{\mathbb{R}}(M)$, which is a $\mathbb{R}-$vector space of dimension $Rank_{\mathbb{R}}(M)$, which proves the result. □

Usually the bound given by Prop. 9 is accurate, i.e., $Im_{\mathbb{R}}(M)$ does have dimension P. A consequence of this fact is that distinguishing the contributions of P planet gears requires P additional assumptions on the modulating function. note that in the case $P = 1$ we simply recover the classical result that amplitude demodulation can be performed only up to an unknown factor, which can be circumvent impsing for instance that the mean value of the modulation is 1.

7.3.3 Application to the analysis of the main gear configurations

Epicyclic gears can be classified depending of the number of planetary gears and we will call "configuration" the number of gears and the position they are given on the planetary carrier. As we will focus here on regular configurations, the number of planets is sufficient to characterize the configuration. The initial motivation for this work was monitoring of a 5-planet epicycle gear, but other configurations are also of interest and will also be studied. Three of them will be considered: 2-planet, 3-planet and 5-planet epicyclic gear systems. For all the above listed configurations, Proposition 9 is applied in order to find out what additional information is required to achieve the decomposition. A specific resolution method is then used for each particular case under study.

2-planets gearbox

Problem introduction For a 2-planets epicycle gear, the two planet gears are opposite to each other as illustrated by Figure 7.7.

Based on the general formulation in Eq. 7.1, the temporal vibration signal expressed at the sensor location can be formulated as

$$s(t_n) = s_{mesh,1}(t_n) s_{planet}(t_n) + s_{mesh,2}(t_n) s_{planet}\left(t_n + \frac{T_c}{2}\right),$$

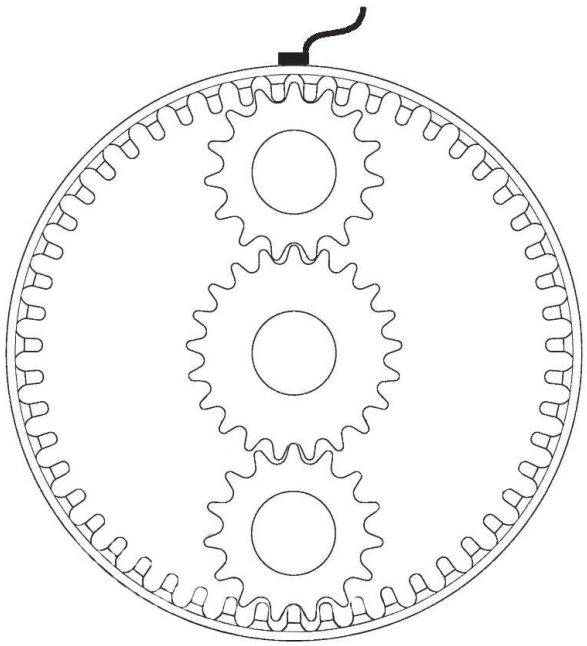


Figure 7.7: General draw of a 2-planets epicyclic gear with equispaced planet gears.

where $s_{planet}(t_n)$ represents the modulation due to the distance to the sensor, $s_{mesh,i}$ is the vibration of the i^{th} planet gear and index i defines each planet gear, i.e. i takes values 1 or 2.

Conditions As it is expressed here, according to Proposition 9 components of the signal $s(t_n)$ cannot be recovered. In order to overcome the problem we propose setting two additional conditions to make the solution unique. The first one, $\mathcal{C}_1$, is the modulation reaching the value 1 when the planetary gear passes by the sensor and the second one, $\mathcal{C}_2$, is the modulation reaching zero when the planetary gear is at the opposite of the sensor. Mathematically, it can be written as:

$$\begin{cases} \mathcal{C}_1 : s_{planet}(0) = 1, \\ \mathcal{C}_2 : s_{planet}(T_c/2) = 0, \end{cases} \tag{7.6}$$

Then, the procedure is quite simple: we parameterize the space of possible solutions by two scalars (see Prop. 9) then write down conditions $\mathcal{C}_1$ and $\mathcal{C}_2$ and obtain two linear equations verified by the parameters. More precisely, let (m_1, m_2) a basis of the space of possible modulation functions allowing reconstructing the signal $s(t_n)$. The true modulation we are looking for takes the shape $s_{planet} = \lambda m_1 + \mu m_2$, which we inject in $\mathcal{C}_1$ and $\mathcal{C}_2$ to obtain the system of equations:

$$\begin{cases} \lambda m_1(0) + \mu m_2(0) = 1 \\ \lambda m_1(T/2) + \mu m_2(T/2) = 0. \end{cases}$$

Solving this system for λ and μ gives the sought modulation s_{planet} and eventually the signals $s_{mesh,1}$ and $s_{mesh,2}$

Simulation A simulation of that system has been done with Matlab where the two planet gears are positioned diametrically opposite the one to the other. The resolution

of the problem is done in two-step: first the estimation algorithm is run on a simulated signal to evaluate its capability to retrieve the modulation and gearmesh of each planet. Then some noise is added to the same simulated signal to evaluate the algorithm reliability to noise, as in real conditions planetary are submerged under the surrounding signals.

In the following simulations we defined the main parameters as, the gearmesh frequency is set to $100Hz$, the modulation frequency is $4Hz$, the sampling frequency is $5kHz$, the number of harmonics of the gearmesh and the modulation are 10 and 8 respectively. In the noisy simulation, a white Gaussian noise is added to the modulated signal, which generates a Signal-to-noise ratio of $5dB$. Figure 7.8 illustrates each generated signal, i.e. the gearmesh, the modulation and the modulated signal while Figure 7.12 gives the same information with in addition the noise signal and the its combination with the modulated signal.

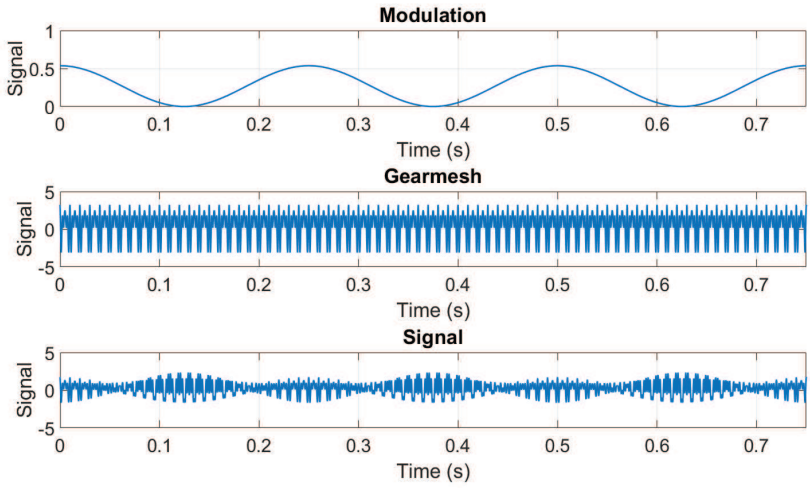


Figure 7.8: Simulation of the temporal signal numerically generated. From top to bottom: Carrier signal, Modulation signal and Global signal set according to Eq.7.2.

Figures 7.9,7.10 and 7.11 below represent the temporal representations of the resulting estimation for each planet gearmesh and modulation. As we can see, there are no differences between the estimated signals and the original ones.

Below, we repeat the same simulation but with an additive white Gaussian noise.

On Figures 7.13, 7.14 and 7.15 it is clearly visible that the estimation is very satisfying: the error is less than 2%.

3-planets gearbox

problem introduction For a 3-planets epicycle gear, planet gears are set such as illustrated in Figure 7.16 and the temporal vibration signal expressed at the sensor location is formulated as

$$s(t_n) = s_{mesh,1}(t_n)s_{planet}(t_n) + s_{mesh,2}(t_n)s_{planet}(t_n + T_c/3) + s_{mesh,3}(t_n)s_{planet}(t_n + 2T_c/3),$$

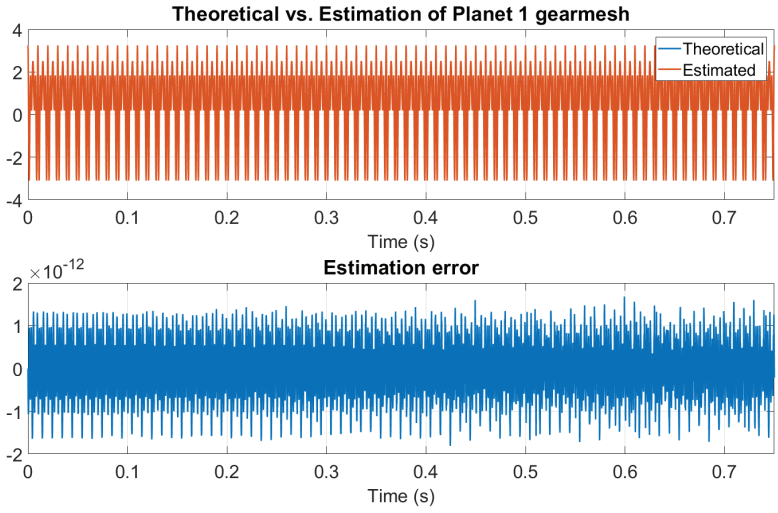


Figure 7.9: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

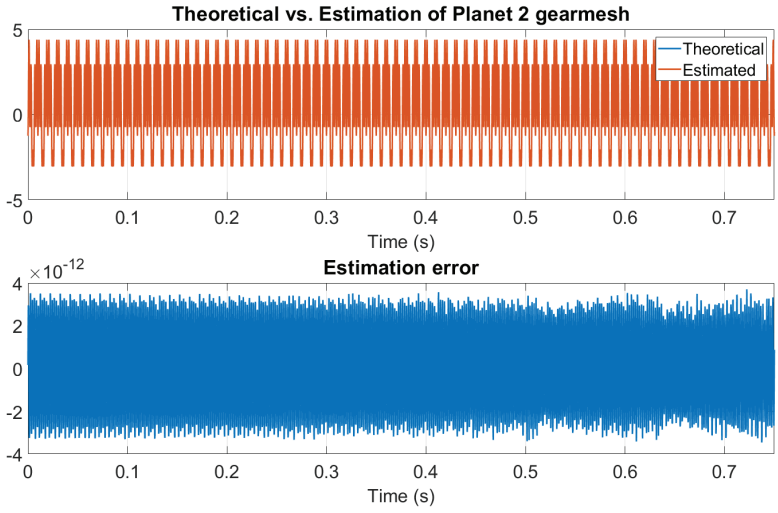


Figure 7.10: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

Conditions As in the previous configuration, some additional information on the amplitude modulation due to the distance to the sensor has to be introduced in order to obtain enough equations. And as there are 3 planets we need 3 conditions. We propose imposing a symmetry condition: the amplitude decay is only due to the distance to the sensor. Mathematically, for each planet gear the conditions can be expressed as:

$$\begin{cases} \mathcal{C}_1 : s_{planet}\left(0\right) = 1 \\ \mathcal{C}_2 : s_{planet}\left(\frac{T_c}{2}\right) = 0 \\ \mathcal{C}_3 : s_{planet}\left(\frac{T_c}{3}\right) = m\left(\frac{2T_c}{3}\right), \end{cases} \tag{7.7}$$

where, we recall, T_c is the period of the carrier rotation. Thus, $\frac{T_c}{2}$ is the instant in the carrier rotation when the meshing point and the sensor are diametrically opposite and $\frac{T_c}{3}$ and $\frac{2T}{3}$ are at one third and two thirds of the ring gear positions with respect to the sensor location respectively.

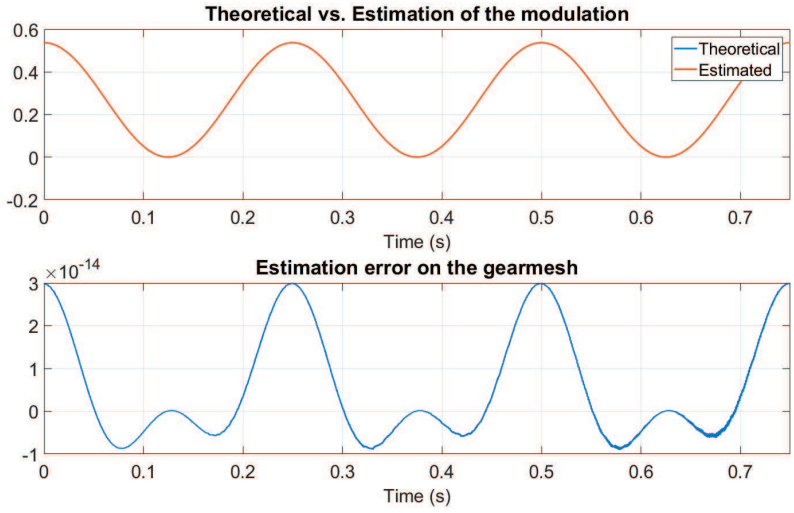


Figure 7.11: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

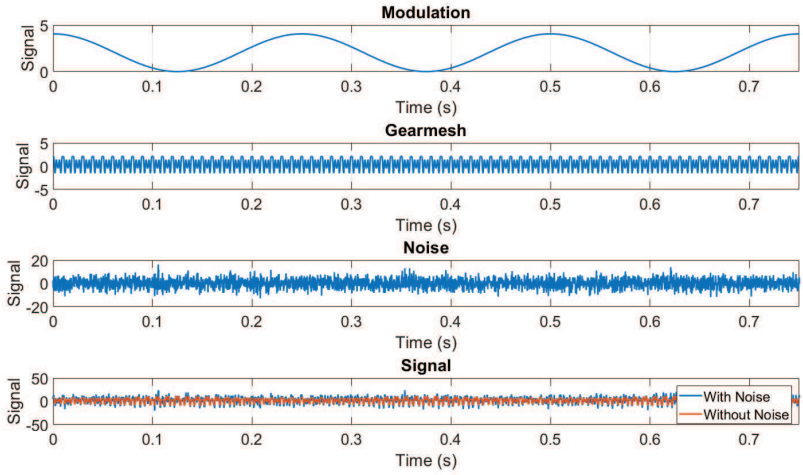


Figure 7.12: Simulation of the temporal signal numerically generated. From top to bottom: Carrier signal, Modulation signal, Noise signal and Overlay of noisy signal set according to Eq.5.1 and the modulated signal only.

The procedure to extract the modulation u from the measured signal is similar to the on used in the 2-planet case: we parameterize the space of possible solutions by two scalars (see Prop. 9) then write down conditions $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}3$ and obtain three linear equations verified by the parameters. More precisely, let (c_1, c_2, c_3) a basis of the space of possible modulation functions allowing reconstructing the signal $s(t_n)$. The true modulation we are looking for takes the shape $s_m = \lambda m_1 + \mu m_2 + \nu m_3$, which we inject into $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ to obtain the system of equations:

$$\begin{cases} \lambda m_1(0) + \mu m_2(0) + \nu m_3(0) = 1 \\ \lambda m_1(T/2) + \mu m_2(T/2) + \nu m_3(T/2) = 0 \\ \lambda m_1(T/3) + \mu m_2(T/3) + \nu m_3(T/3) = \lambda m_1(2T/3) + \mu m_2(2T/3) + \nu m_3(2T/3) = 0 \end{cases} \tag{7.8}$$

Solving this system for λ, μ and ν gives the sought modulation s_c and eventually the

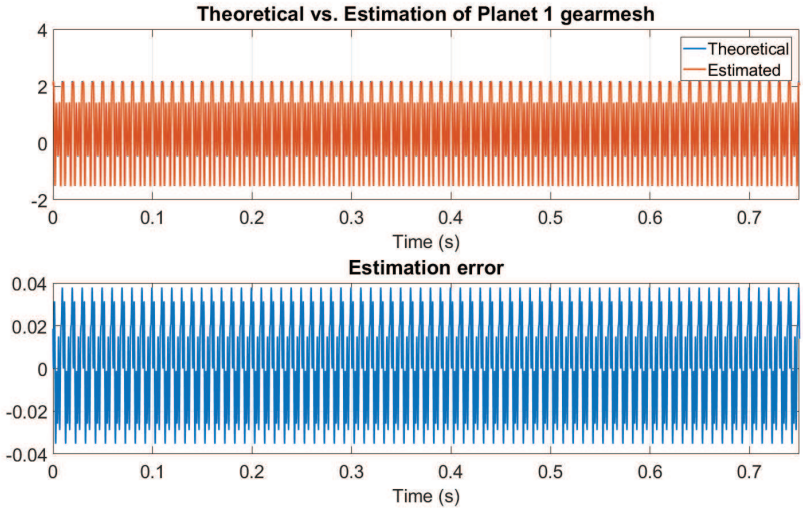


Figure 7.13: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

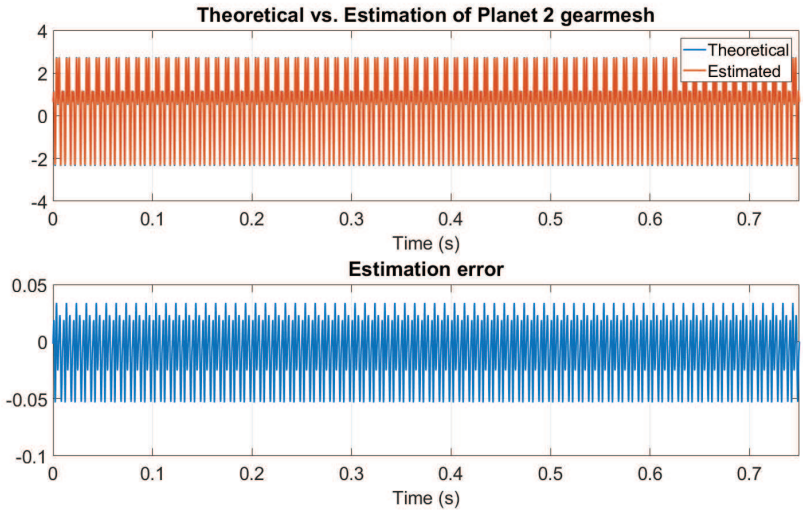


Figure 7.14: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

signals $s_{c,1}, s_{c,2}$ and $s_{c,3}$.

Simulation As in the 2-planet gear epicyclic system, some simulations have been run with exact same parameters and conditions, the only difference being the third planet gear seen as a third couple of modulation/gearmesh in the signal.

Figures 7.187.197.21 below represent the temporal representations of the resulting estimation for each planet gearmesh and modulation. As we can see, there are no differences between the estimated signals and the original ones.

Then we repeat the same simulation but with an additive white Gaussian noise. Figures 7.237.247.207.26 show that the estimation is also very satisfying: even if the error increased a little, it is still less than 3%.

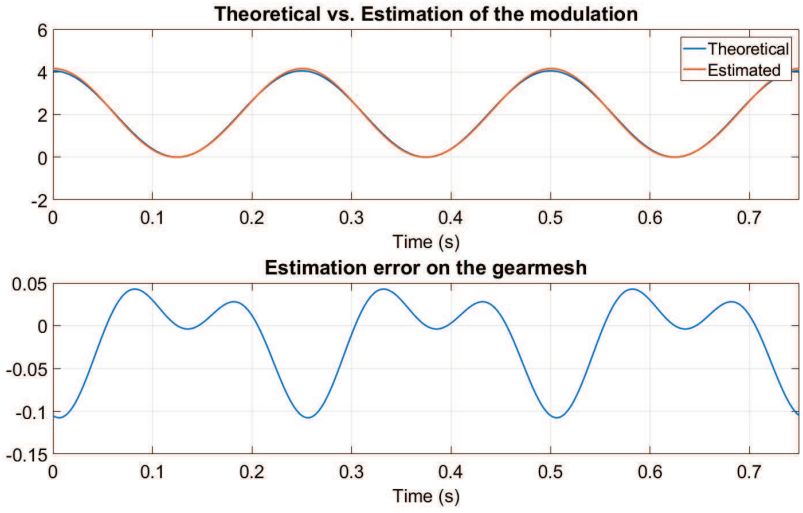


Figure 7.15: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

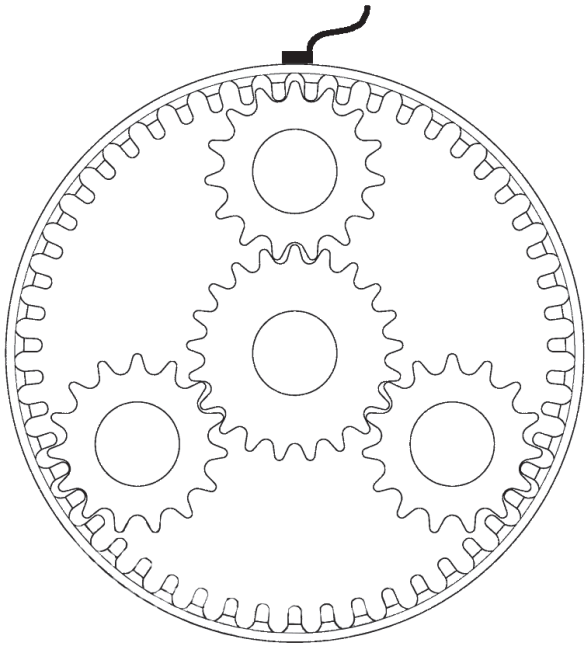


Figure 7.16: General draw of a 3-planets epicyclic gear with equispaced planet gears.

