POLITECNICO DI MILANO

Facoltà di Ingegneria dei Sistemi

Corso di Laurea Specialistica in Ingegneria Matematica

Analysis of Doppler blood flow velocity in carotid arteries for the detection of atherosclerotic plaques

Relatore

Prof. Piercesare SECCHI

Correlatore

Dott.ssa Laura AZZIMONTI

Laureando

Pamela BURATTI (matr. 735303)

Anno Accademico 2010-2011

Abstract

In this thesis, a technique for nearly automated detection of Doppler flow velocity profile is developed. From this, it is possible to estimate the period of the velocity waveform throughout an innovative iterative method, based on Fourier smoothing and non linear least squares.
The estimate of the period enables the reconstruction of a function representing the velocity of the red blood cells, along the common carotid artery and at different distances from the bifurcation. Considering these functional data, it is possible to reduce the dimension of the problem performing a functional principal component analysis. The aim is that of exploring the blood flow along the common carotid artery, before the plaque is reached, and searching for features of the curve that indicate the presence of the plaque downstream through a linear discriminant analysis. This permits to compare the blood velocity field in healthy people with the blood flow of stenotic patients, who will undergo TEA surgery.

Sommario

In questa tesi viene sviluppata una tecnica per la rilevazione quasi automatica del profilo di velocità del flusso sanguigno fornito dall'esame ecodoppler. Una volta estratto il segnale, viene implementato un algoritmo innovativo per la stima del periodo del dato, che sfrutta l'iterazione di smoothing di Fourier e della stima ai minimi quadrati non lineare.
La stima del periodo consente la ricostruzione di una funzione che rappresenta la velocità dei globuli rossi, lungo la carotide comune e a diverse distanze dalla biforcazione. Si ha quindi a disposizione un insieme di dati funzionali, sui quali si applica un'analisi delle componenti principali con il fine di ridurre la dimensione del problema. L'obiettivo è quello di esplorare il flusso di sangue lungo la carotide comune, prima che la placca stenotica sia raggiunta. Quindi, attraverso un'analisi discriminante quadratica, si ricercano quelle caratteristiche della curva che indicano la presenza a valle della placca. Questo permette di confrontare il campo di velocità del sangue di persone sane con quello di pazienti malati. Infatti, individuare quali sono gli aspetti che maggiormente differenziano le due classi potrebbe sia avere fini diagnostici, sia contribuire a migliorare le simulazioni fluidodinamiche.

Contents

List of Figures

List of Tables

Introduction

The human carotid arteries, located on each side of the neck, have the key role of carrying blood to the head. They divide into an external branch supplying the neck, face and other external parts and an internal branch, supplying the brain, eye and other internal part. The carotid bifurcations and the internal carotid arteries are a preferred site of development of atherosclerotic plaques. The growth of a plaque could lead to the hardening of the wall of the vessels, but also to a stenosis, which can cause a lack of blood supply to the brain. Monitoring the carotid system is important in order to prevent stroke, one of the most serious cerebrovascular disease. In fact, over 60-70% of all ischemic cerebral infarctions are caused by arterial embolism [3], typically arising from a carotid artery affected by atherosclerosis. Thus, patients with atherosclerosis of the carotid system are at high risk of stroke. To reduce the risk, it is possible to resort to a highly effective surgical measure called TEA (Carotid Thromboendarterectomy), which is beneficial in individuals with symptomatic high-grade stenosis and, in certain cases, with 60-70% symptomatic stenosis.

To identify those patients who will benefit from TEA, suitable diagnostic procedures are necessary. Among these, the most preferred is the Color Doppler Ultrasound, because it is a totally non-invasive method. The ultrasound system runs in duplex mode: Doppler emissions are interleaved with B-mode emissions. The first provide information on the blood flow along the carotid artery, producing the so-called spectrogram, which shows the distribution of the velocities of the blood in the sample volume chosen within the probed area. B-mode emissions simultaneously form an image of the subcutaneous tissues, providing information on the location of the plaque, its morphology and its grade. Thus, through ultrasounds techniques, it is possible to get the information needed for surgery or for medical man-

agement of stenosis, without the need of additional invasive medical analysis.

This thesis is part of the MACAREN@MOX project (MAthematics for CARotids ENdarterectomy@MOX), a multidisciplinary research project which investigates the the formation of the plaque in carotid arteries by focusing on morphological and hemodynamic factors. The medical equipe of the research team is acquiring different analyses, both in patients who will undergo to the TEA surgical intervention and not. An important part of these analyses consists in Color Doppler ultrasounds images. One of the goals of the MACAREN@MOX project is to analyse this amount of data from a statistical point of view, with the important support of numerical modeling in order to deeply understand the evolution of the plaque.

This thesis addresses a portion of the statistical side of the project and it is aimed to study the Color Doppler Ultrasound signals from a statistical point of view. Doctors conventionally obtain information on blood flow by manual tracings of Doppler profiles. First of all, we develop a technique for nearly automated detection of the Doppler data. From this, we estimate the period of the cardiac cycle throughout an innovative iterative method, based on Fourier smoothing and non linear least squares. Once having an estimate of the period of the data, we are able to reconstruct a function representing the velocity of the red blood cells in time, along the common carotid artery and at different distances from the bifurcation. We will then carry out a functional data analysis, in order to compare the blood velocity field in healthy people and in people with a stenosis who will and will not undergo TEA. In fact, from Doppler Ultrasound obtained holding the transducer at the level of the stenosis, the presence of the plaque is obvious and the medical doctor has to consider its percentage and many other factors before deciding for surgery. Our aim, instead, is to explore the blood flow before the stenosis is reached and to search for features of the curve that indicate the presence of the plaque downstream. This research could be useful both for diagnostic purposes and for increasing the accuracy of numerical simulations, by giving more precise patient specific boundary conditions.
The work is organized according to the following pattern.

- Chapter 1 is a general introduction to all the themes treated in the thesis. First of all, we illustrate the MACAREN@MOX project, presenting the research team, the subjects of the research and its goals.

Then, a synthetic presentation of the carotid arteries and of the features of blood flow in these vessels is presented. In particular, we focus the attention on how a doctor can find the indicators of a stenosis and of the morphological features of the plaque, by observing the Doppler image. The chapter proceeds with the explanation of the data collection protocol, which refers not only to this thesis, but also to the whole MACAREN@MOX project. Instead, the illustration of the population refers only to this work, since it is only a part of the available data. The second part of the chapter is more technical. First, it is explained how to read a Doppler image and what the Doppler spectrum represents. Finally, the last part is a brief introduction to ultrasounds techniques used to estimate blood flow velocities.

- Chapter 2 focuses the attention on the spectrum frame. It is explained how to extract the region of interest, the filter applied in order to remove noise and the rescaling of the axes. The aim of this chapter is to estimate a function representing the velocity of red blood cell in time. This is done by extracting the 95^{th} sample quantile of the histogram of the velocity distribution, available at each time of the cardiac cycle. Since the signal is periodic, it is possible to estimate its period. In order to do this, a new algorithm is settled, which iteratively estimates the period through non linear least squares. Given the the estimate of the period value for each Doppler waveform, Fourier smoothing of the sample quantile is performed for each patients at various distances from the carotid bifurcation. Eventually, all the functions estimated are aligned using landmark registration.

- In Chapter 3, the functional data are explored. After having checked and cleaned the data-set from outliers, we will reduce the dimensionality of the problem by functional principal components analysis. This also allows to detect the most important variability features of the velocity curves. Once the dimensionality is decreased, a linear discriminant analysis is carried out. In fact, detecting the common and different features within our sample of patients, permits to determine a classification tool to distinguish patients coming from three populations: TEA candidates, non-TEA candidates with a low-grade plaque and non-TEA candidates without any plaque.

All the statistical analyses of the thesis have been performed out using

the statistical software environment R [25], version 2.12.2.

Chapter 1

Carotid Stenosis and Doppler Images

This first chapter is an introduction to all the themes treated in this thesis. First of all, we present the aim of our work and we briefly introduce the MACAREN@MOX project, of which this thesis is part. Since we try to estimate a curve representing blood velocity along the carotid artery, giving a synthetic presentation of the carotid arteries and of the features of blood flow within this vessels is compulsory. In particular, the carotid bifurcation is a site where the danger of the creation of a stenotic plaque is high and it has to be monitored in ordered to predict, and acting in time, before serious consequences such as stroke or embolism happen. Thus, we illustrate how a doctor can find the indicators of a stenosis, its extent and its morphological nature in the output of the Doppler ultrasound, an example of which can be seen in Figure 2.1. Later on, we will see how these features can be found also in the results of our statistical analysis. With this theoretical basis, we can then proceed in describing how the data have been acquired, the experimental protocol and the population at our disposal.
Furthermore, in order to clarify some choices made during this work (described in the second chapter, when we show how data are extracted from the Doppler image), we list in detail what the Doppler spectrum represents and how to read properly the image created by an ultrasound scanner (such as the one in Figure 2.1). Finally, we dedicate the last part of the chapter to a brief introduction to ultrasound techniques used to estimate blood velocity (continuous and pulsed wave systems, how they work out and the theoretical signal used to model the data). This last part has been added for sake of

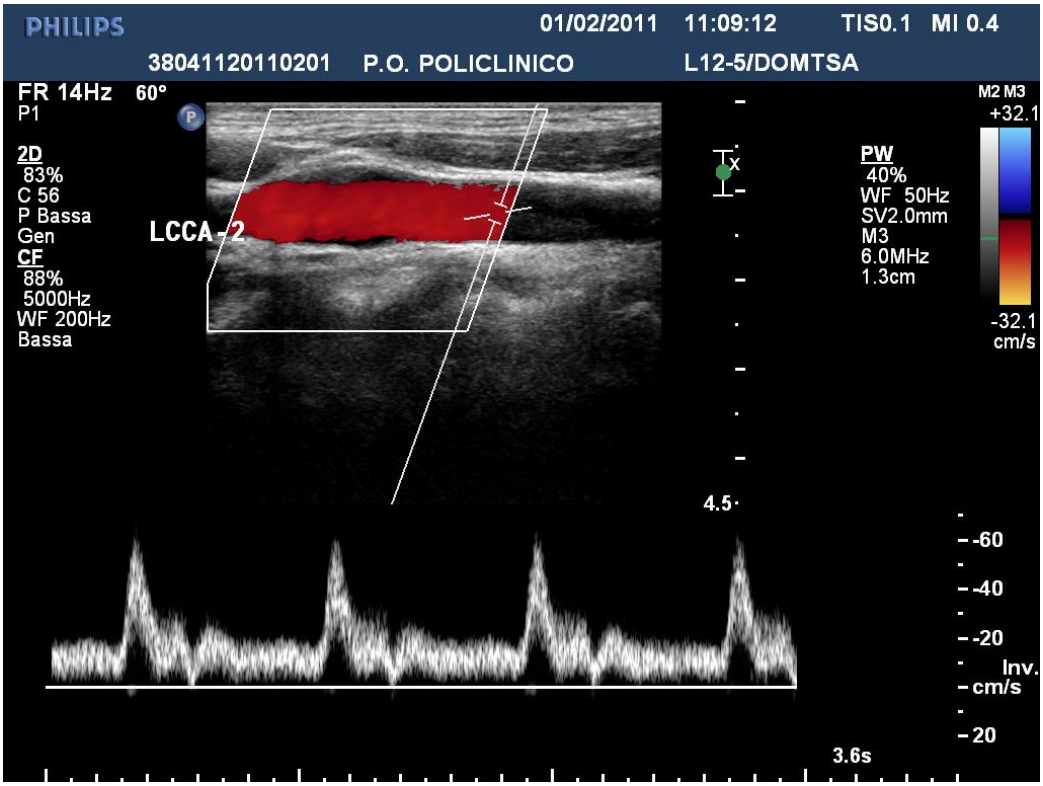

Figure 1.1: Duplex scan showing both B-mode image and spectrogram of the carotid artery.

completeness and it is quite technical; its content is not strictly necessary for the understanding of the methods explained later on in this thesis.

1.1 The MACAREN@MOX project

This thesis is part of a broader project called MACAREN@MOX (MAthematics for CARotids ENdarterectomy@MOX). The research group is composed by a medical equipe and by a team of mathematicians, involving experts in numerical modeling and in statistics. In particular, the clinicians are directed by professor Maurizio Domanin, specialised in vascular surgery at the Fondazione I.R.C.C.S. Cà Granda Ospedale Maggiore Policlinico di Milano, the numerical modeling team is coordinated by professor Christian

Vergara of Università degli Studi of Bergamo and by professor Fabio Nobile at the Laboratory for Modeling and Scientific Computing MOX in the Department of Mathematics "F. Brioschi" at Politecnico di Milano and the statistics team is coordinated by professor Piercesare Secchi, director of the Department of Mathematics at Politecnico di Milano. Finally, doctor Luca Antiga of the group ORBIX deals with the image processing part.
The project focuses on morphological and hemodynamic factors of the carotid bifurcation, which is a preferred site of development and growth of atherosclerotic plaque. It is a wide and multidisciplinary project, involving both clinical and mathematical skills. In particular, for the clinical side, professor Maurizio Domanin and his staff will conduct different analyses in all patients who will undergo to the Tromboendoarteriectomia (TEA, a surgical intervention consisting in the removal of the carotid plaque). These analyses consist in a preoperative static magnetic resonance angiography (CE-MRI), in the acquisition of sequences of phase contrast (PC-MRI) and in Color Doppler Ultrasound acquisitions. After surgery, within the postoperative follow-up of the patients, another magnetic resonance angiography (CE-MRI) is acquired, this time both static and time-resolved. Moreover, a proton magnetic resonance spectroscopy allows to obtain information also on the biochemical characterization of the plaque. Instead, for patients considered healthy or not surgically resectable, just the non-invasive Color Doppler Ultrasounds will be acquired. This will allow to obtain information about the carotid geometry, the velocity of blood flow and the movements of the wall at various instants of the cardiac cycle. Once collected this amount of information, the numerical and statistical part of the team will intervene by reconstructing the three-dimensional geometry of the carotid arteries, processing data from MRI and Color Doppler Ultrasound and performing numerical simulations of blood flow within the reconstructed vessels, taking into account the interaction between blood and the arterial wall.
A dataset will be constructed containing geometric information of the carotid arteries, their fluid dynamics and the biochemical characterization of the plaque (obtained by spectroscopic analysis). This will be done for each patient, that will be about 50 within the end of the project. At the time of this thesis, 22 patients have been analysed.
The aim of the project is to analyse in an integrated way this amount of data from a statistical point of view, in order to find correlations between geometric, hemodynamic and biochemical data. Both statistical and numerical methods will be used, with the purpose to develop new techniques, allowing

the inclusion of the data coming from MRI and Color Doppler Ultrasound in the numerical simulations and resulting in an improvement of their accuracy. In particular, inflow velocities will be estimated through smoothing techniques from Doppler frames and they will then be included in the numerical simulations. This will also allow to obtain from simulations some interesting parameters difficult to measure through non-invasive techniques. Moreover, the parameters measured will be useful in order to being able to choose, among the techniques for removing the plaque, the one ensuring the best results.
Concerning the statistical methods, the project aims to develop new techniques to allow to align and compare functional data, such as complex geometries. Furthermore, the variability of the blood flow profile among patients will be analysed, allowing to make a classification of the population and to identify correlations with the type of plaque.

1.2 Carotid arteries, blood flow and examination techniques

Throughout all the thesis we will treat the blood flow in the common carotid arteries, thus we present in this section a very brief description of these arteries, in order to well understand the data available for our studies. Just the essential information is described, obtained both from literature ([3, 5, 6, 7, 8]) and from team meeting with professor Maurizio Domanin.

1.2.1 Carotid arteries and features of the blood flow

The Carotid arteries are located on the sides of the neck and they have a key role, because through them the blood coming from the heart reaches the brain. There are 2 carotid arteries, the right and left common carotid , which together provide the principal blood supply to the head and neck. Each of the two common carotid arteries (CCA) divides to form internal (ICA) and external (ECA) carotid arteries, which are more superficial. A schematic representation of the arteries supplying the brain can be found in Figure 1.2. The anatomy of the left and the right common carotid artery can be highly variable between different individuals: the major points of variability are the point where the left common carotid artery arises from the aortic arch and

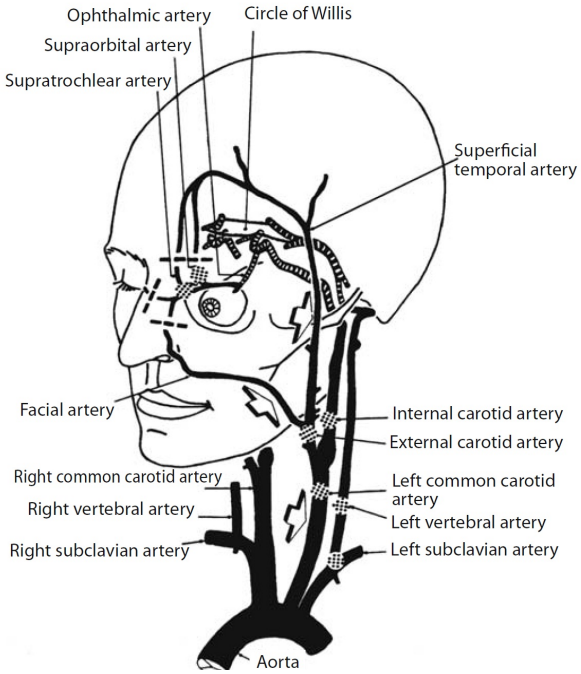

Figure 1.2: Schematic representation of the arteries supplying the brain.

the position of the *carotid bifurcation*, which is usually situated at the level of the fourth or fifth cervical vertebra.
Monitoring the carotid system is important to prevent stroke, the most serious cerebrovascular disease. In fact, over 60-70% of all ischemic cerebral infarctions are caused by arterial embolism [3], typically arising from a carotid artery affected by atherosclerosis. Thus, patients with atherosclerosis of the carotid system are at high risk of stroke. To reduce the risk, it is possible to resort to a highly effective surgical measure called TEA (Carotid Thromboendarterectomy), which is beneficial in individuals with symptomatic high-grade stenosis (meaning a stenosis >70%) and, in certain cases, with 60-70% symptomatic stenosis. To identify those patients who will benefit from TEA, that is to say identifying subjects with carotid stenosis, suitable diagnostic procedures are necessary. Among these, the most preferred is the Color Doppler Ultrasound, because it is a totally non-invasive method, which could be repeated at any time. Ultrasounds run in Duplex mode, not only provide

information on the blood flow along the carotid artery, but also on the location of the plaque, its morphology and its grade. So, through ultrasounds techniques, it is possible to get information needed for surgery or for medical management of stenosis, without the need of additional invasive medical analysis.

We will now briefly describe the blood flow in the carotid arteries, underling some features that can be found later, when we will analyse Doppler images.
Along the common carotid artery (its lumen has a diameter of about 7 mm), the flow is pulsatile and it has a large diastolic component (for sake of completeness, diastole is the relaxing phase when the pressure drops from the peak reached during systolic contraction). In [3], one can read that peak systolic flow velocity ranges from 60 to 100 cm/s under normal conditions, but the velocity is decreased when the lumen is wider. The internal carotid artery has a low-resistance flow and presents a Doppler waveform characterized by a steep systolic upslope, followed by monophasic flow with a fairly large diastolic component, features that characterize a less pulsatile flow, necessary since the ICA has to guarantee a continuous flow of blood to the brain. Instead, flow in the external carotid artery is more pulsatile and it has a smaller diastolic component. Since the common carotid artery has to supply both of them, it presents a mixed waveform. It has to be noticed that pulsatility increases with ages, because it depends on the vessel elasticity.

1.2.2 Examination techniques

The carotid system is quite superficial, thus it can be examined with a high-frequency transducer (7.5 MHz or even 10 MHz), yielding B-mode images with a high spatial resolution, which are useful to the doctor to identify the course of the vessels and their walls. The patient must be in the supine position, while the examiner must move the transducer around to depict the carotid bifurcation as a fork, this would enable a precise localization of the stenotic plaque.
We now underline another kind of variability, since this kind of measurement is highly dependent on the ultrasound equipment and on the examiner. The blood flow is assessed with Doppler ultrasounds following the vessel wall and possible plaques. Usually, a color duplex scanning is performed at the beginning, for initial orientation. At this stage of the medical examination, a

stenosis is suggested by the occurrence of aliasing (color change from red to blue or vice versa), indicating blood recirculation, while occlusion is indicated by the absence of color filling in the lumen of the vessel. After this, Doppler spectra with qualitative determination of flow velocities in longitudinal orientation must be obtained from the common, internal and external carotid arteries. The external one does not raise interest, since it would not be surgically operated even in the presence of plaque; it is however scanned at its origin, to be distinguished by the internal carotid and to identify possible stenosis. Instead, a continuous spectrum is obtained and analysed throughout the common and internal carotid arteries, by sampling at short intervals and using a large sample volume (this fact strongly influence the statistical analysis carried out, as we will clarify later). While scanning the carotid arteries, the examiner probes the interested area placing the transducer on the neck of the patient and, in order to avoid errors in flow velocity measurement, he should try to achieve an insonation angle of less than 60°, by selecting the right transducer.
Under normal conditions, the internal and external carotid arteries are easy to be differentiated. One of the features that allows the examiner to differentiate them is that the external carotid exhibits a more pulsatile flow, with a smaller diastolic component in the Doppler spectrum. Anyway, in case of high grade stenosis at the bifurcation, the external carotid can show a larger diastolic component and its flow profile can become less pulsatile, resembling to that of the internal carotid and making differentiation more difficult. In such cases, the external carotid can be identified by tapping on the temporal artery, because the pulsation will be transmitted to the Doppler waveform, along the diastole, of the external carotid.

1.2.3 Features of a stenosis: degree and morphological aspects

Diagnostic ultrasound of the carotid arteries is prognosis-oriented, meaning that it aims at identifying patients at risk for stroke. TEA surgery is a suitable method for treating high-grade internal carotid artery stenosis, but its main weakness is precisely that TEA could cause what should prevent, i.e. stroke. This is the reason why not all the carotid stenoses are treated surgically: the two risks have to be weighted and evaluated properly before

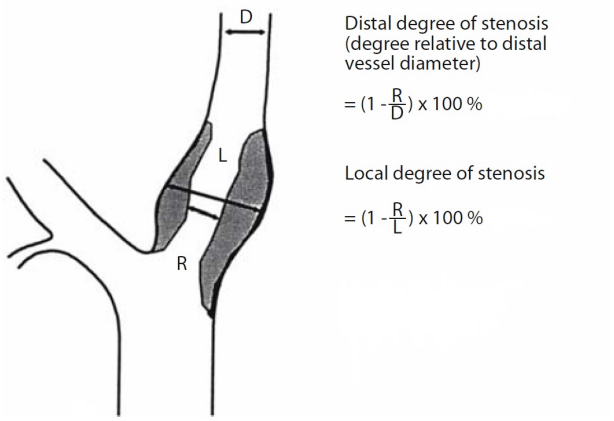

Figure 1.3: Methods of grading a stenosis: local and distal.

proceeding. In [3], it is reported an analysis that shows that TEA significantly reduces the risk of stroke after 5 years in symptomatic individuals affected by a stenosis of degree between 70 and 99%. In individual with 50-69% stenosis, the risk reduction drastically decreases, while it has not any advantage in individuals with a stenosis grade lower than 50%. Finally, TEA is even harmful for patients with a stenosis <30%. In [5] and [6] is stated that randomised studies have shown that surgical removal of the plaque by endarterectomy is superior to antiaggregating medication in the prevention of stroke in case of a stenosis greater than 60% in patients with neurological symptoms, and greater than 70-80% in asymptomatic patients. We will take into account these values at the time of the classification of the carotid arteries of different patients.

But how is defined the degree of stenosis? There exist basically two methods to grade an internal carotid stenosis. The first one identifies the *local* degree of stenosis, which is defined as the ratio of the patient residual lumen to the local vessel lumen without the plaque. This method enables to estimate the plaque thickness and it is more common in the United States. In Europe, instead, the *distal* degree of stenosis is computed from the diameter of the residual lumen in the stenotic area and the diameter of the distal internal carotid artery [3]. Figure 1.3 shows the two methods of stenosis grading.

However, the degree of the stenotic plaque is not the unique determinant of the risk of stroke: the other major factor is the plaque morphology. In [7],

a thickness of more than 2 mm of the intima (the inner wall of an artery) is defined as plaque. There are different stages in the development of a plaque: initially the wall of the artery presents an abnormal thickening, then the size of the plaque increases through lipid and cholesterol inclusion. This disturbs the nutrition of the intima and it can cause a central necrosis of the plaque. Then, different developments can occur and the pulsatile blood flow can even lead to the rupture of the plaque, i.e. embolism. In the carotid system different type of plaques can be distinguished on the basis of their macroscopic appearance:

- flat or fibrous,
- atheromatous or soft,
- calcified or hard,
- ulcerative,
- hemorrhagic.

It has to be underlined that the plaque has an inhomogeneous composition which is reflected in the ultrasound appearance, but it is difficult to succeed in identifying the components on the basis of their different echo levels. Moreover, there are different medical opinions on the connection between sonographic plaque features and the risk of embolism. We thus refer to [3] for an overview of the interpretations of plaque sonographic structures from different authors and for a detailed description of the morphology of a plaque. Again in [3], other important features of a stenotic plaque, besides the composition, are listed:

- Localization:
  - anterior or posterior wall,
  - proximal/distal;
- Extension:
  - circular/ semicircular,
  - diameter of the plaque;
- Surface:
  - clearly delineated/moderately delineated/not delineated,
  - smooth/irregular/ulcer;

- Internal structure:
  - homogeneous/inhomogeneous;
- Echogenicity:
  - hyperechoic (with or without acoustic shadowing)/hypoechoic/ not visualized.

The team treating the numerical part of the MACAREN@MOX project, together with the clinicians, is also developing a procedure to tracing and precisely localize the position of the plaque from the images. Once available, this information should definitely be included in the statistical study carried out in this thesis, with the purpose of understanding at which distance from the plaque the blood flow, and consequently Doppler ultrasound acquisitions, begins to be influenced by the presence of the plaque. To conclude, plaques mainly occur at the level of the carotid bifurcation and along the first 2-3 cm of the internal carotid artery.

1.2.4 How medical doctors quantify a stenosis

In order to try to identify patients with a stenosis looking at the spectrum acquired through Doppler ultrasounds, we should start underlying those features of the Doppler spectrum that allow the doctor to identify a plaque. At the end of our statistical analysis, we will thus be able to compare the characteristics of the reconstructed signal that most discriminate the differences between blood flow of patients with and without a stenosis with the features to which a doctor pays attention.
The flow velocity at the level of the plaque increases in proportion to the degree of stenosis and it reaches its highest values when there is a subtotal occlusion. At the same time, friction occurring at high velocities acts as a decelerating force, decreasing the velocities of the most reflecting blood components. Figure 1.4 (from [8]) shows the relationship between degree of stenosis and intrastenotic peak systolic flow velocity for ICA stenosis: one can recognize the presence of a principal trend, since the velocity increases with the degree of stenosis, but there are also cases of high grade and low velocity, because the heteroskedasticity of the distribution of the velocities increases with the degree of stenosis. Once the plaque has been overtaken, the systolic flow velocity decreases with the stenosis grade.
Stenoses of the common carotid artery are rare respect to those in internal carotid artery. Preferred sites are its origin from the aortic arch and its distal

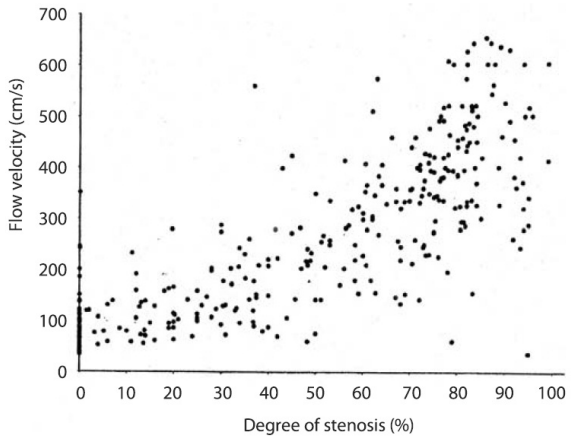

Figure 1.4: ICA stenosis: relationship between the degree of stenosis and instrastenotic peak systolic flow velocity.

segment just in front of the bifurcation. Instead, in case the stenosis occurs in the external branch (again, it rarely happens), the plaque is usually located at its origin, just after the bifurcation. In most cases a stenosis takes shape along the internal carotid artery. Since this is the case that mainly occurs, there are clear parameters of the flow velocities to look at in the diagnostic evaluation. According to [3], there are in particular three important features that allow to quantify a stenosis: the peak systolic flow velocity, the minimum end-diastolic flow velocity and the ratio of the internal carotid artery to common carotid artery velocity (ICA/CCA velocity). A hemodynamically significant stenosis (meaning greater than 50%) is assumed when peak systolic flow velocities reaches values greater than 120 cm/s, while an intermediate or high grade stenosis (>70-80%) is assumed at 180-240 cm/s. Furthermore, usually there are also other factors influencing the velocity, such as contralateral carotid occlusion or other diseases involving other vessels, which leads to higher flow velocities in the carotid system. Therefore, in order to avoid an overestimation of carotid stenosis and false-positive findings, the cut-off velocity for discriminating between low-grade and hemodynamically significant stenosis must be increased to 140-150 cm/s [3]. Also the minimum end-diastolic flow velocity increases with the stenosis grade and velocities of around 40 cm/s suggest a stenosis >50%, while velocities of 80-100 cm/s in-

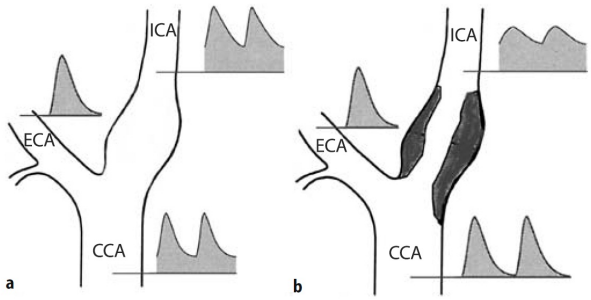

Figure 1.5: Schematic Doppler waveforms of the carotid system in a normal case (a) and a high-grade stenosis case (b).

dicate a high-grade stenosis. Finally, the ratio ICA/CCA can be considered as a kind of normalization with respect to other factors which could influence absolute velocities (hypertension, sclerosis, contralateral occlusion,...) and in [11] this value has been used to distinguish normal arteries (ratio <0.8) from high-grade stenoses (ratio >1.5) carotid arteries. In Figure 1.5, a schematic graphic summary of the Doppler waveforms that one can expects in the carotid system is presented. On the left (part a) there are waveforms expected in a normal patient; along ICA the blood flow is fairly steady, as a result of low peripheral resistance and this is reflected by the fact that the spectrum has a moderate systolic component (little pulsatility) followed by a steady flow that persists throughout diastole. As we explained in previous sections, along ECA the flow is more pulsatile, while in CCA one can find a sort of mixed pulsatility. On the right part of the figure, instead, it can be seen how the Doppler signal changes in presence of a high-grade stenosis at the origin of the internal carotid: along ICA the flow become even less pulsatile, meaning that one can clearly see a larger diastolic component and a reduced peak systolic velocity. In ECA the changes of the flow are not significantly perceptible, while the flow in the CCA resembles the flow in the ECA, because it becomes more pulsatile. Figure 1.6 also helps to visualize the impact of a plaque in the blood flow.
Apart from information about the plaque, from the Doppler spectrum one can obtain other important parameters for evaluating blood flow, such as:

- the peak systolic velocity;
- the peak end-diastolic velocity;

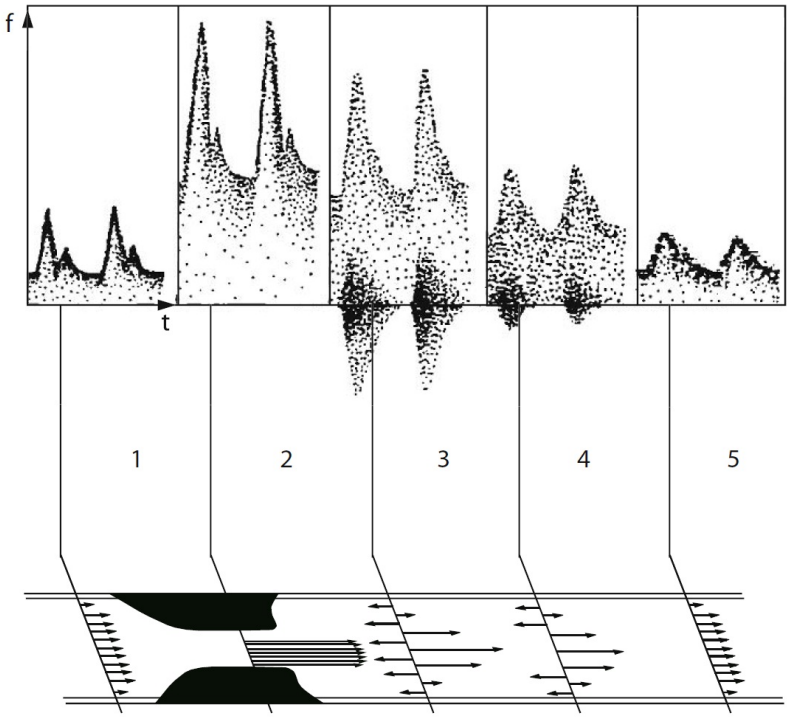

Figure 1.6: Impact on the Doppler waveform of the flow around a stenosis, in ICA. 5 phases can be recognize where pulsatility of the flow changes.

- the average flow velocity;
- the variance, that is to say, the spectral broadening due to disturbed flow.

1.3 Experimental protocol

After having introduced the main features of Doppler spectrum and how they can show the presence of a stenosis, we can now proceed in describing the experimental protocol we have followed to obtain the images that will be used later on in this thesis and within the MACAREN@MOX project.

For each patient, several duplex images of the carotid system have been acquired: all the images and videos have been produced by a scanner combining

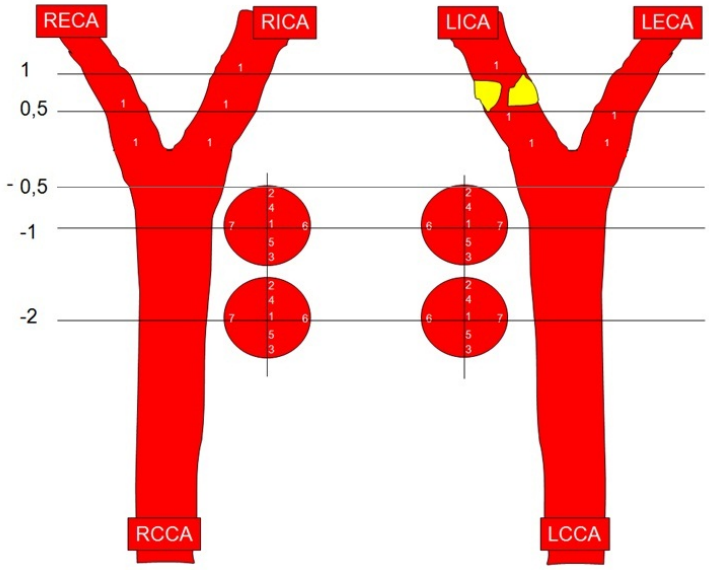

Figure 1.7: Positions where the signal has been acquired for TEA candidates. ICA: Doppler spectrum sampled in the center of the vessel at level 0, 0.5 and 1; ECA: Doppler spectrum sampled in the center of the vessel at level 0 and 0.5. CCA: 7 sample volumes placed at levels -1 and -2 along the longitudinal cross section of the vessel.

B-mode imaging with a 6Mhz pulsed Doppler (Philips Medical Systems, iU 22). Carotid examination is performed by placing the scanner on the neck of the patient, lying in the supine position, along the axis of the carotid vessels. While doing this, a real time B-mode arterial image is generated on a television screen, helping the examiner in the placement of the pulsed Doppler sample volume within the arterial lumen. The sample volume can be placed in different locations in the cross section of the lumen of the vessel and so it is possible to acquire the Doppler ultrasound images in different tansversal points. We remind that Doppler ultrasounds are a non-invasive diagnostic tool and they can be repeated all the times needed. Nevertheless, this kind of exam takes some time and it has organizational costs, so we dispose of a different number of measurements depending whether the patient should undergo TEA or not. For TEA candidates the protocol we settled down requires 19 recordings of the spectrum of blood velocity, while for healthy

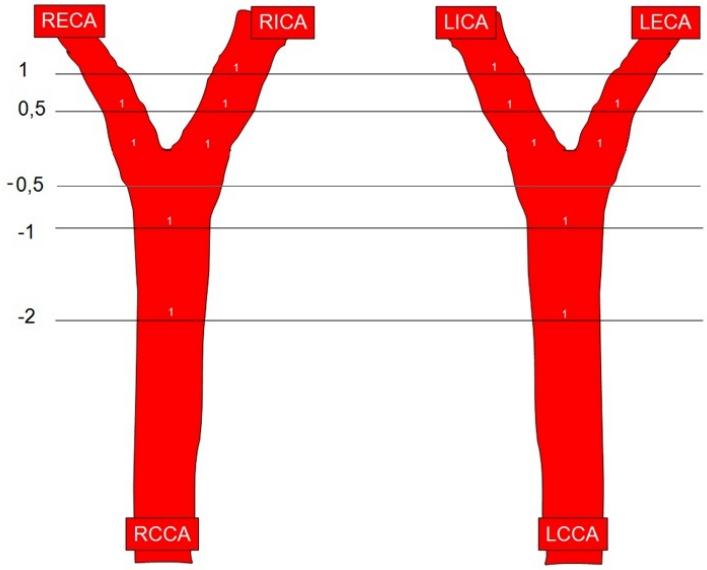

Figure 1.8: Positions where the Doppler signal has been recorded for non-TEA candidates. ICA: spectrum sampled in the center of the vessel at level 0, 0.5 and 1; ECA: spectrum sampled in the center of the vessel at level 0 and 0.5. CCA: sample in the center of the vessel at levels -1 and -2.

patients or patients with a non-high-grade stenosis 7 records are available. Figure 1.7 illustrates the exact positions where the signal has been recorded for TEA candidates: along the ICA one Doppler spectrum sampled from the center of the vessel is acquired after 0.5 cm from the bifurcation (level "0.5") and after 1 cm (level "1"). Since the external carotid is not an usual site for stenosis, only one Doppler image after 0.5 cm from the bifurcation ("0.5") is registered. Regarding the CCA, more images are available, at various levels (meaning distances from the bifurcation) and in various cross sectional points of the vessel. At a distance of 0.5 cm (level "-0.5") before the bifurcation, only one Doppler image is acquired. Instead, at a distance of 1 cm (level "-1") and of 2 cm (level "-2") before the bifurcation, 7 sample volumes have been considered along the longitudinal cross section of the vessel (Figure 1.7 helps understanding what is meant for 'level' and for 'cross section'). Figure 1.8 presents the protocol for non-TEA patients (further in this project, healthy patients or patients with a low or medium-grade stenosis, not enough

extended to be treated with TEA, will be denominated as 'non-TEA'), for whom one Doppler image from the center of the vessel at each level has been acquired. The procedure is repeated first for the right and then for the left carotid arteries, so that for each patient two different classes of Doppler ultrasound acquisitions are available. Finally, only for TEA patients, Doppler images and the values of the peak systolic velocity have been acquired also for the following vessels in the body: vertebral, subclavian, humeral, radial, aorta, common iliac, external iliac, common femoral, popliteal and posterior tibial.

The protocol described regards the whole acquisition of Doppler ultrasounds for the MACAREN@MOX project. Within this thesis only the measurements sampled in the center of the vessel at levels -2, -1 and -0.5 are used. It has to be noticed that there are various factors complicating the valuation, such as body size, an irregular heartbeat or a complicated carotid geometry, so Doppler images from different patients or at different levels do not have the same quality and sometimes they are even not available or not usable for our study. This is the reason why in next section, in the description of the population, sometimes the reader will not find all the 19 images that theoretically the protocol includes.

1.3.1 Population

The patients included in the study treated in this thesis are 22, corresponding to the data collected for this purpose in Policlinico Hospital of Milano from November 2010 and May 2011. Probably, the final project MACAREN@MOX will include a larger number of patients, being a more extended work with a longer duration.
The 22 patients we refer to have a middle age of 67.3 years, the youngest being 42 years old while the oldest 82. There are 14 males and 8 females. Since the number of data is extremely low for a statistical approach, we will consider the right carotid as independent from the left carotid artery. This approach was considered appropriate by a consult with medical staff, because the presence of a plaque in one of the carotid arteries does not modify the blood flow in the other carotid artery. Anyway, there are factors inducing the development of plaques (such as fat, smoking, hypertension, diabetes and so on) which are related to the patient and, thus, they influence both left and right carotid arteries. In this way, it is as if we had at our disposal data

coming from 44 patients. Unfortunately, this happens only for acquisitions of images at distance -2 from the carotid bifurcation. In fact, as one can see in Table 1.1, at level -1 a Doppler acquisition for a patient is missing, reducing the total number to 43, while at level -0.5 the number of Doppler images is 38, because in this site is more difficult to acquire a clear Doppler ultrasound signal.
All the patients have been submitted to ultrasounds at the carotid system and the images of the spectral waveform of blood velocities are available in the points illustrated in the previous section. In Table 1.1, the following details are listed: age, sex, whether the patients is symptomatic (Type S) or asymptomatic (Type A), which Doppler images are available in the three different level of the CCA (-2, -1 and -0.5), the percentage of the stenotic plaque (if there is any) and its nature and, finally, a label stating to which group the patient belongs. In fact, three different groups can be detected, depending on the subjective evaluation of the doctor whether the patient will undergo TEA or not. Label "3" indicates those patients who does not present any plaque in the carotid system (we will refer to them as *healthy* later on), label "2" is linked to patients presenting a stenosis of low or medium grade and who will not undergo the surgery (as already mentioned, we will refer to this people as *non-TEA*) and, finally, patients who will undergo TEA (called *TEA* later on) are indicated with label "1". The main features of the left and right common carotid arteries are reported separately and when patients present a stenosis, it is located along the internal carotid, but at this stage of the project, we do not know exactly at which height of the vessel. Moreover, it should be noticed that a doctor, when deciding if the patient should undergo TEA or not, takes into account more other characteristics, for example if the patient smokes or is affected by other diseases, previous exams to evaluate if the plaque is stable or it is increasing and so on. Trying to consider the whole information available would lead to a less automatic procedure of classification, having to analyse any single case as the doctor does. On the contrary, we would like to develop a method that automatically classify the patients in 2 different groups (TEA, non-TEA), considering uniquely the information derivable from Doppler ultrasounds acquisitions along the common carotid arteries. This could lead to a faster time of analysis of patients by doctors, at least in the early stages, when, before deciding whether a patient should undergo surgical operation or not, the patient is monitored for a period of time that could last months or years, implying therefore high organizational costs for the hospital.

Before starting to practically treat our images, we should explain what actually one can read from a Doppler waveform. This is how we conclude this first chapter, since introducing which information is contained in the Doppler spectrum and treating briefly the fundamental principles of ultrasounds will help in understanding the extraction of data from the images, explained in the 2nd chapter.

1.4 The Doppler spectrum

In this section, we are going to explain which information is contained in the Doppler spectrum, that is to say how the image printed out during an ultrasound scanning should be read. In fact, in a blood vessel, blood components move at different velocities and these velocities are represented in the Doppler spectrum by the scale of greys (or colours): different levels of grey represents the different amplitudes of a range of frequencies, that reflect the distribution of flow velocities in the vessel through the relation $v = \frac{c}{2f_0 \cos\alpha}$, where c is the speed of propagation of ultrasound in blood, f_0 is the emitted frequency and α is the angle between ultrasound beam and direction of blood flow. For the individual velocity values, the corresponding amplitudes are computed and displayed with different shades of grey.
The Doppler spectrum displayed contains the following information on blood flow:

- the vertical axis represents different values of flow velocities;
- the horizontal axis represents the time course of the cardiac cycle;
- on vertical axis, the color or grey intensity, is a density of points (time is fixed along this axis), representing the number of red blood cells moving at that specific velocity in the analysed volume of blood. This kind of information could be represented also as an histogram that varies over time: Figure 1.9 shows the histograms of the density points, once fixed one specific time of the cardiac cycle.

The zero flow line gives another information: flow toward and away from the transducer, represented respectively above and below the baseline.
One of the main features that we will use later on in this thesis, is the fact that the different levels of brightness represent the density of a given velocity

Patient				Left Carotid Artery						Right Carotid Artery					
N.	Type	Sex	Age	CCA			%	Nature	Group	CCA			%	Nature	Group
				-2	-1	-0.5				-2	-1	-0.5			
1	A	M	75	a.	n.a.	a.	80%	n.a.	1	a.	a.	a.	n.a.	n.a.	1
2	A	M	79	a.	a.	n.a.	70%	lipid	1	a.	a.	n.a.	20%		2
3	A	F	65	a.	a.	a.	20%	n.a.	2	a.	a.	a.	20%	n.a.	2
4	A	M	50	a.	a.	a.	n.a.	n.a.	2	a.	a.	a.	n.a.	n.a.	2
5	A	F	63	a.	a.	a.	vs	-	3	a.	a.	a.	vs	-	3
6	A	F	73	a.	a.	a.	70%	lipid	1	a.	a.	a.	30%	lipid	2
7	A	M	53	a.	a.	a.	20%	n.a.	2	a.	a.	a.	20%	n.a.	2
8	A	M	75	a.	a.	a.	30%	n.a.	2	a.	a.	a.	vs	-	3
9	A	M	n.a.	a.	a.	a.	ntd	-	3	a.	a.	a.	ntd	-	3
10	A	M	42	a.	a.	a.	vs	-	3	a.	a.	a.	vs	-	3
11	A	M	44	a.	a.	a.	ntd	-	3	a.	a.	a.	30%	fib	2
12	A	M	76	a.	a.	a.	ntd	-	3	a.	a.	a.	ntd	-	3
13	S	M	49	a.	a.	a.	vs	-	3	a.	a.	a.	vs	-	3
14	S	F	49	a.	a.	a.	ntd	-	3	a.	a.	a.	ntd	-	3
15	S	M	58	a.	a.	a.	vs	-	3	a.	a.	a.	vs	-	3
16	S	M	79	a.	a.	a.	80%	n.a.	1	a.	a.	a.	n.a.	n.a.	1
17	S	M	79	a.	a.	a.	75%	mixed	1	a.	a.	a.	50%	fibcal	2
18	S	M	76	a.	a.	n.a.	75%	n.a.	1	a.	a.	a.	n.a.	n.a.	1
19	S	F	65	a.	a.	n.a.	80%	n.a.	1	a.	a.	a.	70%	n.a.	1
20	S	F	74	a.	a.	n.a.	40%	n.a.	2	a.	a.	a.	75%	n.a.	1
21	S	F	82	a.	a.	a.	n.a.	fib	1	a.	a.	n.a.	80%	fibcal	1
22	S	F	72	a.	a.	a.	30%	n.a.	2	a.	a.	a.	80%	fibcal	1

Table 1.1: Description of the Data Set. Under the symbol %, the percentage of the plaque (when it is present) is reported. The type of patient can be symptomatic (S) or asymptomatic (A). When the percentage of the plaque is not reported, 3 cases can arise: there is a plaque but its dimension is not available (n.a.), there is nothing to detect (ntd) or there is not a stenosis but the presence of a vascular sclerosis (a tissue hardening) without particular lesions (vs). Finally, the nature of the plaque is reported, which could be lipid, fibrotic, mixed or fibrocalcific.

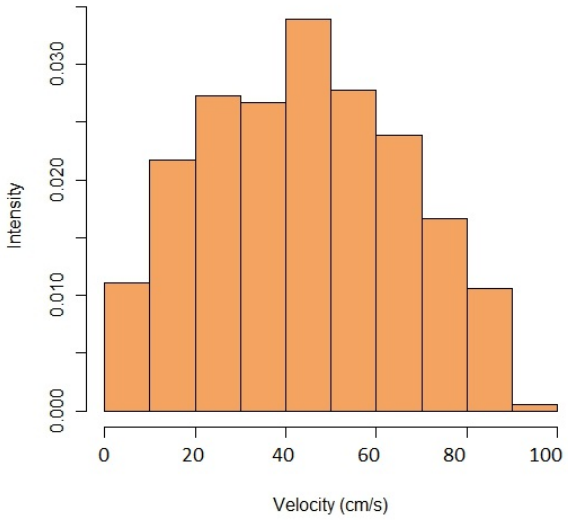

Figure 1.9: Histogram representing the number of red blood cells moving at the velocities specified along horizontal axis, at one specific time along the cardiac cycle.

(or frequency) in the frequency band. Figure 1.10 can help understanding the information contained in the spectrum: it is a three-dimensional Doppler frequency spectrum, showing the distribution of individual frequency shifts (amplitudes), flow direction (above and below the zero-flow line, which is the time axis) and flow velocities (proportional to the Doppler frequency shifts). The height of the boxes correspond to the amplitudes of the respective Doppler frequencies and in 2D images they are represented by different levels of brightness. The black boxes show the averaged flow velocity at a specific point in time.

1.5 Fundamental principles of ultrasounds

In order to estimate the blood velocity, data should be created by using an ultrasound transducer. Figure 1.12 shows some example of the ultrasound transducers used in modern hospitals. The method consists of emitting a si-

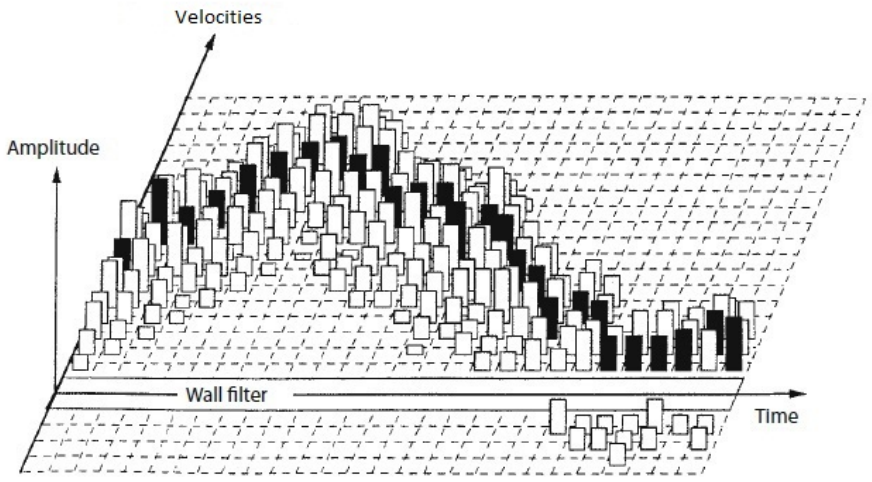

Figure 1.10: 3D Doppler frequency spectrum, showing the distribution of individual frequency shifts (Amplitudes), flow direction (above and below the zero-flow line, which is the time axis) and flow velocities (proportional to the Doppler frequency shifts).

nusoidal pulse and then to subsequently receive the echo signal. The received signal is then sampled at a time corresponding to a selected depth, yielding in this way one temporal sample per each emission. The process is then repeated many times at the pulse repetition frequency (f_{prf}). The velocity of the flow can be derived from the spectrum of the signal sampled at the pulse repetition frequency, as it is explained in the following section. For a deeper understanding of the subject we refer to [1] and [2].

1.5.1 Ultrasound waves

Ultrasound is a mechanical vibration of matter with a frequency above the audible range, which is 20 kHz [1]. No mass is transported during the propagation of the longitudinal wave: the particles of the medium crossed by the ultrasounds just oscillate around their mean positions, instead of being at rest and equally spaced as before the disturbance. The propagation speed

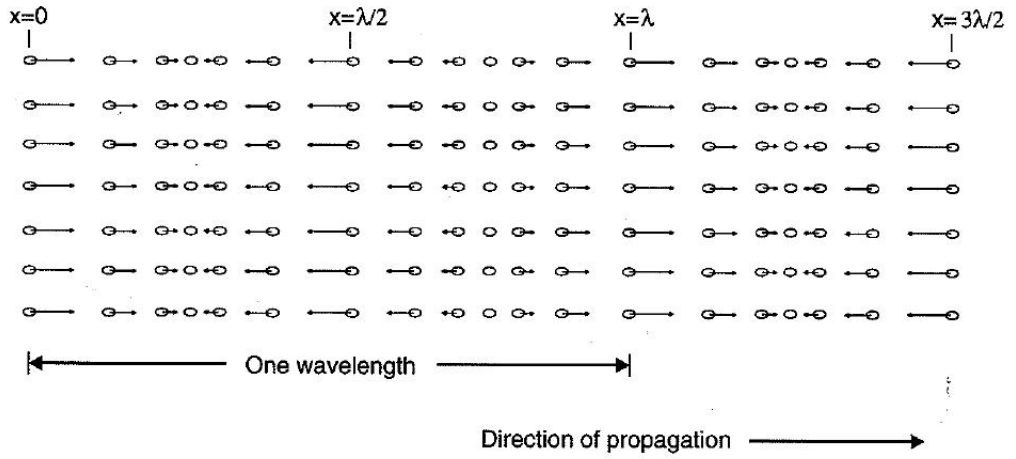

Figure 1.11: Particle displacement for a propagating ultrasound wave.

of the disturbance depends on the medium.
The propagation of the wave could be of different kinds, depending again on the medium used, but in medical ultrasound we can generally assume linearity, at least when tissue with significant attenuation is penetrated, that is to say that the pressure generated by the disturbance is small if compared with the equilibrium pressure.

Given a wave, it is possible to define its acoustic intensity as the average flow of energy through a unit area, normal to the propagation direction, in unit time. The intensity is the average of the rate of work done per unit area by one element of fluid on an adjacent element. However, the measure of intensity is defined in different ways depending on whether the wave is continuous or not. The ultrasound fields used in medical practise are generally pulsed and strongly varying with regard to spatial position, so it is necessary to specify how the intensity is calculated. In table 1.2 (taken from [1]), the highest allowed acoustic field emissions for commercial scanners are listed. These measurements are stated by the United States FDA (Food and Drug Administration) and are based on studies of tissue damage due to ultrasound exposure. In the table, I_{sptp} denotes the spatial peak temporal

averaged intensity, I_{sppa} the spatial peak pulse average intensity and, finally, I_m denotes the maximum intensity.

	I_{spta} (mW/cm^2)		I_{sppa} (W/cm^2)		I_m (W/cm^2)	
Use	In Situ	Water	In Situ	Water	In Situ	Water
Cardiac	430	730	65	240	160	600
Peripheral vessel	720	1500	65	240	160	600
Ophthalmic	17	68	28	110	50	200
Fetal imaging	46	170	65	240	160	600

Table 1.2: Highest know acoustic field emissions for commercial scanners.

1.5.2 Scattering

A wave which propagates through a medium continues straight on the same direction until it encounters different acoustic properties, that is to say, until the wave crosses a new medium. When this happens, part of the wave is transmitted through the new medium, probably changing direction, while part of it is reflected back. This is what should happen when an ultrasound wave sounds out the human body: part of it should be reflected back when encountering a new tissue, so that it would be possible to record it as soon as it reaches the transducer, and finally visualize it as an image on a screen. Thus, in order to visualize the image of a boundary, it is necessary that the reflected wave reaches the transducer, which is quite difficult when talking about human body, because the transducers must have limited dimensions to be able to move around, and the wave, when reflected, changes its direction. Moreover, boundaries are rarely found.
So, what is it that makes possible to recognize different tissues, when watching the images recorded by the transducer? The solution is a phenomenon called scattering, which consists in the fact that the ultrasound wave is

forced to deviate from its straight trajectory, every time it encounters non-uniformities in the medium through which it propagates. In short, a scattered wave is created whenever, in the propagating medium, small changes in density, compressibility or absorption are encountered. Then, the scattered wave radiates in all directions, making possible for the transducer to receive a back scattered field. From this field it is possible to extract the information needed, such as the blood velocities.
This backscattered signal is, of course, weaker than the signal reflected by the boundaries, but such boundaries are encountered just in few cases, such as diaphragm, blood vessels walls and organ boundaries. Instead, the scattered wave is emanated from a lot of different contributors, so that it can be characterized in statistical terms, with an amplitude distribution which typically follows a zero-mean Gaussian distribution.
This does not mean that new measurements of the backscattered signal generate new values: if the structure probed by the transducer is stationary, the same signal will always result and, what it is more important is that slight shifts in position will lead to high correlated signals. Here lies the idea to detect blood velocities: by analysing the correlation between successive measurements of moving blood cells, it is possible to trace the shifts in position.

The strength of the returned signal is described in terms of the power of the scattered signal, which could depend on the position between the ultrasound emitter and receiver. This is the case of muscle tissue for example. Instead, in the case of blood, only one transducer is used for transmission and reception, and only the backscattered signal is considered. The signal power generated by a single scatterer is $P_s = I_i \sigma_{sc}$, where the power P_s is generated when a beam of intensity I_i intersect the scattering cross section σ_{sc},which is dependent on the material and indicates how strongly scattering the material is.