5-planets gearbox

problem introduction For a 5-planets epicycle gear, planet gears are set such as illustrated in Figure 7.16 and the temporal vibration signal expressed at the sensor location is formulated as

$$\begin{aligned} s(t_n) =& s_{mesh,1}(t_n)s_{planet}(t_n) + s_{mesh,2}(t_n)s_{planet}(t_n + T_c/5) + s_{mesh,3}(t_n)s_{planet}(t_n + 2T_c/5) \\ &+ s_{mesh,4}(t_n)s_{planet}(t_n + 3T_c/5) + s_{mesh,5}(t_n)s_{planet}(t_n + 4T_c/5), \end{aligned}$$

Conditions For such a configuration, as there are more signals to estimate from the mixture, more conditions are logically needed than for the previous 3-planet config-

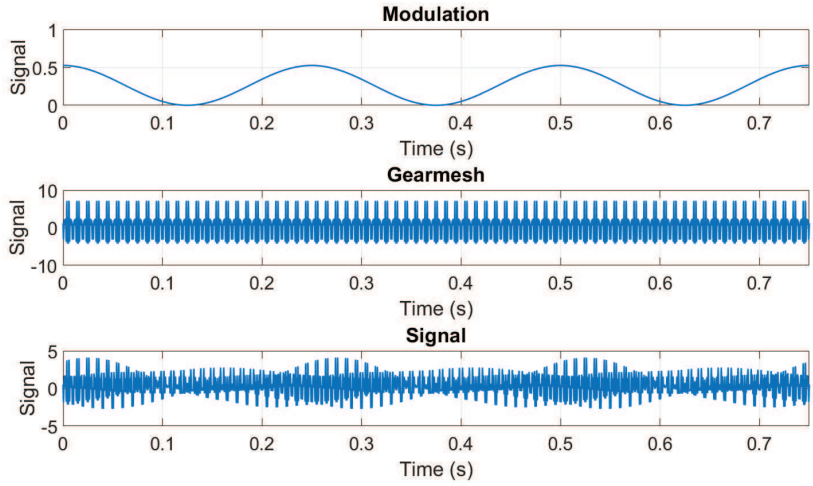


Figure 7.17: Simulation of the temporal signal numerically generated. From top to bottom: Carrier signal, Modulation signal and Global signal set according to Eq.7.2.

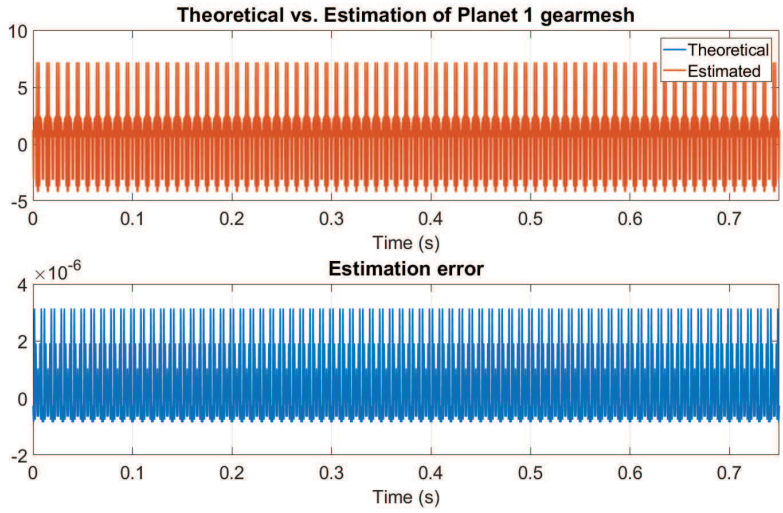


Figure 7.18: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

uration. We can still use the two conditions set for the 2-planet configuration then for the 3-planet configuration, namely: the modulation takes value one at the sensor location and is null at the sensor opposite point. We add two symmetry conditions: $s_n(T_c/5) = s_n(4T_c/5)$ and $s_n(2T_c/5) = s_n(3T_c/5)$, and end up with 4 conditions.

Remark 39. *The symmetry is now encoded by two equations* $s_n(T_c/5) = s_n(4T_c/5)$ *and* $s_n(2T_c/5) = s_n(3T_c/5)$. *This only encodes symmetry between points* $T_c/5$ *and* $4T_c/5$ *on one hand, and between* $2T_c/5$ *and* $3T_c/5$ *on the other hand. So an idea could be to use other points (than* $T_c/5$ *and* $2T_c/T$*) to add as many conditions as we need. Actually it does not work because the conditions we ad that way are redundant. Indeed, our search space for the sequence* $s_{planet}(t_n)$ *contains* $s_{planet}(t_n)$ *and all linear combinations of* $s_{planet}(t_n), s_{planet}(t_n + T_c/5), s_{planet}(t_n + 2T_c/5), s_{planet}(t_n + 3T_c/5), s_{planet}(t_n + 4T_c/5)$ *is symmetrical. If* s_{planet} *is symmetrical, so are all combinations of the form* $\alpha s_{planet}(t_n) + \beta s_{planet}(t_n + T_c/5) + \gamma s_{planet}(t_n + 2T_c/5) + \gamma s_{planet}(t_n + 3T_c/5) + \beta s_{planet}(t_n + 4T_c/5)$. *In*

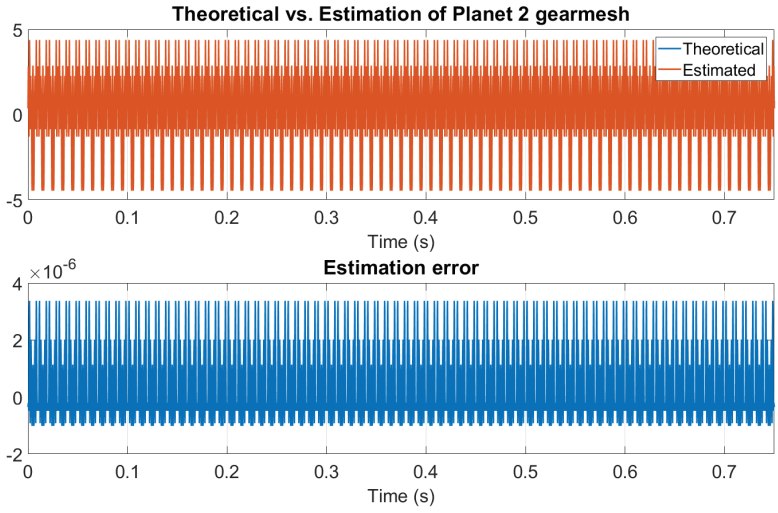


Figure 7.19: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

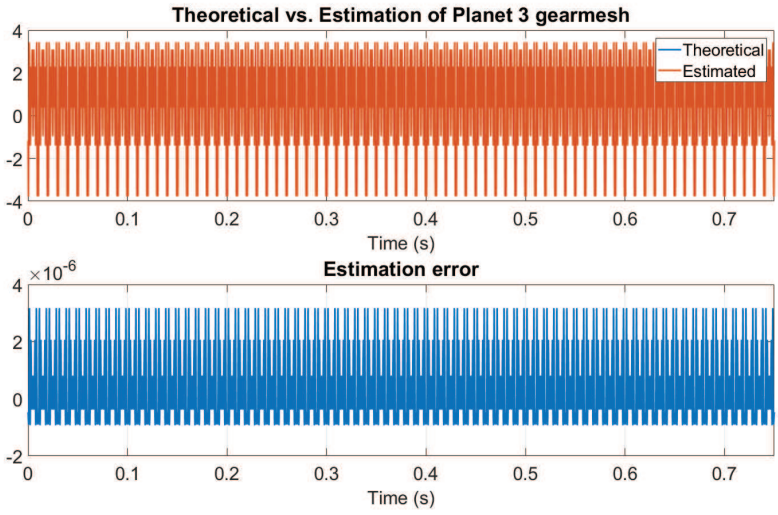


Figure 7.20: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

other words, the space of symmetric candidates for s_{planet} is 3-dimensional. This means that regardless of the number of symmetry conditions we add, the remaining space of candidates still has dimension 3. Thus, once the space of solutions has been reduced from 5 to 4 using 2 symmetry conditions, all the information present in symmetry of the modulation function has already been extracted.

Due to Remark 39 above we don't have enough information to recover the original signals. More precisely, the space of candidate signals has dimension 1. Let us obtain a parametric description of this space of candidates, denoting by an arbitrary real α the value of s_{planet} at time $3T_c/10$ then setting the additional condition $s_m(3T_c/10) = \alpha$. When α describes $\mathbb{R}$, the resolution solution will describe the space of remaining

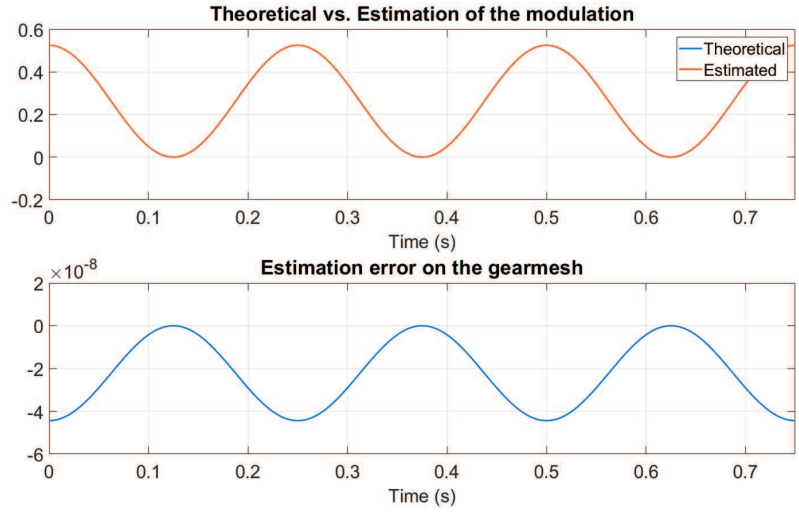


Figure 7.21: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

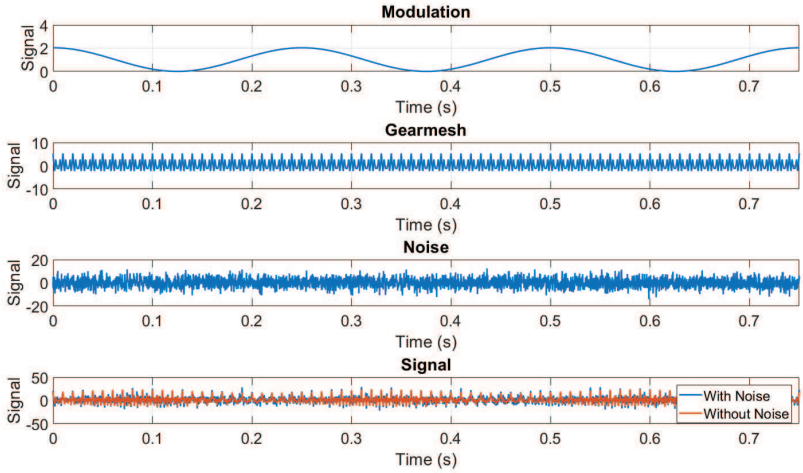


Figure 7.22: Simulation of the temporal signal numerically generated. From top to bottom: Carrier signal, Modulation signal, Noise signal and Overlay of noisy signal set according to Eq.5.1 and the modulated signal only.

candidate signals. In brief, we have now the five conditions we need:

$$\begin{cases} \mathcal{C}_1 : s_{planet}\left(0\right) = 1 \\ \mathcal{C}_2 : s_{planet}\left(\frac{T}{2}\right) = 0 \\ \mathcal{C}_3 : s_{planet}\left(\frac{T}{5}\right) - s_{planet}\left(\frac{4T}{5}\right) = 0 \\ \mathcal{C}_4 : s_{planet}\left(\frac{2T}{5}\right) - s_{planet}\left(\frac{3T}{5}\right) = 0 \\ \mathcal{C}_5 : s_{planet}(3T/10) = \alpha \end{cases} \tag{7.9}$$

Let m_1, m_2, m_3, m_4, m_5 be the time counterparts of a basis of the image of M_s as an $\mathcal{R}$-linear operator: the sought modulation can we written under the shape $s_m = \sum_{i=1}^{5} \lambda_i m_i$. Introducing this expression into the conditions (7.9) we obtain. the following set of

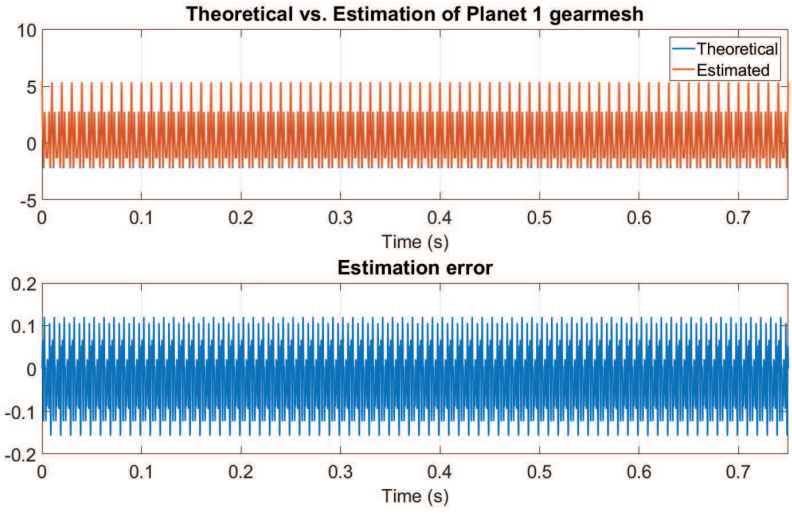


Figure 7.23: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

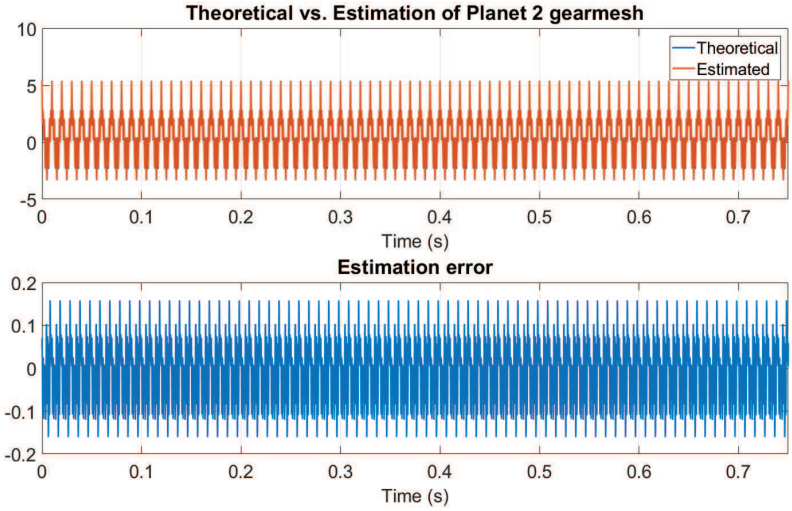


Figure 7.24: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

equations:

$$\sum_{i=1}^{5} \lambda_i m_i\left(0\right) = 1$$

$$\sum_{i=1}^{5} \lambda_i m_i\left(\frac{T}{2}\right) = 0$$

$$\sum_{i=1}^{5} \lambda_i \left[m_i\left(\frac{T}{5}\right) - m_i\left(\frac{4T}{5}\right)\right] = 0$$

$$\sum_{i=1}^{5} \lambda_i \left[m_i\left(\frac{2T}{5}\right) - m_i\left(\frac{3T}{5}\right)\right] = 0$$

$$\sum_{i=1}^{5} \lambda_i m_i(3T/10) = \alpha$$

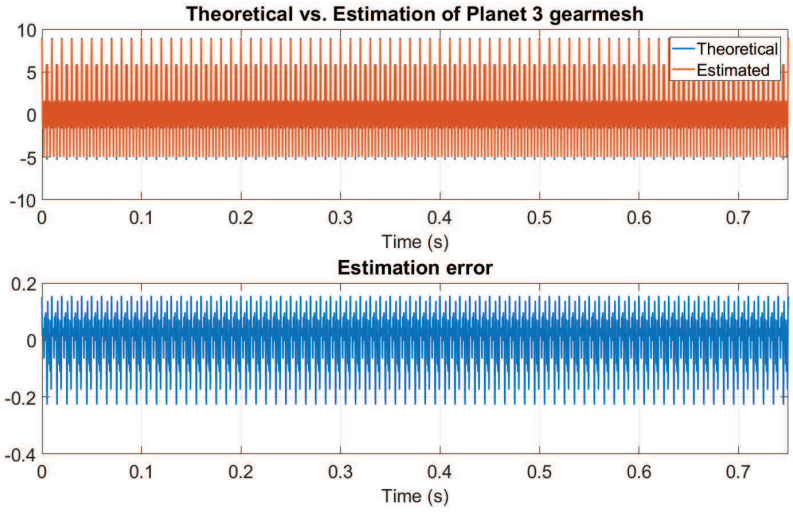


Figure 7.25: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

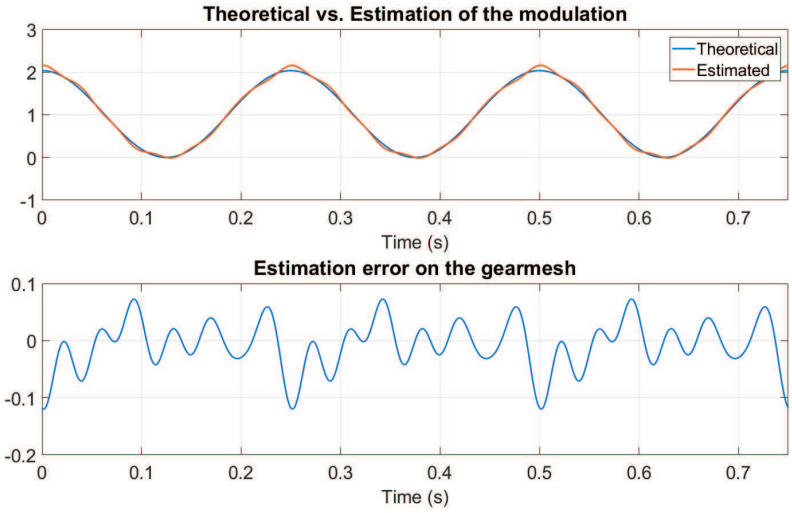


Figure 7.26: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

Using the vector notation $\Lambda = [\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5]^T$, the set of conditions above can be re-written under the following shape:

$$C\Lambda = [1000\alpha]^T$$
$$\Leftrightarrow \Lambda = C^{-1}\,[1000\alpha]^T\,.$$

We noticed that the value of parameter α has a great influence on the estimation of the signals: when the chosen value moves off its true value, the shape of estimated signals changes a lot. An idea to find the true value of α is imposing that the value of the modulation decreases when the distance to the sensor increases. The time value of the modulation can be obtained as $B\Lambda$, with $B = [m_1^T, m_2^T, m_3^T, m_4^T, m_5^T]$ the matrix the columns of which are our basis of the space of solutions. Denoting by D the $(N-1)\times N$ the difference operator returning the time difference between 0 and $T_c/2$ and its opposite on the rest of the vector (i.e. returning the vector of all variables we want to be positive)

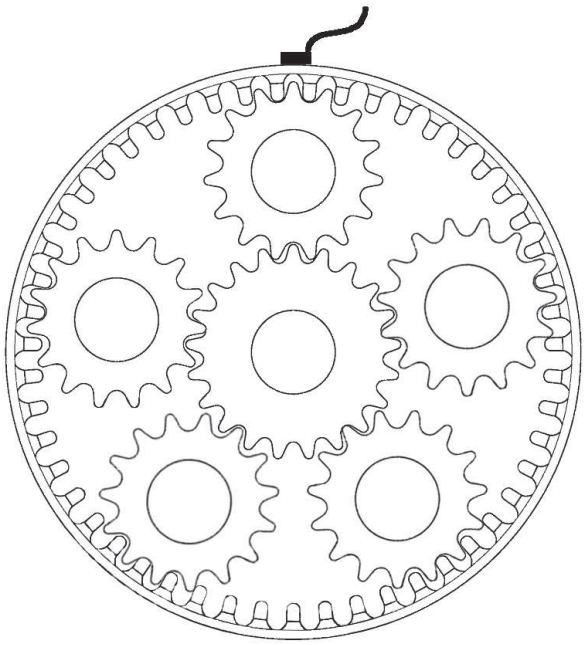


Figure 7.27: General draw of a 5-planets epicyclic gear with equispaced planet gears.

we end up with the following condition:

$$DB\Lambda \geqslant 0$$

or equivalently:

$$DBC^{-1}\left[1000\alpha\right]^T \geqslant 0$$

Such a condition takes the form of an inequality, not an an quality, and thus only allows deriving an interval where the parameter α lives. An estimate of the signal can then be obtained taking for instance the middle of the interval.