Finally, another phenomenon has to be mentioned: the ultrasound wave propagating in tissue will be attenuated. In tissue, attenuation is due both to scattering, which will spread energy in all directions, and to absorption, which means conversion into thermal energy and the dependence between attenuation, distance traveled and frequency can often be considered as linear.

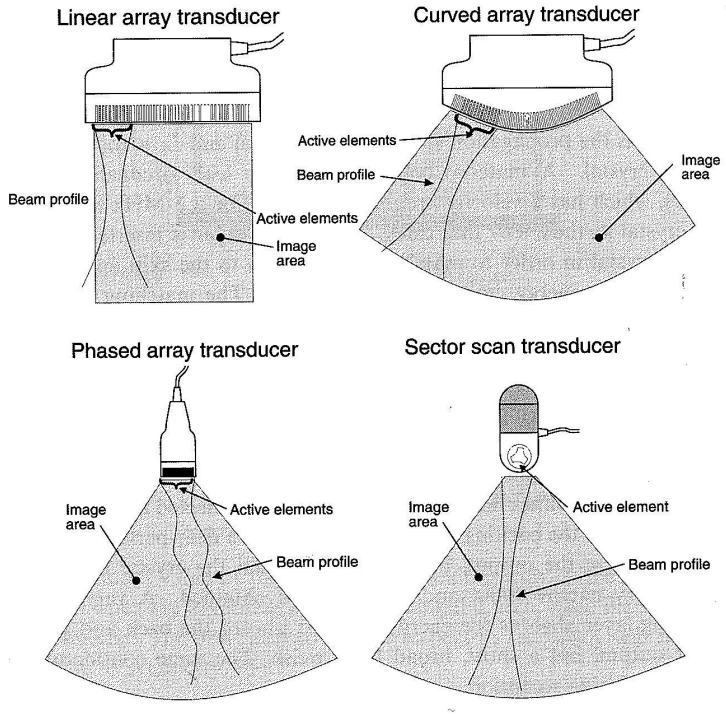

Figure 1.12: Different ultrasound transducers for acquiring B-mode images.

1.5.3 Continuous wave systems

The simplest, non-invasive, way to detect blood velocity consists in the continuous wave (CW) system. The idea behind the continuous wave system is to insonate part of the body by emitting a continuous sinusoidal wave and then compare the received signal with the emitted one, in order to detect the change in frequency. Thus, continuous wave Doppler ultrasound uses two transducers, with one continually transmitting and the other continually recording the ultrasonic wave. The received signal $r_s(t)$ is multiplied by a quadrature signal of frequency f_0 (which is the frequency of the emitted signal), in order to find the Doppler shift. Applying the Fourier transform we get

$$r_s(t)exp\{j2\pi f_0 t\} \leftrightarrow R_s(f - f_0).$$

As suggested in [1], the emitted signal can be described as

$$e(t) = cos(2\pi f_0 t)$$

while the received signal is

$$r_s(t) = a\, cos(2\pi f_0 \alpha(t - t_0))$$

with a frequency αf_0 proportional to the center frequency f_0 of the emitted signal. To give an idea of how, by measuring the frequencies of the received signal, one could estimate the blood velocity in the direction of the ultrasound beam, it is enough to look at how α and αt_0 are approximated [1]:

$$\alpha \approx 1 - \frac{2v_z}{c}$$

$$\alpha t_0 \approx \frac{2d_0}{c}$$

denoting with v_z the velocity along the direction z (the direction of the ultrasound beam) and d_0 the initial position of the scatterer. The received signal is then multiplied by a quadrature signal. Afterwards, a band pass filter is used, in order to remove both the higher frequency signal (at twice the emitted frequency) and the component coming from the stationary tissue, which would cover the contribution from the blood. The latter is often done by subtracting out the mean of the signal prior to the filtering. The remaining signal is thus:

$$m_f(t) \approx \frac{a}{2} exp\left(j2\pi f_0 \frac{2v_z}{c} t\right) exp(-j2\pi f_0 \alpha t_0)$$

where the second exponential term represents the delay caused by the round trip time between emission and reception. There is another delay that should be considered, which sum up the depth in tissue and the speed of sound due to the ultrasound propagation velocity, but this delay is usually negligible.

Detectors

Detectors are the way the acquired information is then presented. There are different techniques, such as

Audio output , which is useful to give a sensation of flow direction.

Zero Crossing Detectors means detecting the most dominant signal in the flow by counting the number of times the signal crosses its mean value.

Spectral Display. Since the frequency content of the signal received by the transducer corresponds to the velocity distribution of the blood, a display of the distribution of velocities can be made by taking the Fourier transform of the received signal and showing the results (a so-called sonogram). In order to do this, the received signal is divided into segments and the power density spectrum is calculated for each of these segments. Then, the spectra are displayed side by side, so that the evolution of the velocity distribution can be observed. The intensity at a point on the screen indicates the amplitude of the spectrum, proportional to the number of blood scatterers moving at a particular velocity. Eventually, we state the result that with this method the velocity direction is preserved (see [1] for more details).

In the following sections we are going to introduce pulsed wave systems and explain why they are preferred to CW systems. In fact, the limitation of CW Doppler is that the signals from all moving reflectors along the path of the ultrasound beam are detected, with their respective frequency shift. Consequently, CW systems cannot differentiate flow signals from two closed vessels, when one lies behind the other along the beam path. However, aliasing errors are more easily controlled for a CW system, so continuous wave measurements are still used in modern scanners, as a supplement to the pulsed techniques, expecially for the detection of high flow velocities.

Stationary echoes

Stationary echoes are the signals coming from vessel boundaries and tissues around them. As one could perceive by intuition, these factors are larger than the signal coming from the blood and it is necessary to remove them, to avoid a disturbance in the measurements of blood velocities. A solution to reduce these disturbances is by inserting a low pass filter after the multipliers, even if this procedure is not without consequences: depending on the inserted filter, the lowest velocity that can be estimated by the system and how quickly new measurements can be taken would be modified. In any case, it is necessary to remove the stationary echoes, especially for small vessels,

where the blood signal could be totally hidden by them.
The order of the filter that should be applied can be computed by the ratio between the amplitude of the stationary echoes to the amplitude of the blood signal.

Finally, we mention another source of distortion for the estimate of the power spectral density of the blood signal, which consists in the partial insonation of the vessel and the attenuation phenomenon. In fact, the whole vessel is not equally insonated (especially large vessels) and part of the vessel receives more ultrasound energy than other, leading to disturbances in the PSD of the blood.

1.5.4 Pulsed wave systems

In the previous section, we have listed some of the drawbacks of investigating the blood flow by a CW Doppler system. A reason why a pulsed wave system should be preferred to a continuous wave system could be, for example, that two vessels can be inadvertently interrogated at the same time, resulting in a distortion of the spectrum. Moreover, no unique mapping from Doppler spectrum to velocity profile exists, consequently it would be difficult to differentiate a normal flow pattern from a pathological pattern.
By using a pulsed wave systems, instead, these problems are avoided, but the classic Doppler effect can not be used anymore, because of the problem of attenuation. As explained in [1], in fact, ultrasound pulses emitted into the body, experience attenuation during propagation through the tissue, which increases with frequency. By using a Doppler system then, the higher frequency part of the pulse spectrum would get progressively more attenuated than the lower part, hiding some essential information.

In this section, we are going to explain briefly the steps of a pulsed wave system, but first we would like to underline the fact that scatterers move from time to time, shifting their position, therefore consecutively received signals are shifted in time compared with the proceeding and preceding RF (Radio Frequency) line.
An ultrasound pulse, emitted from a transmitter, propagates into the tissue and blood and, interacting with them, it causes the emission of a backscattered signal, which is then received by the same transducer, amplified and multiplied by the center frequency of the emitted pulse. Finally, the received signal is low-pass filtered, to reduce the echoes signals. A sampling and analog-to-digital conversion is then performed, and for each pulse emitted

only one sample is acquired. Figure 1.13 might help to understand the way in which sampling is done: here the depth in tissue is fixed and the signals shown in the left side of the picture result from a sequence of pulses. Each line corresponds to a single pulse, and the different pulses are emitted at a pulse repetition frequency, f_{prf}. On the right side, instead, there is the resulting sampled signal, produced by taking into account the amplitude of each pulse after a fixed time (indicated by the dashed line in the left graph). If the sampling is done T_s seconds after the pulse emission, the depth in tissue is determined by

$$d_0 = \frac{T_s c}{2}.$$

In the case that the probed tissue was stationary, a constant sampled value would result. As blood is not stationary, we obtain changing values.
Often a B-mode image is presented along with the sonogram in a duplex system, in order to show on the image the area of investigation or range gate. Acquiring the B-mode image requires gaps in the sampled data used to estimate the spectrum, and there exist algorithms developed in order to optimize this procedure. This is an important aspect, because removing a part of the data (used to create the B-mode image) it means to increase the pulse delay, and the longer the pulse delay is, the lower will be the peak flow velocity that can be detected.
The combination of two-dimensional real-time imaging with pulsed Doppler is known as *duplex ultrasound* and it provides flow information from a sample volume at a defined depth. Duplex scanning enables calculation of blood flow velocity from the Doppler frequency shift, because the angle of incidence between the ultrasound beam and the vessel axis can be measured in the B-mode image.

1.5.5 Data Model

In this section, we expose the data model, mantaining the same notation as in [4]. The sinusoid which is emitted from the ultrasound system has the center frequency f_c and can be expressed as

$$p(t) = sin(2\pi f_c t).$$

This signal will reflect on the blood scatterer, positioned at the depth d. The reflected, received and filtered signal is described in [1]. Here, we can

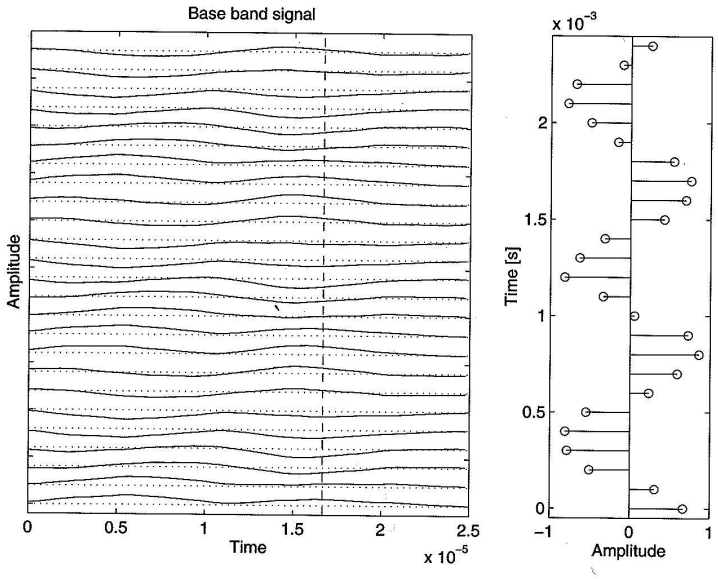

Figure 1.13: Sampling for a pulsed wave system: the left graph shows the different received signals and a single pulse correspond to each line. The right graph instead shows the resulting sampled signal.

summarize the results: denoting with c the speed of sound, f_s the sampling frequency of the system, k the index proportional to the depth in the probed vessel and with a_v the amplitude of the beamformed echo of the blood scatterer, the received signal is:

$$r(k) = a_v sin\left(2\pi f_c\left(k\frac{1}{f_s} - \frac{2d}{c}\right)\right) \tag{1.1}$$

Typically, samples are acquired over several consecutive transmissions; in the meanwhile, the blood scatterer moves in the depth direction from its initial position d_0. Its position at the transmission time index l (*slow time*) can be expressed as

$$d(l) = 2d_0 + l\frac{v}{f_{prf}} \tag{1.2}$$

where v is the axial velocity of the blood scatterer. The index l represents the so called *slow time*, the time at which the pulses are emitted, represented along the vertical axis in Figure 1.13. With *fast time*, instead, we will denote

the time along the horizontal axis in Figure 1.13, that is to say the time it takes to each emitted wave to reach a certain depth and to go back to the receiver. Later on, the fast time will be indicated by index k.
Inserting (1.2) into (1.1), we obtain the 2-D function

$$r(k,l) = a_v sin\left(2\pi f_c\left(k\frac{1}{f_s} - \frac{2d_0}{c} - l\frac{2v}{f_{prf}c}\right)\right). \tag{1.3}$$

In order to detect the direction of the moving blood, an analytic signal over fast-time should be generated, by taking the discrete Hilbert transform, $\mathcal{H}_k\{r(k,l)\}$, of the received signal in the fast-time direction:

$$z(k,l) = r(k,l) + j\mathcal{H}_k\{r(k,l)\} \tag{1.4}$$

Then, we can write

$$\begin{aligned} z(k,l) &= a_v \exp\left\{j2\pi f_c\left(k\frac{1}{f_s} - \frac{2d_0}{c} - l\frac{2v}{f_{prf}c}\right)\right\} \\ &= a_v \exp\left\{-j\frac{4\pi f_c d_0}{c}\right\}\exp\left\{j2\pi f_c\left(k\frac{1}{f_s} - l\frac{2v}{f_{prf}c}\right)\right\} \end{aligned} \tag{1.5}$$

denoting all the fixed terms with A_v:

$$A_v = a_v \exp\left\{-j\frac{4\pi f_c d_0}{c}\right\}.$$

If there is a distribution of scatterers within the resolution cell of the ultrasound system, the total complex signal can be written

$$y(k,l) = \int A_v \exp\left\{j2\pi f_c\left(k\frac{1}{f_s} - l\frac{2v}{f_{prf}c}\right)\right\} dv + n(k,l) \tag{1.6}$$

where $n(k,l)$ is a noise process, assumed to be a white, complex, circular symmetric stochastic process independent of the blood signal, with zero mean and variance σ_n^2.
We introduce two new variables, just to simplify the notation:

$$\begin{aligned} \phi &\triangleq \frac{f_c}{f_s} \\ \psi &\triangleq \frac{2vf_c}{f_{prf}c}. \end{aligned}$$

Allowing the measured signal to be written as

$$y(k,l) = \int A_v \exp\left\{j2\pi\left(\phi k - \psi l\right)\right\} d\psi + n(k,l) \tag{1.7}$$

with the corresponding slow-time Power Spectral Density (PSD) of the signal $y(k,l)$ formed as

$$P_y(\psi) = |A_\psi|^2 + \sigma_n^2 \tag{1.8}$$

This shows that estimating the power density spectrum would give an indirect estimate of the velocity distribution in the interrogated blood volume. Following this model, one is only able to measure the velocity component along the ultrasound beam direction. This is a significant limitation, since most of the vessels are parallel to the skin surface and, also, the flow is usually not parallel to the vessel surface. Thus, an angle correction has to be added to the model:

$$f_d = f_0 - f_r = \frac{2 f_0\, v\, \cos\alpha}{c} \tag{1.9}$$

where f_d is the Doppler frequency shift, f_0 and f_r the emitted and the received frequency, v is the mean flow velocity of the reflecting red blood cells and α the angle between ultrasound beam and direction of blood flow (Figure 1.14). For angles of about 90°, the cosine function yields values around zero, at which there is no Doppler frequency shift. The blood flow velocity is then computed as

$$v = \frac{c}{2 f_0 \cos\alpha} \tag{1.10}$$

The accuracy of velocity measurements increases with the acuity of the angle. Usually, an angle around 60°is used, while larger angles result in high errors in the velocity estimate.

1.5.6 The minimum and maximum detectable velocity

The signal received from a single moving blood scatterer depends on its velocity and on the pulse repetition frequency. It will have the same shape as the emitted pulse, but its perceived time scaled and, thus, frequency will be different from the RF pulse. The time shift between the individual lines is

$$t_s = \frac{2 v_z}{c} T_{prf}$$

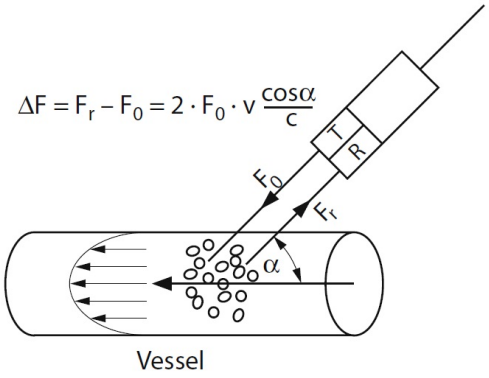

Figure 1.14: Effect of the angle of incidence on the Doppler measurement. T trasmitter, R receiver, F_0 emitted frequency, F_r reflected frequency.

and it increases linearly with the line number i. Then, a single sinusoidal component is received, as explained in the previous chapters. It is worth noting that T_{prf} corresponds to the time which is the inverse of the central frequency f_0. The frequency of the received signal is $\frac{2v_z}{c}f_0$, that is to say a scaled version of f_0. Thus, the center frequency of the pulse is transformed to $f_p = \frac{2v_z}{c}f_0$ and the spectrum of the received signal has the spectral shape of the pulse, with a scaled frequency axis. This is under the assumption that a sufficient number of lines are acquired to sample a whole pulse duration.

At least one period of the waveform needs to be observed for detecting the velocity and for distinguishing the signal from that of a stationary structure. The lowest possible velocity is thus found from

$$NT_{prf} = \frac{c}{2v_{min}f_0} \tag{1.11}$$

making the minimum detectable velocity equal to

$$v_{min} = \frac{cf_{prf}}{2Nf_0} \tag{1.12}$$

and the minimum frequency $f_{min} = \frac{f_{prf}}{N}$. If fewer lines are acquired, only part of the pulse is seen and this corresponds to weighting the pulse with a rectangular window, of which the spectrum should be convolved with the

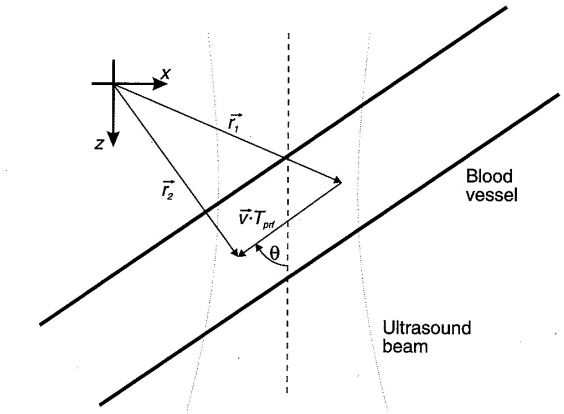

Figure 1.15: Coordinate system for blood particle moving through an ultrasound beam.

pulse spectrum. The pulse spectrum is narrow for very slow velocities and the resulting spectrum is nearly solely determined by the windows spectrum. However, the window has no effect if the whole pulse is sampled.

The maximum velocity is determined by the pulse repetition frequency because aliasing occurs for frequencies above $\frac{f_{pfr}}{2}$. The relation is

$$\frac{f_{prf}}{2} \leq \frac{2v_{max}}{c} f_0$$

so that

$$v_{max} = \frac{c}{2} \frac{f_{prf}}{2f_0}. \tag{1.13}$$

More precisely v_{max} is a bit lower since aliasing of the components above f_0 in the pulse spectrum takes place.

Chapter 2

Data Extraction, Period Estimation and Fourier Smoothing

Up to now, it has been described the amount of data at our disposal, but we need to extract the information we want to use from the images acquired through Doppler ultrasound. In this chapter, thus, we present the extraction of the region of interest from the whole image, from which we cut just the box containing the Doppler waveform. Once having the spectrum in a grey scale, some changes have to be taken before proceeding, like a threshold filter (section 2.1.2) in order to remove the noise and a rescaling of the axes, to express measures in cm/s and in seconds instead of number of pixels. After this little changes, it is possible to concentrate on the spectrum. In section 2.2, we analyse the spectrum of blood velocities for a fixed time, with the purpose of estimate the probability density it represents. This attempt was a first exploration of the data and a possible way to reduce the noise without the use of the initial filter.
In section 2.1.3 we return to look at the spectrum moving in time, trying to estimate a function representing the velocity of the red blood cells in time. We will focus on the reason why, later on, we will keep on treating only the 95^{th} quantile of the histogram of the data available at each fixed time. Nevertheless, we will save also other statistical sample indexes such as the mode, the median, the mean and the interquartile interval (IQR), since they can still give some information about the flow. The following step is the smoothing of the points extracted, so that we can handle functional data.

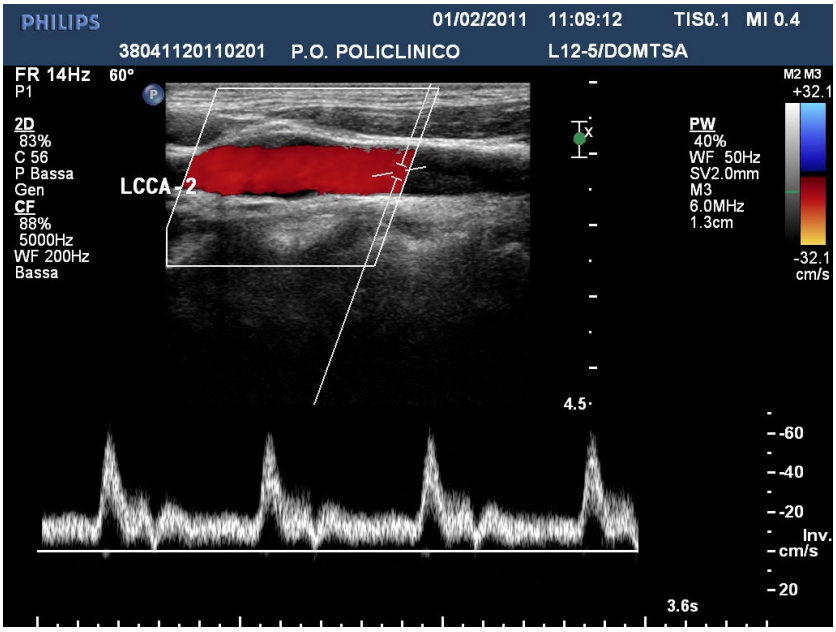

Figure 2.1: Duplex scan showing both B-mode image and spectrogram of the left common carotid artery of patient number 10 of our study, at -2 cm before the bifurcation.

Since the signal is periodic (the blood flow depends on the heartbeat), we first settle a new procedure to iteratively estimate the period of the signal and, once obtained, we can use Fourier smoothing on the data. Finally, the alignment of the functions coming from each patients is performed.

2.1 Extraction of the data

2.1.1 Extraction of the Region of Interest

We present in this section the procedure we followed for processing each duplex image of the study. In fact, of the whole Doppler image (for example the one in Figure 2.1, that refers to patient number 10 of our study, left CCA, level -2), we are interested just in the spectrum waveform at the bottom of the picture. Thus, when importing the figure, we cut it in order to maintain only the frame containing the information needed. The cutting procedure

is simple and constant for all the images, but one has to pay attention to quantities reported along the axes: the box containing the spectrum has always the same dimensions, but whilst the time axis (the horizontal axis) is constant for all the pictures (about 4 seconds are showed), along the vertical axis, which represents velocity values, things may change from patient to patient. This is why in Figure 2.2, the original axes with their ticks are still maintained: we need them to properly scale the measurements which will be computed, otherwise expressed in pixels. We have manually detected the horizontal ticks and inserted the scale values, but this procedure could be automated by the use of a Sobel filter[1], as explained in [9]. Once having saved these landmarks, only the spectrum is kept, cut at the level of the zero flow line (ZFL, section 1.4 of Chapter 1), as Figure 2.3 shows. In fact, under this line, usually there is the backflow of blood after systolic peak. This backward flow is absent in the common carotid artery, because it is a low-resistance flow that has to reach the brain continuously [3], thus the flow is steady throughout diastole and we do not loose any substantial information by cutting the black zone under the ZFL.

2.1.2 Threshold filter

The spectrum showed in Figure 2.2 is clear and well defined. Unfortunately, not all the Doppler images have the same quality, because factors like fat tissue or arrhythmias can interfere in the acquisition process, resulting in a very noisy spectrum, like the one shown at the top of Figure 2.4. Thus, filtering the image before proceeding is necessary. In [9] and [10] a threshold filter is proposed, which automatically computes the threshold as the pixel intensity value such that the 25% of pixels in the image results over that level. After this first filter, in order to remove outliers, a median filter is applied, which consists in running through each column of the image (thus along a fixed time of the cardiac cycle) pixel by pixel, replacing each value with the median of the first few preceding and following entries [12]. This is a completely automated procedure which works well with most of the noisy images and that we would recommend in case of treating a large amount of images. Anyway, we did not apply this kind of threshold filter, because of the limited number of frames at our disposal. In fact, among the Doppler

[1]Sobel filtering is often used in image processing, particularly within edge detection algorithms. It consists in a discrete differentiation operator that computes an approximation of the gradient of the image intensity function.

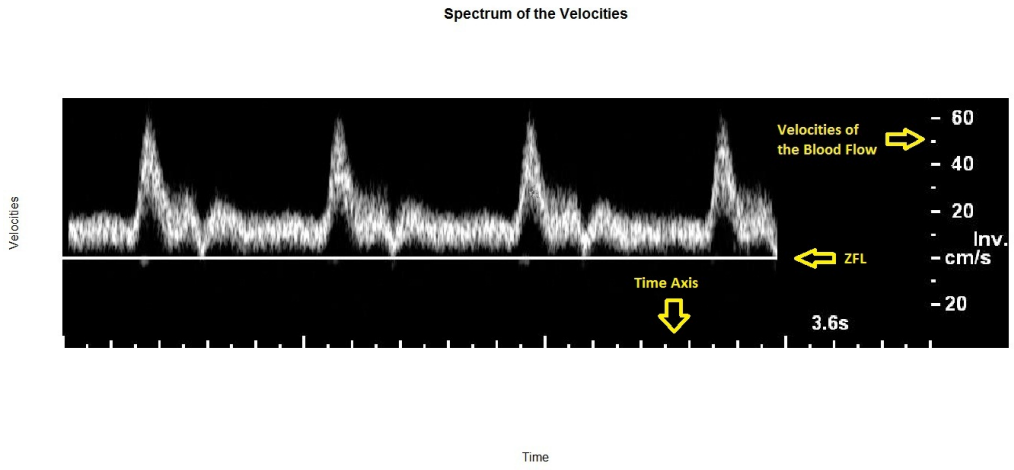

Figure 2.2: Clipping of the frame containing the spectrogram from the Doppler Image.

images of the patients in our study few extremely noisy pictures occur, on which the filter described in [9, 10] is not strong enough, but we can not afford to further reduce the number of patients wasting these cases. Figure 2.4 shows the filters just described applied to one of these cases: on top there is the original frame, in the center the same frame filtered and at the bottom the final result, after the median filter. It is clear that the noise is still too much. Moreover, the final median filter applied slightly modifies the histograms of the velocity intensities (the few black pixels under the velocity flow are completely removed).
The filter we have chosen and applied, instead, is more immediate: for each frame we identify the pixel with the largest intensity, which is for sure on the Doppler velocity profile. From this, we compute the threshold value just as a percentage of the highest intensity and we filter out the whole image by setting to zero all the values under the threshold. Usually it is enough to choose a percentage of 20%, but for images really noisy as the one in Figure 2.4, the threshold could also reaches values of 90% of the highest pixel intensity. Figure 2.5 shows the result of this procedure. The frame obtained has been shown to doctors for visual evaluation, and it was confirmed that we are not loosing any information by filtering the Doppler spectra.

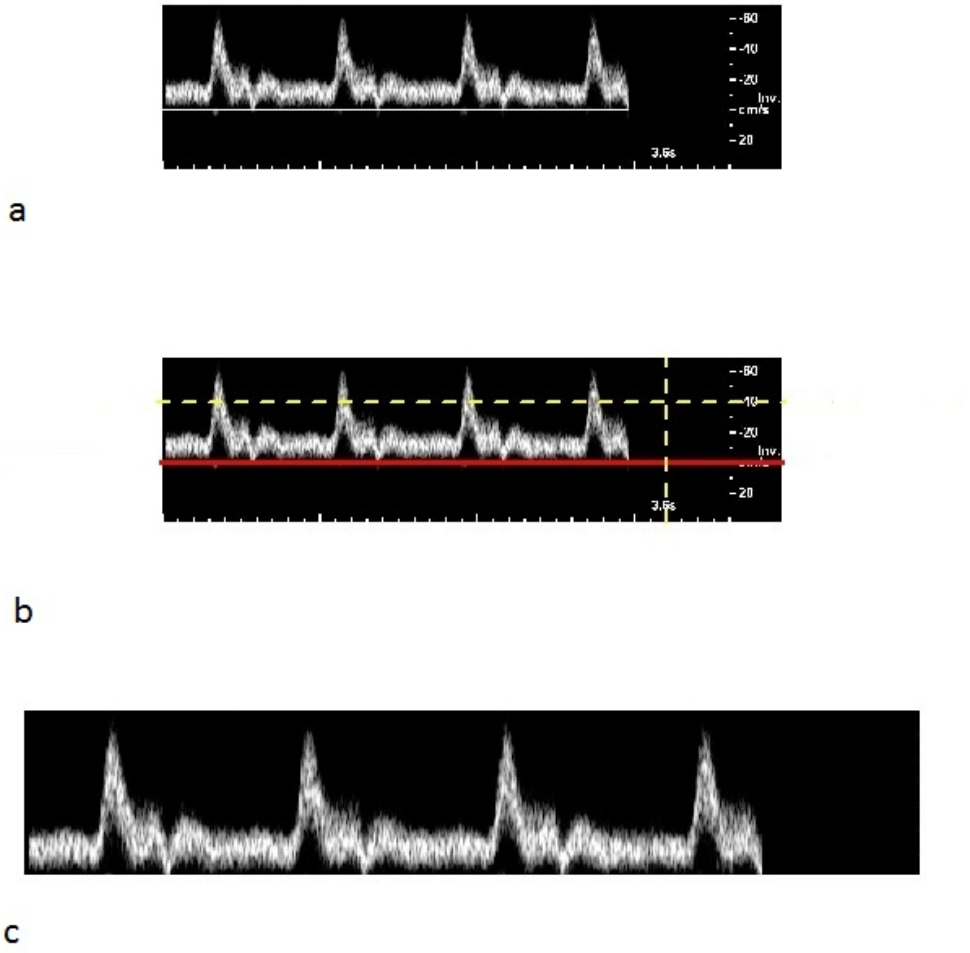

Figure 2.3: Sequence showing the steps of the clipping of the frame with the spectrum waveform. In (a) the axes are still maintained in order to detect the zero-flow line (red line in (b)) and the landmarks for the rescaling of the units from number of pixels to cm/s or seconds (yellow dashed lines). Finally, the frame showing the Doppler spectrum of the velocities cut (c).

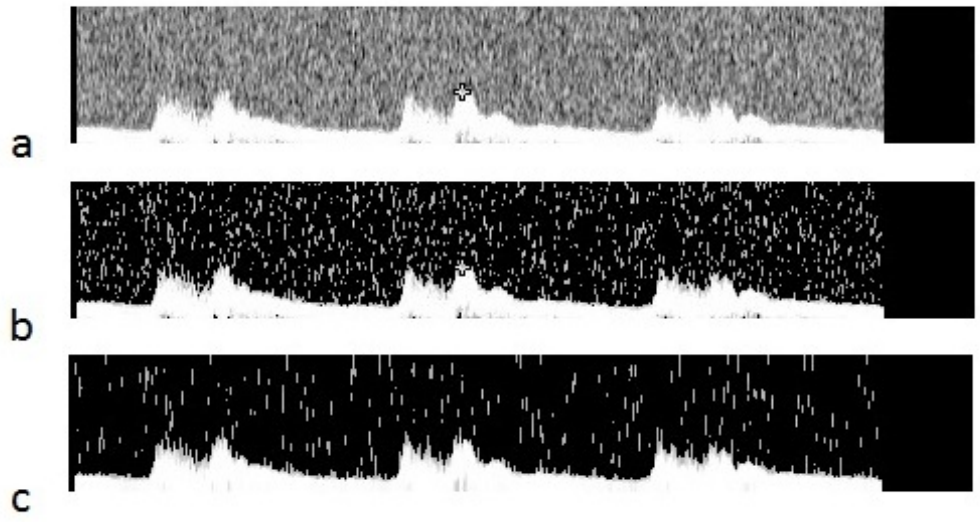

Figure 2.4: Example of a noisy image (patient 17, Right CCA -2) treated with an automated threshold filtered (as illustrated in [10]): on top (a) there is the original frame, while (b) shows the same frame filtered out by automatically computing the threshold as the pixel intensity value such that the 25% of pixels in the image results over that level. At the bottom (c) there is the final result, after the median filter, applied in order to remove possible outliers.

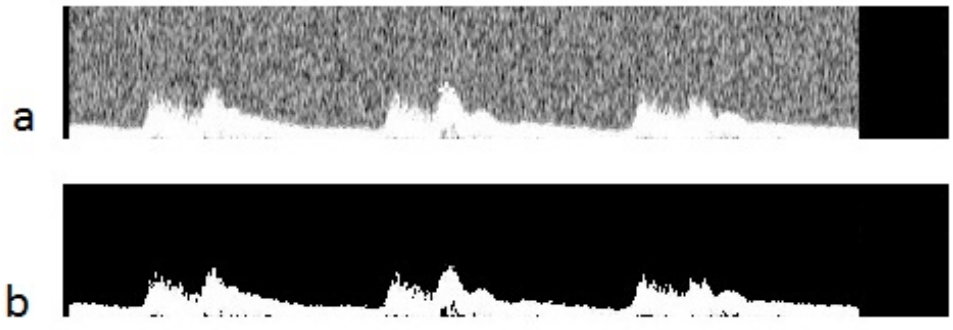

Figure 2.5: Example of a noisy image (patient 17, Right CCA -2) treated with the filter we developed, which simply computes the threshold value as a percentage of the highest intensity among the pixels and sets to zero all the pixel values under the threshold. On top there is the original frame, whilst at the bottom the image filtered out, with a threshold value of 90% the highest pixel intensity.

2.1.3 Detection of the Doppler flow velocity

The analysis of the clinician of the Doppler flow velocity is based on manual identification of some important points, such as the systolic and diastolic peaks. The doctor considers only the highest points of the spectrum of velocities and can read the values of the peaks by manually positioning the cursor on the point of interest on the screen. In order to automatize this procedure, in [10] an edge detection method is explained, consisting of scanning the image, for each time-column, from the top to the bottom, searching for the first pixel with an intensity greater than a predefined threshold, assumed as the contour of the velocity profile. Instead, to automatically detect the Doppler velocity profile from the Doppler tracings, for each image separately, we calculate the frequency histogram at a fixed time for each time step and, from this, we extract some statistical indexes: 95^{th} sample quantile, mode, mean, median, variance and the interquartile range $q_{95} - q_{25}$. Once having extracted these sample indexes, it is possible to continue analysing all of them, but the 95^{th} quantile has actually more importance, and it is the one on which we will present the results. In fact, the reason why medical doctors look at the

highest point of the flow profile lies in how the Doppler signal is sampled: the sample volume is fixed and it has a diameter between 2.0 and 3.8 mm [9]. When the lumen of the vessel is narrow (maybe because the presence of a plaque), the sample volume catches effectively the whole velocity profile and looking to a high level quantile means looking to the highest velocities reached by the blood flow. Figure 2.6 can help to visualize why one should look to a high quantile: other indexes such as the mean or the median depend on the dimension of the vessel respect to the sample volume and they do not represent the maximum speed reached by the blood flow in the section of interest. In a parabolic flow profile, as it should be theoretically when the CCA is approximated with a cylinder, considering the quantile has the strong physical sense of looking to the highest speed, that is to say where the signal is higher. Finally, the 95^{th} sample quantile, and not the maximum sampled velocity value or any higher order quantile, is chosen, in order to avoid outliers generated by noise.

In Figure 2.7 the 95^{th} sample quantile, the sample mean and the sample mode extracted from the images are presented superimposed to the original velocity spectrum. This figure has been subjected to medical advice, that agrees with the choice of considering the 95^{th} sample quantile as representative of the blood flow profile. Figure 2.8 shows the sample IQR and the sample mean $\pm 1.5 * \sqrt{variance}$. From the latter it is possible to see how the variance represents the spectral broadening due to disturbed flow, and this is the reason why it has been computed for each frame.

2.2 Density estimation at fixed time

The Doppler spectrum of velocities can also be read as an histogram that varies over time. As a matter of fact, once fixed one specific time of the cardiac cycle, the grey intensities represent the number of red blood cells moving at that specific velocity in the sample volume and, thus, a density of points. From this, the idea of estimating the probability density function at fixed time arises and one could ask if the extraction of the statistical indexes (95^{th} quantile, mean, mode and so on) by the estimated probability density function instead of directly using the sample indexes could be a method to reduce noise. In order to answer this question, we performed a local polynomial fitting with kernel weights, which allows to estimate the density at each

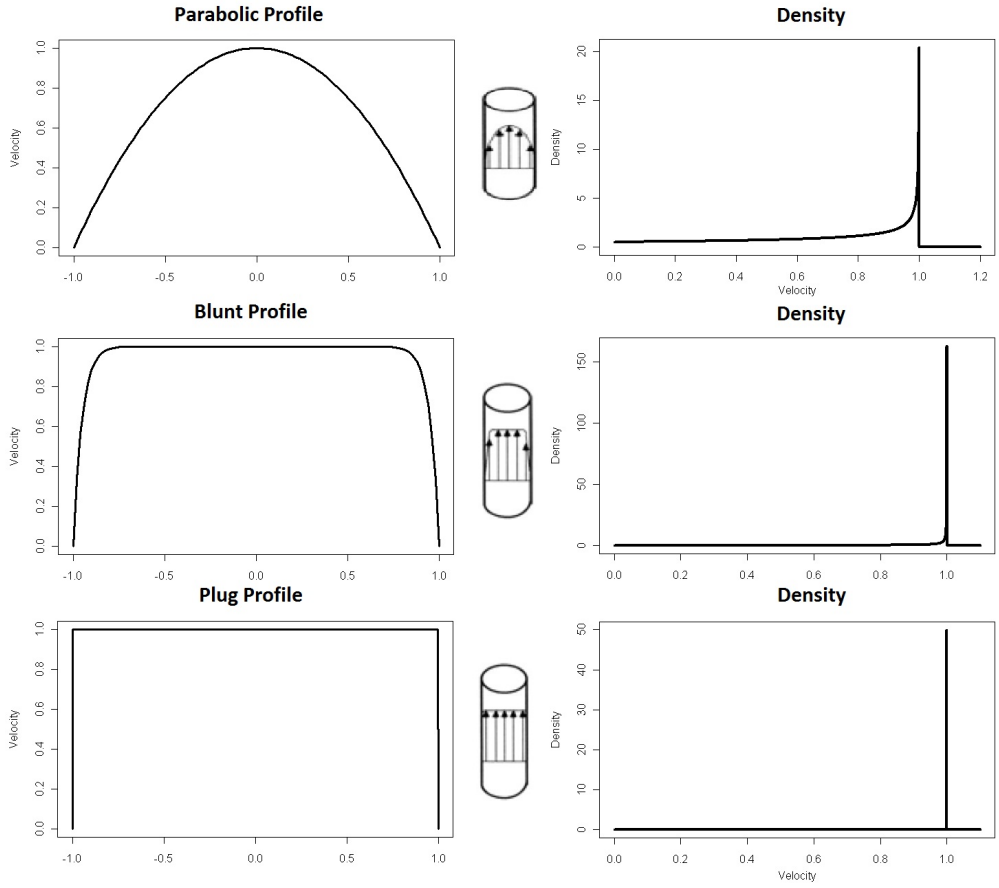

Figure 2.6: Different kind of velocity profiles and their spectra. When blood enters a narrow lumen, parabolic flow changes into plug flow and then return the original parabolic profile.

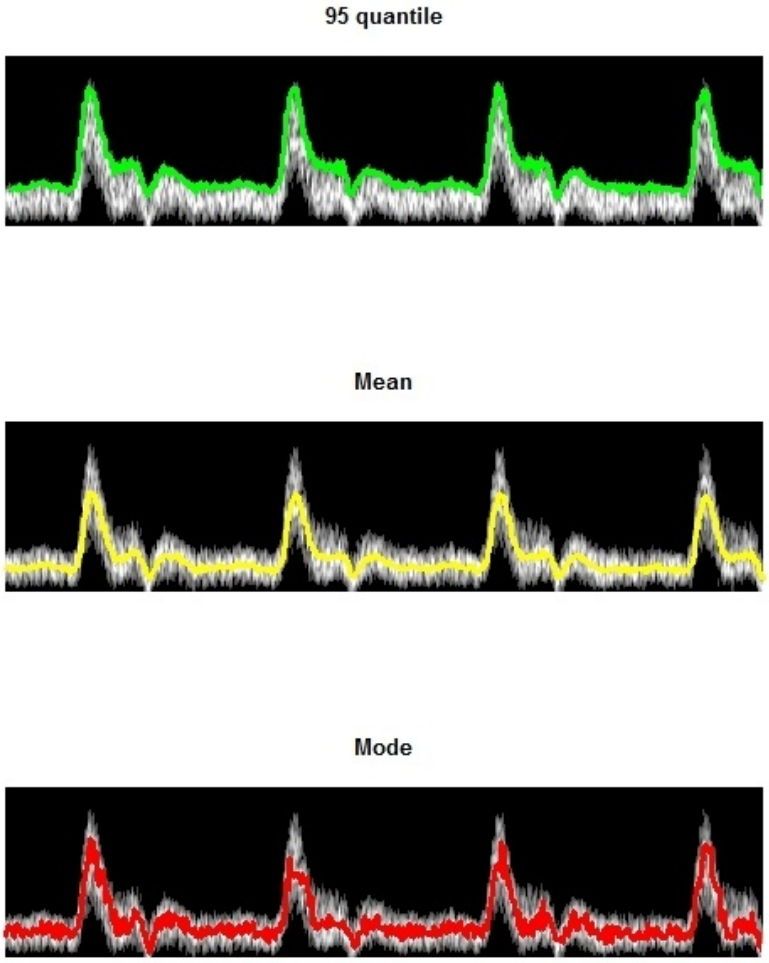

Figure 2.7: 95^{th} sample quantile, sample mean and sample mode extracted from the histograms are presented superimposed to the original velocity spectrum.

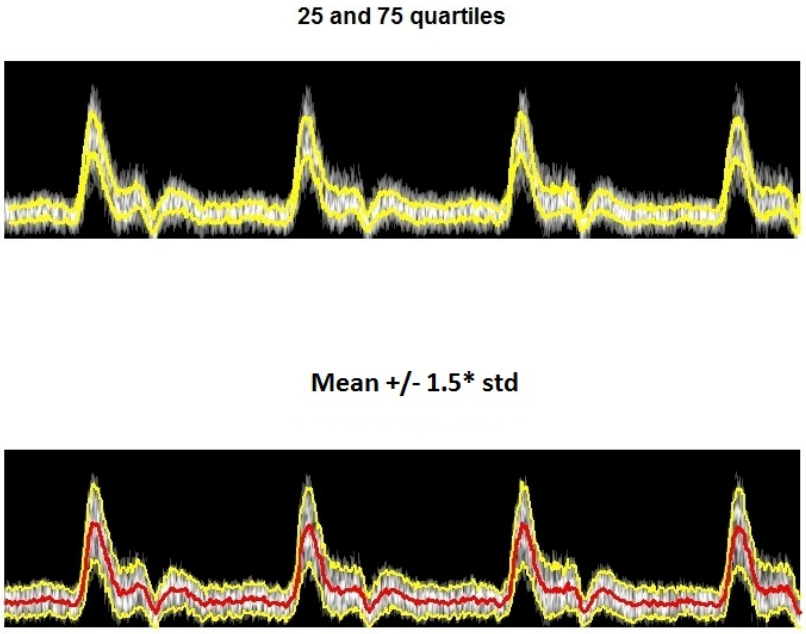

Figure 2.8: 25^{th} and 75^{th} sample quantiles, in the top panel, and sample mean $\pm 1.5 * \sqrt{variance}$, in the bottom panel, extracted from the histograms are presented superimposed to the original velocity spectrum.

time and we noticed that there are not substantial differences in the sample statistical indexes extracted directly from the histograms rather than from the estimated density function. Thus, we will not recur to the density estimation further in this thesis, since it does not add value to the analysis we will carry on, but we now present anyway the local polynomial smoothing applied.

Local polynomial smoothing is a localized least squares fitting method. The idea behind these kind of methods is substantially that the value of the function estimate at a point t should be influenced mostly by the observations near t. These techniques are widely explained in [**?**, 20], while a brief but clear enough explanation can be found in [13], from which we take our notation. Local polynomial smoothing estimates the values of the function y at the argument t by minimizing the weighted sum of the squared errors, with

respect to the coefficients c_l of the polynomial. The objective function to be minimized is thus:

$$\sum_{j=1}^{n} w_h(t_j, t) \left[y_j - \sum_{l=0}^{L} c_l (t - t_j)^l \right]^2 , \tag{2.1}$$

where the weights $w_h(t_j, t)$ are chosen such to consider relatively closed points, in order to include only those values t_j fairly close to the target value t. The parameter h is called *bandwidth parameter* and it indicates how far from t the values t_j are still considered. The localizing weights $w_h(t_j, t)$ are constructed through a *kernel* function, built such that it has the most of its mass concentrated at zero and it decays quickly. In particular, we will use a Gaussian kernel function:

$$kern(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2),$$

defining the weights values as

$$w_h(t_j, t) = kern(\frac{t_j - t}{h}).$$

The bandwidth parameter h controls the balance between bias and variance: a small value of h generates an estimate $\hat{y}(t)$ close to the true value $y(t)$, but with a high variability since few observations are used. On the contrary, large values of h decrease the variance of the estimate but increase the bias, because the values used cover a wider region. Different data-driven techniques for choosing h have been developed and they can be found in [22] and [21]. We will use a rule-of-thumb developed by Silverman [20], that chooses the value of the bandwidth for a Gaussian kernel density estimator proportional to the standard deviation.

To perform a local polynomial smoothing on the spectra of velocities, first of all each column of the spectrum has to be extracted and smoothed down independently from the other columns. So, for each time, we have a vector of velocity observations, each one with an intensity value representing the number of red blood cells in the sample volume moving at that specific velocity. In equation 2.1, we set $L = 0$, recovering in this way the Nadaraya-Watson estimator of $y(t)$:

$$\hat{y}(t) = \sum_{j}^{n} S_j(t) y_j,$$

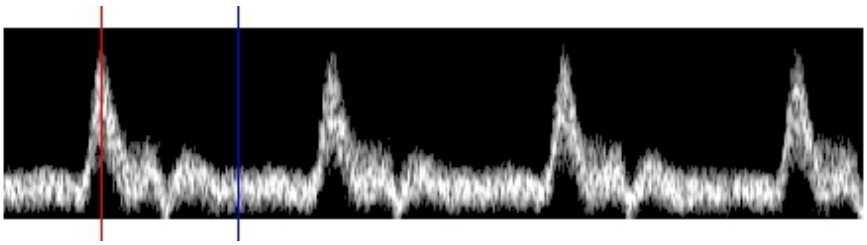

Figure 2.9: Choice of 2 times of the cardiac cycle in order to perform a kernel density estimation. The first time (red line) is $t = 0.369$ seconds and it corresponds to a systolic peak, while the second one (blue line), $t = 0.9s$, corresponds to a generic time during diastole.

with the weights values normalized to have a unit sum

$$S_j(t) = \frac{kern[(t_j - t)/h]}{\sum_r kern[(t_r - t)/h]}.$$

Figure 2.9 shows an example of two fixed times, chosen such that the first (red line) corresponds to a systolic peak and the second one (blue line) to a random time during diastole. Kernel density smoothing can be performed using the function `locpoly`, contained in the `R` package `KernSmooth`, by setting the degree to zero and choosing the gaussian kernel. The bandwidth has been set to $h = 3.0$ (meaning that each value is obtained weighting the intensities of the neighbour velocities not farer than 3 cm/s), estimated through the function `density`, which allows to choose the method for the estimation of the bandwidth (in our case, Silverman's method). Figure 2.10 shows the probability densities estimated for the two time points fixed. Increasing the value of the bandwidth h would lead to a smoother estimation, but also to a considerable decrease of the maximum intensity reached (indicated with a vertical dashed line in the figure), meaning that the density we are estimating reduces the number of particles moving at the central velocities. We do not desire this effect, which can be clearly seen in Figure 2.11, because we do not want to modify the distribution of the velocities of the red blood cells, thus we should keep on using a bandwidth value around 3. Finally, in Figures 2.10 and 2.11, the 95^{th} quantile and the mode extracted from the estimated density function are shown with vertical dashed lines.

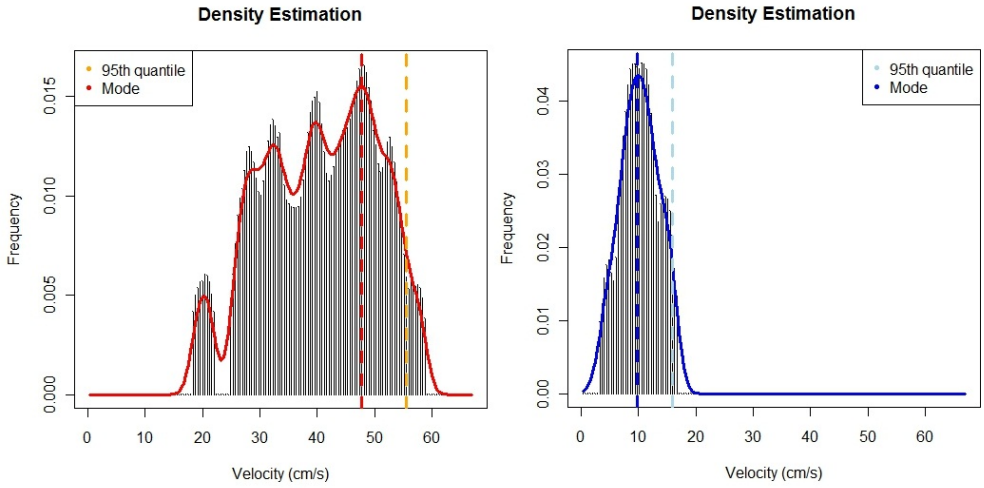

Figure 2.10: Probability density functions estimated for the two times indicated in Figure 2.9, superimposed to the intensity histograms. In this case, the bandwidth value has been set to 3.

2.3 Fourier Smoothing of Doppler Velocity

In this section we are going to explain the procedure developed to estimate the period of the data and how to pass from the data extracted from the Doppler waveform to a smooth function. First, we briefly introduce the Fourier smoothing and the non linear least squares estimate. Using these methods, we will be able to represent our data through smooth periodic functions and, thus, to register them and to compare the curves of the blood velocities from various patients.

The Fourier basis system

For each Doppler image, we will treat the data extracted, and in particular the 95^{th} sample quantile, as functional data, meaning that we consider them as a discrete acquisition of n pairs (t_j, y_j), where y_j is generated by a function $x = x(t_j)$, of which we suppose the existence, plus a term of noise ϵ_j [13]:

$$y_j = x(t_j) + \epsilon_j.$$

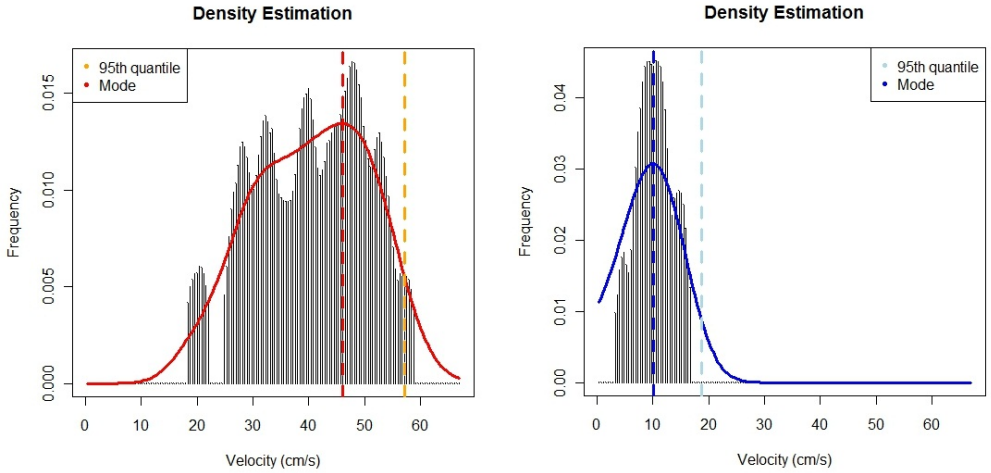

Figure 2.11: Probability density functions estimated for the two times indicated in Figure 2.9, superimposed to the intensity histograms. In this case a bandwidth $h = 10$ was chosen: a considerable decrease of the maximum intensity reached can be seen, with respect to the density estimations shown in Figure 2.10.

We desire the function x to be *smooth*, that is to say continuous and that one or more derivatives of x exist and are continuous. In our specific case, the grid $\{t_j\}$ over which data are collected is not the same for all the records: the number n of pairs (t_j, y_j) varies from frame to frame, while the distance between two consecutive time points Δt is always the same and it corresponds to one pixel.
The function x can be represented as a weighted sum of K basis functions Φ_k, orthogonal to each other and belonging to the same space as x (i.e. $\mathcal{C}'$ for previous definition). Following [13], we can thus estimate x as:

$$\hat{x}(t) = \sum_{k=1}^{K} \hat{c}_k \Phi_k(t) \tag{2.2}$$

or, in matrix notation:

$$\hat{x} = \hat{\mathbf{c}}'\mathbf{\Phi} = \mathbf{\Phi}'\hat{\mathbf{c}}, \tag{2.3}$$

where $\hat{\mathbf{c}}$ is the vector of the K coefficients $\hat{c}_k$ and $\mathbf{\Phi}$ is the vector whose elements are the basis functions. The number K of basis functions determines

the degree of smoothness of the function x: when K is large, more computation is required and x almost interpolates the data (we have interpolation if $K = n$). Instead, a low value of K leads to a more smoothed function. Later on, we will also make use of the first derivative of x, which can be estimated as:

$$D\hat{x}(t) = \sum_{k=1}^{K} \hat{c}_k D\Phi_k(t) = \hat{\mathbf{c}}' D\mathbf{\Phi}(t). \tag{2.4}$$

Our data, i.e. the statistical indexes extracted from the Doppler spectrum, are periodic, since blood flow follows the periodicity of the cardiac cycle. Moreover, for each patient, the curvature of the extracted data tends to be of the same order. These features suggest the use of probably the best known basis functions: the Fourier basis. Consequently, we can write equation (2.2) as a linear combination of sines and cosines:

$$\hat{x} = c_0 + c_1 \sin \omega t + c_2 \cos \omega t + c_3 \sin 2\omega t + c_4 \cos 2\omega t + \ldots, \tag{2.5}$$

having defined the basis through

$$\begin{aligned} \Phi_0(t) &= 1, \\ \Phi_1(t) &= \sin \omega t, \\ \Phi_2(t) &= \cos \omega t, \\ &\vdots \\ \Phi_{2r-1}(t) &= \sin r\omega t, \\ \Phi_{2r}(t) &= \cos r\omega t. \end{aligned}$$

The period of the basis is $T = 2\pi/\omega$, determined by the parameter ω, which we will estimate from the data.
After having estimated the period, the coefficients c_k will be estimated using least squares, that is to say fitting the model to the data by minimizing the sum of the squared errors. For this purpose, we used the package `fda` [26] implemented in `R` and described in [14]. We refer to [13] for more details about smoothing functional data by least squares, but we spend some further word about the choice of the number of basis K. As usual, a trade-off between bias and variance arises, since if K is large there is a good fit of the data, but at the same time the variance could be high and there is the risk of fitting the noise that one would desire to ignore by recurring to smoothing methods. On the other hand, if K is too small the variance is low, but the

risk is to loose some important feature of the function to estimate, leading to high bias. To control both variance and bias at the same time, a possibility is to check the mean-squared error $MSE[\hat{x}(t)] = E[(\hat{x} - x(t))^2]$, which links together bias and variance. We decided to use a scaled version of this index, the *generalised cross-validation* measure (GCV [17]), which can be expressed as:

$$GCV(K) = (\frac{n}{n-K})(\frac{MSE}{n-K}).$$

In particular, in order to choose a suitable value for K, it is possible to compute the values of the GCV sequentially (in our case, it has been done through `smooth.basis`, included in the package `fda`), applying consecutively a Fourier smoothing to the data by letting K increasing. Then, a widely used criterion is to search for an elbow in the plot of the GCV against the increasing number of functions K (Figure 2.12). Repeating this procedure independently for each frame, we had that the optimal number of basis functions suggested from this method would be an odd number between 11 and 15 (in the presented figure, for example, the optimal value would be $K = 13$). We thus choose to use $K = 15$ for all the patients and all the levels.