Simulation As for the two previous study-cases, the same simulations have been run.

On Figure 7.29 and 7.30 we can see that the estimation is not as good as the 2-planets nor the 3-planets systems, but knowing how complex is such as gearbox, the results are quite satisfying.

However, unlike the two previous simulations, this time there is an issue regarding the reproducibility of the estimation. Indeed for some reason, sometimes the algorithm does not estimate the proper signal but its opposite. We presently do not have any appropriate explanation for this issue, but we guess that it is due to the fact that we could not define enough conditions on the signal.

On the noisy simulation, we observe the exact same phenomenon.

We observe in the noisy simulation that when the sign is correct the estimation is satisfying.

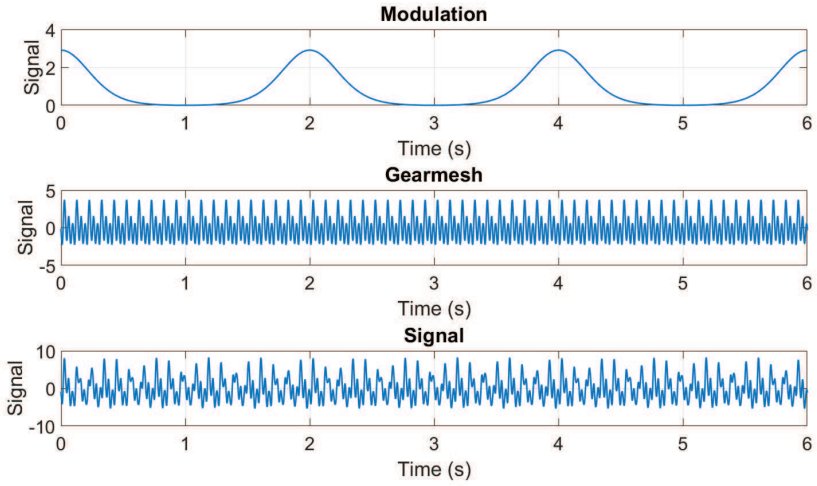


Figure 7.28: Simulation of the temporal signal numerically generated. From top to bottom: Carrier signal, Modulation signal and Global signal set according to Eq.7.2.

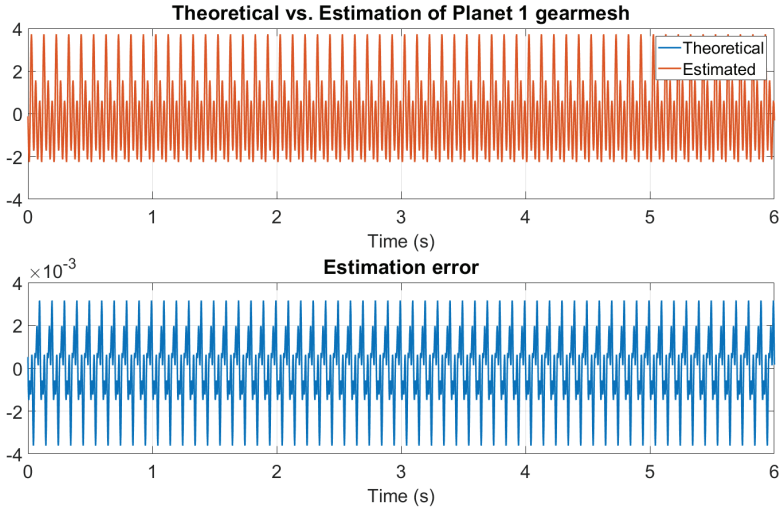


Figure 7.29: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

7.4 Discussion on the applicability to more complex models

7.5 Conclusion

In this Chapter we have presented the epicyclic gearing system. We have first recalled generalities about its functioning and especially detailed information regarding the mechanism that produce vibrations. Some models found in the literature have been briefly presented and a new empirical modeling have been proposed in order to define vibration signal originating from epicyclic gearing in a similar way as fixed-shaft gearbox. Based on this new modeling and on the previous multi-carrier amplitude demodulation, we adapted the demodulation to epicyclic gearing vibration in order to estimated

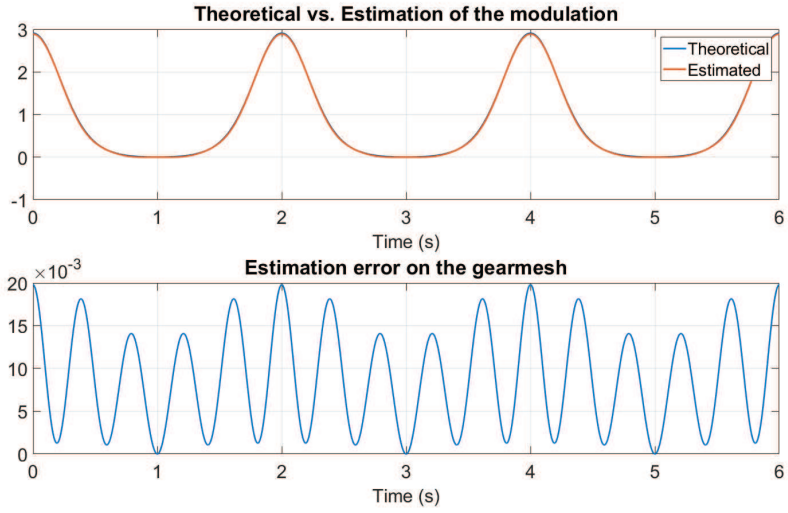


Figure 7.30: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

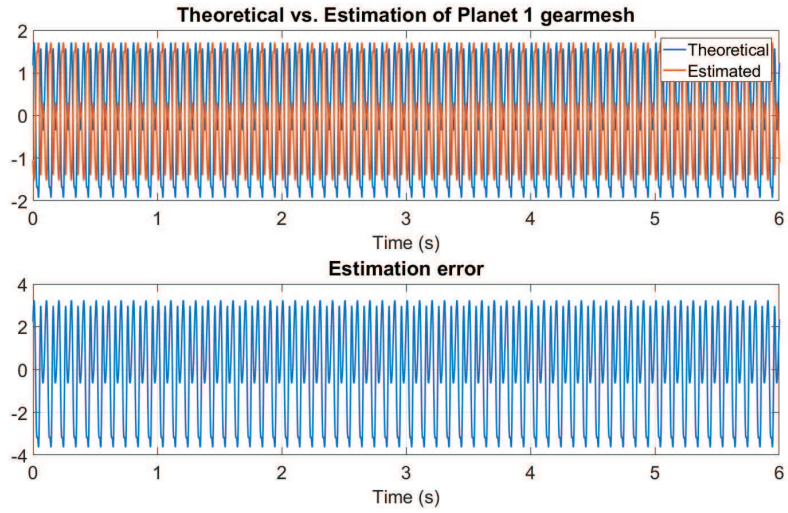


Figure 7.31: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

all couples of gearmesh/modulation generated by each planet passage in front of a single sensor. Three different systems have been studied and simulated in order to test the method. For both 2-planets and 3-planets systems, the estimation is very satisfying even in noisy situation as the relative error does not exceed 3% of the global signal energy. For the 5-planets system, we have not been able to find enough independent conditions that would enable us to estimate properly the signals. Future work can go into three directions:

1. We designed a method for the exact case then applied it in the presence of noise without modification although it does not solve the optimal decomposition problem. The latter could be addressed using optimization algorithms similar to those used in Chapter 6.

2. We only considered one sensor although using several of them could bring more

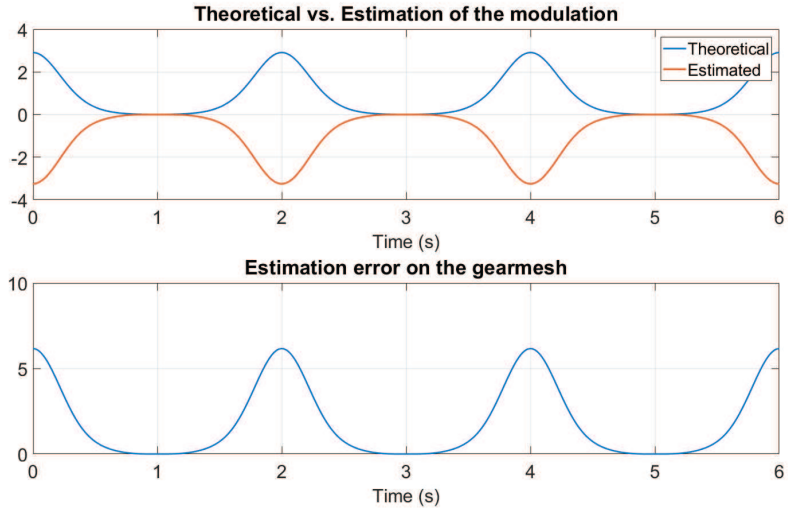


Figure 7.32: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

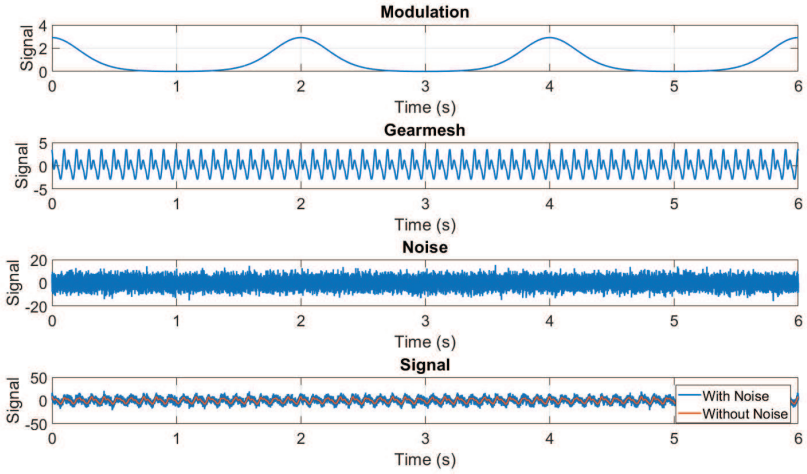


Figure 7.33: Simulation of the temporal signal numerically generated. From top to bottom: Carrier signal, Modulation signal and Global signal set according to Eq.7.2.

information helping the decomposition.

3. We used a simplified model where the signals from the 5 planet gears have the frequency of the meshing, i.e., they are not affected by planet and sun rotations. The framework we propose can be extended to this case: the matrix representation of the spectrum simply has to include all combinations of the meshing, sun and gear frequencies. Only one case is problematic: if two different combinations of these frequencies overlap.

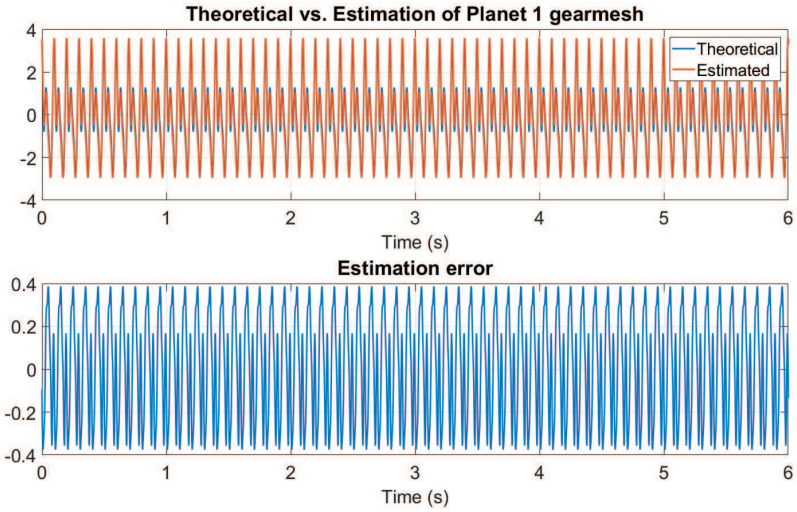


Figure 7.34: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

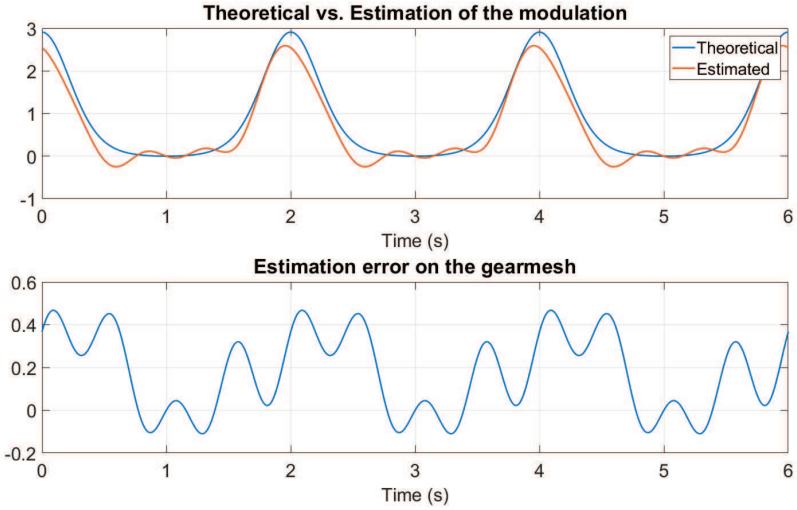


Figure 7.35: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

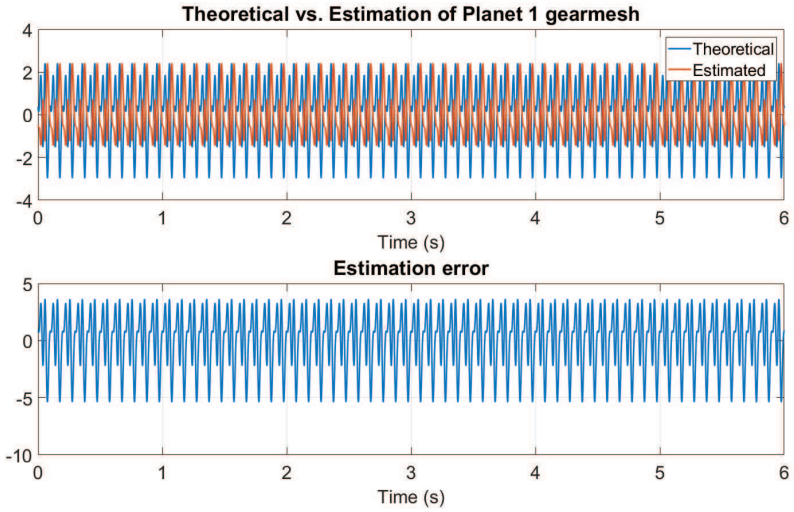


Figure 7.36: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

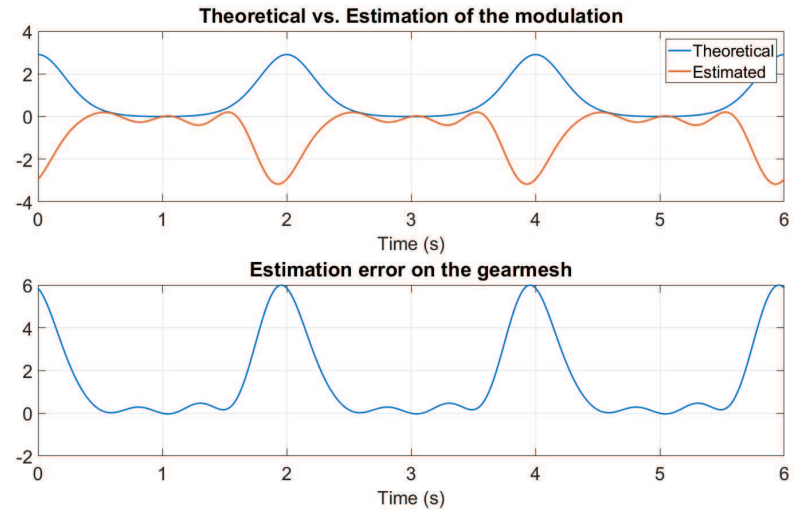


Figure 7.37: Overlay of the estimated signal with the theoretical one (top) and estimation error representation on the temporal representation (bottom).

Part III

Application to gearboxes

Chapter 8

Applicability to fault detection

In the present Chapter, a possible use of the multi-carrier amplitude demodulation is going to be presented. We recall that the starting point of this whole work was about gearbox monitoring, and more especially how to enhance incipient fault detection so as to be able to predict how long will the system be able to be operating and thus to plan the system's maintenance. The impact of gearbox faults on the spectrum is going to be briefly recall, then we will present how we apply the proposed demodulation algorithm on real data. Two points are going to be studied, first by comparing the classical demodulation technique with our optimal demodulation algorithm and second we will compare with some usual indicators how optimal demodulation allows to detect fault earlier than other techniques.

8.1 Fault detection using the multi-carrier amplitude demodulation

Remainder on faults We have seen in Chapter 1 that there are several class of faults that may happen to gearbox. Some are usual and come with the system getting older, such as wear, but other may happen suddenly and are the sign of a malfunctioning. Both class of fault bring interesting information about the system and are worth be detected. On one hand, in order to be able to better plan maintenance, it is important to evaluate which stage of wear the system has reach. On the other hand, sudden changes in any monitored signal due to incipient fault that may lead to failure are very important as they may severely damage other pieces of the system, which could get the whole system to break.

In the case of power transmission systems, sizing of gearbox is designed such that it should be one of latest system to be damaged. That is why unexpected fault in gearbox must be detected as early as possible.

We have previously seen that gear faults mainly affect the spectrum, which justifies that many fault detection methods are based on spectral considerations, and so is our method.

Proposed method In this work we got interested in one specific signal processing method, namely demodulation, as it is widely used in the telecommunication but less extended to mechanical systems. When applied on vibration signal, it allows to enhance

fault detection of mechanical systems. An optimal amplitude demodulation algorithm has been developed in the present work which has proved to be able to estimate product components for simulated signals. Here we propose to adapt this technique for real vibration signal originating from test-rig.

The multi-carrier demodulation performed on real test-rig dataset differs a little from the demodulation done in the simulation, in the sense that there are some preliminary processing techniques to be applied. The obtained algorithm is performed in several steps, including basic and classical preprocessing steps. First the signal is usually ordertracked. With a tachometer signal (an encoder can be enough but less precise), the vibration, or any signal under study is rearranged with respect to the gear rotation. This has the effect of making the spectrum lines more ”peaky", indeed the demodulation techniques , and more specifically multi-carrier demodulation are based on the Fourier coefficient estimation, which means that the more precise is the spectrum computation the better the estimation.

Once all those steps are done, the proper multi-carrier demodulation algorithm may start. First frequencies of interest are selected, as well as the number of considered harmonics. The vibration spectrum is then rewritten in the matrix tool. Thanks to a SVD (or any low-rank approximation technique), two vectors are estimated, representative of the gearmesh and modulations.

Remark 40. *As we select the frequencies of interest, i.e. the rotation frequencies of the input and output shafts, our demodulation techniques performs a highly selective filter, which allows to remove a huge amount of background noise along with uninteresting frequencies, i.e. bearing frequencies, electrical component ...*

Remark 41. *To enhance even more the shape of the spectrum lines, it is possible to cut the signal in order to have a integer number of the gearbox great period.*

The amplitude demodulation computed here has been described in Algorithm 1. We remember that the signal estimation is performed up to a scale factor. In order to avoid extreme magnitude scales, the modulation signal average is set to one.

Eventually we will have an estimation of the gearmesh vibration along with the modulations. Then recovering each gear component is mere using a band-pass filter.

8.2 Test rig presentation

CETIM dataset The evaluation of fault detection performance has been done on a publicly available dataset. The test bench has been done at the GIPSAlab and instrumented by CETIM. It is a fatigue test where the gearbox has been running for 12 days with an acquisition taken on everyday. The gearbox reducer is a fixed-shaft gearing systems made of two gears which have 20 and 21 teeth respectively. The acquisitions are made at a sampling rate of 20 kHz. The rotation frequencies of both gears are $f_1 = 16.75Hz$ and $f_2 = 17Hz$ respectively.

Remark 42. *It can be noticed that as the numbers of teeth are very close, the rotation frequencies are very similar. This particularity of the test rig turns the identification and the separation of the two gears component to be very tricky. Indeed at low frequencies they cannot be distinguished as the sampling rate is too low.*

In this dataset the only available measurement is the vibration given by a single-axis accelerometer.

UNSW dataset A second dataset have been used to test the developed technique. Within the partnership with the condition monitoring research lab of the UNSW, some tests have been run with their fixed-shaft gearbox test rig. The reducer is made of two gears of 19 and 52 teeth respectively.

Figure 8.1: The spur gear test rig at the University of New South Wales (UNSW).

The test-rig is equipped with several sensors. Vibrations are measured on two different points: on the input and output shafts on the brake side. A tachometer and an encoder are also mounted on each rotating shaft.

8.3 Real data experiments

All running test have been done using the Cetim dataset. In this section we will specifically look at three points: the results of the multi-carrier demodulation on the vibration, the comparison of the estimation of modulations between the classical and multi-carrier demodulation and eventually we will compare which technique allows to better detect incipient faults.