Non linear least squares estimates

In order to represent the velocity profile through Fourier smoothing, it is necessary to estimate the period of the data, which, just in case of the CCA, corresponds also to the period of the cardiac cycle. Instead, for ICA and ECA the frequency of the pulses of the blood flow diverges from the heartbeat and having a procedure to estimate the period could provide a useful information. For this purpose, we will apply nonlinear least-squares estimates of the parameters of a nonlinear model, consisting in a cosine function (in next section, it will be clear the reason of the choice of this model). Non linear regression allows us to adapt data to a curvilinear function, containing the period as a parameter, which we want to estimate. The estimation of the model parameters is done by the Gauss-Newton iterative method, illustrated in [15] and implemented in function `nls` of `R`. It is necessary to provide initial values of the parameter, that will be corrected at each iteration of the algorithm until convergence is reached. Since the procedure developed to estimate the period, described in next section, is fast and it works well, we will use non linear least squares estimates for estimating the periods of all the frames, regardless of if the data come from the CCA or from the internal

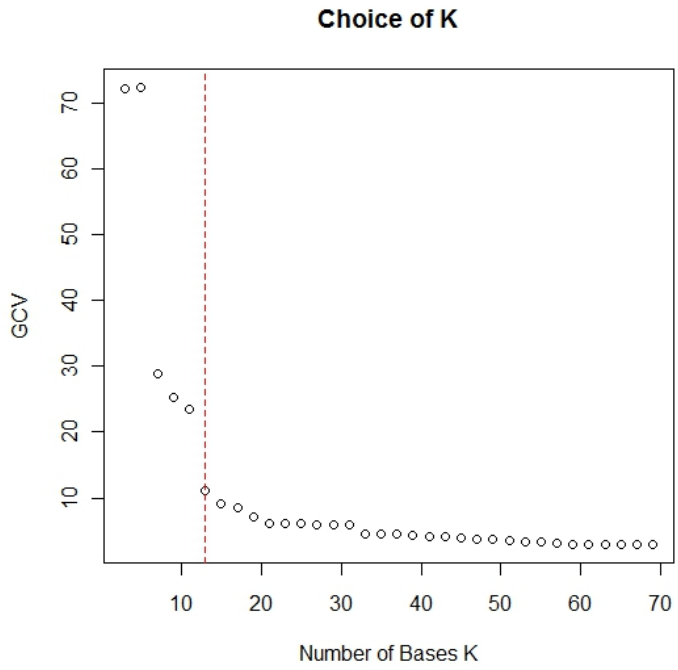

Figure 2.12: Plot of GCV values obtained by letting K increasing. The points correspond only to odd values for K and the dashed line indicates the elbow that corresponds, in this case, to the value $K = 13$.

Carotid arteries.

2.3.1 Estimating the blood flow period

In this section, we illustrate the new procedure we developed to estimate the period of the velocity blood flow. This step is necessary, because using directly the heartbeat period, which is recorded directly by the doctor, could lead to mistakes because it is not granted that the period of the data extracted from the Doppler images is exactly the same of the heartbeat period. We will thus estimate the period from each Doppler frame of our project. Since for each patient three images at three different levels of the CCA (-2,-1 and -0.5) are available, afterwards it will be possible to carry out a control of the estimated periods. Anyway, the images at the three levels of the CCA

have been registered at different moments of the same medical examination and meanwhile the conditions of the patient could have slightly changed. So, we do not expect to estimate exactly the same period of the blood flow along the CCA, but we do expect something similar, at least. What should be almost identical, instead, are the periods estimated from the position indexes (such as the 95^{th} sample quantile, the mode and the sample mean) and from the dispersion indexes (sample variance and IQR) extracted from the same Doppler frame.
This method differs from the procedures usually applied to estimate the period of sinusoidal data. In fact, it is typically estimated using the FFT or by observation of the zero-crossing times [27]. Instead, our method is based on a first Fourier smoothing, maintained very close to the original data since the period used is the whole range of time values. Then, we use non-linear least squares to estimate the period of the function obtained.
The algorithm implemented in order to estimate the period, applied to each statistical sample index, is the following:

Step 1 Initialization. A Fourier basis is generated, to specify which one needs to define two informations: the number of basis functions and the period T. In this first step of the algorithm, we choose a number of basis $K = 25$ (maintained constant for all the indexes and all the images), which is high enough to catch all the features of the data. Regarding the period, the default value of the range of time values t spanned by the data is given. Defined the basis functions, the data are then smoothed down computing the coefficients $\hat{c}_k$ in (2.2) through least squares estimates.

Step 2 Detection of the first peak. The data now are represented by a smooth function x, of which we need to detect the position of the first peak. This is done by detecting the first time at which the second derivative of x is negative and the first derivative is zero. A further check is added, verifying that the value of x at the obtained point is over a certain threshold (such as `max(data)/2`). It is possible to estimate the derivatives of the r basis function by using

$$\begin{aligned} D \sin r\omega t &= r\omega \cos r\omega t, \\ D \cos r\omega t &= -r\omega \sin r\omega t \end{aligned} \tag{2.6}$$

The coefficients for the Fourier expansion of Dx are already available,

since they are $(0, c_1, -\omega c_2, 2\omega c_3, -2\omega c_4, \ldots)$, whilst for D^2x the coefficients are $(0, -\omega^2 c_1, -\omega^2 c_2, -4\omega^2 c_3, -4\omega^2 c_4, \ldots)$.

Step 3 Estimation of the period and registration. Translating the function x so that the first peak corresponds to the origin on the time axis, it is possible to adapt it to a simple periodic non-linear model, which also have the first peak in zero:

$$x \sim a + b\cos(\omega t). \tag{2.7}$$

The parameters of this model are estimated through non-linear least squares, using a Gauss-Newton iterative method. The parameters a and b have been added to the model for a scale reason. The method converges for a wide variety of initial values a_0 and b_0. On the contrary, one needs to pay attention to the initial value of the parameter ω, which is the most important, being connected to the estimation of the period through the relation $T = \dfrac{2\pi}{\omega}$. Its starting point ω_0 has been chosen, for each frame, as the range of time values available (i.e. the number of data) divided by the number of peaks in the frame.

Figure 2.13 shows the three steps of the algorithm, for the 95^{th} sample quantile index only: at the top both the data and the first smoothing are plotted, in the center a visual check that the right peak has been detected and at the bottom there are the smooth function x, translated so that its first peak is in zero, and the cosine model fitted out. With the estimate of the period, it is then possible to perform a final smoothing of the data, in order to obtain a function (of the 95^{th} quantile moving in time for instance) instead of discrete points. This will be done using the estimate of the period to create a Fourier basis, so that, when fitting the data, all the measurements obtained from consecutive cardiac cycles will be averaged among them. In fact, the original Doppler waveform contains variability among consecutive cardiac cycles, due to the acquisition procedure, and averaging them by a periodic Fourier smoothing is a good way to resume in just one cycle the original information.

2.3.2 Results

Since the period of blood flow in the CCA should be similar to the heartbeat, we have a way to test the validity of our method. The typical resting

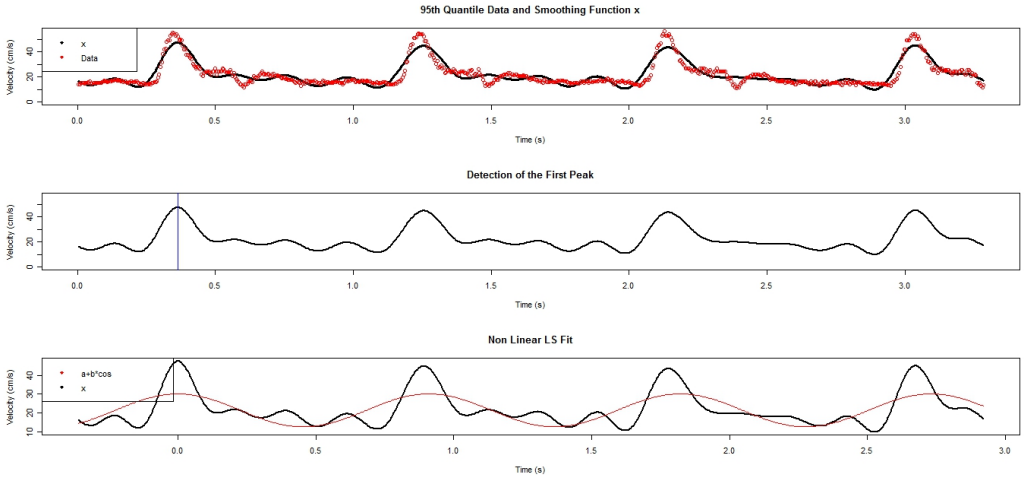

Figure 2.13: The 3 steps of the algorithm which allows the estimation of the period, for the 95^{th} sample quantile of patient 10, left CCA, level -2. In the top panel both the data, as extracted from the Doppler frame, and the first smoothing are plotted, in the central panel a visual check that the right peak has been detected and at the bottom there are the smooth function x, translated so that its first peak is in zero, and the cosine model fitted out.

heart rate in adults is 60-80 beats per minute[2], but it depends on the physical condition of the person. For example, in a trained individual, it could reaches lower frequencies of 50-60 bpm. Thus, when estimating periods of the blood flow in carotid arteries, we should expect values around 0.75 and 1.2 seconds.

After having extracted different statistical indexes from each frame, we should compare the estimates of the period of these data. So, using the 95^{th} quantile, the mean and the mode, we estimate the function's period for each index separately. This results in having, for each image, three different values for the period of the blood velocity. Comparing them, the differences are of the order of the hundredth of second and this suggests that the method is robust. Figure 2.14 shows the periods estimated from the 95^{th} quantile and the mode curves at -2: when it is possible to see only the red point (period of

[2]Resting Heart Rate, American Heart Association.

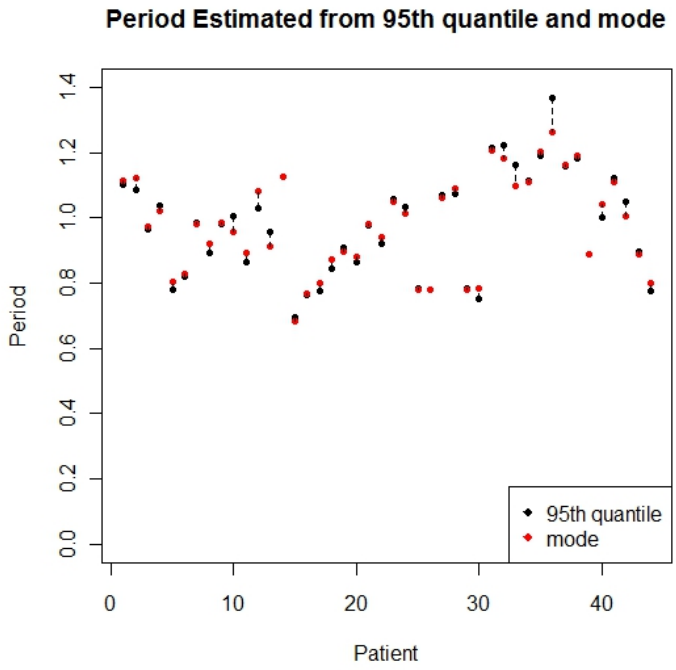

Figure 2.14: Estimation of the period of the blood velocity at -2 from the 95^{th} quantile (black dots) and from the mode (red dots). The distances between the two periods are shown with a black dashed line.

the mean) it is because the two periods are exactly the same; the difference between the 2 estimates for patient 36 is the largest one and it is equal to $\hat{T}_{q_{95}} - \hat{T}_{Mode} = 0.09s$. This procedure is repeated for all the images, at various distances from the carotid bifurcation and the results are shown in Table 2.1.

It is also possible to check how the period of the blood flow profile evolves along the common carotid artery: sometimes the value of the period at -2 is different from the one at -1 and -0.5, as Figure (2.15) clearly shows. We can justify this by the fact that the acquisition at -2, -1 and -0.5 have not been taken at the same moment, and some condition in the patient (such as breathing) could have been changed, influencing the heartbeat and thus the period of the signal. Moreover, the period estimates at -2 and -1 are more similar to each other (black and red line in the figure) than to values

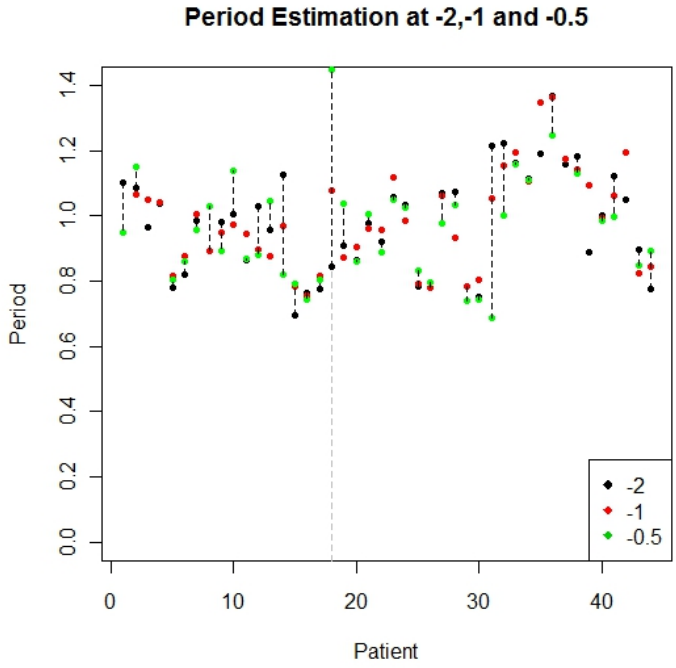

Figure 2.15: Estimation of the period of the blood velocity profile at a distance of -2 (black) from the carotid bifurcation, -1 (red) and -0.5 (green). The dashed line connects the periods at -0.5 with the periods at level -2.

at -0.5 (green line), which are more deviated: a probable explanation is the closeness of the carotid bifurcation, after which blood splits in two flows (ICA and ECA) with two different pulsatilities; furthermore, in this zone, the acquisition of the image is more difficult (as medical staff showed us) and sometimes the examiner has to try different times, moving the transducer to find the right location. Consequently, the estimates of the periods along the CCA of the same patient may differ from each other because the original Doppler images are different. To justify this we show the Doppler acquisitions for the right CCA of patient number 9 (the one for which the periods are more different, as the dashed vertical line in Figure 2.15 underlines), looking at which one can see at one glance that the periods are not the same at the three positions (Figure 2.16).

Finally, it has to be noticed that the estimation of the periods along the left

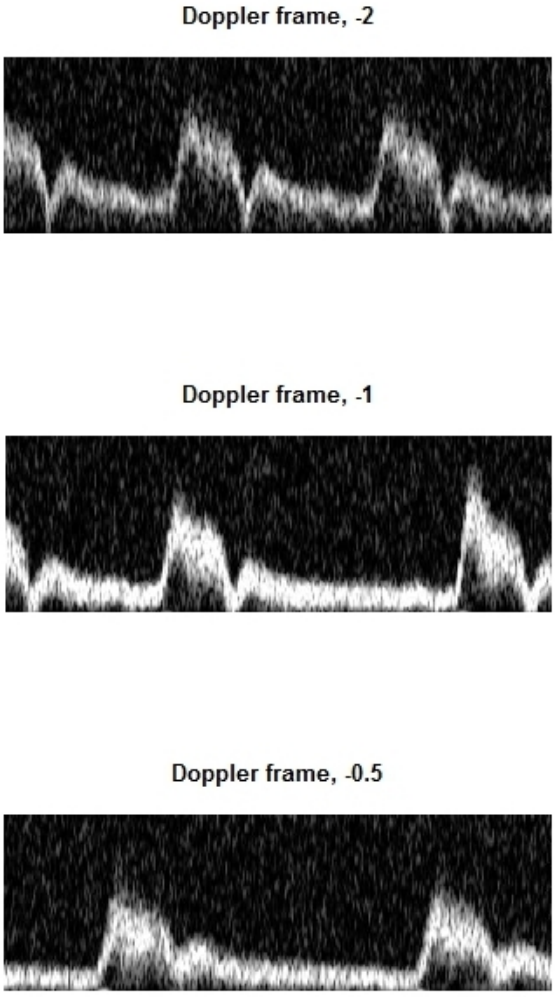

Figure 2.16: Doppler acquisitions at the three levels of CCA for patient 9, right carotid artery: the periods of the heartbeat are clearly not the same.

CCA are very similar to the estimations along the right CCA (looking at each level separately), and there are only few cases when the difference is greater than 0.1 s. Again, for these cases it is possible to notice a difference between the waveforms of the left and right carotid arteries already from the original frame. To strengthen the fact that the periods estimated along the left CCA are very similar to the periods of the right CCA, a paired Mann-Whitney U test [16] has been performed. It is a paired two samples case, where the null hypothesis is that the distribution of the periods of the left CCA minus the periods of the right CCA ($\hat{\omega}_{Left} - \hat{\omega}_{Right}$) is symmetric around zero. The p-values obtained from the test are $p = 46.28\%$ at -2, $p = 94.67\%$ at -1 and $p = 35.91\%$ at -0.5.

2.3.3 Final Smoothing

We now present the final smoothing, which creates the function that should properly represent the data. This is done by creating a Fourier basis with a period equal to the one estimated for each frame and a number of bases $K = 15$. The smooth function that we obtain has the property of being periodic and it is generated by fitting the data through least squares, using the package `fda` in `R`. It has to be noticed that the basis functions are periodic and they repeat themselves with exactly the same shape at every period. Thus, when fitting the data, all the measurements obtained from consecutive cardiac cycles are automatically averaged and this permits to reduce variability coming from the acquisition of the Doppler images and to take into account possible respiratory variations. The final smooth functions, obtained for each statistical index listed before in this chapter, presented as in Figure 2.17 and 2.18, have been shown to a clinician and he agreed that the smoothing of the 95^{th} quantile represents properly the information contained in the Doppler spectrogram. Thus, hereafter, we will carry on the analysis just on this index.

2.4 Curve registration

Now that the 95^{th} sample quantiles evolving in time for each patient are in functional form, we would like to proceed in the analysis of the data, comparing the curves coming from different patients. But before doing that, a registration of the data is necessary and, in particular, a transformation of

Patient	Left Carotid Artery			Right Carotid Artery		
	-2	-1	-0.5	-2	-1	-0.5
N.	T	T	T	T	T	T
1	1.103	n.a.	0.973	1.085	1.082	1.162
2	0.965	1.051	n.a.	1.039	1.042	n.a.
3	0.779	0.809	0.805	0.822	0.896	0.871
4	0.984	1.011	0.955	0.893	0.930	1.010
5	0.983	0.978	0.894	1.006	0.974	1.130
6	0.864	0.940	0.874	1.029	0.893	0.881
7	0.957	0.879	0.991	1.126	0.947	0.860
8	0.697	0.786	0.762	0.763	0.756	0.755
9	0.778	0.809	0.805	0.844	1.080	1.383
10	0.910	0.858	1.035	0.865	0.908	0.860
11	0.976	0.956	1.030	0.922	1.022	0.889
12	1.059	1.106	1.032	1.034	1.002	1.008
13	0.783	0.798	0.805	0.779	0.788	0.780
14	1.065	1.059	0.981	1.083	0.940	1.032
15	0.783	0.787	0.758	0.764	0.804	0.749
16	1.211	1.052	0.721	1.166	1.153	1.006
17	1.140	1.172	1.160	1.113	1.117	1.135
18	1.270	1.247	n.a.	1.297	1.356	1.309
19	1.159	1.173	n.a.	1.186	1.138	1.136
20	0.889	1.082	n.a.	1.023	1.016	0.978
21	1.118	1.065	0.999	1.059	1.199	n.a.
22	0.892	0.821	0.841	0.795	0.855	0.879

Table 2.1: Estimates of the velocity period, for all the patients and at various distances from the carotid bifurcation.

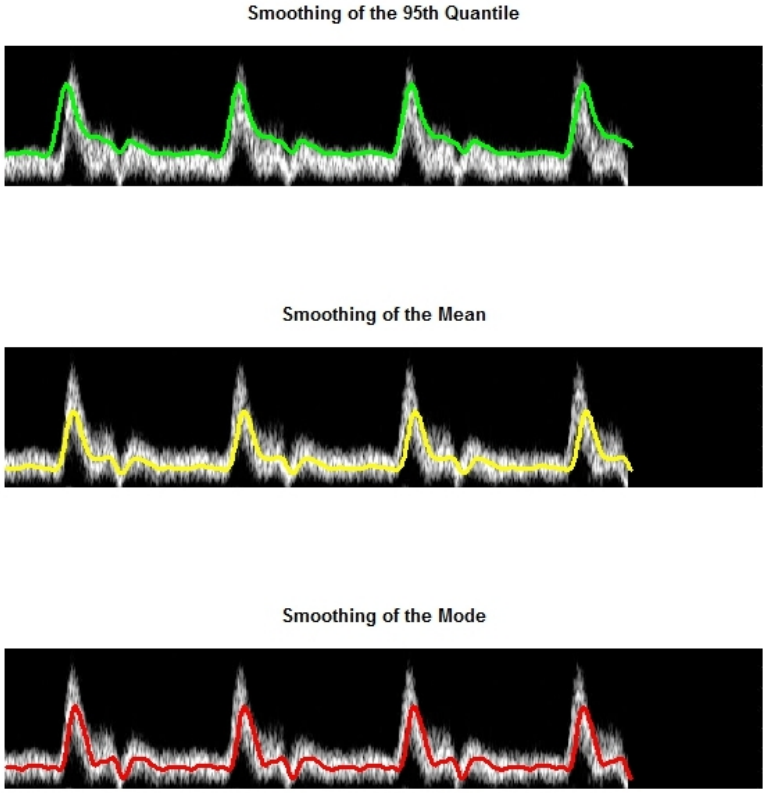

Figure 2.17: Final results of the smoothing of the 95^{th} quantile, the mean and the mode, obtained after having estimated the period of these indexes. Results for patient 10, left CCA, level -2.

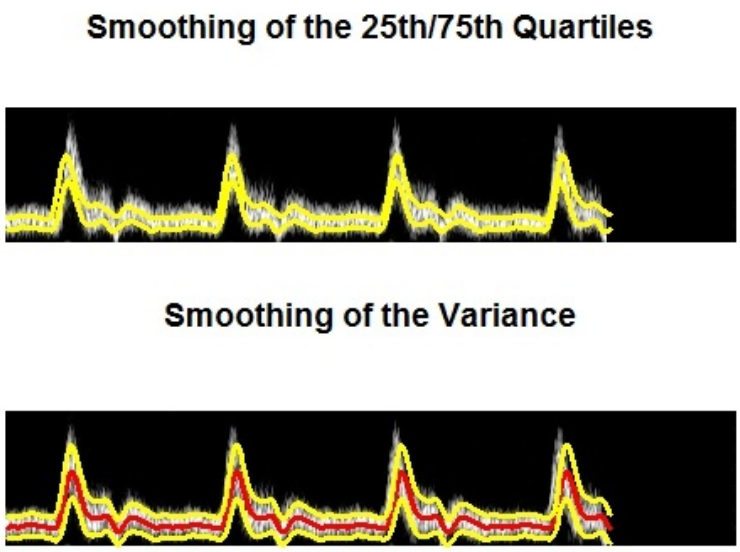

Figure 2.18: Final results of the smoothing of the 75^{th} and 25^{th} quantiles (top panel) and, at the bottom, of $mean \pm 1.5 * \sqrt{variance}$, obtained after having estimated the period of these indexes. Results for patient 10, left CCA, level -2.

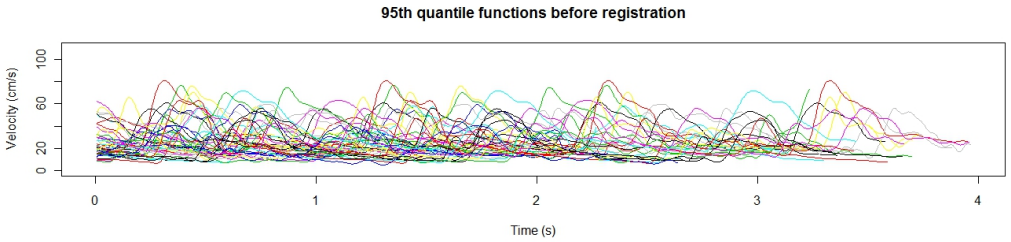

Figure 2.19: Functions representing the 95^{th} sample quantile for each patient at level -2, plotted before being registered.

the argument time. In fact, observing the whole functional observations on the same plot, as in Figure 2.19, it is possible to see basically two kind of variation: in phase and in amplitude. We will not modify the variation in amplitude, since it is the first feature exhibiting the presence of a stenosis. Instead, we will correct the phase variability, that is to say, the variations in the timings of the systolic peaks, because this kind of variability depends on the acquisition of the Doppler image by the examiner, who, obviously, does not begin recording the velocity signal at the same time of the cardiac cycle for each patient. We will also correct the fact that each curve has a different length (a different time interval).
Thus, in order to compare the functions coming from all the patients, the time scale has to be transformed, so that all the peaks of curves happen at exactly the same time, for all the patients and for all the levels. This is achievable by applying a landmark registration [13]. First of all, the number of observations and the number of cardiac cycles vary from image to image. Performing a Fourier smoothing implies averaging the information coming from different cardiac cycles and thus, it will be enough to represent only one cardiac cycle for each patient. The interval $\mathcal{T}$ over which the functions are to be registered is $[0, \bar{\omega}]$, where $\bar{\omega}$ has been chosen as the mean value of the estimated periods of the 95^{th} sample quantiles. In this way, all the functions will be represented with a uniform time axis $t \in (0, \bar{\omega})$, where t is the argument value of $x(t)$. The transformation performed in order to pass from each patient time grid to the uniform one is simple: the Fourier basis is evaluated in order to have the first systolic peak p_i (of which the position has already be detected during the estimation of the period) in the center of the range $\mathcal{T}$, that is to say from $p_i - \omega_i/2$ to $p_i + \omega_i/2$, where ω_i

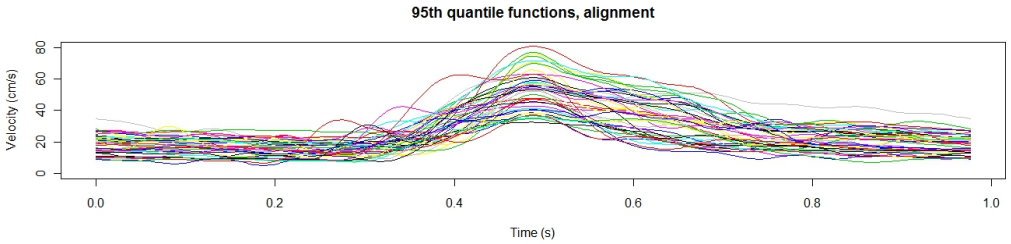

Figure 2.20: Functions representing the 95^{th} quantile registered functions at level -2. For each patient only one period of the cardiac cycle is represented.

is the period estimated for the patient i, with a time transformation of $\frac{\omega_i}{\bar{\omega}}t$. Figure 2.20 shows the registered functions over one period, while Figure 2.21 shows the registered functions over three periods in the top panel, while in the bottom panel the registered functions are shown with 2 different colours: the curves corresponding to TEA patients are red, while all the other are shown in black.

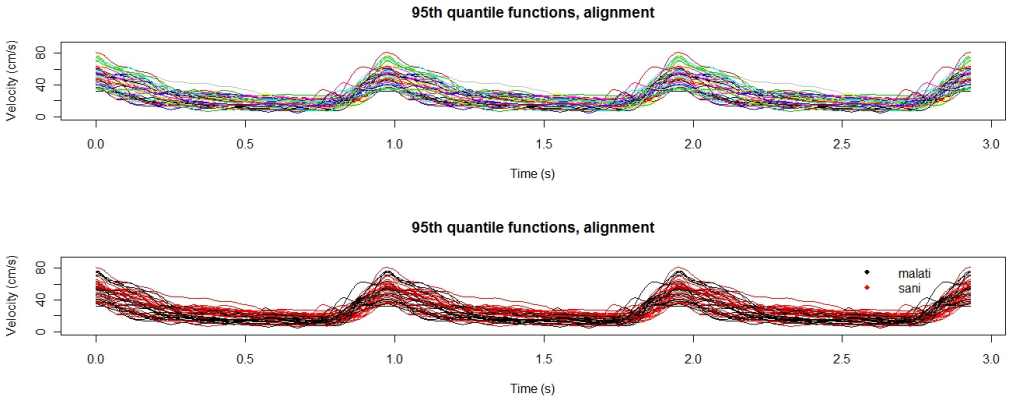

Figure 2.21: Functions representing the 95^{th} quantile registered functions at level -2. For each patient 3 periods of the cardiac cycle are represented. In the bottom panel, functions corresponding to TEA candidate patients are shown in red.