Some results of multi-carrier demodulation After performing the multi-carrier demodulation, we obtain the gearmesh on one hand and the two modulations on the other hand. Let us now display the two components $s_m(t_n)$ and $s_c(t_n)$ for a healthy run and for a run where a fault has been diagnosed on one gear. The outputs of the optimal decomposition are represented in Figure 8.2. The two modulations have been separated with a specific filter.

(a)

(b)

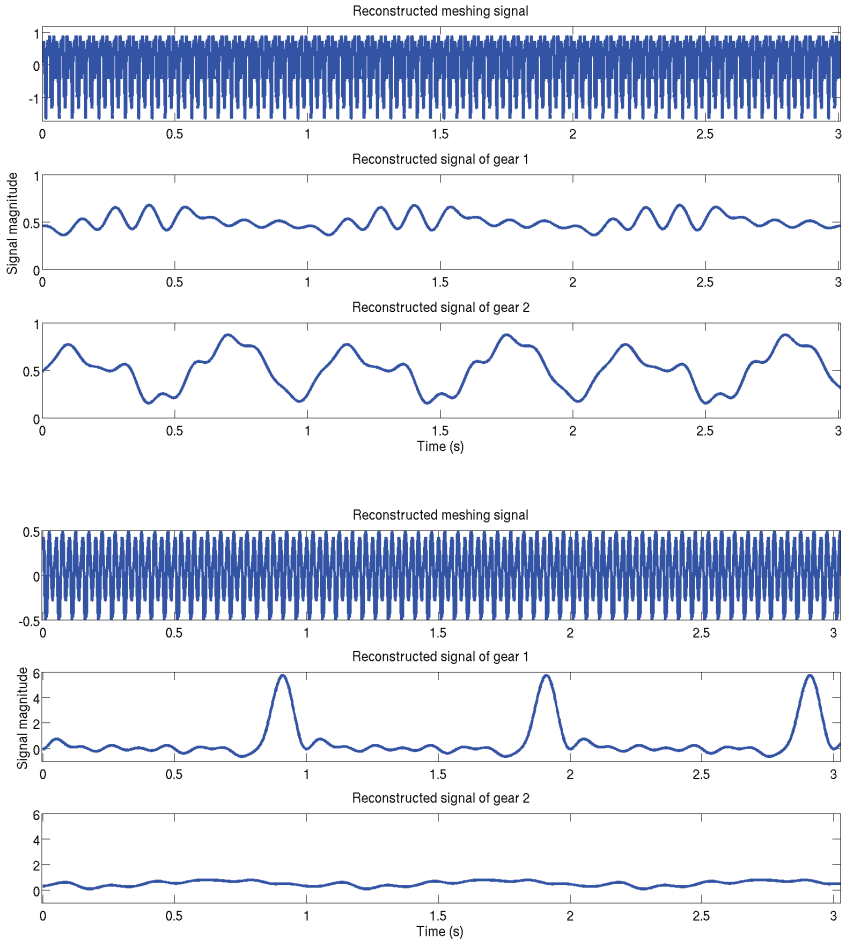


Figure 8.2: Temporal representation of the reconstructed signals (meshing, gear 1 and gear 2) for a healthy run (a) and a faulty run (b).

We see that considering each component of the signal separately makes monitoring easier. When a failure appears, it becomes visible on the concerned recovered signal, as in Figure 8.2. Then, common fault indicators can be calculated directly on s_{gear1} and s_{gear2}, allowing early detection of the fault as well as its localization.

Comparison with demodulation In Chapter 5 the proposed multi-carrier demodulation technique has been compared with the classical demodulation using Hilbert transform. There we have seen that in term of statistical study, the proposed multi-carrier demodulation is a better estimator for both carrier and modulation amplitude parameters. In this section we propose to compare those two methods on the real dataset presented above in order to quantify the improvement obtained by using the multi-carrier demodulation.

On Figure 8.3, the measurement has been done at the beginning of the test. It is visible that with the classical demodulation there is an additional low frequency component that has been filtered out with the multi-carrier demodulation. However, in Figure 8.4, the fault is clearly visible and its energy is so strong that both methods gives approximately the same results, even if the estimation given by multi-carrier demodulation is much more regular than the other.

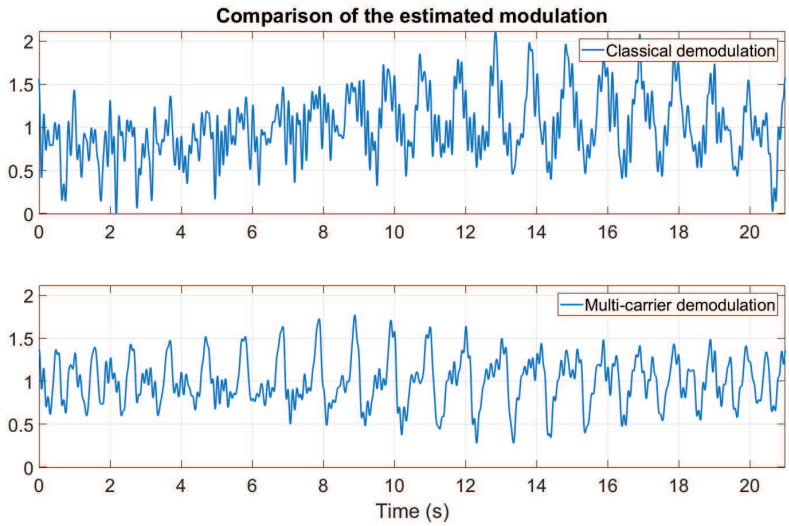


Figure 8.3: Temporal representations of the modulation given by the classical demodulation (top) and multi-carrier demodulation (bottom).

Early fault detection For system maintenance purpose, we are generally interested in detecting incipient fault. As a matter of fact, if a fault is detected in its early stages, corrective and preventive action can be taken to avoid any significant machine failure.

In this section we want to compare several usual fault indicators calculated on diverse signals. Here the we compare the results of four basic indicators (i.e. Root Mean Square (RMS), kurtosis, Peak-to-Peak and FM4) computed on three signals, i.e. raw signal, demodulated with classical and multi-carrier techniques. This will allows us to study the capability of our method to detect incipient faults.

By analyzing all Figures 8.5, 8.6, 8.7 and 8.8, the first thing that has to be noticed is that the multi-carrier does not always give the best response to incipient fault. When looking to the FM4 indicator, it is possible to say that the best indication is given with the raw signal. However, with all indicators, it is also important to remark that multi-carrier demodulation avoid some false detection that classical demodulation do not.

8.4 Conclusion

In this Chapter we have applied the amplitude multi-carrier demodulation technique we have developed. It shows to have interesting properties when we just compare the quality of modulation estimation. However, even if opposite to classical demodulation it avoids easily false alarms, it is not always the best signal on which indicator are the most efficient. The multi-carrier method gives similar results to those of classical demodulation, but both are sometimes worse than the simple raw signal.

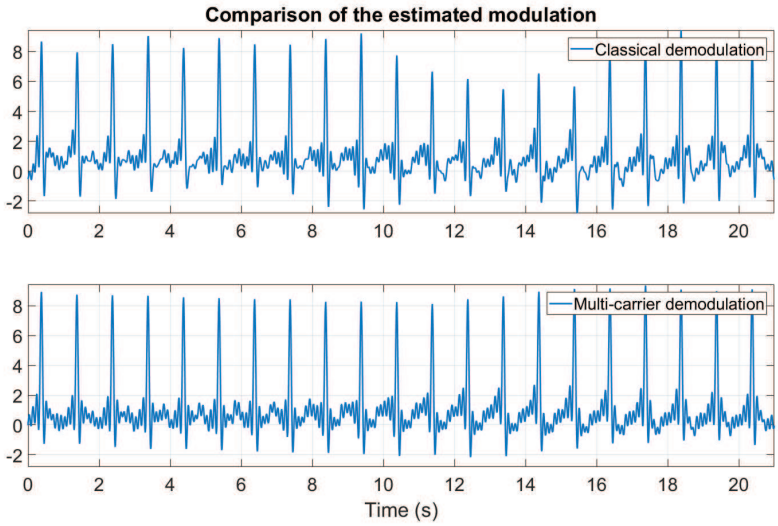


Figure 8.4: Temporal representations of the modulation given by the classical demodulation (top) and multi-carrier demodulation (bottom).

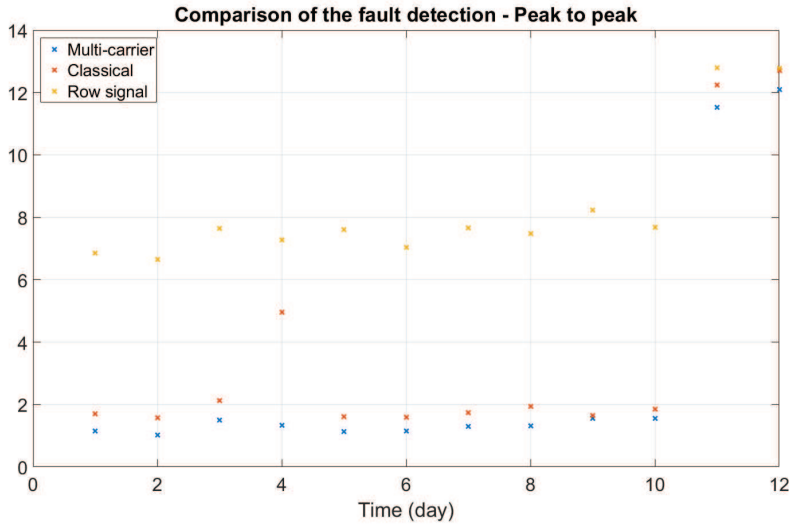


Figure 8.5: Computation of the RMS indicator on the three signals every day.

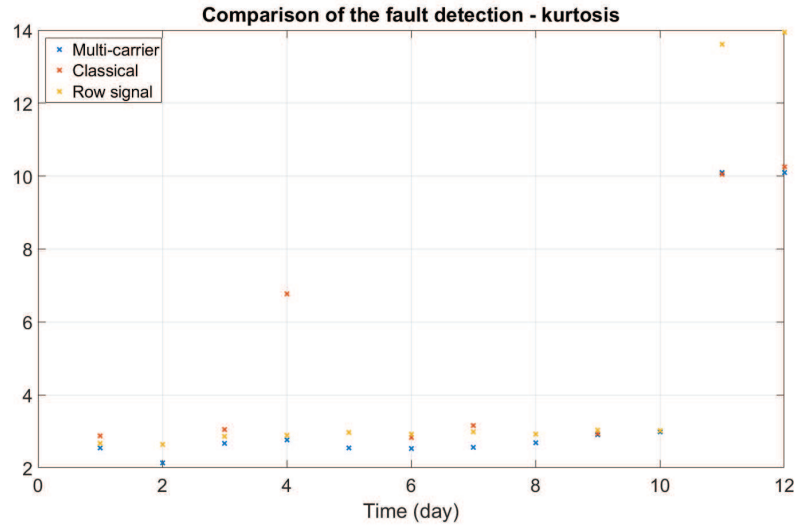


Figure 8.6: Computation of the kurtosis indicator on the three signals every day.

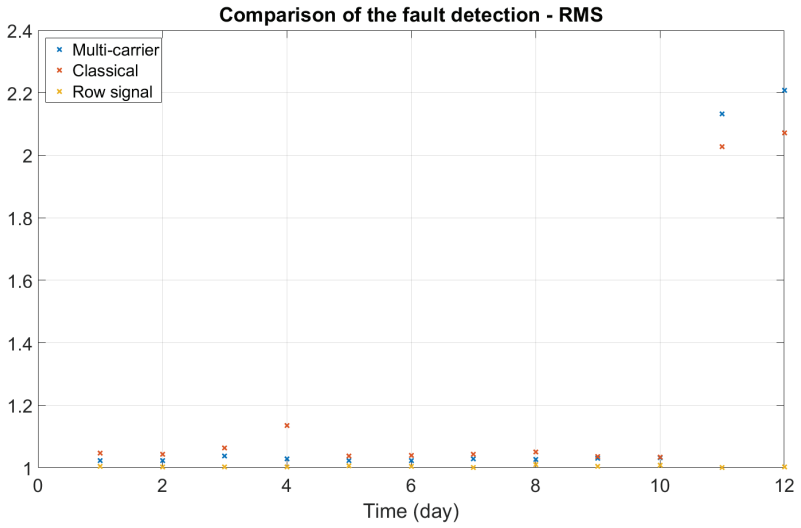


Figure 8.7: Computation of the PP indicator on the three signals every day.

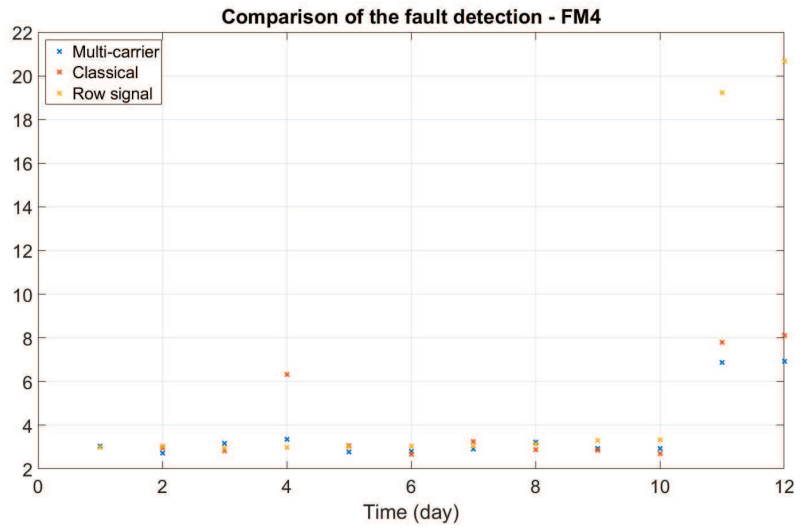


Figure 8.8: Computation of the FM4 indicator on the three signals every day.

Chapter 9

Signal model testing

9.1 About another possible use of optimal demodulation

9.1.1 How to test product model validity?

In the previous Chapters we introduced a new optimal demodulation technique we called multi-carrier demodulation. Unlike traditional methods, the multi-carrier demodulation takes into consideration the whole signal, avoiding the band pass filtering step and ensuring a smaller loss of information in the process. Thanks to the matrix representation of spectrum tool that has also been developed, a link has been made between low-rank matrices an the operator theory.

Thus, one possible way to use the multi-carrier demodulation method we proposed in the previous Chapters is to verify if a signal can be represented by a product. As we have shown that a multiplicative signal can be exactly decomposed into its two components, a first interesting point is that we can verify whether or not a signal can be decomposed into the two components of a product.

In the case of gearbox vibration signal, we have seen that usually, they are empirically modeled by a product (see Chapter 2). An interesting question is to evaluate how representative is that modeling for vibration signals.

9.1.2 Fixed-shaft gear vibration model testing

In order to test the multiplicative model of gearbox vibration, the multi-carrier demodulation is run on the vibration signal originating from the CETIM dataset. The decomposition has been made considering 15 harmonics for the carrier signal and 9 harmonics for the modulation signal. The temporal representation of the resulting decomposition is displayed in Figure 9.1.

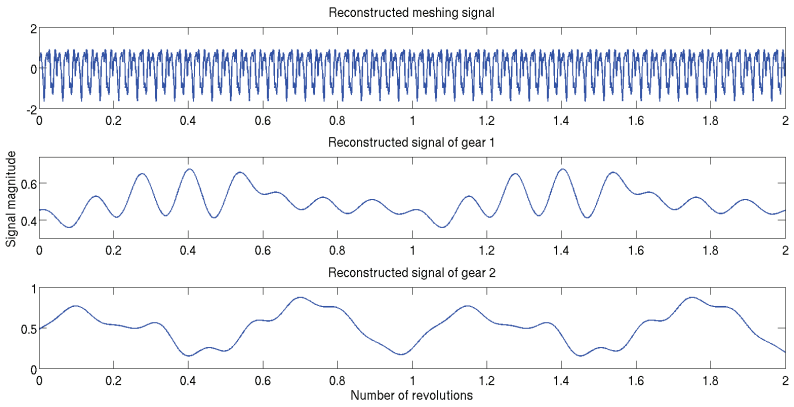


Figure 9.1: temporal representation of the estimated signals with the multi-carrier amplitude demodulation technique. From top to bottom: Gearmesh, Gear 1 and Gear 2.

First opinion on the decomposition Obviously we have been able to obtain an estimation of the signals, but as we do not have the theoretical ones, we cannot asses how right they are. A way do verify how close we are to the original signal is to multiply the estimated carrier and modulation and combined them in the same way as the model, and then to compare with the original signal. This comparison is illustrated in Figure 9.2 for the temporal representation and in Figure 9.3 for the overlay of the spectra.

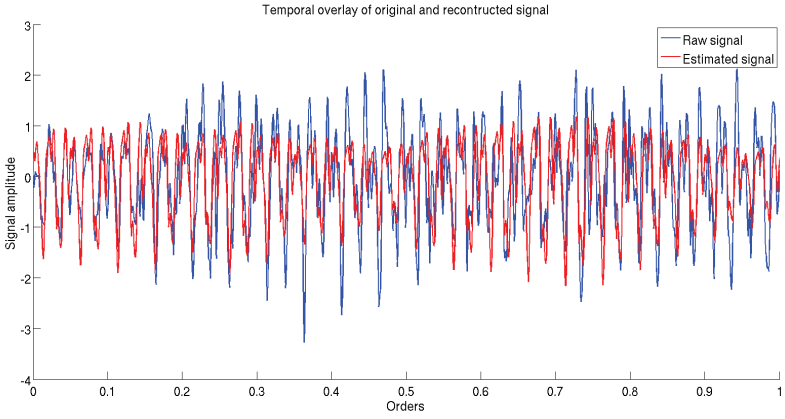


Figure 9.2: Overlay of the temporal representation of the raw signal and the estimated one.

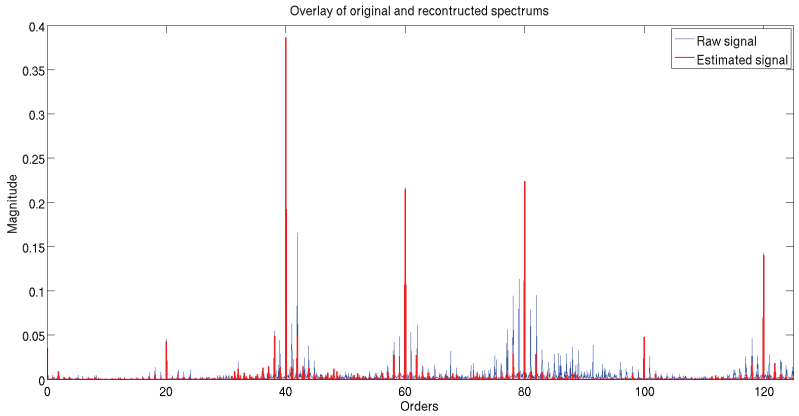


Figure 9.3: Overlay of both spectra of the raw signal and the estimated one.

It is clearly visible that the estimation is not perfect. This may have several reasons:

1. an arbitrary number of carrier harmonics has been chosen which may not be the optimal one,

2. with this algorithm, the condition of no overlapping between the sidebands is done, even if there is no evidence that it is actually the case,

3. there might be other phenomenons at play that are not well considered and above all they may interact with the vibration components we are interested in for the estimation.

Nevertheless, we considered that even not perfect, it was a good approximation of the vibration signal for that dataset and we first validated it.

Second opinion on the decomposition But after some discussions with researcher having a more mechanical approach, we have noticed that sometimes the modulations originating from the gears rotation were periodically crossing the zero line, as illustrated by Figure 9.4.

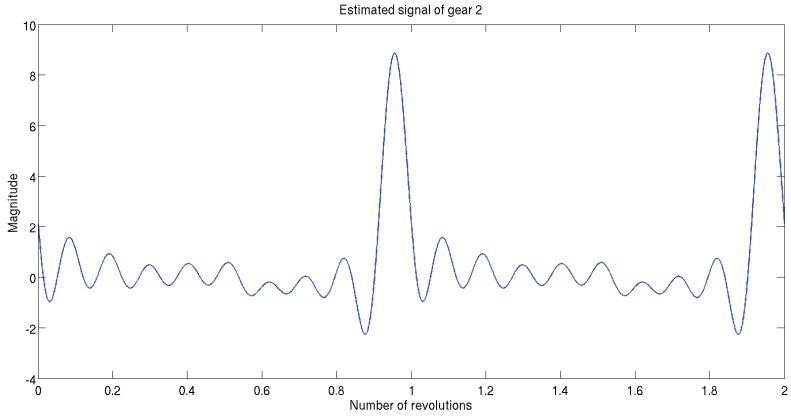


Figure 9.4: Temporal representation of the estimated signal corresponding to a faulty gear.