Chapter 3

Dimension Reduction and Classification of Functional Doppler Spectra

At this point of our work, we are confident with the fact that the data extracted from the Doppler spectra can be interpreted as realizations of random functions. Furthermore, after the estimate of the period performed in the previous chapter, the quantitative information concerning each patient of the dataset is well described by smooth functions, meaning that we are working with functional data. In this chapter, we will thus proceed with the exploratory analysis on these functional data, through functional principal components analysis (in section 3.1), with the aim of reducing the dimensionality of the problem, by visualizing the most important variability features of the data. In order to do inference instead, we will perform linear discriminant analysis (LDA, in section 3.2), after having checked that the hypotheses requested by these methods are fulfilled.
Detecting the common and different features within our sample of patients will permit to determine a classification tool to distinguish patients coming from the three populations: TEA candidates, non-TEA patients with a low-grade plaque and healthy people.

3.1 Functional principal components analysis

Functional principal components analysis (FPCA) is an important technique for exploring data and finding the features that characterize the functions at our disposal. Moreover, it can give extremely clear results, including an indication on the complexity of the data. As a matter of fact, FPCA guarantees an informative analysis of the variability structure of the data, more comprehensible than a direct examination of the variance-covariance function. In this section, we will follow [13] in order to introduce the FPCA approach, which will provide a deep exploration of the functions representing blood flow in the carotid arteries.

3.1.1 Defining functional principal components analysis

Let $x_i(s)$ be N realizations ($i = 1, \dots, N$) of a random function $X(s)$, centred with respect to the mean and measured on a continuous scale indexed by s. Functional principal components are defined similarly to the multivariate principal components [18], but instead of being vectors, in this contest they are continuous functions $\beta(s)$ (also called *eigenfunctions*), which allow to visualize the most important variability features of the analysed phenomenon. In fact, the aim is to find a set of principal components $\beta(s)$ that maximize the variance along each component and are orthogonal to each other.
The principal component scores corresponding to the weight β can be expressed as:

$$f_i = \beta x_i = \int \beta(s) x_i(s) ds.$$

The first principal component $\beta_1(s)$ is determined by maximizing the mean square

$$\frac{1}{N} \sum_{i=1}^{N} f_{i1}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\int \beta_1(s) x_i(s) ds \right)^2, \tag{3.1}$$

subject to $\int \beta_1(s)^2 ds = \|\beta_1(s)\|_{L^2}^2 = 1$. The maximization of 3.1 identifies the main cause of variability of the functions, while the constraint is necessary to make the problem well defined, since without it the inner product between functions and eigenfunctions could be arbitrarily large.
The second and following eigenfunctions $\beta_m(s)$ are consequently determined

as the functions that maximize

$$\frac{1}{N}\sum_{i=1}^{N} f_{im}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\int \beta_m(s)x_i(s)ds\right)^2,$$

this time subject to 2 constraints: that $\beta_m(s)$ has unit norm

$$\int \beta_m(s)^2(ds) = \|\beta_m(s)\|_{L^2}^2 = 1$$

and that β_m is orthogonal to the principal components already determined:

$$\int \beta_m(s)\beta_k(s)ds = 0, \quad k < m.$$

Finding next principal components means searching for variability indexes that permit to add new information to that already available with the first principal component.
Anyway, principal components have also another important meaning of dimensionality reduction. The aim is that of determining a set of K orthogonal functions $\xi_k(s)$ $(k = 1, \ldots, K)$ so that the expansion of each curve $x(s)$ in terms of these basis functions approximates the curve as closely as possible. Since the basis functions are chosen orthonormal, the expansion will be:

$$x_i(s) \approx \hat{x}_i(s) = \sum_{k=1}^{K} \omega_{ik}\xi_k(s),$$

where $\omega_{ik} = \int \xi_k(s)x_i(s)ds$. It is possible to show [13] that the orthonormal basis $\{\xi_1, \ldots, \xi_K\}$ minimizing the integrated squared error

$$\sum_{i=1}^{N} \|x_i - \hat{x}_i\|_{L^2}^2 = \sum_{i=1}^{N}\int [x(s) - \hat{x}(s)]^2 ds,$$

is exactly the set of the first K functional principal components of the dataset.

3.1.2 FPCA on carotid arteries velocities

Data cleaning

Before applying the FPCA to the data set of smooth functions representing the blood velocity along the CCA that we have constructed, we should

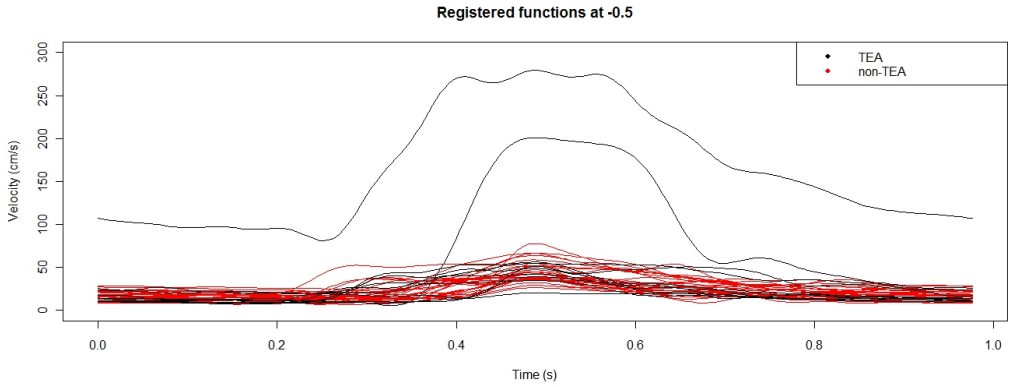

Figure 3.1: Registered functions representing the 95^{th} sample quantile at level -0.5.

first check these curves, in order to detect and remove possible outliers. For this purpose, we present in Figure 3.1 and 3.2 the registered functional data at level -2, -1 and -0.5, where the black functions correspond to TEA patients, while the red one to all the other patients. One can immediately see that among the TEA candidates there are two patients whose blood velocity at -0.5 reaches really high values of 200 and 250 cm/s. So high values are not typical in the common carotid artery (we remind that within the MACAREN@MOX project only patient with a stenosis along the internal carotid artery -and not along the CCA- are included), whilst velocity values at the height of the plaque are even higher ($\sim$ 400 cm/s). What we conclude is that, probably, this two patients have a plaque at the beginning of the ICA, that is to say at the level of the bifurcation. This is the reason why velocities at -0.5 are so high: the blood flow is already influenced by the presence of the plaque. The two anomalous velocities are those registered for patient 19 (right carotid) and 21 (left carotid). Hereafter we will remove these two cases from the dataset, because they could influence the scores of the FPCA, at least at level -0.5. Consequently, the number of functional data available becomes 36 at -0.5, 41 at -1 and 42 at a distance of -2 from the bifurcation, having removed 2 TEA candidates.

We will now show in details the results of the FPCA, performed after having removed the two outliers. In particular, we will analyse separately

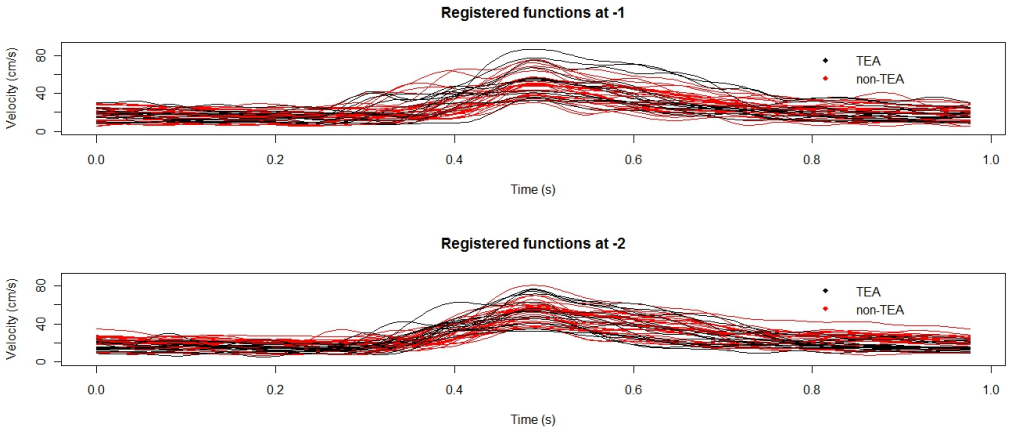

Figure 3.2: Registered functions representing the 95^{th} sample quantile at level -1 (top panel) and -2 (bottom panel).

data concerning different distances from the bifurcation, -2, -1 and -0.5. After having visualized the eigenfunctions and selected, among them, those able to describe with enough accuracy the variability of the problem, we will proceed with a deep analysis of the scores, that is to say the values obtained from the inner product in $\mathcal{L}^2$ between each eigenfunction and each functional data $x_i(s)$. Interpreting the components is not always a straightforward matter, we will thus follow some techniques illustrated in [13]. First of all, for each level of the CCA, it is possible to see the percentage of variability explained by the first principal components, shown in Figure 3.3. In Table 3.1 the proportion of variance explained by the first six principal components is reported, which for the first three is 90.67%, 92.31 % and 88.67% respectively for levels -0.5, -1 and -2. Going on maintaining only the first 3 components would thus be a good choice, in order to considerably reduce the dimensionality of the data. Nevertheless, we keep on treating with the first 4 principal components, even if the fourth one explains only a low percentage of the total variability. This choice is motivated by the fact that, when we will perform the discriminant analysis (later on in this chapter), we will see that the fourth principal components can improve the discrimination criterion in some cases.

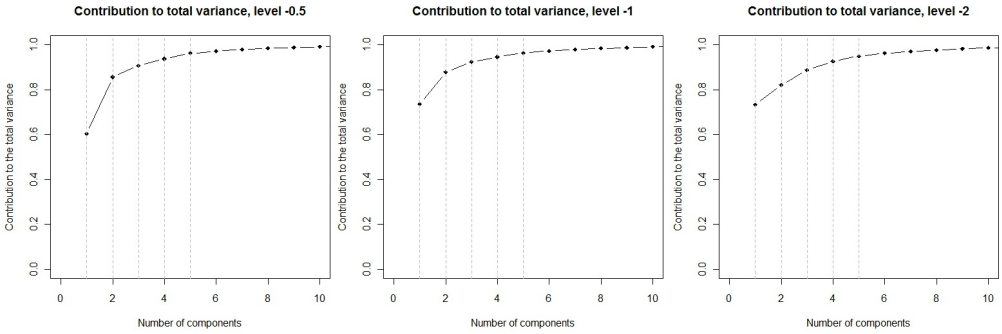

Figure 3.3: Percentage of variability explained by the first 10 principal components, for each level of the CCA: in the left panel the results at distance -0.5 cm before the carotid bifurcation, in the central panel at -1 cm and in the right panel at a distance of -2 cm. In this case, the right carotid of patient 19 and the left carotid of patient 21 have already been removed.

Plotting components as perturbations of the mean

A method to display the eigenfunctions is to plot the sample mean function and the functions obtained by adding and subtracting a suitable multiple of the principal component in question. Figures 3.4, 3.5 and 3.6 show the principal component curves estimated for the three levels. For each figure, the solid black line represents the overall mean function, while the dotted and dashed lines ($+$ and $-$) show the effects of adding and subtracting each principal component multiplied by its own eigenvalue (i.e. the standard deviation), as suggested in [19]. What is worth to notice is that, for the three distances from the bifurcation, the principal components represent almost the same type of variability.

We will now spend some word on how interpreting the estimated principal components. The effect of the first principal component of variation $\hat{\beta}_1$ is principally to add or subtract a constant to the blood velocity approximately throughout the whole cardiac cycle, apart from the acceleration zone (the increasing slope just before the systolic peak). Thus, $\hat{\beta}_1$ describes the variability due to a scale factor of the velocity for all three distances -0.5, -1 and -2. By a fluid dynamic point of view, the first principal component represents the diversity along the vertical axis between the curves of the

Level	Principal Component					
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
-0.5	60.38%	25.24%	5.05%	3.02%	2.48%	0.99%
- 1	73.64%	14.25%	4.42%	2.29%	1.82%	0.81%
- 2	73.39%	8.70%	6.58%	3.97%	2.31%	1.19%

Table 3.1: Proportion of variance explained by the first six principal components, 2 patients removed.

blood velocity of various patients, i.e. a vertical translation that is transmitted nearly uniformly throughout the cardiac cycle (horizontal axis), apart from the first systole phase. The second principal function $\hat{\beta}_2$ describes the variability between the two slopes of the velocity curve. Here, high scores represent velocity functions with a very steep left slope of the peak and a more gentle downward phase. Instead, low scores correspond to curves with a wider systolic time, where the peak is less sharp, and a steep descending phase. Moreover, for levels -2 and -0.5, $\hat{\beta}_2$ represents also a variability along the diastole. The third principal component $\hat{\beta}_3$ describes the variability of the velocity along the cardiac cycle, underling a contrast between the systolic and the diastolic time lags: low values of the scores indicate a higher value of blood velocity at the systolic peak and a lower value along the diastolic relaxing time. On the contrary, score with high values describe a less pulsatile velocity, higher during diastole and lower during systole. It can thus be interpreted as variability in the shape of the peak: higher and narrower against lower and wider and, from a fluid dynamic point of view, this component points out the pulsatility of the blood flow, differentiating flows in which there is much difference between the systolic and diastolic peak from flows where, instead, the jump is smaller. Eventually, $\hat{\beta}_4$ catches the variability at the systolic peak and, at the same time, along the descending slope, but at this level, a fluid dynamic interpretation becomes difficult.

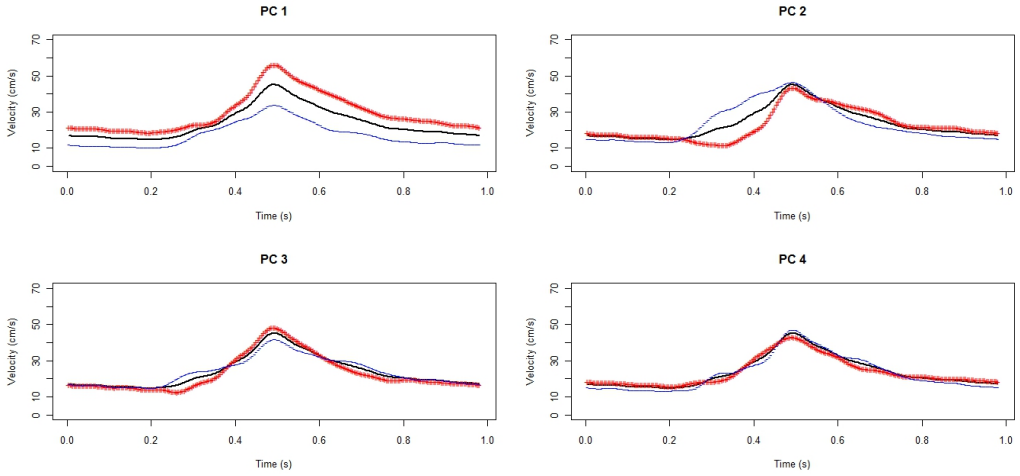

Figure 3.4: First 4 estimated principal component curves at level -0.5. The solid black line represents the overall mean function, while the dotted and dashed lines show the effects of adding (+) and subtracting (−) each principal component multiplied by the standard deviation of the corresponding score.

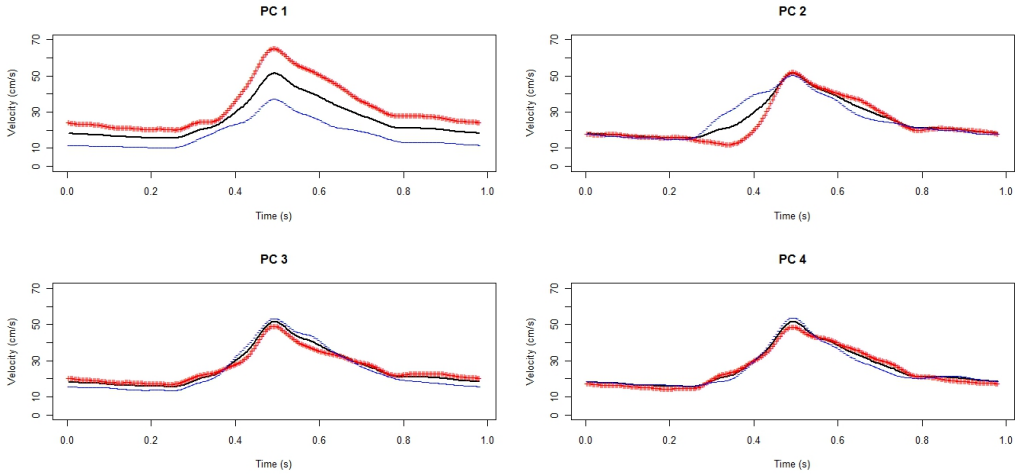

Figure 3.5: First 4 estimated principal component curves at level -1. The solid black line represents the overall mean function, while the dotted and dashed lines show the effects of adding (+) and subtracting (−) each principal component multiplied by the standard deviation of the corresponding score.

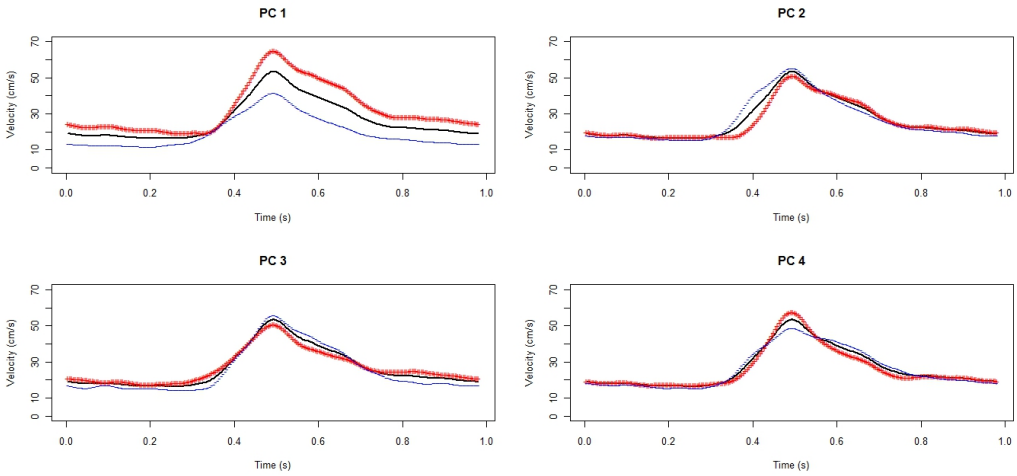

Figure 3.6: First 4 estimated principal component curves at level -2. The solid black line represents the overall mean function, while the dotted and dashed lines show the effects of adding (+) and subtracting (−) each principal component multiplied by the standard deviation of the corresponding score.

FPCA on the whole data set

For seek of completeness, we spend a brief subsection describing how the results of the FPCA would have been, if we had performed it on the whole data set, that is to say without removing the two outliers we detected. Basically, the main difference is that, including all the patients, the principal components at levels -2, -1 and -0.5 do not longer represent the same phenomena of variability, but rather the components at level -0.5 differ from the one at levels -1 and -2. This fact can be seen in Figures 3.7, 3.8 and 3.9, which show the principal component curves estimated for the three levels as perturbations of the mean. The first eigenfunction $\hat{\beta}_1$ describes the variability due to a scale factor of the velocity along the whole cardiac cycle, but the ascending slope. In particular, at -0.5 this scale factor is especially pronounced over the systolic peak, because of the two functions reaching very high velocity values. Keeping the attention at -0.5, the second principal component $\hat{\beta}_2$ can be interpreted as variability in the shape of the peak: higher and narrower against lower and wider. Thus, it shows the variability between the systolic and the diastolic time lags. Finally, the third principal function

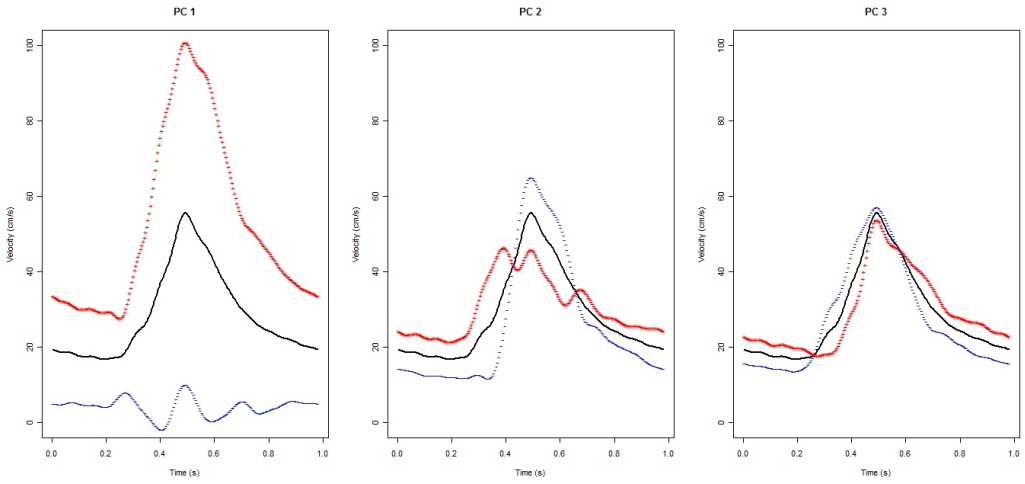

Figure 3.7: First three estimated principal component curves at level -0.5. The solid black line represents the overall mean function, while the dotted and dashed lines show the effects of adding (+) and subtracting (−) each principal component multiplied by the standard deviation of the corresponding score.

$\hat{\beta}_3$ underlines the contrast between the side slopes. Instead, concerning the distances -1 and -2, it seems like $\hat{\beta}_2$ and $\hat{\beta}_3$ switch roles between them with respect to the one just described: $\hat{\beta}_2$ points out the contrast between the side slopes, while $\hat{\beta}_3$ highlights the variability between systole and diastole.

There is another difference with respect to the FPCA performed on the cleaned data-set, again concerning only level -0.5, where the percentage of variability explained by the first principal components, shown in Figure 3.10, strongly changes. Table 3.2, which reports the proportion of variance explained by the first six principal components of the FPCA performed on the whole data-set, quantifies and detects the reason of the difference. In fact, one can see that the first principal component now explains about 90% of the total variability, with respect to the 60% explained in the previous (and exact) case. This change was predictable, since the first eigenfunction reproduce the vertical-translation variability of the velocity and, without removing the outliers, the variance is obviously higher. Instead, results concerning level -2 and -1 are very similar to the results obtained on the cleaned data-set, since in these two levels there are not velocity curves reaching so high values.

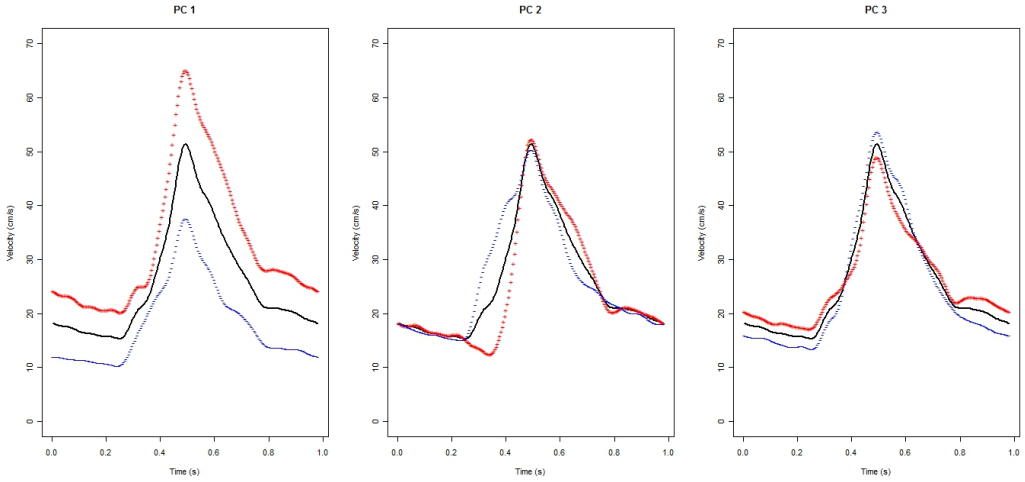

Figure 3.8: First three estimated principal component curves at level -1. The solid black line represents the overall mean function, while the dotted and dashed lines show the effects of adding (+) and subtracting (−) each principal component multiplied by the standard deviation of the corresponding score.

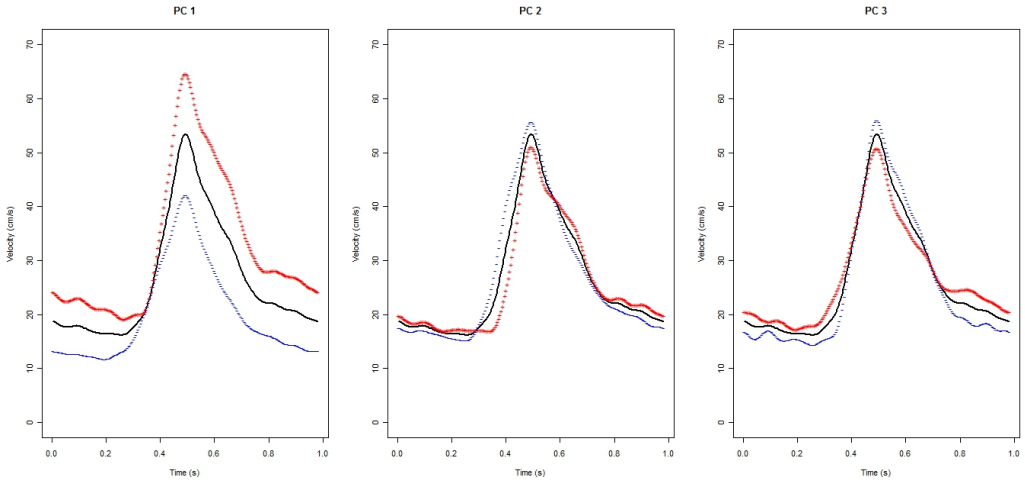

Figure 3.9: First three estimated principal component curves at level -2. The solid black line represents the overall mean function, while the dotted and dashed lines show the effects of adding (+) and subtracting (−) each principal component multiplied by the standard deviation of the corresponding score.

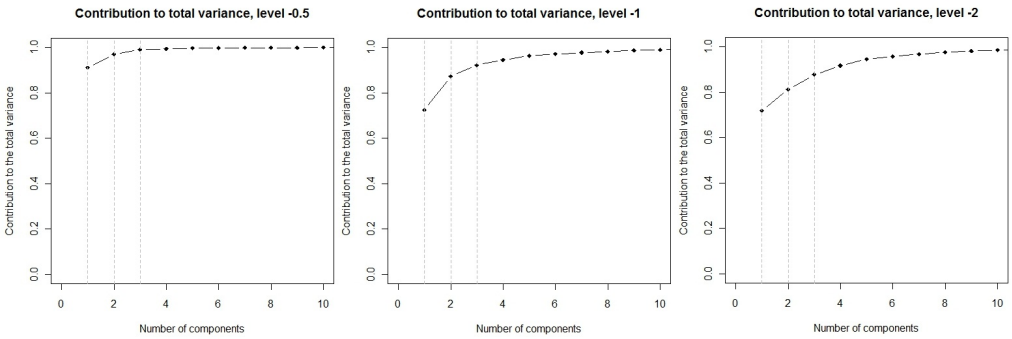

Figure 3.10: Percentage of variability explained by the first 10 principal components, for each level of the CCA: in the left panel the results at distance -0.5 cm before the carotid bifurcation, in the central panel at -1 cm and in the right panel at a distance of -2 cm.

To conclude this parenthesis, we would like to underline that, since we are analysing data at each level independently, we do not expect the same results for the three distances, but what seems anomalous in the results of the FPCA performed on the non-cleaned data-set is that they are quite similar at levels -2 and -1, while at -0.5 they differ. Thus, removing the two outliers from the analysis is a wise choice.

Principal component scores

The analysis of the principal component scores can reveal differences among groups of data. In fact, we know that the sample of patients we are using can be divided into three groups: TEA candidate, non-TEA (that have a low grade stenosis) and healthy patients. We will thus proceed with a first descriptive analysis of the scores, trying in this way to detect differences among the sets of data, in order to prepare the data classification, treated in next section.
First, we should define what are the scores in FPCA: they represents the new coordinates of the functions corresponding to the observations, in the orthogonal basis obtained by the principal components. In other words, the score corresponding to the ith observed curve x_i and the kth estimated eigenfunction $\hat{\beta}_k$ is defined as the component along $\hat{\beta}_k$ of the ith observed function

Level	Principal Component					
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
-0.5	91.02%	5.91%	2.11%	0.30%	0.27%	0.12%
- 1	72.46%	14.85%	4.88%	2.31%	1.81%	0.80%
- 2	71.92%	9.32%	6.48%	3.99%	2.82%	1.18%

Table 3.2: Proportion of variance explained by the first six principal components.

x_i, centred around the sample mean $\bar{x}$ [19]:

$$\int (x_i(s) - \bar{x}(s))\hat{\beta}_k(s)ds.$$

Hereafter we will keep on maintaining two different classification at the same time, the first, in two groups, defined in this way:

- ***TEA***: all the patients with a high grade stenosis, candidates to undergo TEA surgery;
- ***non-TEA***: all other patients, that is to say both healthy patients and those with a low-grade stenosis, who will not undergo surgery;

and one second discrimination in three groups:

- ***TEA***: same as before, all the patients with a high grade stenosis, who will undergo TEA (label 1 in Table 1.1);
- ***mild***: patients with a low grade stenosis, who are periodically monitored, but the plaque is not extended enough to be surgically removed (label 2 in Table 1.1);
- ***healthy***: patients with no plaque (label 3 in Table 1.1).

Thus, *non-TEA* group includes both *mild* and *healthy* groups; before choosing this clustering, we tried to understand to which group the *mild* curves most

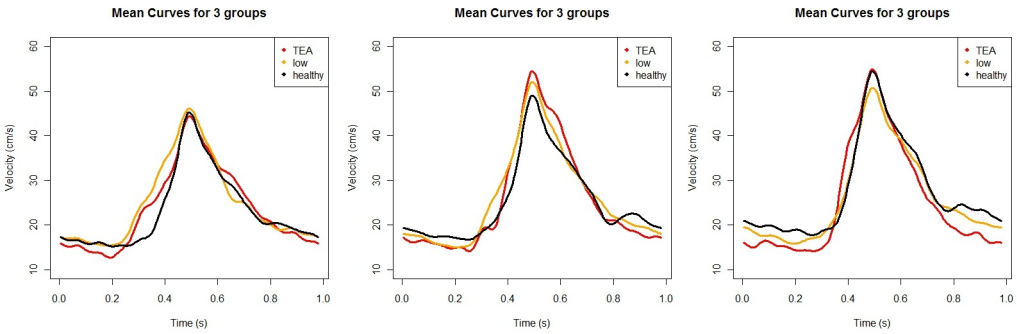

Figure 3.11: Mean curves for the three groups: the red line represents the mean function of *TEA*, the orange line the overall mean of *mild* patients and the black one the mean taken over *healthy* patients.

resemble, for instance looking at the plot of the 3 groups mean curves, which is reported in Figure 3.11: the red line represents the mean function of *TEA*, the orange line the overall mean of *mild* curves and the black one the mean taken over *healthy* patients. Unfortunately, it is not that clear to which group mild stenosis data should belong, because at some time of the cardiac cycle they match with the healthy data, but other times with high-grade stenosis (and sometimes they lie exactly in the middle). This is the reason why we maintained the classification in 3 groups, despite the low number of data available. We report in Table 3.3 the sample mean and the standard deviation of the scores of the first principal components, for the 2 groups (*TEA* and *non-TEA*) at the top of each sub-table and, at the bottom, for the 3 groups (*TEA*, *mild* and *healthy*), at each level of the CCA. It is possible to notice that the sample means for the group *TEA* have symmetric values with respect to mean values for the group *non-TEA*: when the first has a positive scores mean, the second is negative and the other way around. Anyway, since the standard deviation is very high, we prefer to deduce nothing up to now. In order to add more information to our scores analysis, instead, we should have a look to the boxplots of the scores, that allow to see the median and the interquartile range. In Figure 3.12 we show the boxplots for the first 4 principal components, each of them split into two groups (red for TEA and yellow for non-TEA). In the top panel the scores refer to level -0.5: the 3^{rd} and 4^{th} components have a median clearly distinguished for the

Level -0.5								
	Sample Mean				Std			
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
TEA	-1.03	-3.53	-8.24	-13.00	102.89	88.37	24.57	13.11
non-TEA	0.34	1.17	2.74	4.33	95.78	52.79	28.67	22.16
TEA	-1.03	-3.53	-8.24	-13.00	102.89	88.37	24.57	13.11
mild	9.50	9.50	-1.38	5.95	109.64	59.40	36.08	26.12
healthy	-5.95	21.22	5.58	3.22	88.21	37.77	23.15	19.82

Level -1								
	Sample Mean				Std			
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
TEA	7.91	4.49	-24.27	-0.48	159.71	48.81	27.47	22.00
non-TEA	-3.27	-1.85	10.04	0.20	105.98	56.32	24.97	21.71
TEA	7.91	4.49	-24.27	-0.48	159.71	48.81	27.47	22.00
mild	4.48	-17.08	-1.78	3.22	118.17	64.67	22.99	24.12
healthy	-9.58	10.50	19.65	-2.25	98.49	47.03	22.83	19.99

Level -2								
	Sample Mean				Std			
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
TEA	-16.71	-27.87	-18.71	-2.09	111.49	42.25	29.21	22.16
non-TEA	7.49	12.49	8.38	0.94	102.77	24.88	28.99	25.63
TEA	-16.71	-27.87	-18.71	-2.09	111.49	42.25	29.21	22.16
mild	-4.99	8.71	7.15	-5.57	105.49	25.83	36.68	30.18
healthy	17.64	15.56	9.39	6.23	102.81	24.48	22.13	20.75

Table 3.3: Sample means and the standard deviations of the scores of the first three principal components. The three tables contain values of level -0.5, -1 and -2, from the top to the bottom. Moreover, in each table the two first rows are for the the 2 groups (*TEA*/*non-TEA*), while the last 3 rows for the 3 groups (*TEA*/*mild*/*healthy*).

2 groups. The central panel refers to level -1, where, again, it is mostly the 3^{rd} components that underlines a variation between the 2 groups. Instead, at level -2 (bottom panel), the median line varies between the 2 groups for all the first three principal components. Since we said that the third component is fluid dynamically interpreted as pulsatility of the flow, the fact that this component for all the levels permits to distinguish between the two classes of patients means that particular attention should be kept on this feature. Probably, the pulsatily, and thus the gap between the peaks of systolic and diastolic phases, could be a good index to diagnose the presence of a plaque in carotid arteries.
Figure 3.13 shows the boxplots for the first 4 principal components, this time split into three groups (red for *TEA*, orange for *mild* and yellow for *healthy*). The middle group sometimes seems to be linked to *TEA* curves and other times to the *healthy* one. Boxplots add the information concerning median and IQR, but to support the observations just listed down, we should add statistical tests on the mean and the variance, which can confirm differences between the groups and, moreover, they can be used to check the hypotheses behind LDA.

Statistical Tests

First, we check that data have a normal distribution, especially because the sample is composed of a low number of curves (36 at -0.5, 41 at -1 and 42 at -2). Afterwards we will perform a t-test and a Mann-Whitney U test, to control if the means of the 2 or 3 groups are different and, finally, we will pay attention to the dispersion of the scores of the groups, performing a Bartlett test and a Box's M test to test for homogeneity of variance-covariance matrices.

Shapiro test The Shapiro-Wilk test verifies normality checking two different estimators of the variance, the first not parametric based on the slope of the QQ-plot and the second one parametric, that is the classic sample variance. The p-values obtained for the univariate and multivariate version of the test are here reported, for the two groups TEA and non-TEA:

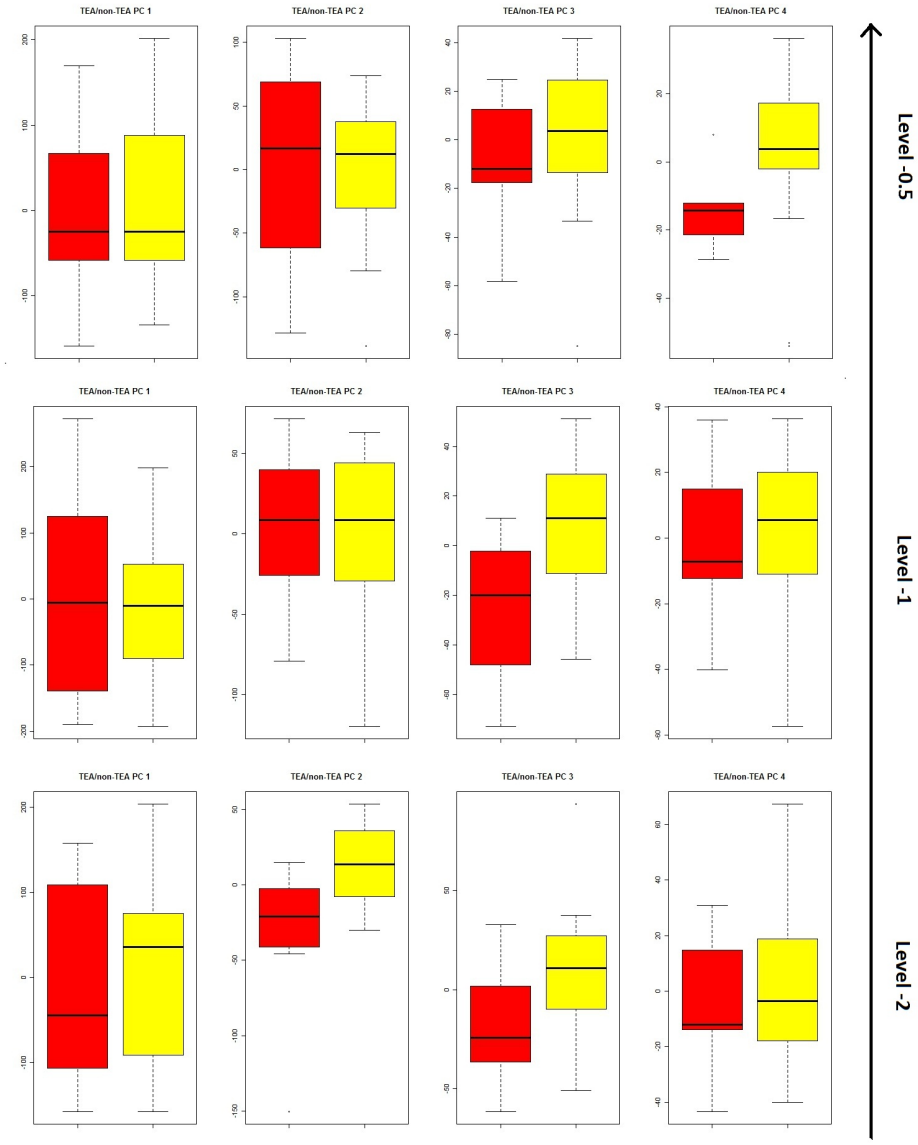

Figure 3.12: Boxplots of the scores of the first 6 principal components. Each of them is split into two groups (red for TEA and yellow for non-TEA). In the top panel the scores refer to level -0.5, in the central panel to level -1 and, in the bottom panel to level -2.

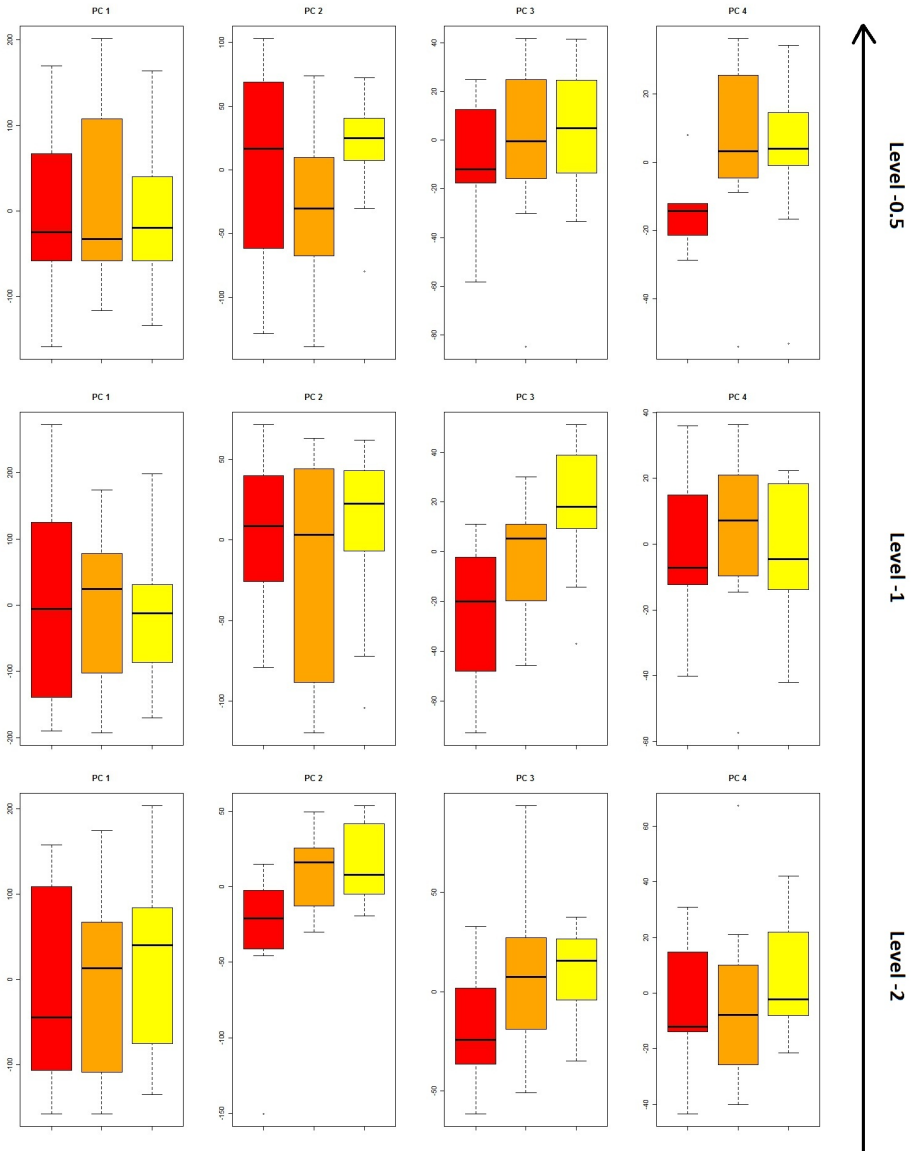

Figure 3.13: Boxplots of the scores of the first 6 principal components. Each of them is split into three groups (red for *TEA*, orange for *mild*. In the top panel the scores refer to level -0.5, in the central panel to level -1 and, in the bottom panel to level -2.

Level -0.5					
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$
TEA	98.94%	21.41%	33.77%	14.30%	22.00%
non-TEA	2.07%	7.94%	7.04%	1.20%	1.0%
Level -1					
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$
TEA	24.72%	69.04%	27.40%	39.34%	80.68%
non-TEA	61.71%	3.19%	35.61%	18.09%	44.48%
Level -2					
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$
TEA	14.26%	3.60%	89.85%	32.07%	2.92%
non-TEA	7.74%	24.28%	9.60%	31.51%	27%

Regarding levels -2 and -1, even if there is some low percentage, the p-values are all above the 3%, thus we cannot refuse the null hypothesis of normality of the scores. On the contrary, for level -0.5, there are values that make us doubt the normality of the data. Consequently, later on we will not take for grant normality for this level.

T test and Mann-Whitney U test From the boxplots it was clear that, for some components, the means of the two populations TEA and non-TEA are not equal. Supposing normality of the scores we can check if the means of the two groups are the same, which is the null hypothesis of the t-test, while the alternative hypothesis is that the difference between the means of the two groups is different from zero. We also provide p-values of the Mann-Whitney U test, which does not require the normal distribution of the scores, but, as one can see from the following tables, the two test always agree with the order of the p-value percentage. In general, these tests confirm and refine what one can see from the boxplots.

Level -0.5				
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
t-test	97.23%	88.27%	28.23%	0.91%
u-test	100%	80.16%	21.84%	0.68%
Level -1				
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
t-test	82.63%	72.05%	0.14%	92.85%
u-test	96.57%	92.01%	0.09%	74.17%
Level -2				
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
t-test	51.23%	0.55%	1.04%	69.88%
u-test	57.23%	0.05%	1.09%	74.72%

At the three levels of the CCA, the principal components displaying a difference in the mean of the two groups is only one and it varies at each level. At -0.5 we can refuse the null hypothesis of equal means only for the scores of the fourth principal component, at -1 for those of the third and at -2 for the second and third principal components. Concerning the division into three groups instead, we can apply t-test and u-test between 2 of the 3 sets at one time. The results obtained when performing the test between *TEA-mild* and between *TEA-healthy*confirm previous observations. On the other hand, when testing if the mean of *healthy* can be considered equal to the mean of *mild*, one cannot refuse the null hypothesis at level -2, where the lowest p-value is 15.6%. This support the choice of merging the patients with a low-degree plaque together with the healthy patients. Instead, at levels -1 and -0.5, the lowest p-values are respectively 2.28% and 2.25%, thus one can refuse the null hypothesis.

Bartlett test and Box's M test Bartlett-test checks the homogeneity of the variances of two groups and it gives a quantitative information of what is usually seen in the heights of the box in a boxplot. From the p-values of the univariate test, performed for the scores of each principal component, and reported below, one can see that they are not so low to refuse the null hypothesis, thus we can consider that the two groups have similar variances.

Level -0.5				
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Bartlett-test	80.40%	5.73%	60.79%	10.44%
Level -1				
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Bartlett-test	9.24%	58.39%	70.49%	95.80%
Level -2				
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Bartlett-test	73.97%	2.46%	97.57%	56.47%

In order to test for homogeneity of covariance matrices between groups, it is also possible to use the Box's M statistic [23], the null hypothesis of which is the equality of the covariance matrices. It is a test statistic based on a likelihood-ratio test, which thus requires normality of data.

Level -0.5		
	2 groups	3 groups
Box's M-test	78.0%	17.36%
Level -1		
	2 groups	3 groups
Box's M-test	97.34%	17.36%
Level -2		
	2 groups	3 groups
Box's M-test	98.56%	5.69%

From the p-values, we can conclude that covariance matrices are not significantly different, both in the case of division in two groups and in the case of three groups.

3.2 Discriminant Analysis

After having reduced the dimensionality of the problem through FPCA, we would like to check if it is possible to distinguish patients with a stenosis in the ICA from other patients, using data extracted from the CCA. Doctors claim that, in most cases, it is difficult to detect the presence of a plaque located in the ICA already from the blood flow in the CCA and this is the reason why introducing a statistical analysis of the curves could result useful.

Furthermore, the analysis of the scores of the FPCA just done suggests that there are features of the blood flow functions influenced by the presence of a plaque downstream. We will thus proceed with a discriminant analysis, not really meant, at least at this stage of the project, to build a classification criterion for new patients data, also because the number of data is still too low in order to assign to the results a predictive importance. Moreover, finding out which are the characteristics that differ between a high-grade or low-grade stenosis, could help in providing useful informations for numerical simulations, by adding the variability between the groups of patients.
Therefore, we will now apply a linear discriminant analysis LDA [18] on the scores of the first four principal components at each of the three levels along the CCA, even if at -0.5 the normality of data can be questioned.

3.2.1 Linear Discriminant Analysis

LDA is a method to find a linear combination of features which characterize two or more classes of objects. It considers a set of observations (training set) with known class. The assumptions that have to be fulfilled in order to apply LDA are that

- the conditional probability density functions of observations of each class are normally distributed;
- the covariance matrices are homoschedastic, i.e. the class covariances are identical and with full rank.

The classification problem is then to find a good predictor for the class of any sample of the same distribution (not necessarily from the training set) given only an observation. Further details about LDA can be found in [18] and [24]. We will present the results through confusion matrices, that is to say by listing, for each class (on rows), the number of scores classified with the right and wrong label (columns). A way to determine the quality of the classification is the rate of error AER, which can be approximated by the APER (APparent Error Rate), consisting in the ratio between the sum of the classification errors and the total number of data. However, this is only an underestimate of the effective rate of error, in particular in cases when the sample is small. In order to have a better estimate and a better evaluation of the classification criterion, one should recur to the Cross-Validation (CV) technique, which is an iterative method that classifies

each of the n observations of the sample following the discriminant criterion provided by other $n-1$ observations. All the results presented afterwards refers to a cross validated LDA.

Results

The assumption of homogeneity among the variance-covariance matrices is fulfilled for the three levels and for both kind of division in two or three groups. This is the reason why we go on with the LDA and we do not recur to the quadratic discriminant analysis QDA. Instead, as already mentioned when we wrote the Shapiro test p-values, normality for level -0.5 cannot be taken for grant. For this level, we will anyway report the results of the LDA, even if they cannot be fully trusted.
Starting from the classification in two groups (TEA/non-TEA), we should decide the number K of principal components scores on which apply LDA. This can be done exploring LDA classification errors computed for an increasing number of principal component scores considered, as the following table shows. Only levels -2 and -1 have been considered to take a decision, since we can consider the scores as normal distributed.

Level -1								
	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$
[TEA|non-TEA]	0%	0%	13.79%	10.34%	13.79%	13.79%	27.59%	27.59%
[non-TEA|TEA]	100%	100%	50%	50%	58.33%	58.33%	50%	58.33%
Level -2								
	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$
[TEA|non-TEA]	0%	6.9%	13.79%	13.79%	13.79%	13.79%	13.79%	13.79%
[non-TEA|TEA]	100%	69.23%	53.85%	38.46%	38.46%	38.46%	38.46%	38.46%

In the table, the conditional percentage of error have be reported: the first row refers to the error of assigning to group TEA patients who indeed belong to group non-TEA , while in the second row there are the percentages of patients belonging to group TEA, but assigned in group non-TEA. The number of principal components minimizing both errors is $K=4$, thus we will perform LDA only on the first 4 principal components scores. Before providing the results obtained by classifying each score following the criterion obtained by the standard LDA, we could have a look to the scatterplot of the scores, in order to see if there is any pair of scores clearly showing a neat

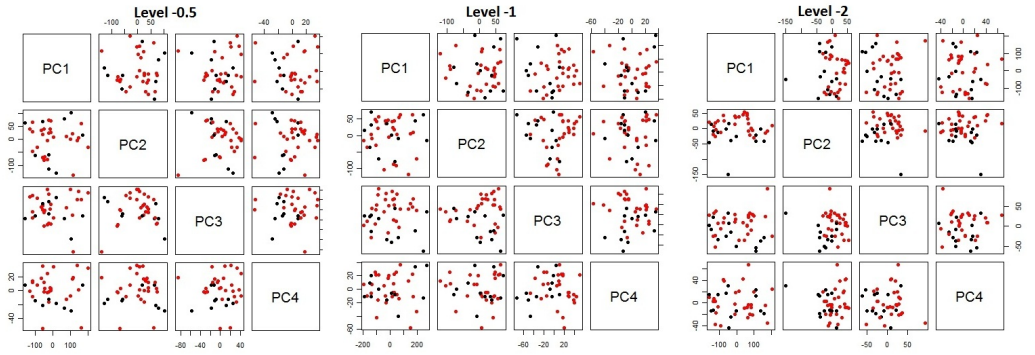

Figure 3.14: Scatterplots of the first 4 principal component scores, for the three levels -0.5, -1 and -2. The scores referring to TEA patients are black, whilst the non-TEA red. At a first sight it seems like there is not a couple of scores which clearly shows the division in 2 groups, except for level -2, component 2 and 3.

distinction between the two groups. Figure 3.14 show the scatterplots of the first 4 principal component scores, for the three levels -0.5, -1 and -2. The scores referring to TEA patients are black, whilst the non-TEA are red. At a first sight it seems like there is not a couple of scores which clearly shows a neat division in 2 groups, even if at level -2 two distinguished zones can be detected for the scores of the second and third components. In Tables 3.4, 3.5 and 3.6 the results for the three levels are reported in four forms. The first row contains two tables, one reporting the absolute values obtained and the other the relative percentages. In the second row of tables, one can find the conditional percentages, which can be of two types, depending if one wants to condition with respect to the true class ([`Assigned` | `True`]), to the left, or with respect to the assigned class ([`True` | `Assigned`]), to the right.

It is possible to see that, in general, correctly classifying a TEA patients is more difficult than correctly classifying a non-TEA patient, that is to say that the test has a high specificity. In our case, *specificity* is the ability to correctly identify the non-TEA patients and its values are uniform for the three levels of the CCA, since they are 0.86, 0.90 and 0.89 respectively at -2, -1 and -0.5. Instead, the *sensitivity*, which is the ability to correctly identify the TEA patients, changes among the levels: at -2 sensitivity is 0.62, at -1 it decreases to 0.5 and it becomes null at -0.5. Focusing on the con-

	Absolut ASSIGNMENT		Relative ASSIGNMENT	
	TEA	non-TEA	TEA	non-TEA
TEA TRUE	8	5	19.05%	11.90%
non-TEA TRUE	4	25	9.52%	59.52%

	Conditional [Assigned|True] ASSIGNMENT		Conditional [True|Assigned] ASSIGNMENT	
	TEA	non-TEA	TEA	non-TEA
TEA TRUE	61.54%	38.46%	66.67%	16.67%
non-TEA TRUE	13.79%	86.21%	33.33%	83.33%

Table 3.4: Level -2, results of classical LDA performed on the first 4 principal component scores.

	Absolut ASSIGNMENT		Relative ASSIGNMENT	
	TEA	non-TEA	TEA	non-TEA
TEA TRUE	6	6	14.63%	14.63%
non-TEA TRUE	3	26	7.32%	63.41%

	Conditional [Assigned|True] ASSIGNMENT		Conditional [True|Assigned] ASSIGNMENT	
	TEA	non-TEA	TEA	non-TEA
TEA TRUE	50%	50%	66.67%	18.75%
non-TEA TRUE	10.34%	89.66%	33.33%	81.25%

Table 3.5: Level -1, results of classical LDA performed on the first 4 principal component scores.

	Absolut		Relative	
	ASSIGNMENT		ASSIGNMENT	
	TEA	non-TEA	TEA	non-TEA
TEA TRUE	0	9	0%	25%
non-TEA TRUE	3	24	8.33%	66.67%

	Conditional [Assigned|True]		Conditional [True|Assigned]	
	ASSIGNMENT		ASSIGNMENT	
	TEA	non-TEA	TEA	non-TEA
TEA TRUE	0%	100%	0%	27.27%
non-TEA TRUE	11.11%	88.89%	100%	72.73%

Table 3.6: Level -0.5, results of classical LDA performed on the first 4 principal component scores.

ditional error `[Assigned | True]` (i.e. `[Assigned non-TEA | TEA True]` and `[Assigned TEA | non-TEA True]`), one can see that at level -2, the probability of misclassifying a TEA patient is 38.46%, while the probability of misclassifying a non-TEA patient is 13.79%. At level -1 the probability of misclassifying a TEA patient increases to 50%, while the probability of misclassifying a non-TEA patient decreases to 10.34%. Finally, at level -0.5, the classification tool is totally unable to catch TEA patients. We could list different reasons justifying this bad behaviour of the LDA, for instance that we cannot assume normal distributed scores. Also, it has to be considered the fact that 5 mm before the carotid bifurcation, not only the acquisition of Doppler signal is more complicated and noisy, but also the blood flow is more turbulent than other levels of the CCA. Moreover, the sample we are using is small and it becomes even smaller from level -2 (43 cases) to level -0.5 (only 36 cases). Because of this fact, we should find another way to judge the goodness of our classification (at least at -2 and -1). The estimates of the AER of the CV-LDA at -2 and -1 are 21.43% and 24.39% (33.33% at -0.5), but how much significant are these values? A way to answer this question is to follow the idea of the *gap statistics* in [24]. Generating a large enough number of random vectors of labels *TEA* and *non-TEA* (each vector as long as the number of patients available at each level), applying these labels randomly to the patients and finally estimating the error rate of the

LDA classification, it is possible to plot the distribution of the classification error. From this, we can compare where the error of the true vector of labels is located with respect to the whole distribution.
In order to draw the distribution of the misclassification error, we thus have generated 10000 random vectors of labels to be linked randomly to the patients of the study. For each of these 10000 combinations between the patients of the study and the simulated permutation of labels, we performed CV-LDA to the first four components scores of the FPCA and we computed and recorded the estimate of the error rate. Having at disposal 10000 values of APER, we can thus estimate the probability density function of the error rate distribution. Figure 3.15 shows the results: for each level two panels are reported, the histogram of the 10000 simulated APER values and the density estimation of the error distribution. The vertical red line indicates the APER value obtained when using the true labels. Results concerning level -2, on the left of the figure, are very good, since the number of cases when the APER of the simulated labels has a value lesser or equal than the real APER are only 0.15% of the total. In the centre there are the histogram and the probability density function of the error rate at level -1, where the number of simulated APER lesser or equal the real one is 0.69% of the total. Finally, the last two panels on the right part of the figure refers to level -0.5. Here the classification tool does not work well, since it is unable to catch the TEA patients, and the number of simulated APER better or equal the real APER value is of 9.95%, which is a quite high value with respect to the percentages obtained for the other levels. This fact confirms that at level -0.5 the shapes of the curves are mixed up and one should not trust the Doppler frame acquired at this level, since the signal could be distorted due to the increasing turbulence of the flow. Anyway, from the simulated error distribution, we can be quite satisfied of the results of the LDA performed, since the error rate estimated from the classification criterion lies in the extreme left side of the distribution.