Even if this observation may seem hardly worth mentioning, it has to be precised that when we are considering mechanical systems, there is a physical reality that has to be taken into account. Here we can directly link the vibration due to the gears rotation to the force that is applied on those gears. When we have a closer look into the system dynamic, we see that during the rotation, when the tooth of the driving pinion starts to push the tooth of the driven gear, this one generates a reaction force.

9.2 A quest for understanding

We have previously demonstrate in Chapter 5 that it is always possible to break a product of two signals down into its two components, even in highly noisy signals. From this observation, the methodology cannot be questioned for the wrong reconstruction of the two modulation signals. If the problem does not find its origins in the methodology, the only possible explanation is that the product model is not representative of the gearbox vibration.

In order to verify this hypothesis, we looked more deeply in the vibration signal and found that even though the established model may be an interesting approximation at first sight, our experimental investigation has shown several inconsistencies.

First, there are sidebands present in the spectrum at the crossed frequency $f_c \pm (f_{m1} \times f_{m2})$. This is actually justified by the fact that a model

$$s(\theta) = s_c(\theta)(1 + s_{m1}(\theta)s_{m2}(\theta))$$

may be more physically rigorous, i.e. the combined effect of the two gears must have a periodicity which corresponds to the event of the same two teeth meshing again. This part has been studied afterward and will be presented in a future work, in collaboration with the UNSW condition monitoring lab.

Secondly, irregularities of the sideband patterns can be observed in every experimental spectra we have, i.e. the ratio between the sidebands of different gearmesh harmonics is not constant, even after a precise order-tracking using an encoder with a high number of pulses. There are several hypotheses regarding the explanation for this difference.

1. The transfer function. The transfer function between excitation and actual measurement is the obvious suspect for this distortion. However, even after applying a cepstral long-pass lifter (aimed at removing the transfer function âĂIJscalingâĂİ of the spectrum), the differences between sideband patterns remain strong.

2. Simultaneous effect of tooth stiffness and geometric profile error A more subtle hypothesis instead involves the consideration of two effects resulting in the modulation of the gearmesh. Under this hypothesis, tooth-stiffness-induced vibrations (load dependent) and vibrations induced by profile error (independent of load) are acting as two parallel models, with different carriers and modulations, but coincident frequencies. Under this assumption the gearmesh harmonics show different combined patterns of sidebands because of the additive effect of two different gearmesh carriers and modulations.

We studied how different were the ratio between sidebands of different gearmesh harmonics for gearbox vibration signal. The difference between the ratios is evaluated graphically. The four first gearmesh harmonics are normalized and both left and right

sidebands are displayed. Two points are to be verified, first the symmetry for a single gearmesh harmonic between right and left sideband, and second the fact that sideband patterns should repeat identically at each carrier harmonics.

9.2.1 Vibration signal: 20Hz-20Nm test

On the test-rig of the UNSW lab we investigated the lines repartition in the spectrum. In Figure 9.5 we see the ordertracked (OT) spectrum of the vibration signal in the low orders.

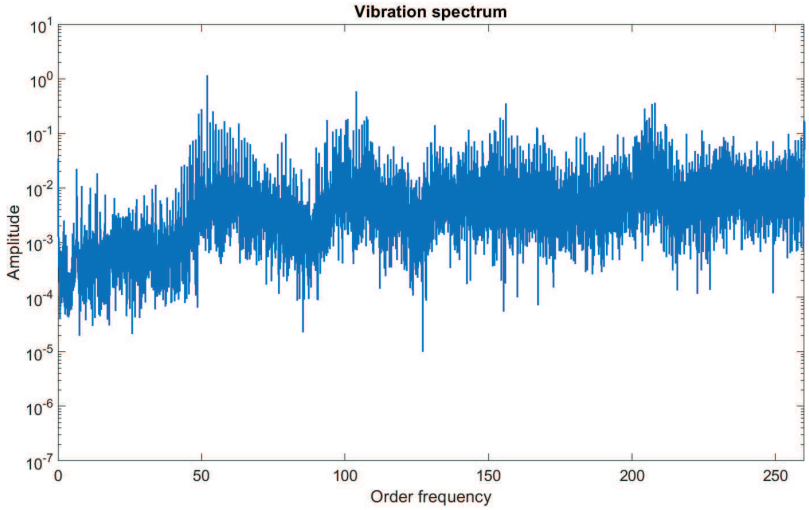


Figure 9.5: Spectrum representation of the ordertracked vibration signal for the four first harmonics.

On the spectrum it is clearly visible that there are many sidebands for all the gearmesh frequencies, but it is hard to say how symmetric they are regarding the gearmesh harmonic neither if the ratio between sidebands of different gearmesh harmonics is similar. Figure 9.6 shows the repartition of the sidebands for the four first harmonics of the signal.

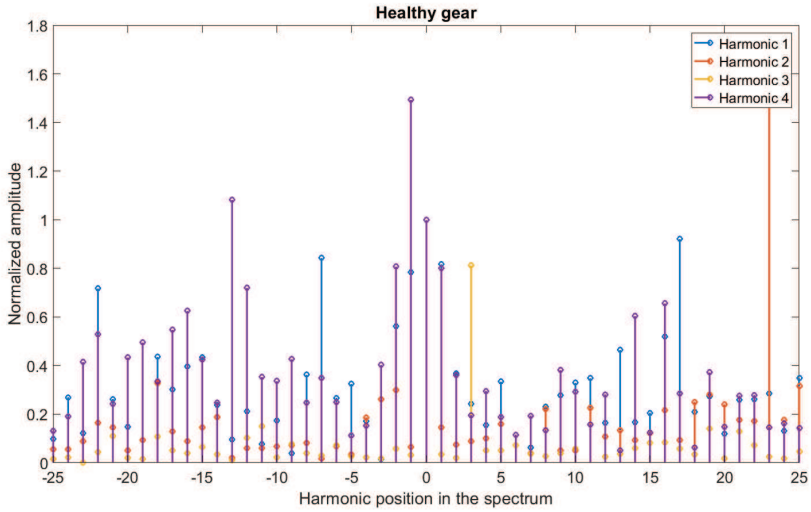


Figure 9.6: Display of the ratio between sidebands and the gearmesh harmonics.

This results shows that the patterns are massively different even disregarding their phase, which should also coincide after normalisation by the carrier harmonics. Two possible explanations for such behaviour were suggested: amplifications due to the system transfer function, or dominant frequency-modulation effects. This complete lack of element tends to validate that the vibration cannot be modeled as a product. Nevertheless, there are some elements that may interfere and hinder the identification of a clue that would go in the direction of the model validity. For example, we know that in a gearbox system, the transfer function is messing a lot in the hole spectrum but not with the same intensity for all frequencies. That is why we will estimate the transfer function and remove it from the signal. Figure 9.7 show the previous spectrum of Figure 9.5 with the estimated transfer function overlaid.

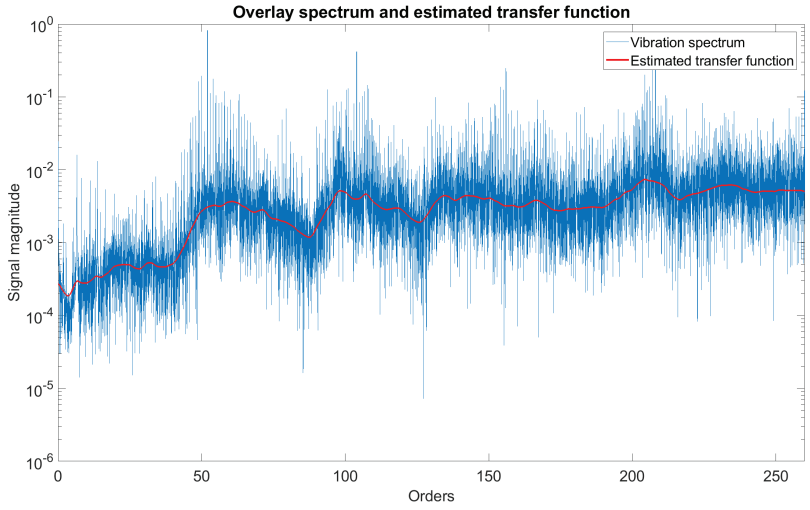


Figure 9.7: Overlay of the previous spectrum (blue) with the estimated transfer function (red).

After removing the transfer function, the spectrum is flatter, giving hope that the symmetry and ratio will become more visible. Figure 9.8 shows the new spectrum.

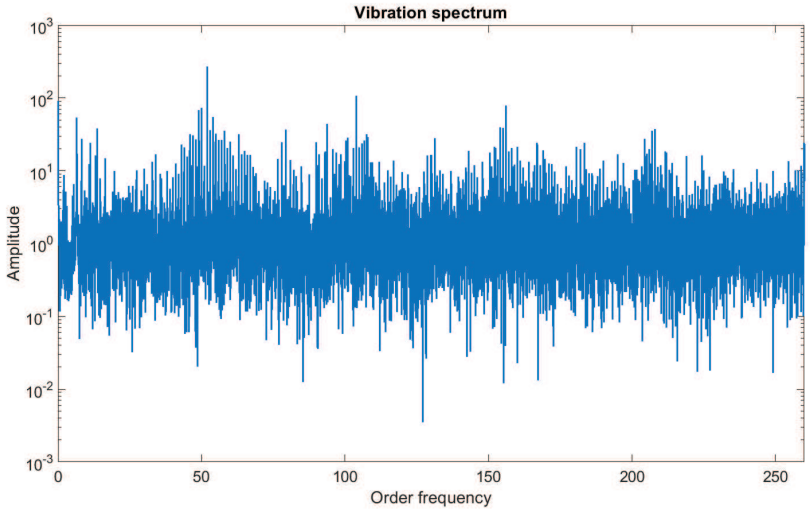


Figure 9.8: Spectrum representation of the vibration signal after removing the estimated transfer function.

We compute and display the new sidebands repartition in Figure 9.9.

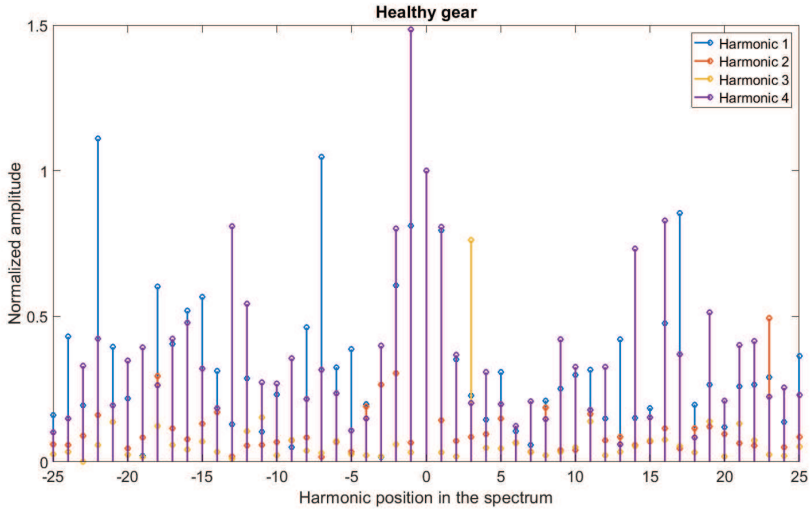


Figure 9.9: Display of the ratio between sidebands and the gearmesh harmonics after removing the transfer function.

Despite the good result in terms of spectral liftering, the problems observed in Figure 9.6 continue to be as severe in the normalised spectrum harmonics reported in Figure 9.9. There is no real impact of the removing of the transfer function on the sidebands.

9.2.2 Vibration signal: low speed and no load test

A possible explanation of the differences in the sideband patterns could be found in the different roles played by two different root-cause mechanisms resulting in gear vibration: geometric and static transmission error. The first is due to profile irregularities, whereas the second is due to the angular dependence of the gear-meshing compliance under load. An additional test was therefore executed at very low load ($1.5Nm$, just enough to maintain contact between the gear teeth) and speed ($2Hz$), where geometric transmission errors were expected to dominate. In order to have comparable displays of the sideband harmonics with the previous test, we repetead the exact same analysis. Figure 9.10 represents the spectrum of the gearbox vibration and Figure 9.11 the sidebands.

Remark 43. *As the rotation speed of the gearbox is very slow, in the low frequencies of the spectrum there is a huge interference coming from some electromagnetic interference within the test rig. It should not be taken into account though we looked at the harmonics 2 to 5.*

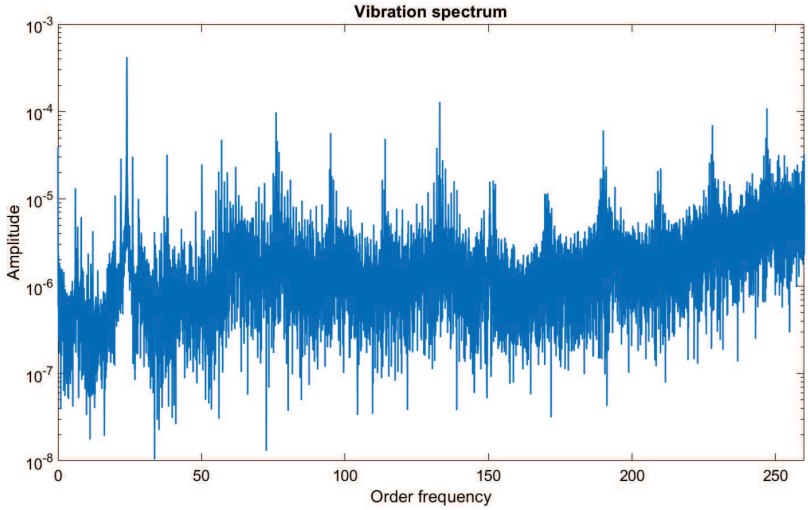


Figure 9.10: Spectrum representation of the ordertracked vibration signal for the four first harmonics.

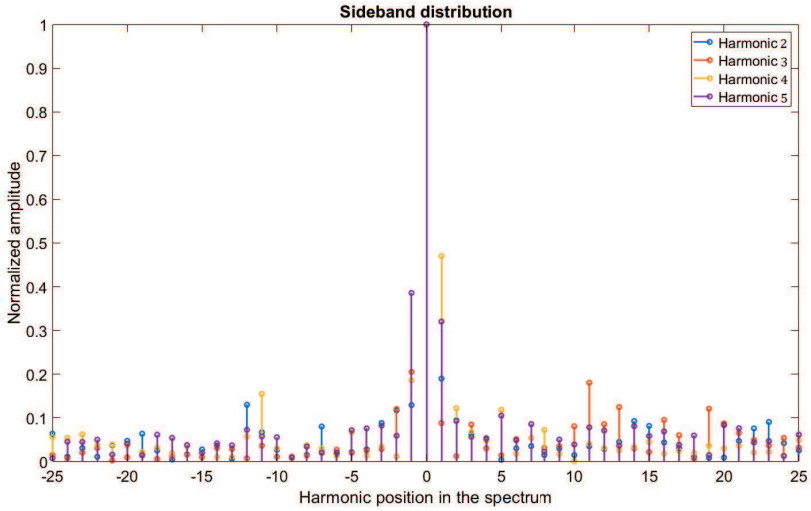


Figure 9.11: Spectrum representation of the ordertracked vibration signal for the four first harmonics.

As for the faster and loaded case, we notice that there is no clear multiplicative pattern recognizable within the representation of the sidebands. Thus, once again we estimate and remove the transfer function, Figure 9.12, and then repeat the process: Figure 9.13 and Figure 9.14 illustrate the transfer function-less spectrum and the associated harmonics representation.

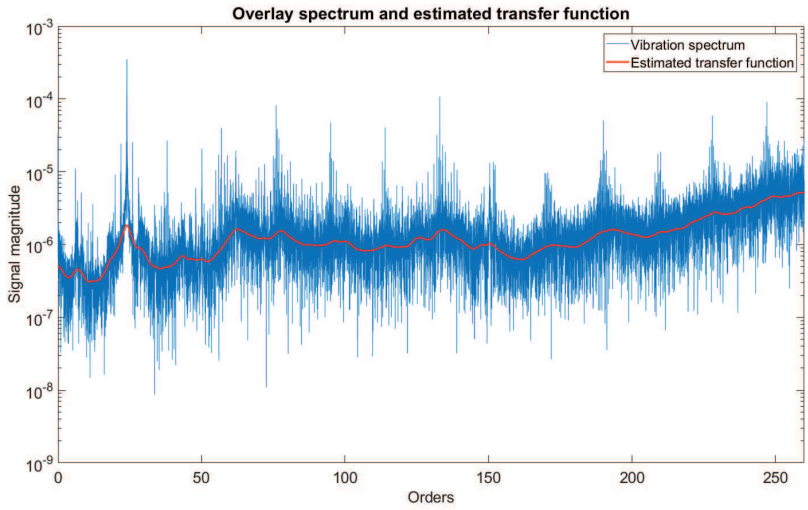


Figure 9.12: Spectrum representation of the ordertracked vibration signal for the four first harmonics.

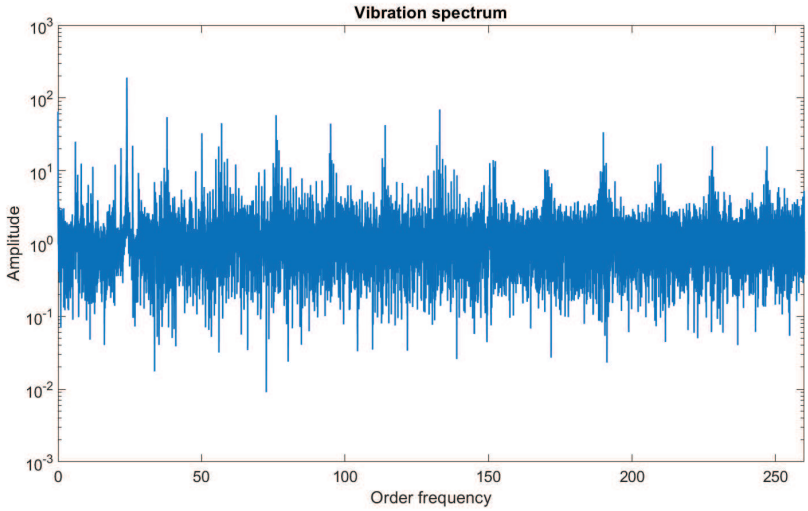


Figure 9.13: Spectrum representation of the ordertracked vibration signal for the four first harmonics.

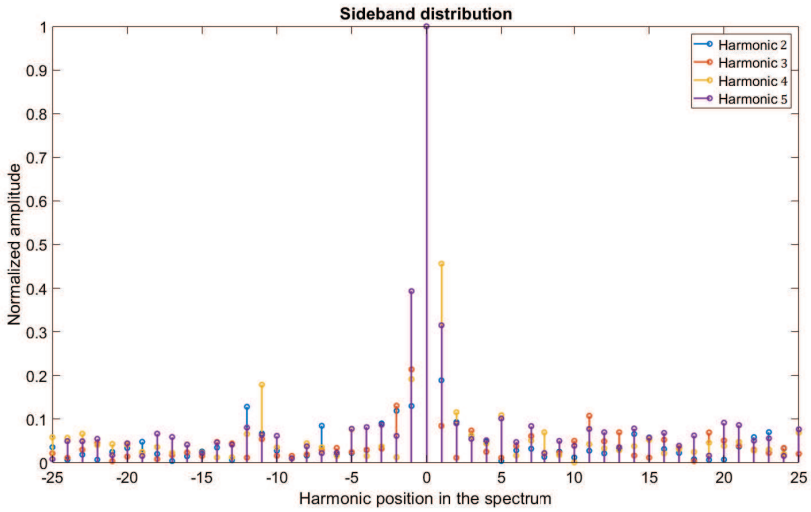


Figure 9.14: Spectrum representation of the ordertracked vibration signal for the four first harmonics.

The patterns shown in Figure 9.14 are much more consistent, even if discrepancies are still present, thus supporting the idea of a potential two-mechanism root-cause of the observed pattern inconsistency.

9.2.3 Transmission error signal: low speed and no load test

In this example we are interested in the transmission error (TE) of the gearbox. The TE is computed as the difference between the two encoders (or tachometers) located on both the input and output shafts. The running test that has been done is a dry wear test, which means that there were no lubricant in the gearbox. The gear is considered healthy at the very beginning and faulty later as some pitting has been found on the gears surface.

We display the sideband repartition in both healthy in Figure 9.15 and faulty case in Figure 9.16.

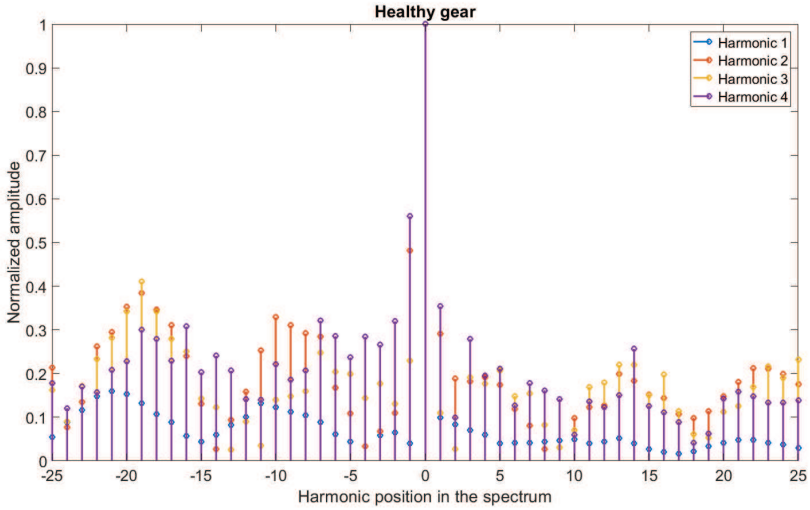


Figure 9.15: Display of the ratio between sidebands and the gearmesh harmonics after removing the transfer function.

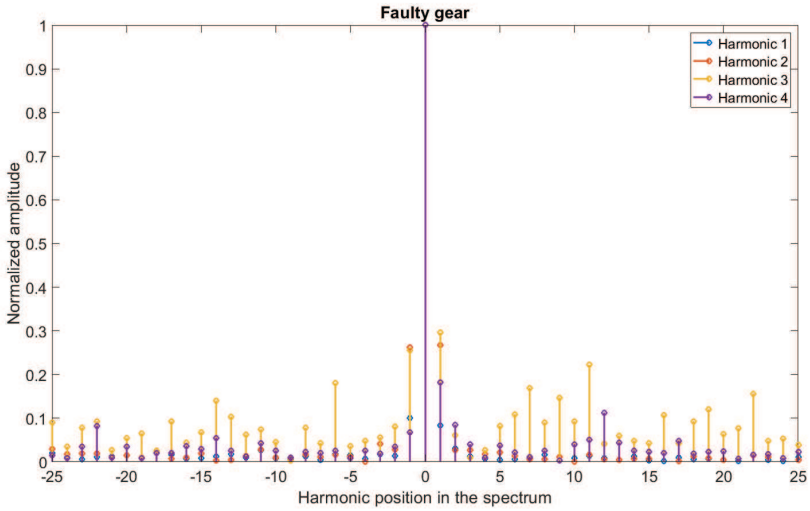


Figure 9.16: Display of the ratio between sidebands and the gearmesh harmonics after removing the transfer function.

In the case of the healthy gear there is no major indication that the sidebands follows the multiplicative rules. As the data has been taken at the very beginning of the test, i.e. during the first minute, the run-in period may not have occurred yet, which causes an uncertainty regarding the vibrating behavior of the gearbox. Unexpectedly, we notice that in the case of the faulty gear, it is reasonable to say that the ratio between sidebands and gearmesh harmonics is quite identical. It is so far the only situation where the hypothesis of a product could be done.

9.3 Conclusion

We were interested in this Chapter in knowing whether or not a fixed-shaft gearbox vibration could be reasonably described by the multiplicative model extensively used in

the literature. After a first good impression, we observed an incompatibility between the mechanical properties of the system and the estimated signal characteristics. Then we led an investigation looking for any clue that may suggest the model is representative of some aspect of the gearbox operation. Our conclusion is that for accelerations there were no situation which gave acceptable characteristics of the multiplicative modeling, even when the transfer function is removed. Whereas when looking to the transmission error in very specific operating conditions, for some reason, multiplicative modeling can be considerate accurate.

Chapter 10

Conclusions and future prospects

10.1 Conclusions

Gearboxes are vital systems in rotating machinery. A reliable monitoring system is critically needed in industries to provide early warning of damage or malfunction in order to avoid sudden failures and breakdowns. It is also of major interest for economic purposes as it allows to better plan the maintenance of mechanical systems. In aircraft engines, power transmissions systems are supposed to be the system that last the longest in time which means that unexpected troubles may have important consequences on the rest of the system.

In this PhD work, we have proposed a new approach for the analysis of gearbox vibration signals. Vibration signals originating from gearboxes are very complex and hard to represent through simple modelings, though they are usually represented as a multiplication between a high frequency signal called carrier and some modulations. In this case the carrier stands for the gearmesh signal, i.e. the periodic contact between the teeth of the two gears together and the modulations are identified to the rotation periods of each gears. This amplitude modulated signal model has been widely used in condition monitoring to estimate the modulations. In this work we have been interested in a new technique that allows to represent product signals as optimization problems. This formulation as the striking property to enable a matrix representation of the signal spectrum which let us link the study of this estimation problem with rank one approximation problems.

Based on the optimization framework, an amplitude demodulation algorithm has been developed and has shown to be the best approximation possible for product signals. This new result has shown promising uses. In this work we have basically used the amplitude multi-carrier demodulation for fault detection and mostly to verify how representative is the product of gear signals. For fault detection we have seen that the optimal amplitude demodulation does not enhance significantly the detection but it may improve false alarms. Regarding the representativity of the multiplicative model for gearbox vibration signal, we have experimentally shown that there are some inconsistencies. For example, sometimes the estimated modulations becomes negative, which is impossible with mechanical systems. Moreover we have noticed that there are some peaks in the spectrum that do not belong to the amplitude modulation model.

In this work we have also extended the optimization framework developed for amplitude demodulation to phase and amplitude modulated signals. This situation may appear when there is some fluctuation in the rotation speed, which is usually the case.

We have obtain necessary and sufficient conditions to the solvability of the exact problem. In the case where the signal is noisy, or when the small fluctuation hypothesis is not fulfilled, the exact conditions are not enough. This is why we proposed some algorithms based on gradient descent estimation. Those techniques have been proven to be interesting in the case of simulated signals but have not been tested on real signals.

The case of epicyclic gearboxes has also been studied. In that case the matrix formulation of the spectrum can also be used, which allows us to perform a theoretical separation of each planet gear contribution to the global signal. We have proposed some algorithms that separate each planet gear contribution to the signal but just for specific gearbox configuration.

10.2 Future prospects

Future research is suggested to address the following topics:

1. A first very interesting topic is to pursue studies on the epicyclic gearboxes. Indeed, in the case of five planetary gears we have not been able to propose enough conditions on the system to separate planet contributions. A better understanding of the system would help to define proper conditions that would allow the problem resolution. A theoretical study with symbolic computation and homological algebra tools may help get a better understanding of the mathematical objects at play.

2. As we have seen that the traditional modeling of fixed-shaft gears as a multiplication between the gearmesh and the gears rotation signals is not exact, a work could be done on a more representative model along with some further study on more experimental data to compare their spectrum under several speed and load conditions.

3. It would be interesting to compare all fault detection method and indicators on more dataset. In this work, we did not have in our possession enough vibration signals originating from different test-rig to have representative results.

Appendix

Appendix 1 : Solving inhomogeneous linear systems

10.2.1 Notation & basic homological algebra

Let us introduce a few notations and recall basic results of *homological algebra* (see, e.g., [76]) which will be of constant use in what follows.

In what follows, $\mathbb{K}$ will denote a field (e.g., $\mathbb{K}=\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$) and $A\in\mathbb{K}^{r\times s}$ a $r\times s$ matrix with entries in $\mathbb{K}$. Associated with $A\in\mathbb{K}^{r\times s}$, we can consider the two $\mathbb{K}$-linear maps:

$$\begin{array}{rcl@{\qquad}rcl} .A:\mathbb{K}^{1\times r} & \longrightarrow & \mathbb{K}^{1\times s} & A.:\mathbb{K}^{s\times 1} & \longrightarrow & \mathbb{K}^{r\times 1}\\ \lambda & \longmapsto & \lambda\,A, & \eta & \longmapsto & A\,\eta. \end{array} \tag{10.1}$$

If $B\in\mathbb{K}^{s\times t}$ is such that $A\,B=0$, then we can consider the following so-called *complexes* of $\mathbb{K}$-vector spaces

$$\mathbb{K}^{1\times r}\xrightarrow{.A}\mathbb{K}^{1\times s}\xrightarrow{.B}\mathbb{K}^{1\times t}, \tag{10.2}$$

$$\mathbb{K}^{r\times 1}\xleftarrow{A.}\mathbb{K}^{s\times 1}\xleftarrow{B.}\mathbb{K}^{t\times 1}, \tag{10.3}$$

i.e., linear maps with zero consecutive compositions, i.e.:

$$\begin{aligned} &\mathrm{im}_{\mathbb{K}}(.A):=\mathbb{K}^{1\times r}\,A\subseteq\ker_{\mathbb{K}}(.B):=\{\mu\in\mathbb{K}^{1\times s}\mid\mu\,B=0\},\\ &\mathrm{im}_{\mathbb{K}}(B.):=B\,\mathbb{K}^{t\times 1}\subseteq\ker_{\mathbb{K}}(A.):=\{\eta\in\mathbb{K}^{s\times 1}\mid A\,\eta=0\}. \end{aligned}$$

The *defect of exactness* of (10.2) (resp., (10.3)) is defined by

$$\begin{aligned} &H(\mathbb{K}^{1\times s}):=\ker_{\mathbb{K}}(.B)/\mathrm{im}_{\mathbb{K}}(.A)\\ &(\text{resp., } H(\mathbb{K}^{s\times 1}):=\ker_{\mathbb{K}}(A.)/\mathrm{im}_{\mathbb{K}}(B.)), \end{aligned}$$

where E/F stands for the *quotient* of a $\mathbb{K}$-vector space E by a $\mathbb{K}$-subvector space F. E/F is the $\mathbb{K}$-vector space defined by the *residue classes* $\pi(e)$ for all $e\in E$, where $\pi(e_1)=\pi(e_2)$ if $e_1-e_2\in F$, endowed with the operations:

$$\forall\,e_1,\,e_2\in E,\ \forall\,k\in\mathbb{K},\ \left\{\begin{array}{l}\pi(e_1)+\pi(e_2):=\pi(e_1+e_2),\\ k\,\pi(e_1):=\pi(k\,e_1).\end{array}\right.$$

The complex (10.2) (resp., (10.3)) is said to be *exact* at $\mathbb{K}^{1\times s}$ (resp., $\mathbb{K}^{s\times 1}$) if we have

$$H(\mathbb{K}^{1\times s})=0\quad(\text{resp., } H(\mathbb{K}^{s\times 1})=0),$$

i.e., if $\ker_{\mathbb{K}}(.B)=\mathrm{im}_{\mathbb{K}}(.A)$ (resp., $\ker_{\mathbb{K}}(A.)=\mathrm{im}_{\mathbb{K}}(B.)$).

An exact sequence of the form $0 \longrightarrow \mathbb{K}^{1\times s} \xrightarrow{.B} \mathbb{K}^{1\times t}$ means that $\ker_{\mathbb{K}}(.B) = 0$, i.e., that $.B$ is injective, whereas the exact sequence $\mathbb{K}^{1\times r} \xrightarrow{.A} \mathbb{K}^{1\times s} \longrightarrow 0$ means that $\mathrm{im}_{\mathbb{K}}(.A) = \ker_{\mathbb{K}}(0) = \mathbb{K}^{1\times s}$, i.e., that $.A$ is surjective. Similar comments hold for linear maps of the form $(A.)$.

A standard result on the duality of $\mathbb{K}$-vector spaces and $\mathbb{K}$-linear maps asserts that if (10.2) (resp., (10.3)) is an exact sequence of finite-dimensional $\mathbb{K}$-vector spaces, then so is (10.3) (resp., (10.2)). In homological algebra, we say that the *functor* $\hom_{\mathbb{K}}(\,\cdot\,,\mathbb{K})$ is *exact*, where $\hom_{\mathbb{K}}(E,\mathbb{K})$ denotes the $\mathbb{K}$-vector space formed by all the $\mathbb{K}$-linear maps (forms) from a $\mathbb{K}$-vector space E to $\mathbb{K}$. See, e.g., [76].

Similarly a standard result in linear algebra asserts that if (10.2) (resp., (10.3)) is an exact sequence of finite-dimensional $\mathbb{K}$-vector spaces, then, for any $q \in \mathbb{N}$, so is

$$\mathbb{K}^{q\times r} \xrightarrow{.A} \mathbb{K}^{q\times s} \xrightarrow{.B} \mathbb{K}^{q\times t}, \tag{10.4}$$

$$\left(\text{resp.,}\ \mathbb{K}^{r\times q} \xleftarrow{A.} \mathbb{K}^{s\times q} \xleftarrow{B.} \mathbb{K}^{t\times q}\right), \tag{10.5}$$

where $.A : \mathbb{K}^{q\times r} \longrightarrow \mathbb{K}^{q\times s}$ denotes the natural extension of (10.1) defined by $(.A)(\Lambda) := \Lambda\, A$ for all $\Lambda \in \mathbb{K}^{q\times r}$, and similarly for $.B$, $A.$ and $B.$ respectively. In homological algebra, we say that the *tensor functor* $\mathbb{K}^{q\times 1} \otimes_{\mathbb{K}} \cdot$ (resp., $\cdot \otimes_{\mathbb{K}} \mathbb{K}^{1\times q}$) is exact, where $E \otimes_{\mathbb{K}} F$ stands for the *tensor product* of two finite-dimensional $\mathbb{K}$-vector spaces E and F. For more details, see, e.g., [76].

If $0 \longrightarrow \mathbb{K}^{1\times r} \xrightarrow{.A} \mathbb{K}^{1\times s} \xrightarrow{.B} \mathbb{K}^{1\times t} \longrightarrow 0$ is a short exact sequence of finite-dimensional $\mathbb{K}$-vector spaces, then the *Euler-Poincaré characteristic* asserts that:

$$t - s + r = 0.$$

More generally, the Euler-Poincaré characteristic of a long exact sequence of finite-dimensional $\mathbb{K}$-vector spaces

$$0 \longrightarrow \mathbb{K}^{1\times r_n} \xrightarrow{.A_n} \mathbb{K}^{1\times r_{n-1}} \xrightarrow{.A_{n-1}} \cdots \xrightarrow{.A_1} \mathbb{K}^{1\times r_0} \longrightarrow 0$$

asserts that $\sum_{i=0}^{n}(-1)^i\, r_i = 0$. See e.g., [76]. Now, since the duality, i.e., the functor $\hom_{\mathbb{K}}(\,\cdot\,,\mathbb{K})$, preserves the exactness of long exact sequences of finite-dimensional $\mathbb{K}$-vector spaces, the same result holds for a long exact sequence of the form:

$$0 \longleftarrow \mathbb{K}^{r_0\times 1} \xleftarrow{A_1.} \mathbb{K}^{r_1\times 1} \xleftarrow{A_2.} \cdots \xleftarrow{A_n.} \mathbb{K}^{r_n\times 1} \longleftarrow 0.$$

Such an exact sequence always *splits* ([76]), i.e., there exist $B_i \in \mathbb{K}^{r_{i-1}\times r_i}$ for $i = 0,\ldots,n$, such that:

$$B_0 = B_{n+1} = 0,\ B_i\, A_i + A_{i+1}\, B_{i+1} = I_{r_i},\ B_{i+1}\, B_i = 0.$$

10.2.2 A standard result of linear algebra

Let $\mathbb{K}$ be field (e.g. $\mathbb{K} = \mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$), $A \in \mathbb{K}^{r\times s}$ a $r \times s$ matrix with entries in $\mathbb{K}$, and $y \in \mathbb{K}^{r\times t}$. We first state a standard result on the existence of solutions $x \in \mathbb{K}^{s\times t}$ of the following $\mathbb{K}$-linear inhomogeneous system:

$$A\, x = y. \tag{10.6}$$

Let us consider the following $\mathbb{K}$-vector space:

$$\ker_{\mathbb{K}}(.A) := \{\lambda \in \mathbb{K}^{1\times r} \mid \lambda\, A = 0\}.$$

Let $B \in \mathbb{K}^{q\times r}$ be a matrix whose rows form a basis of $\ker_{\mathbb{K}}(.A)$. In other words, B is a *full row rank matrix* (i.e., $\mu\, B = 0$ yields $\mu = 0$ since the rows of B are $\mathbb{K}$-linearly independent) which satisfies:

$$\ker_{\mathbb{K}}(.A) = \mathrm{im}_{\mathbb{K}}(.B) := \mathbb{K}^{1\times q}\, B.$$

In particular, we have $q = \dim_{\mathbb{K}}(\ker_{\mathbb{K}}(.A))$.

Then, we have:

$$\forall\, \lambda \in \ker_{\mathbb{K}}(.A) :\ \lambda\, y = \lambda\, A\, x = 0.$$

Hence, $B\, y = 0$ is a necessary condition for the solvability of (10.6). The next theorem shows that it is also sufficient.

Theorem 2. *For a fixed $A \in \mathbb{K}^{r\times s}$ and a fixed $y \in \mathbb{K}^{r\times t}$, the system (10.6) is solvable, i.e., (10.6) admits a solution $x \in \mathbb{K}^{s\times t}$, iff the following compatibility condition holds:*

$$B\, y = 0. \tag{10.7}$$

Then, all the solutions of (10.6) are given by

$$\forall\, z \in \mathbb{K}^{u\times t}, \quad x = E\, y + C\, z, \tag{10.8}$$

where $C \in \mathbb{K}^{s\times u}$ is a matrix whose columns form a basis of $\ker_{\mathbb{K}}(A.) := \{\eta \in \mathbb{K}^{s\times 1} \mid A\,\eta = 0\}$, i.e., C is a full column rank matrix *(i.e., $C\,\theta = 0$ yields $\theta = 0$) which satisfies*

$$\ker_{\mathbb{K}}(A.) = \mathrm{im}_{\mathbb{K}}(C.) := C\, \mathbb{K}^{u\times 1},$$

and $E \in \mathbb{K}^{s\times r}$ is a generalized inverse *of A, namely:*

$$A\, E\, A = A.$$

Proof 1. *By definition of B and C, we have the following exact sequences of $\mathbb{K}$-vector spaces*

$$0 \longrightarrow \mathbb{K}^{1\times q} \xrightarrow{.B} \mathbb{K}^{1\times r} \xrightarrow{.A} \mathbb{K}^{1\times s},$$

$$\mathbb{K}^{r\times 1} \xleftarrow{A.} \mathbb{K}^{s\times 1} \xleftarrow{C.} \mathbb{K}^{u\times 1} \longleftarrow 0,$$

where $(.A)(\lambda) := \lambda\, A$ for all $\lambda \in \mathbb{K}^{1\times r}$, $(A.)(\eta) := A\,\eta$ for all $\eta \in \mathbb{K}^{s\times 1}$, and similarly with B and C. Since the functor $\hom_{\mathbb{K}}(\,\cdot\,, \mathbb{K})$ (duality) is exact, we get the following long exact sequences of $\mathbb{K}$-vector spaces:

$$0 \longrightarrow \mathbb{K}^{1\times q} \xrightarrow{.B} \mathbb{K}^{1\times r} \xrightarrow{.A} \mathbb{K}^{1\times s} \xrightarrow{.C} \mathbb{K}^{1\times u} \longrightarrow 0,$$

$$0 \longleftarrow \mathbb{K}^{q\times 1} \xleftarrow{B.} \mathbb{K}^{r\times 1} \xleftarrow{A.} \mathbb{K}^{s\times 1} \xleftarrow{C.} \mathbb{K}^{u\times 1} \longleftarrow 0. \tag{10.9}$$

The Euler-Poincaré characteristic then yields:

$$q - r + s - u = 0.$$

Moreover, since $\mathbb{K}^{t\times 1}\otimes_{\mathbb{K}}\cdot$ *and* $\cdot\otimes_{\mathbb{K}}\mathbb{K}^{1\times t}$ *are two exact functors, we obtain the following long exact sequences:*

$$0\longrightarrow \mathbb{K}^{t\times q}\xrightarrow{.B}\mathbb{K}^{t\times r}\xrightarrow{.A}\mathbb{K}^{t\times s}\xrightarrow{.C}\mathbb{K}^{t\times u}\longrightarrow 0,$$

$$0\longleftarrow \mathbb{K}^{q\times t}\xleftarrow{B.}\mathbb{K}^{r\times t}\xleftarrow{A.}\mathbb{K}^{s\times t}\xleftarrow{C.}\mathbb{K}^{u\times t}\longleftarrow 0. \tag{10.10}$$

The long exact sequence (10.10) shows that (10.6) is solvable iff $y\in \mathrm{im}_{\mathbb{K}^{1\times t}}(A.):=A\,\mathbb{K}^{s\times t}$*, i.e., iff* $B\,y=0$*. Since the long exact sequence (10.9)* splits *([76]), there exist matrices* $D\in\mathbb{K}^{r\times q}$*,* $E\in\mathbb{K}^{s\times r}$*, and* $F\in\mathbb{K}^{u\times s}$ *such that:*

$$\begin{gathered} B\,D=I_q,\ D\,B+A\,E=I_r,\ E\,A+C\,F=I_s,\ F\,C=I_u,\\ E\,D=0,\ F\,E=0. \end{gathered} \tag{10.11}$$

Then, E *is a generalized inverse of* A*, i.e.,* $A\,E\,A=A$*.*

Let us now suppose that $y\in\mathbb{K}^{r\times t}$ *is such that* $B\,y=0$*. Using the second identity of (10.11) and* $A\,C=0$*, we get* $y=A\,E\,y=A\,(E\,y+C\,z)$ *for all* $z\in\mathbb{K}^{u\times t}$*, which shows that (10.8) are solutions of (10.6). Finally, if there exists* $x^\star\in\mathbb{K}^{s\times t}$ *satisfying (10.6), i.e.,* $A\,x^\star=y$*, then we have*

$$A\,(x^\star-E\,y)=A\,x^\star-A\,E\,A\,x^\star=0,$$

which shows that $x^\star-E\,y\in\ker_{\mathbb{K}^{1\times t}}(A.)=C\,\mathbb{K}^{u\times t}$*. Thus, there exists* $z\in\mathbb{K}^{u\times t}$ *such that* $x^\star-E\,y=C\,z$*, which shows that all the solutions of (10.6) are of the form of (10.8).*

Appendix 2: Polynomial computation

We can define $\mathcal{C}\begin{pmatrix}\mathbf{c}_R\\ \mathbf{c}_I\\ \mathbf{m}_R\\ \mathbf{m}_I\end{pmatrix}=\mathcal{C}(\vec{X})$, and $\vec{\nabla}\mathcal{C}$ the previously computed gradient.