In order to evaluate the classifier from a point of view of the forecasting ability, on the other hand, it is interesting to look at the conditional probabilities [`True` | `Assigned`]. What we can see is that the forecasting ability of the classifier is higher for detecting non-TEA patients than for detecting TEA. As it is possible to read in the confusion matrices 3.4, 3.5 and 3.6, the probabilities of correctly forecasting a non-TEA curve ([`non-TEA True` | `non-TEA Assigned`]) is 83.33% at level -2, 81.25% at level -1 and 72.75% at -0.5. On the contrary, the probability of correctly forecasting a TEA patient

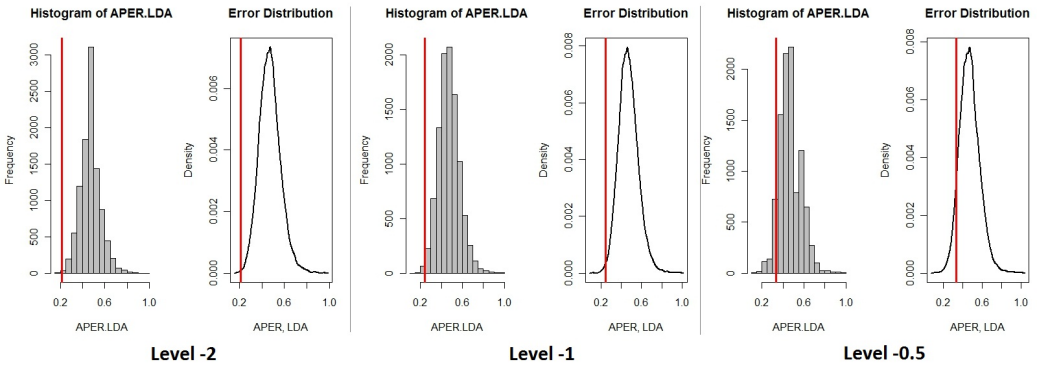

Figure 3.15: Simulation of the classification (in 2 groups) error distribution. On the left, the two panels contain the histogram of the 10000 computed APER values and the density estimation of the error distribution, while the vertical red line indicates the APER value obtained when using the true labels. In the centre, histogram and probability density function of the error rate at level -1 and, on the right, at level -0.5.

([`TEA True` | `TEA Assigned`]) is acceptable at levels -2 and -1, where it is 66.67%, but it is null at level -0.5. Again, so lousy results are due to the low number of observations available at this stage of the project, but also to the problems related to the acquisition of ultrasounds image at this position. Usually, to evaluate this type of conditional probabilities, two kinds of errors are defined: *false positives* and *false negatives*. The *false positives* are the patients belonging to group non-TEA but assigned to group TEA. This error is high for all the levels considered, since it is 33.33% at -2 and -1 and even 100% at level -0.5. The *false negatives* instead are the patients belonging to group TEA but assigned to group non-TEA, with a value of 16.67% at -2, 18.75% at -1 and 27.27% at -0.5.
Finally, we also tried to perform LDA not only using the scores of the first 4 eigenfunctions together, but exploring the combinatorial tree, but we did not obtained any improved result, as the scatterplot already suggests.

Results of LDA in 3 groups

We now report the results obtained by searching for 3 groups, maintaining again only the first 4 principal component scores. First of all, Figure

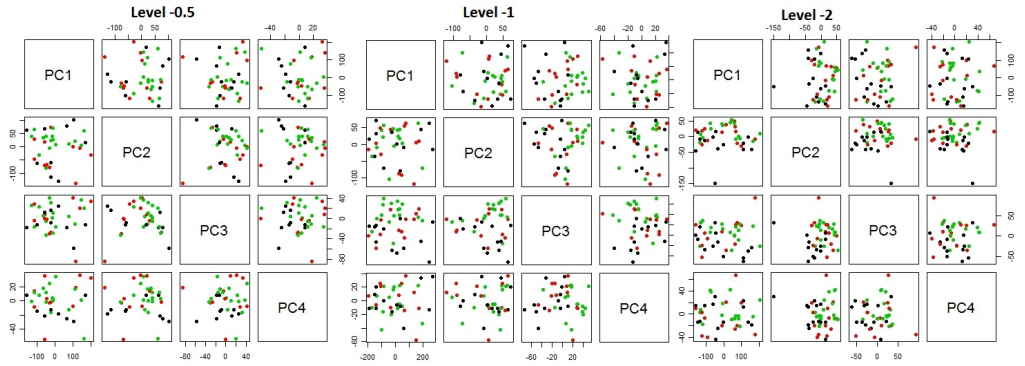

Figure 3.16: Scatterplots of the first 4 principal component scores, for the three levels -0.5, -1 and -2 and coloured depending on which of the 3 group they belong: black for TEA, green for mild and red for healthy. At a first sight it seems like there is not a couple of scores which clearly shows the division in 3 groups.

3.16 shows the scatter plot of the scores, coloured depending on which of the 3 group they belong: black for TEA, green for mild and red for healthy. Tables 3.7, 3.8 and 3.9 report the results for the three levels, again in four forms: absolute values, relative percentages and the two kinds of conditional percentages. Also in this case we do not trust the results regarding level -0.5, for which the classification tool totally ignore the existence of group TEA. At levels -2 and -1, the group of mild stenosis is the most difficult to be found, because sometimes it is confused with group TEA and other times with the group of healthy patients: the percentages of scores correctly classified are the lowest among the groups (23.08% at -2 and 38.46% at -1). In the case of three groups, talking about sensitivity and specificity is not totally correct, because of the presence of the mild cluster. We can anyway compute the specificity of the test as the number of correctly identified non-TEA patients (that is to say the ratio between the number of classified healthy and the total number of true healthy), which increases from a value of 0.63 at level -2, to 0.75 at -1 and 0.94 at level -0.5. Defining the sensitivity as the ratio between the patients classified as TEA over the total number of TEA, we obtain on the contrary that its value goes from a value of 0.77 at -2, 0.5 at -1 and it is null at -0.5. What is worth to notice is that the mistakes of the discriminant analysis are usually between two neighbouring classes: for instance, at level

-2, among the true TEA patients, only 3 are wrongly classified and they are assigned to the neighbour class mild, while no one is classified as healthy. From a forecasting point of view, we should check the conditional probabilities `[True | Assigned]`, which can be found in the fourth table in 3.7, 3.8 and 3.9. In this case, the probability that a patient classified as TEA is really a TEA candidate is 71.43% at level -2, at -1 it is 54.55% and at level -0.5 it is null. Instead, the probability that a patient classified as healthy does not really present any plaque is 58.82% at -2, 66.67% and 55.56% at -0.5. We should anyway consider cautiously these values, especially because the sample size is very low and we are now trying to divide the data in three groups. As in the case of two groups, a method to evaluate the quality of the discriminant analysis could be a random permutation of the labels TEA, mild and healthy with respect to the curves. In this way we are able to draw the simulated distribution of the APER and to check where the real APER is located with respect to the whole distribution, which can be found in Figure 3.17. We can be quite satisfied of how LDA performs, since the error rate estimated from the classification criterion using the true labels lies in the extreme left side of the distribution. In particular, the number of cases when the APER of the simulated labels has a value lesser or equal than the real APER are 2.62% of the total at level -2 and 1.77% at level -1. Concerning level -0.5, the percentage of APER with a better value than the real one is 37.85% and the same considerations as before are valid: at this level the variability between the curves is so high that it is impossible to distinguish the velocity flow in 2 or 3 groups.

3.2.2 Final Comments

To conclude this chapter, we should spend some word on the results of the functional PCA and of the discriminant analysis, in order to clarify how one should proceed when having at disposal a larger amount of data. First of all, detecting and cleaning the data set from outliers is usually a good practice. In our case, we found two outliers at level -0.5 and we decided to remove the corresponding frames from all the levels, in order to avoid that a possible plaque located at the bifurcation could influence the velocity profile, clearly as at level -0.5, or more covertly at other levels, farer from the site of the stenosis. When disposing of more curves, the number of the outliers could increase and, in case they became a larger amount, the possibility of including them in the analysis should be considered.

	Absolut			Relative		
	ASSIGNMENT			ASSIGNMENT		
	TEA	*mild*	healthy	TEA	*mild*	healthy
TEA TRUE	10	3	0	23.81%	7.14%	0%
mild TRUE	3	3	7	7.14%	7.14%	16.67%
healthy TRUE	1	5	10	2.38%	11.90%	23.81%

	Conditional [Assigned|True]			Conditional [True|Assigned]		
	ASSIGNMENT			ASSIGNMENT		
	TEA	*mild*	healthy	TEA	*mild*	healthy
TEA TRUE	76.92%	23.08%	0%	71.43%	27.27%	0%
mild TRUE	23.08%	23.08%	53.85%	21.43%	27.27%	41.18%
healthy TRUE	6.25%	31.25%	62.50%	7.14%	45.45%	58.82%

Table 3.7: Level -2, results of classical LDA performed on the first 4 principal component scores.

	Absolut			Relative		
	ASSIGNMENT			ASSIGNMENT		
	TEA	*mild*	healthy	TEA	*mild*	healthy
TEA TRUE	6	4	2	14.63%	9.76%	4.88%
mild TRUE	4	5	4	9.76%	12.20%	9.76%
healthy TRUE	1	3	12	2.44%	7.32%	29.27%

	Conditional [Assigned|True]			Conditional [True|Assigned]		
	ASSIGNMENT			ASSIGNMENT		
	TEA	*mild*	healthy	TEA	*mild*	healthy
TEA TRUE	50%	33.33%	16.67%	54.55%	33.33%	11.11%
mild TRUE	30.77%	38.46%	30.77%	36.36%	41.67%	22.22%
healthy TRUE	6.25%	18.75%	75%	9.09%	25%	66.67%

Table 3.8: Level -1, results of classical LDA performed on the first 4 principal component scores.

	Absolut			Relative		
	ASSIGNMENT			ASSIGNMENT		
	TEA	*mild*	healthy	TEA	*mild*	healthy
TEA TRUE	0	4	5	0%	11.11%	13.89%
mild TRUE	1	3	7	2.78%	8.33%	19.44%
healthy TRUE	0	1	15	0%	2.78%	41.67%

	Conditional [Assigned|True]			Conditional [True|Assigned]		
	ASSIGNMENT			ASSIGNMENT		
	TEA	*mild*	healthy	TEA	*mild*	healthy
TEA TRUE	0%	44.44%	55.56%	0%	50%	18.52%
mild TRUE	9.09%	27.27%	63.64%	100%	37.50%	25.93%
healthy TRUE	0%	6.25%	93.75%	0%	12.50%	55.56%

Table 3.9: Level -0.5, results of classical LDA performed on the first 4 principal component scores.

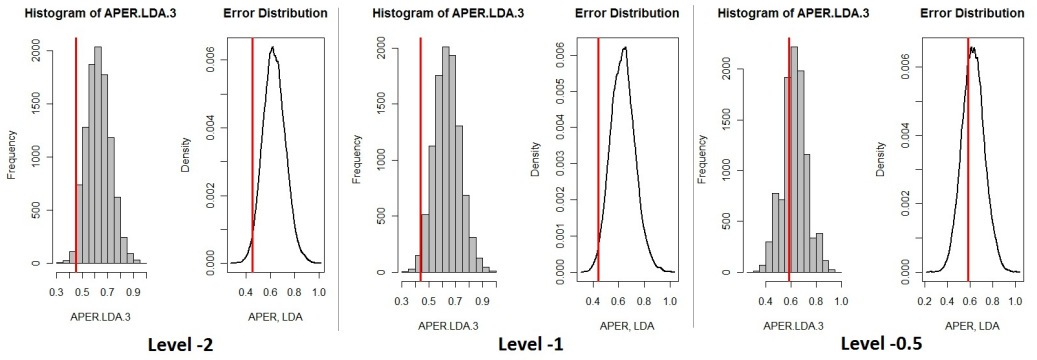

Figure 3.17: Simulation of the classification error distribution, when trying to detect three groups. On the left, the two panels contain the histogram of the 10000 computed APER values and the density estimation of the error distribution, while the vertical red line indicates the APER value obtained when using the true labels. In the centre, histogram and probability density function of the error rate at level -1 and, on the right, at level -0.5.

About the discriminant analysis, the sample used in this thesis would suggest that one should not include in the study the Doppler acquisitions at level -0.5, for several reasons. First of all they are more difficult to be recorded, because of the closeness to the bifurcation. In fact, at this point the blood flow is turbulent and, in some cases, it is already divided into two flows, heading towards ICA and ECA and the Doppler image acquisition strongly depends on the orientation of the transducer. Moreover, the results of the LDA at -0.5 are quite poor and totally unable to recognize different groups of patients. On the other hand, at level -2 and -1 the results of the LDA, when searching for 2 groups, are better. Since the sample is small, we do not assign them a predictive power, but we are confident that, by increasing the sample, things could work out. Furthermore, results at level -2 and -1 are very similar up to now, but one would expect that they improve from -2 to -1, since the plaque (if there is) becomes closer and the blood flow should reflect this, underling a difference between the TEA patients and the non-TEA. Finally, dividing the patients in three groups could have more sense when the number increase. Probably the classification algorithm would continue assigning the class of patients with a low-grade to the other two class, and this could be read as a support tool for doctors, when deciding whether the patient should undergo surgery or not.

Conclusions

In this work, the characteristics of the blood flow in the carotid artery extracted by Color Doppler ultrasounds images were taken into consideration, with the aim of detecting variability features capable of discriminate between different types of flow within the carotid arteries. Thanks to close cooperation between the Policlinico Hospital of Milan and the Laboratory for Modeling and Scientific Computing MOX, not only a large number of Doppler frames have been provided, but also ideas, knowledge and a strong incentive for the execution of this thesis.

We paid particular attention in treating and filtering the Doppler data at disposal, since it is a very difficult kind of data to deal with, because of the high variability to which it is subject to. In fact, it has to be considered that not only the morphology of the stenosis plaque, its type and its location are a source of variability, but also the acquisition procedure is highly sensitive and the Doppler images could result distorted or dirty. Keeping all these factors in mind, we extracted the sample quantile of the velocity distribution at each available time of the cardiac cycle.

The innovative algorithm developed in order to estimate the period of the velocity flow is satisfying, because it is quick and robust to different initial values. This is an important aspect, because the periodicity of the blood flow inside the carotid arteries need not be equal to the frequency of the heartbeat. Moreover, since more acquisitions along the CCA and the ICA are available for each patient, estimating the blood velocity period at different points of the vessel could be useful in fluid dynamic simulations.

Data dimensional reduction performed in the last chapter was necessary to proceed with the discriminant analysis. Before performing LDA, the con-

jecture was that, getting closer to the bifurcation, the discrimination between patients with a high grade stenosis and patients with a low-grade stenosis or even without a plaque would have become more evident. On the contrary, the results show that 2 cm before the bifurcation, it is easier to detect the presence of a plaque downstream in the ICA than 5 mm before the bifurcation (called level -0.5 in the thesis). This fact would suggest that one should be cautious in drawing conclusions by extracting information from Doppler acquisitions at level -0.5, because of the closeness to the bifurcation. It is true that at this point the plaque is closer, but blood flow is also more turbulent and, in some cases, it is already divided into two flows, heading towards ICA and ECA. The Doppler image can thus be very noisy and the velocity waveform different from that of few mm before (at 1 cm before the bifurcation for instance). On the other hand, by analysing data acquired 2 and 1 cm before the carotid bifurcation, it is possible to distinguish different groups of patients. The analysis, however, is not yet exhaustive; indeed this is not its ambition. The task of this work is rather to illustrate the possible sites of investigation and to stimulate future research. Moreover, the sample available for this thesis is small and not yet complete; we are confident that, by increasing the number of patients of the study, the results of the discriminant analysis could improve.

Finally, to corroborate the results of this work, it would be useful to repeat the analysis on a larger sample for each class of patients. Moreover, with the help of the medical team of the research group, it would be interesting to give a fluid dynamic interpretation to the components that more discriminate among the groups of velocity functions, in order to obtain different numerical simulations for each class. Also, adding to the data-set the information about the exact position of the plaque in the internal carotid artery could improve the results of the analysis, since it would allow to merge the data taking into account not only the distance from the bifurcation, but also the distance from the stenotic plaque.

Appendix

Below it is reported the code implemented in R [25] for the estimation of the period of the blood flow velocity.

```
# Estimate of the period and position of the first peak

library(KernSmooth)
library(fda)
library(ReadImages)

rm(list=ls(all=TRUE))

#K <- 15  #number of basis functions
#gap <- 3 #number of peaks of the Doppler waveform

Periodo <- function(){

T <- length(data)

## STEP 1 : First Fourier smoothing
rangeval <- c(1,T)
basisfourier = create.fourier.basis(rangeval,K,T)
eval.base <- eval.basis(seq(1, T, 1), basisfourier, Lfdobj=0)

smoothing <- Data2fd(1:T, data, basisfourier)
pixelList = smooth.basis(1:T, data, basisfourier)
coeff <- smoothing$coefs[,1]
data.smooth <- colSums(coeff*t(eval.base))

par(mfrow=c(3,1))
```

```
plot((1:T)/800*3.6,data.smooth,type='l',lwd=3,ylim=c(0,max(data)),
     xlab='Time (s)', ylab='Velocity (cm/s)',
     main='95th Quantile Data and Smoothing Function x')
points((1:T)/800*3.6,data,col='red')
legend("topleft",c("x","Data"),pch=18,col=1:2)

## STEP 2 : Detection of the first peak
Ddata.smooth <- colSums(coeff*t(eval.Dbase))
D2data.smooth <- colSums(coeff*t(eval.D2base))

threshold <- c()
threshold[1] <- max(data.smooth)/3*2
criterion1 <- data.smooth > threshold
criterion2 <- D2data.smooth < 0
idx <- which((criterion1+criterion2)== 2)
idx <- idx[idx < T/gap]
V <- which.min(abs(Ddata.smooth[idx]))
picco <- idx[V]

plot((1:T)/800*3.6,data.smooth,type='l',lwd=2,col=1,ylim=c(0,max(data)),
     xlab='Time(s)', ylab='Velocity, (cm/s)',
     main='Detection of the First Peak')
abline( v = picco/800*3.6,col='blue')

## STEP 3 : translation in t=0 of the first peak
start <- -picco[1]+1
end <- T-picco[1]
t <- seq(start,end,1)

    # Preparation of the data frame necessary for nls:
    Dati <- data.frame(Time = t,Velocity = data.smooth)
dimnames(Dati)[[1]]<- 1:T

## Estimate of the model
#initial points:
a <- a_0
b <- b_0
w <- w_0
```

```
model <- nls(Velocity ~ a+b*cos(w*Time), Dati,
             start= list(w = w,a = a,b = b),trace=TRUE)
summary(model)
plot(t/800*3.6,Dati$Velocity,type='l',lwd=2,xlab='Time (s)',
     ylab='Velocity (cm/s)', main='Non Linear LS Fit')#(1:T)
lines(t/800*3.6, predict(model), col=2)
legend("topleft",c("a+b*cos","x"),pch=18,col=c("red","black"))

w <- coef(model)[1]
a <- coef(model)[2]
b <- coef(model)[3]

Hat_periodo <- 2*pi/w

W <- Hat_periodo
p <- picco
result=list(W=W,p=p,T=T,a=a,b=b)
result
}
```

Ringraziamenti

Desidero innanzitutto ringraziare il Professor Piercesare Secchi per i preziosi insegnamenti durante i mesi di lavoro per questa tesi, le idee, le conoscenze e i numerosi spunti di approfondimento. Inoltre, ringrazio sentitamente la Dott.ssa Laura Azzimonti per le numerose ore dedicate alla mia tesi e per essere sempre stata disponibile a risolvere i miei dubbi durante la stesura del lavoro. Intendo poi ringraziare il Dottor Maurizio Domanin, per avere fornito testi e dati indispensabili per la realizzazione della tesi e per la sua disponibilità ad illustrare gli aspetti medici della ricerca e a monitorare i risultati dell'analisi statistica. Un sentito ringraziamento anche a tutti i ricercatori del progetto MACAREN@MOX, in primis al Dr. Christian Vergara, non solo per il suo contributo scientifico al progetto, ma anche per la disponibilità a coordinare gli incontri del team e la gestione dei dati.

Rimanendo in ambiente universitario, vorrei ringraziare tutte le persone incontrate in questi anni di studio al Politecnico di Milano. Grazie ai compagni di corso di *ingegneria matematica*, sia a quelli conosciuti fin dal primo anno (che mi sembra ieri!), sia ai giovani ingegneri con il quale ho trascorso l'ultimo semestre (in particolare grazie a Marco Nanni e Giorgio Bertolini per avere animato le lezioni di finanza...). Nominarvi tutti sarebbe impossibile e non ci provo nemmeno. Spero che, anche se ormai ciascuno procederà per la sua strada, i nostri sporadici aperitivi (un grazie all'instancabile organizzatore Stefano Bosia) continueranno a lungo. Grazie anche agli amici di BEST Milano, che hanno contribuito a rendere indimenticabili questi anni. Vorrei dedicare un ringraziamento anche agli studenti del POLIMI che ho conosciuto in Svezia, che mi hanno fatto respirare un po' di Milano anche a Lund e, in quest'ultimo anno, un po' di Lund anche a Milano! Ripensando ai due anni trascorsi al nord, c'è una persona in particolare che terrei a ringraziare: un immenso grazie alla dott.ssa Francesca Fogal, per il suo lavoro

e per i preziosi consigli.

Voglio poi ricordare tutte le persone che mi sono sempre state vicine. Prima di tutti i miei genitori: grazie mamma e grazie papà, per il vostro sostegno. Grazie alla mia fantastica sorella Giada, che fin dalle elementari ha tentato di farmi capire di passare meno tempo sui libri e giocare di più! Un grazie anche ai miei nonni, siete davvero speciali.
Non mi dimenticherei mai di ringraziare Tommaso Brambilla, che in tutte le situazioni ha la capacità di generare un sorriso sulla mia faccia. Sei un vero amico (oltre che un ottimo compagno di malefatte!). Grazie anche a Heide Holste e Thorben Heins, per tutte le avventure che abbiamo passato e che passeremo insieme, che siano a -40 o a +40℃! Poi un grazie a Martina Poretti, all'egregio dottor Luca Moroni, a Davide Floriello (soprattutto per le tue brillanti domande!), a Emma, Paolo, Mario, Gloria e Claudio. E come non ringraziare Emanuela (intendo proprio te, Sbaddu!): dovremmo riprendere le abitudini di un tempo, quando nemmeno un metro di neve ci ostacolava l'ape del venerdì. Infine (ma mi sto sicuramente dimenticando qualcuno!), grazie a Luca, Manuel, Vanessa e Filippo...per tempo o per i km che ci separano non si incontra spesso, ma ogni volta è sempre come se non ci si vedesse dal giorno prima! Un grandissimo grazie anche a Gemmina e Danilo (e non solo per la grappa alle fragoline di bosco :) e al super asso della montagna Stefano.

Dulcis in fundo, grazie ad Andrea, perchè mi incoraggi sempre in tutto... Grazie per avermi sostenuta (e sopportata!!) nella stesura di ben due tesi specialistiche, ma soprattutto... grazie di esistere.

Bibliography

[1] Jensen J.A., 1996. *Estimation of Blood Velocities Using Ultrasound*. Cambridge University Press.

[2] Samira Hirji, 2006. *Real-Time and Interactive Virtual Doppler Ultrasound*. University of Western Ontario, London (Ontario, Canada).

[3] Schäberle W., 2004. *Ultrasonography in Vascular Diagnosis*. Springer.

[4] Jakobsson A., Gran F. and Jensen J.A., 2009. *Adaptive Spectral Doppler Estimation*. IEEE Transactions on Ultrasonics, Ferroelectric Freq Control (vol.56, pages 700-714).

[5] North American Symptomatic Carotid Endarterectomy Trial Collaborators, 1991. *Beneficial effect of carotid endoarterectomy in symptomatic patients with high grade carotid stenosis*. The New England Journal of Medicine (325: 445-453).

[6] Rothwell P.M., Eliasziw M., Gutnikov G.A., Phila D, Fox J.A., Taylor D.W., Mayberg M.R., Warlow C.P., Barnett HJM., 2003. *Analysis of pooled data from the randomised controlled trials of endarterectomy for symptomatic carotid stenosis*. The Lancet (361: 107-116).

[7] Li R, Cai J, Tegeler C et al., 1996. *Reproducibility of extracranial carotid atherosclerotic lesions assessed by B-mode ultrasound: the ARIC Study*. Ultrasound Med Biol (22:791-799).

[8] Moneta GL, Edwards JM, Papanicolaou G, Hatsukami T, Taylor LM, Strandness DE, Porter JM., 1995. *Screening for asymptomatic internal carotid artery stenosis: duplex criteria for discriminating 60% to 99% stenosis*. J Vasc Surg 21: 989-994.

[9] Magagnin V., Delfino L., Cerutti S., Turiel M. and Caiani E., 2007. *Nearly automated analysis of coronary Doppler flow velocity from transthoracic ultrasound images: validation with manual tracings*. Med Biol Eng Comput.

[10] Caiani E.G., Asquer G., Meraviglia E., Cerutti S., Turiel M., Bailliart O., Cholley B., Capderou A., Vaida P., 2004. *Semiautomated analysis of Doppler images for quantification of changes in mitral inflow pattern during parabolic flight*. Computers in Cardiology (321-324).

[11] Blackshear WM, Phillips DJ, Chikos PM, Harley JD, Thiele BL, Strandness DE Jr., 1980. *Carotid artery velocity patterns in normal and stenotic vessels*. STROKE journal of the American Stroke Association (11(1):67-71).

[12] G.R. Arce., 2005. *Nonlinear Signal Processing: A Statistical Approach*. Wiley, New Jersey, USA.

[13] Ramsay J.O. and Silverman B.W., 2005. *Functional Data Analysis*. Springer.

[14] Ramsay J.O., Giles Hooker and Spencer Graves, 2009. *Functional Data Analysis with R and MATLAB*. Springer.

[15] Bates DM. and Watts DG., 1988. *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons, Inc., New York.

[16] Wilcoxon F., 1945. *Individual comparisons by ranking methods*. Biometrics Bulletin 1 (6): 80-83.

[17] Florencio I. Utreras, 1986. *On Generalized Cross-Validation for Multivariate Smoothing Spline Functions*. SIAM J. Sci. and Stat. Comput. 8, pp. 630-643.

[18] R.A. Johnson, D.W. Wichern., 2007. *Applied Multivariate Statistical Analysis*. VI edition, Pearson, Prentice Hall.

[19] L.M. Sangalli, P. Secchi, S. Vantini, A. Veneziani., 2009. *A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery*. Journal of the American Statistical Association, vol. 104, issue 485, pages 37-48.

[20] B. W. Silverman., 1986. *Density estimation for statistics and data analysis*. Chapman and Hall.

[21] S. J. Sheather, M. C. Jones., 1991. *A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation*. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 53, No. 3, pp. 683-690.

[22] Scott D. W., 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

[23] Box G.E.P., 1949. *A general distribution theory for a class of likelihood criteria*. Biometrika 36: 317-346.

[24] Hastie Trevor, Tibshirani Robert and Friedman J. H., 2001 *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag.

[25] R Development Core Team, 2011. *R: A Language and Environment for Statistical Computing*. Vienna, ISBN: 3-900051-07-0, url: http://www.R-project.org

[26] J. O. Ramsay and Hadley Wickham and Spencer Graves and Giles Hooker, 2010. *fda: Functional Data Analysis*. R package version 2.2.5, url: http://CRAN.R-project.org/package=fda

[27] Brian M. Sadler and Stephen D. Casey, 2000. *Sinusoidal frequency estimation via sparse zero crossings*. Journal of the Franklin Institute, Vol. 337, Issues 2-3, Pages 131-145.