When some little variations $d\vec{X}$ are introduced in the cost function, we have

$$(\vec{X}+d\vec{X})=\min_{\alpha}\mathcal{C}\left(\vec{X}+\alpha\vec{\nabla}\mathcal{C}\right).$$

This new approach allows another formulation of the problem, where the optimal descent step is the solution of the fourth degree polynomial.

$$\begin{aligned}\left|\left|\mathbf{M}-(\,u+\alpha\vec{\nabla}u\,)(\,v+\alpha\vec{\nabla}v\,)^*\right|\right|_{Fro}^2 &= ||\mathbf{M}_R||^2+||\mathbf{M}_I||^2\\ &\quad-2\left(\mathbf{m}_R+\alpha\vec{\nabla}v_1\right)^*\mathbf{M}_R^*\left(\mathbf{c}_R+\alpha\vec{\nabla}u_1\right)\\ &\quad-2\left(\mathbf{m}_I+\alpha\vec{\nabla}v_2\right)^*\mathbf{M}_R^*\left(\mathbf{c}_I+\alpha\vec{\nabla}u_2\right)\\ &\quad+2\left(\mathbf{m}_I+\alpha\vec{\nabla}v_2\right)^*\mathbf{M}_I^*\left(\mathbf{c}_R+\alpha\vec{\nabla}u_1\right)\\ &\quad-2\left(\mathbf{m}_R+\alpha\vec{\nabla}v_1\right)^*\mathbf{M}_I^*\left(\mathbf{c}_I+\alpha\vec{\nabla}u_2\right)\\ &\quad+\left|\left|\mathbf{c}_R+\alpha\vec{\nabla}v_1\right|\right|^2\left|\left|\mathbf{m}_R+\alpha\vec{\nabla}v_1\right|\right|^2\\ &\quad+\left|\left|\mathbf{c}_R+\alpha\vec{\nabla}v_1\right|\right|^2\left|\left|\mathbf{m}_I+\alpha\vec{\nabla}v_2\right|\right|^2\\ &\quad+\left|\left|\mathbf{c}_I+\alpha\vec{\nabla}u_2\right|\right|^2\left|\left|\mathbf{m}_R+\alpha\vec{\nabla}v_1\right|\right|^2\\ &\quad+\left|\left|\mathbf{c}_I+\alpha\vec{\nabla}u_2\right|\right|^2\left|\left|\mathbf{m}_I+\alpha\vec{\nabla}v_2\right|\right|^2\end{aligned}$$

All the computation steps are going to be developped separately below, in order to clarify the process.

$$\begin{aligned}-2\left(\mathbf{m}_R+\alpha\vec{\nabla}v_1\right)^*\mathbf{M}_R^*\left(\mathbf{c}_R+\alpha\vec{\nabla}u_1\right) &= -2\left(\mathbf{m}_R^*\mathbf{M}_R^*\mathbf{c}_R+\mathbf{m}_R^*\mathbf{M}_R^*\alpha\vec{\nabla}u_1\right.\\ &\quad\left.+\,\alpha\vec{\nabla}v_1^*\mathbf{M}_R^*\mathbf{c}_R+\alpha\vec{\nabla}v_1^*\mathbf{M}_R^*\alpha\vec{\nabla}u_1\right)\\ &= -2\left(\alpha^2\vec{\nabla}v_1^*\mathbf{M}_R^*\vec{\nabla}u_1+\alpha\left(\mathbf{m}_R^*\mathbf{M}_R^*\vec{\nabla}u_1+\vec{\nabla}v_1^*\mathbf{M}_R^*\mathbf{c}_R\right)\right.\\ &\quad\left.+\,\mathbf{m}_R^*\mathbf{M}_R^*\mathbf{c}_R\right)\end{aligned}$$

$$\begin{aligned}-2\left(\mathbf{m}_I+\alpha\vec{\nabla}v_2\right)^*\mathbf{M}_R^*\left(\mathbf{c}_I+\alpha\vec{\nabla}u_2\right) &= -2\left(\mathbf{m}_I^*\mathbf{M}_R^*\mathbf{c}_I+\mathbf{m}_I^*\mathbf{M}_R^*\alpha\vec{\nabla}u_2\right.\\ &\quad\left.+\,\alpha\vec{\nabla}v_2^*\mathbf{M}_R^*\mathbf{m}_R+\alpha\vec{\nabla}v_2^*\mathbf{M}_R^*\alpha\vec{\nabla}u_2\right)\\ &= -2\left(\alpha^2\vec{\nabla}v_2^*\mathbf{M}_R^*\vec{\nabla}u_2+\alpha\left(\mathbf{m}_I^*\mathbf{M}_R^*\vec{\nabla}u_2+\vec{\nabla}v_2^*\mathbf{M}_R^*\mathbf{c}_I\right)\right.\\ &\quad\left.+\,\mathbf{m}_I^*\mathbf{M}_R^*\mathbf{c}_I\right)\end{aligned}$$

$$\begin{aligned}+2\left(\mathbf{m}_I+\alpha\vec{\nabla}v_2\right)^*\mathbf{M}_I^*\left(\mathbf{c}_R+\alpha\vec{\nabla}u_1\right) &= +2\left(\mathbf{m}_I^*\mathbf{M}_I^*\mathbf{c}_R+\mathbf{m}_I^*\mathbf{M}_I^*\alpha\vec{\nabla}u_1\right.\\ &\quad\left.+\,\alpha\vec{\nabla}v_2^*\mathbf{M}_I^*\mathbf{c}_R+\alpha\vec{\nabla}v_2^*\mathbf{M}_I^*\alpha\vec{\nabla}u_1\right)\\ &= +2\left(\alpha^2\vec{\nabla}v_2^*\mathbf{M}_I^*\vec{\nabla}u_1+\alpha\left(\mathbf{m}_I^*\mathbf{M}_I^*\vec{\nabla}u_1+\vec{\nabla}v_2^*\mathbf{M}_I^*\mathbf{c}_R\right)\right.\\ &\quad\left.+\,\mathbf{m}_I^*\mathbf{M}_I^*\mathbf{c}_R\right)\end{aligned}$$

$$
\begin{aligned}
-2\left(\mathbf{m}_R+\alpha\vec{\nabla}v_1\right)^*\mathbf{M}_I^*\left(\mathbf{c}_I+\alpha\vec{\nabla}u_2\right) &= -2\left(\mathbf{m}_R^*\mathbf{M}_I^*\mathbf{c}_I+\mathbf{m}_R^*\mathbf{M}_I^*\alpha\vec{\nabla}u_2\right.\\
&\quad\left.+\,\alpha\vec{\nabla}v_1^*\mathbf{M}_I^*\mathbf{c}_I+\alpha\vec{\nabla}v_1^*\mathbf{M}_I^*\alpha\vec{\nabla}u_2\right)\\
&= -2\left(\alpha^2\vec{\nabla}v_1^*\mathbf{M}_I^*\vec{\nabla}u_2+\alpha\left(\mathbf{m}_R^*\mathbf{M}_I^*\vec{\nabla}u_2+\vec{\nabla}v_1^*\mathbf{M}_I^*\mathbf{c}_I\right)\right.\\
&\quad\left.+\,\mathbf{m}_R^*\mathbf{M}_I^*\mathbf{c}_I\right)
\end{aligned}
$$

$$
\begin{aligned}
\left|\left|\mathbf{c}_R+\alpha\vec{\nabla}v_1\right|\right|^2\left|\left|\mathbf{m}_R+\alpha\vec{\nabla}v_1\right|\right|^2 &= \left(||\mathbf{c}_R||^2+\alpha^2\left|\left|\vec{\nabla}u_1\right|\right|^2+2\alpha\mathbf{c}_R^*\vec{\nabla}u_1\right)\\
&\quad\times\left(||\mathbf{m}_R||^2+\alpha^2\left|\left|\vec{\nabla}v_1\right|\right|^2+2\alpha\mathbf{m}_R^*\vec{\nabla}v_1\right)\\
&= ||\mathbf{c}_R||^2||\mathbf{m}_R||^2+\alpha^2||\mathbf{c}_R||^2\left|\left|\vec{\nabla}v_1\right|\right|^2\\
&\quad+2\alpha||\mathbf{c}_R||^2v_1^*\vec{\nabla}v_1+\alpha^2\left|\left|\vec{\nabla}u_1\right|\right|^2||\mathbf{m}_R||^2\\
&\quad+\alpha^4\left|\left|\vec{\nabla}u_1\right|\right|^2\left|\left|\vec{\nabla}v_1\right|\right|^2+2\alpha^3\left|\left|\vec{\nabla}u_1\right|\right|^2v_1^*\vec{\nabla}v_1\\
&\quad+2\alpha||\mathbf{m}_R||^2u_1^*\vec{\nabla}u_1+4\alpha^2u_1^*\vec{\nabla}u_1\mathbf{m}_R^*\vec{\nabla}v_1\\
&= \alpha^4\left|\left|\vec{\nabla}u_1\right|\right|^2\left|\left|\vec{\nabla}v_1\right|\right|^2+2\alpha^3\left|\left|\vec{\nabla}u_1\right|\right|^2v_1^*\vec{\nabla}v_1\\
&\quad+\alpha^2\left(||\mathbf{c}_R||^2\left|\left|\vec{\nabla}v_1\right|\right|^2+\left|\left|\vec{\nabla}u_1\right|\right|^2||\mathbf{m}_R||^2+4u_1^*\vec{\nabla}u_1\mathbf{m}_R^*\vec{\nabla}v_1\right)\\
&\quad+\alpha\left(2||\mathbf{c}_R||^2v_1^*\vec{\nabla}v_1+2||\mathbf{m}_R||^2u_1^*\vec{\nabla}u_1\right)\\
&\quad+||\mathbf{c}_R||^2||\mathbf{m}_R||^2
\end{aligned}
$$

$$
\begin{aligned}
\left|\left|\mathbf{c}_R+\alpha\vec{\nabla}u_1\right|\right|^2\left|\left|\mathbf{m}_I+\alpha\vec{\nabla}v_2\right|\right|^2 &= \left(||\mathbf{c}_R||^2+\alpha^2\left|\left|\vec{\nabla}u_1\right|\right|^2+2\alpha\mathbf{c}_R^*\vec{\nabla}u_1\right)\\
&\quad\times\left(||\mathbf{m}_I||^2+\alpha^2\left|\left|\vec{\nabla}v_2\right|\right|^2+2\alpha\mathbf{m}_I^*\vec{\nabla}v_2\right)\\
&= ||\mathbf{c}_R||^2||\mathbf{m}_I||^2+\alpha^2||\mathbf{c}_R||^2\left|\left|\vec{\nabla}v_2\right|\right|^2\\
&\quad+2\alpha||\mathbf{c}_R||^2v_2^*\vec{\nabla}v_2+\alpha^2\left|\left|\vec{\nabla}u_1\right|\right|^2||\mathbf{m}_I||^2\\
&\quad+\alpha^4\left|\left|\vec{\nabla}u_1\right|\right|^2\left|\left|\vec{\nabla}v_2\right|\right|^2+2\alpha^3\left|\left|\vec{\nabla}u_1\right|\right|^2v_2^*\vec{\nabla}v_2\\
&\quad+2\alpha||\mathbf{m}_I||^2u_1^*\vec{\nabla}u_1+4\alpha^2u_1^*\vec{\nabla}u_1\mathbf{m}_I^*\vec{\nabla}v_2\\
&= \alpha^4\left|\left|\vec{\nabla}u_1\right|\right|^2\left|\left|\vec{\nabla}v_2\right|\right|^2+2\alpha^3\left|\left|\vec{\nabla}u_1\right|\right|^2v_2^*\vec{\nabla}v_2\\
&\quad+\alpha^2\left(||\mathbf{c}_R||^2\left|\left|\vec{\nabla}v_2\right|\right|^2+\left|\left|\vec{\nabla}u_1\right|\right|^2||\mathbf{m}_I||^2+4u_1^*\vec{\nabla}u_1\mathbf{m}_I^*\vec{\nabla}v_2\right)\\
&\quad+\alpha\left(2||\mathbf{c}_R||^2v_2^*\vec{\nabla}v_2+2||\mathbf{m}_I||^2u_1^*\vec{\nabla}u_1\right)\\
&\quad+||\mathbf{c}_R||^2||\mathbf{m}_I||^2
\end{aligned}
$$

$$\begin{aligned}
\left|\left|\mathbf{c}_I+\alpha\vec{\nabla}u_2\right|\right|^2\left|\left|\mathbf{m}_R+\alpha\vec{\nabla}v_1\right|\right|^2 &= \left(||\mathbf{c}_I||^2+\alpha^2\left|\left|\vec{\nabla}u_2\right|\right|^2+2\alpha\mathbf{c}_I^*\vec{\nabla}u_2\right)\\
&\quad\times\left(||\mathbf{m}_R||^2+\alpha^2\left|\left|\vec{\nabla}v_1\right|\right|^2+2\alpha\mathbf{m}_R^*\vec{\nabla}v_1\right)\\
&= ||\mathbf{c}_I||^2||\mathbf{m}_R||^2+\alpha^2||\mathbf{c}_I||^2\left|\left|\vec{\nabla}v_1\right|\right|^2\\
&\quad+2\alpha||\mathbf{c}_I||^2v_1^*\vec{\nabla}v_1+\alpha^2\left|\left|\vec{\nabla}u_2\right|\right|^2||\mathbf{m}_R||^2\\
&\quad+\alpha^4\left|\left|\vec{\nabla}u_2\right|\right|^2\left|\left|\vec{\nabla}v_1\right|\right|^2+2\alpha^3\left|\left|\vec{\nabla}u_2\right|\right|^2v_1^*\vec{\nabla}v_1\\
&\quad+2\alpha||\mathbf{m}_R||^2u_2^*\vec{\nabla}u_2+4\alpha^2u_2^*\vec{\nabla}u_2\mathbf{m}_R^*\vec{\nabla}v_1\\
&= \alpha^4\left|\left|\vec{\nabla}u_2\right|\right|^2\left|\left|\vec{\nabla}v_1\right|\right|^2+2\alpha^3\left|\left|\vec{\nabla}u_2\right|\right|^2v_1^*\vec{\nabla}v_1\\
&\quad+\alpha^2\left(||\mathbf{c}_I||^2\left|\left|\vec{\nabla}v_1\right|\right|^2+\left|\left|\vec{\nabla}u_2\right|\right|^2||\mathbf{m}_R||^2+4u_2^*\vec{\nabla}u_2\mathbf{m}_R^*\vec{\nabla}v_1\right)\\
&\quad+\alpha\left(2||\mathbf{c}_I||^2v_1^*\vec{\nabla}v_1+2||\mathbf{m}_R||^2u_2^*\vec{\nabla}u_2\right)\\
&\quad+||\mathbf{c}_I||^2||\mathbf{m}_R||^2
\end{aligned}$$

$$\begin{aligned}
\left|\left|\mathbf{c}_I+\alpha\vec{\nabla}u_2\right|\right|^2\left|\left|\mathbf{m}_I+\alpha\vec{\nabla}v_2\right|\right|^2 &= \left(||\mathbf{c}_I||^2+\alpha^2\left|\left|\vec{\nabla}u_2\right|\right|^2+2\alpha\mathbf{c}_I^*\vec{\nabla}u_2\right)\\
&\quad\times\left(||\mathbf{m}_I||^2+\alpha^2\left|\left|\vec{\nabla}v_2\right|\right|^2+2\alpha\mathbf{m}_I^*\vec{\nabla}v_2\right)\\
&= ||\mathbf{c}_I||^2||\mathbf{m}_I||^2+\alpha^2||\mathbf{c}_I||^2\left|\left|\vec{\nabla}v_2\right|\right|^2\\
&\quad+2\alpha||\mathbf{c}_I||^2v_2^*\vec{\nabla}v_2+\alpha^2\left|\left|\vec{\nabla}u_2\right|\right|^2||\mathbf{m}_I||^2\\
&\quad+\alpha^4\left|\left|\vec{\nabla}u_2\right|\right|^2\left|\left|\vec{\nabla}v_2\right|\right|^2+2\alpha^3\left|\left|\vec{\nabla}u_2\right|\right|^2v_2^*\vec{\nabla}v_2\\
&\quad+2\alpha||\mathbf{m}_I||^2u_2^*\vec{\nabla}u_2+4\alpha^2u_2^*\vec{\nabla}u_2\mathbf{m}_I^*\vec{\nabla}v_2\\
&= \alpha^4\left|\left|\vec{\nabla}u_2\right|\right|^2\left|\left|\vec{\nabla}v_2\right|\right|^2+2\alpha^3\left|\left|\vec{\nabla}u_2\right|\right|^2v_2^*\vec{\nabla}v_2\\
&\quad+\alpha^2\left(||\mathbf{c}_I||^2\left|\left|\vec{\nabla}v_2\right|\right|^2+\left|\left|\vec{\nabla}u_2\right|\right|^2||\mathbf{m}_I||^2+4u_2^*\vec{\nabla}u_2\mathbf{m}_I^*\vec{\nabla}v_2\right)\\
&\quad+\alpha\left(2||\mathbf{c}_I||^2v_2^*\vec{\nabla}v_2+2||\mathbf{m}_I||^2u_2^*\vec{\nabla}u_2\right)\\
&\quad+||\mathbf{c}_I||^2||\mathbf{m}_I||^2
\end{aligned}$$

When combining all the previous intermediary steps, the problem can be explained as below:

$$
\begin{aligned}
\left|\left|\mathbf{M}-(u+\alpha\vec{\nabla}u)(v+\alpha\vec{\nabla}v)^*\right|\right|_{Fro}^2 = \alpha^4 &\left(\left|\left|\vec{\nabla}u_1\right|\right|^2\left|\left|\vec{\nabla}v_1\right|\right|^2+\left|\left|\vec{\nabla}u_1\right|\right|^2\left|\left|\vec{\nabla}v_2\right|\right|^2\right.\\
&\left.+\left|\left|\vec{\nabla}u_2\right|\right|^2\left|\left|\vec{\nabla}v_1\right|\right|^2+\left|\left|\vec{\nabla}u_2\right|\right|^2\left|\left|\vec{\nabla}v_2\right|\right|^2\right)\\
+\alpha^3 2&\left(\left|\left|\vec{\nabla}u_1\right|\right|^2 v_1^*\vec{\nabla}v_1+\left|\left|\vec{\nabla}u_1\right|\right|^2 v_2^*\vec{\nabla}v_2\right.\\
&\left.+\left|\left|\vec{\nabla}u_2\right|\right|^2 v_1^*\vec{\nabla}v_1+\left|\left|\vec{\nabla}u_2\right|\right|^2 v_2^*\vec{\nabla}v_2\right)\\
+\alpha^2&\left(-2\vec{\nabla}v_1^*\mathbf{M}_R^*\vec{\nabla}u_1-2\vec{\nabla}v_2^*\mathbf{M}_R^*\vec{\nabla}u_2\right.\\
&+2\vec{\nabla}v_2^*\mathbf{M}_I^*\vec{\nabla}u_1-2\vec{\nabla}v_1^*\mathbf{M}_I^*\vec{\nabla}u_2\\
&+||\mathbf{c}_R||^2\left|\left|\vec{\nabla}v_1\right|\right|^2+\left|\left|\vec{\nabla}u_1\right|\right|^2||\mathbf{m}_R||^2+4u_1^*\vec{\nabla}u_1\mathbf{m}_R^*\vec{\nabla}v_1\\
&+||\mathbf{c}_R||^2\left|\left|\vec{\nabla}v_2\right|\right|^2+\left|\left|\vec{\nabla}u_1\right|\right|^2||\mathbf{m}_I||^2+4u_1^*\vec{\nabla}u_1\mathbf{m}_I^*\vec{\nabla}v_2\\
&+||\mathbf{c}_I||^2\left|\left|\vec{\nabla}v_1\right|\right|^2+\left|\left|\vec{\nabla}u_2\right|\right|^2||\mathbf{m}_R||^2+4u_2^*\vec{\nabla}u_2\mathbf{m}_R^*\vec{\nabla}v_1\\
&\left.+||\mathbf{c}_I||^2\left|\left|\vec{\nabla}v_2\right|\right|^2+\left|\left|\vec{\nabla}u_2\right|\right|^2||\mathbf{m}_I||^2+4u_2^*\vec{\nabla}u_2\mathbf{m}_I^*\vec{\nabla}v_2\right)\\
+\alpha 2&\left(-\mathbf{m}_R^*\mathbf{M}_R^*\vec{\nabla}u_1-\vec{\nabla}v_1^*\mathbf{M}_R^*\mathbf{c}_R-\mathbf{m}_I^*\mathbf{M}_R^*\vec{\nabla}u_2-\vec{\nabla}v_2^*\mathbf{M}_R^*\mathbf{c}_I\right.\\
&+\mathbf{m}_I^*\mathbf{M}_I^*\vec{\nabla}u_1+\vec{\nabla}v_2^*\mathbf{M}_I^*\mathbf{c}_R-\mathbf{m}_R^*\mathbf{M}_I^*\vec{\nabla}u_2-\vec{\nabla}v_1^*\mathbf{M}_I^*\mathbf{c}_I\\
&+||\mathbf{c}_R||^2 v_1^*\vec{\nabla}v_1+||\mathbf{m}_R||^2 u_1^*\vec{\nabla}u_1\\
&+||\mathbf{c}_R||^2 v_2^*\vec{\nabla}v_2+||\mathbf{m}_I||^2 u_1^*\vec{\nabla}u_1\\
&+||\mathbf{c}_I||^2 v_1^*\vec{\nabla}v_1+||\mathbf{m}_R||^2 u_2^*\vec{\nabla}u_2\\
&\left.+||\mathbf{c}_I||^2 v_2^*\vec{\nabla}v_2+||\mathbf{m}_I||^2 u_2^*\vec{\nabla}u_2\right)\\
&+||\mathbf{M}_R||^2+||\mathbf{M}_I||^2\\
&-2v_1^*\mathbf{M}_R^*\mathbf{c}_R-2v_2^*\mathbf{M}_R^*\mathbf{c}_I+2v_2^*\mathbf{M}_I^*\mathbf{c}_R-2v_1^*\mathbf{M}_I^*\mathbf{c}_I\\
&+||\mathbf{c}_R||^2||\mathbf{m}_R||^2+||\mathbf{c}_R||^2||\mathbf{m}_I||^2\\
&+||\mathbf{c}_I||^2||\mathbf{m}_R||^2+||\mathbf{c}_I||^2||\mathbf{m}_I||^2
\end{aligned}
$$

Bibliography

[1] Antoni, J. (2006). The spectral kurtosis: A useful tool for characterising non-stationary signals. *Mechanical Systems and Signal Processing*, 20(2):282–307.

[2] Antoni, J. and Randall, R. B. (2006). The spectral kurtosis: Application to the vibratory surveillance and diagnostics of rotating machines. *Mechanical Systems and Signal Processing*, 20(2):308–331.

[3] Aragonda, H. and Seelamantula, C. S. (2015). Demodulation of narrowband speech spectrograms using the riesz transform. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1824–1834.

[4] BADAOUI, M., ANTONI, J., GUILLET, F., DANIÈRE, J., and VELEX, P. (2001). Use of the Moving Cepstrum Integral To Detect and Localise Tooth Spalls in Gears. *Mechanical Systems and Signal Processing*, 15(5):873–885.

[5] Barszcz, T. and Randall, R. B. (2009). Application of spectral kurtosis for detection of a tooth crack in the planetary gear of a wind turbine. *Mechanical Systems and Signal Processing*, 23(4):1352–1365.

[6] Blunt, D. and Keller, J. (2006). Detection of a fatigue crack in a UH-60A planet gear carrier using vibration analysis. *Mechanical Systems and Signal Processing*, 20(8):2095–2111.

[7] Bogert, B., Healy, M. J. R., and Tukey, J. (1963). The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross- Cepstrum and Saphe Cracking. In *Proceedings of the Symposium on Time Series Analysis*.

[8] Bonnardot, F. (2004). *Comparaison entre les analyses angulaire et temporelle des signaux vibratoires de machines tournantes. Etude du concept de cyclostationnarité floue*. PhD thesis.

[9] Borghesani, P., Pennacchi, P., and Chatterton, S. (2014). The relationship between kurtosis- and envelope-based indexes for the diagnostic of rolling element bearings. *Mechanical Systems and Signal Processing*, 43(1-2):25–43.

[10] Boungou, D., Guillet, F., El Badaoui, M., Lyonnet, P., and Rosario, T. (2015). Fatigue damage detection using cyclostationarity. *Mechanical Systems and Signal Processing*, 58:128–142.

[11] Braun, S. (2011). The synchronous (time domain) average revisited. *Mechanical Systems and Signal Processing*, 25(4):1087–1102.

[12] Cai, Y., He, Y., Li, A., Zhao, J., and Wang, T. (2010). Application of Wavelet to Gearbox Vibration Signals for Fault Detection. *Advanced Computer Control (ICACC), 2010 2nd International Conference on*, 192:441–444.

[13] Capdessus, C. (1992). *Aide au diagnostic des machines tournantes par traitement du signal (Aid to the diagnosis of rotating machines by signal processing)*. PhD thesis.

[14] Chen, Z. and Shao, Y. (2011). Dynamic simulation of spur gear with tooth root crack propagating along tooth width and crack depth. *Engineering Failure Analysis*, 18(8):2149–2164.

[15] Cheon, G.-J. and Parker, R. G. (2004). Influence of manufacturing errors on the dynamic characteristics of planetary gear systems. *KSME international journal*, 18(4):606–621.

[16] Cohen, L. (1995). *Time-frequency analysis*, volume 778. Prentice Hall PTR Englewood Cliffs, NJ:.

[17] Cover, T. M., Hart, P. E., et al. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

[18] Cramér, H. (1946). *Mathematical methods of statistics*. NewYork: Princeton Univ. Press.

[19] D'Elia, G., Mucchi, E., and Cocconcelli, M. (2017). On the identification of the angular position of gears for the diagnostics of planetary gearboxes. *Mechanical Systems and Signal Processing*, 83:305–320.

[20] D'Elia, G., Mucchi, E., and Dalpiaz, G. (2013). On the Time Synchronous Average in Planetary Gearboxes. *Surveillance 7 International Conference*, pages 1–11.

[21] Djurović, I., Wang, P., Simeunović, M., and Orlik, P. V. (2018). Parameter estimation of coupled polynomial phase and sinusoidal fm signals. *Signal Processing*, 149:1–13.

[22] El Badaoui, M., Guillet, F., and Daniere, J. (2004). New applications of the real cepstrum to gear signals, including definition of a robust fault indicator. *Mechanical Systems and Signal Processing*, 18(5):1031–1046.

[23] Endo, H. and Randall, R. (2007). Enhancement of autoregressive model based gear tooth fault detection technique by the use of minimum entropy deconvolution filter. *Mechanical Systems and Signal Processing*, 21(2):906–919.

[24] Endo, H., Randall, R. B., and Gosselin, C. (2009). Differential diagnosis of spall vs. cracks in the gear tooth fillet region: Experimental validation. *Mechanical Systems and Signal Processing*, 23(3):636–651.

[25] Feng, Z., Chu, F., and Zuo, M. J. (2011). Time–frequency analysis of time-varying modulated signals based on improved energy separation by iterative generalized demodulation. *Journal of Sound and Vibration*, 330(6):1225–1243.

[26] Feng, Z. and Zuo, M. J. (2012). Vibration signal models for fault diagnosis of planetary gearboxes. *Journal of Sound and Vibration*, 331(22):4919–4939.

[27] Forrester, B. D. (1989). Use of the Wigner-Ville distribution in helicopter transmission fault detection. In *Proceedings of the Australian symposium on signal processing and applications (ASSP), Adelaide, Australia*.

[28] Frazier, R., Samsam, S., Braida, L., and Oppenheim, A. (1976). Enhancement of speech by adaptive filtering. In *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 251–253. IEEE.

[29] Friedlander, B. and Francos, J. M. (1995). Estimation of amplitude and phase parameters of multicomponent signals. *IEEE Transactions on Signal Processing*, 43(4):917–926.

[30] Gianfelici, F., Biagetti, G., Crippa, P., and Turchetti, C. (2005). Am-fm decomposition of speech signals: an asymptotically exact approach based on the iterated hilbert transform. In *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, pages 333–338. IEEE.

[31] Gianfelici, F., Biagetti, G., Crippa, P., and Turchetti, C. (2007). Multicomponent am–fm representations: an asymptotically exact approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):823–837.

[32] Grossmann, A. and Morlet, J. (1984). Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM journal on mathematical analysis*, 15(4):723–736.

[33] Gu, X. Y. and Velex, P. (2011). A lumped parameter model to analyse the dynamic load sharing in planetary gears with planet errors. In *Applied Mechanics and Materials*, volume 86, pages 374–379. Trans Tech Publ.

[34] He, D., Wang, X., Li, S., Lin, J., and Zhao, M. (2016). Identification of multiple faults in rotating machinery based on minimum entropy deconvolution combined with spectral kurtosis. *Mechanical Systems and Signal Processing*, 81:235–249.

[35] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 454, pages 903–995. The Royal Society.

[36] Kahraman, A. (1994). Load sharing characteristics of planetary transmissions. *Mechanism and Machine Theory*, 29(8):1151–1165.

[37] Kaiser, J. F. (1990). On a simple algorithm to calculate the'energy'of a signal. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 381–384. IEEE.

[38] Kaplan, A. and Haenlein, M. (2019). Siri, siri, in my hand: WhoâĂŹs the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25.

[Kay] Kay, S. M. *Fundamentals of statistical signal processing: Estimation theory*. Prentice Hall PTR Upper Saddle River, NJ:.

[40] Khawaja, T. S., Georgoulas, G., and Vachtsevanos, G. (2008). An efficient novelty detector for online fault diagnosis based on least squares support vector machines. In *2008 IEEE AUTOTESTCON*, pages 202–207. IEEE.

[41] Khazaee, M., Ahmadi, H., Omid, M., and Moosavian, A. (2012). An appropriate approach for condition monitoring of planetary gearbox based on fast fourier transform and least-square support vector machine. *International Journal of Multidisciplinary Sciences and Engineering*, 3(5):22–26.

[42] Koukoura, S., Carroll, J., Weiss, S., and McDonald, A. (2017). Wind turbine gearbox vibration signal signature and fault development through time. In *Signal Processing Conference (EUSIPCO), 2017 25th European*, pages 1380–1384. IEEE.

[43] Lejeune, G., Lacoume, J.-L., Marchand, P., Durnerin, M., Martin, N., Liénard, J., Silvent, A., Mailhes, C., Castanié, F., Prieur, P., and Goullet, G. (1997). Cyclostationnarités d'ordre 1 et 2 : application à des signaux vibratoires d'engrenages. *Seizième Colloque Gretsi*, pages 323–326.

[44] Li, S. and Kahraman, A. (2013). A tribo-dynamic model of a spur gear pair. *Journal of Sound and Vibration*, 332(20):4963–4978.

[45] Li, Z., Yan, X., Tian, Z., Yuan, C., Peng, Z., and Li, L. (2013). Blind vibration component separation and nonlinear feature extraction applied to the nonstationary vibration signals for the gearbox multi-fault diagnosis. *Measurement*, 46(1):259–271.

[46] Liang, X., Zuo, M. J., and Hoseini, M. R. (2015). Vibration signal modeling of a planetary gear set for tooth crack detection. *Engineering Failure Analysis*, 48:185–200.

[47] Ligata, H., Kahraman, A., and Singh, A. (2008). An experimental study of the influence of manufacturing errors on the planetary gear stresses and planet load sharing. *Journal of Mechanical Design*, 130(4):041701.

[48] Lin, J. and Parker, R. G. (1999). Analytical characterization of the unique properties of planetary gear free vibration. *Journal of vibration and acoustics*, 121(3):316–321.

[49] LIN, J. and ZUO, M. (2003). Gearbox Fault Diagnosis Using Adaptive Wavelet Filter. *Mechanical Systems and Signal Processing*, 17(6):1259–1269.

[50] Liu, R., Yang, B., Zio, E., and Chen, X. (2018). Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing*, 108:33–47.

[51] Mallat, S. A. (2008). *A Wavelet Tour of Signal Processing*.

[52] Mark, W. D. (1977). Analysis of the vibration excitation of of gear systems: Basic theory. *Journal of Acoustical Society of America*, 63:1409–1430.

[53] Martin, H. (1989). Statistical moment analysis as a means of surface damage detection. In *Proceedings of the 7th International Modal Analysis Conference*, pages 1016–1021.

[54] McDonald, G. L. and Zhao, Q. (2017). Multipoint optimal minimum entropy deconvolution and convolution fix: Application to vibration fault detection. *Mechanical Systems and Signal Processing*, 82:461–477.

[55] McFadden, P. D. (1986). Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing vibration. *Journal of vibration, acoustics, stress, and reliability in design*, 108(2):165–170.

[56] McFadden, P. D. (1987). A revised model for the extraction of periodic waveforms by time domain averaging. *Mechanical Systems and Signal Processing*, 1(1):83–95.

[57] McFadden, P. D. and Smith, J. D. (1985). 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: Journal of Mechanical Engineering Science*, 199(1):65–70.

[58] Mehra, R. K. and Peschon, J. (1971). An innovations approach to fault detection and diagnosis in dynamic systems. *Automatica*, 7(5):637–640.

[59] Meyer, Y. (1990). Ondelettes et opérateurs I: Ondelettes (Hermann, Paris, 1990).

[60] Molina Vicuña, C. (2014). Vibration characteristics of single-stage planetary gear transmissions. *Ingeniare. Revista chilena de ingeniería*, 22:88–98.

[61] Molla, M. K. I. and Hirose, K. (2007). Single-mixture audio source separation by subspace decomposition of hilbert spectrum. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):893–900.

[62] Mouyan, Z., Zhenming, C., and Unbehauen, R. (1991). Separation of periodic signals by using an algebraic method. In *Circuits and Systems, 1991., IEEE International Sympoisum on*, pages 2427–2430. IEEE.

[63] Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848.

[64] Nguyen, P. H. and Kim, J.-M. (2015). Multifault diagnosis of rolling element bearings using a wavelet kurtogram and vector median-based feature analysis. *Shock and Vibration*, 2015.

[65] Özgüven, H. N. and Houser, D. R. (1988). Mathematical models used in gear dynamicsâĂŤa review. *Journal of sound and vibration*, 121(3):383–411.

[66] Palácios, R. H. C., da Silva, I. N., Goedtel, A., and Godoy, W. F. (2015). A comprehensive evaluation of intelligent classifiers for fault identification in three-phase induction motors. *Electric Power Systems Research*, 127:249–258.

[67] Parey, A., El Badaoui, M., Guillet, F., and Tandon, N. (2006). Dynamic modelling of spur gear pair and application of empirical mode decomposition-based statistical analysis for early detection of localized tooth defect. *Journal of sound and vibration*, 294(3):547–561.

[68] Parker, R., Vijayakar, S., and Imajo, T. (2000a). Non-linear dynamic response of a spur gear pair: modelling and experimental comparisons. *Journal of Sound and vibration*, 237(3):435–455.

[69] Parker, R. G., Agashe, V., and Vijayakar, S. M. (2000b). Dynamic response of a planetary gear system using a finite element/contact mechanics model. *Journal of Mechanical Design*, 122(3):304–310.

[70] Potamianos, A. and Maragos, P. (1994). A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation. *Signal Processing*, 37(1):95–120.

[71] Prieto, M. D., Cirrincione, G., Espinosa, A. G., Ortega, J. A., and Henao, H. (2013). Bearing fault detection by a novel condition-monitoring scheme based on statistical-time features and neural networks. *IEEE Transactions on Industrial Electronics*, 60(8):3398–3407.

[72] Prueter, P. E., Parker, R. G., and Cunliffe, F. (2011). A study of gear root strains in a multi-stage planetary wind turbine gear train using a three dimensional finite element/contact mechanics model and experiments. In *ASME Power Transmission and Gearing Conference, Washington, DC, p. DETC2011–47451*.

[73] Rafiee, J., Arvani, F., Harifi, A., and Sadeghi, M. (2007). Intelligent condition monitoring of a gearbox using artificial neural network. *Mechanical systems and signal processing*, 21(4):1746–1754.

[74] Randall, R. B. (1982). A new method of modeling gear faults. *Journal of Mechanical Design*, 104(2):259–267.

[75] Randall, R. B. (2016). A history of cepstrum analysis and its application to mechanical problems. *Mechanical Systems and Signal Processing*, pages 1–16.

[76] Rotman, J. J. (2008). *An introduction to homological algebra*. Springer Science & Business Media.

[77] Samuel, P. D. and Pines, D. J. (2005). A review of vibration-based techniques for helicopter transmission diagnostics. *Journal of Sound and Vibration*, 282(1-2):475–508.

[78] Santhanam, B. and Maragos, P. (2000). Multicomponent am-fm demodulation via periodicity-based algebraic separation and energy-based demodulation. *IEEE Transactions on Communications*, 48(3):473–490.

[79] Seshadrinath, J., Singh, B., and Panigrahi, B. K. (2014). Vibration analysis based interturn fault diagnosis in induction machines. *IEEE Transactions on Industrial Informatics*, 10(1):340–350.

[80] Shao, Y. and Mechefske, C. K. (2009). Gearbox vibration monitoring using extended kalman filters and hypothesis tests. *Journal of Sound and Vibration*, 325(3):629–648.

[81] Stewart, R. M. (1977). *Some Useful Data Analysis Techniques for Gearbox Diagnostics*. University of Southampton.

[82] Tallian, T. E. (1992). The failure atlas for hertz contact machine elements. *Mechanical Engineering*, 114(3):66.

[83] Toomer, G. J. (1998). Ptolemy's almagest. *Ptolemy's Almagest: translated into English and annotated by GJ Toomer. With a foreword by Owen Gingerich. Princeton: Princeton University Press, rev. ed., 712 pp., 219 figures, 38 tables; ISBN: 9780691002606*.

[84] Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer.

[85] Velex, P. and Maatar, M. (1996). A mathematical model for analyzing the influence of shape deviations and mounting errors on gear dynamic behaviour. *Journal of Sound and Vibration*, 191(5):629–660.

[86] Wang, D. (2016). K-nearest neighbors based methods for identification of different gear crack levels under different motor speeds and loads: Revisited. *Mechanical Systems and Signal Processing*, 70:201–208.

[87] Wang, W. and McFadden, P. (1996). Application of wavelets to gearbox vibration signals for fault detection. *Journal of sound and vibration*, 192(5):927–939.

[88] Wang, W. and Wong, A. K. (2002). Autoregressive model-based gear fault diagnosis. *Journal of vibration and acoustics*, 124(2):172–179.

[89] Wang, W. J. and McFadden, P. D. (1993a). Early detection of gear failure by vibration analysis–ii. interpretation of the time-frequency distribution using image processing techniques. *Mechanical Systems and Signal Processing*, 7(3):205–215.

[90] Wang, W. J. and McFadden, P. D. (1993b). Early detection of gear failure by vibration analysis i. calculation of the time-frequency distribution. *Mechanical Systems and Signal Processing*, 7(3):193–203.

[91] Wasserstein, R. L., Lazar, N. A., et al. (2016). The asaâĂŹs statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133.

[92] Wiggins, R. (1978). Minimum entropy deconvolution: Geoexploration.

[93] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

[94] Xiong, Y., Chen, S., Dong, X., Peng, Z., and Zhang, W. (2017). Accurate measurement in doppler radar vital sign detection based on parameterized demodulation. *IEEE Transactions on Microwave Theory and Techniques*, 65(11):4483–4492.

[95] Xue, S., Entwistle, R., Mazhar, I., and Howard, I. (2016). The spur planetary gear torsional stiffness and its crack sensitivity under quasi-static conditions. *Engineering Failure Analysis*, 63:106–120.

[96] Yesilyurt, I. (2004). The application of the conditional moments analysis to gearbox fault detection - A comparative study using the spectrogram and scalogram. *NDT and E International*, 37(4):309–320.

[97] Yu, Z., Sun, Y., and Jin, W. (2016). A novel generalized demodulation approach for multi-component signals. *Signal Processing*, 118:188–202.

[98] Yuksel, C. and Kahraman, A. (2004). Dynamic tooth loads of planetary gear sets having tooth profile wear. *Mechanism and Machine Theory*, 39(7):695–715.

[99] Zakarjsek, J. J., Townsend, D. P., and Decker, H. J. (1993). An analysis of gear fault detection methods as applied to pitting fatigue failure data. Technical report.

[100] Zakrajsek, J. (1994). A review of transmission diagnostic research at NASA Lewis research center. Technical Report December.