UNIVERSITÀ DEGLI STUDI DI PALERMO

Dottorato in Statistica, Statistica Applicata e Finanza Quantitativa
Dipartimento di Scienze Statistiche e Matematiche (DSSM)
Settore Scientifico Disciplinare SECS/S01-Statistica

STATISTICAL METHODS FOR THE DISCRIMINATION OF FOUR FORMS OF DIPLEGIA

IL DOTTORE

CURCURU' GIUSEPPE

IL COORDINATORE

PROF. MARCELLO CHIODI

IL TUTOR

PROF. ALBERTO LOMBARDO

CICLO XXIV

ANNO CONSEGUIMENTO TITOLO 2015

To my beloved parents, Marta and Vincenzo

Content

List of figures

List of Tables

ACKNOWLEDGEMENTS

I express my gratitude to Dott. Prof. Adriano Ferrari (Faculty of Medicina e Chirurgia, University of Modena e Reggio Emilia, IRCSS Arcispedale Santa Maria Nuova, Reggio Emilia), to Ing. Alberto Ferrari (DEI, University of Bologna) and to Dott. Rita Neviani (IRCCS Arcispedale Santa Maria Nuova, Reggio Emilia).

I am grateful to Ing. Marco Cocconcelli (Dipartimento di Scienza e metodi per l'Ingegneria (DISMI), University of Modena e Reggio Emilia) for his practical support.

I am also grateful to Prof. Alberto Lombardo (Tutor) and to Prof. Marcello Chiodi (Coordinator) for their useful observations and support.

SUMMARY

Cerebral Palsy (CP) is a lesion of the central nervous system that determines a more or less extended loss of brain tissue. As a result, motor functions can be altered. The incidence of infantile cerebral palsy is of 2-3 cases per 1,000 births. CP is not a homogeneous disorder. Actually, the disease can have different degrees of severity and may occur in many different forms. In medical literature, on the basis of the *topographic localization* of the disturbances, different definitions and classifications of cerebral palsy are proposed. In particular, the term *diplegia* is used to identify the effects of the lesion on the lower limbs.

This research work is based on the classification of *diplegia* into four forms proposed and validated by Ferrari A. e al. (University of Modena-Reggio Emilia)[3][4][5]. Such a classification can be considered as an effective support for the therapy. Actually, since each form is characterized by a different degree of severity, a correct identification would allow medical staff to activate appropriate therapeutic protocols. However, up to date, due to the unreliability or absence of objective data and methodologies, the identification of the *diplegia forms* has been completely entrusted to the professional skills of the specialists. To overcome the inevitable subjectivity of the medical evaluation, *gait analysis* allows the measurement and the quantitative evaluation of the kinematics and dynamics of the motion. Therefore, in LAMBDA laboratory (laboratory for the analysis of the movement of the disabled child, Santa Maria Nuova Hospital, Reggio Emilia), kinematic motion data have been acquired on three *gait cycles* for 91 diplegic patients belonging to the four identified groups. The laboratory is equipped with an optoelectronic system with eight cameras (Vicon, UK) by means of which *motion capture* tests can be carried out. *Motion capture* makes possible the recording of the movement of one or multiple subjects

through a series of infrared cameras, and then its reproduction in a digital environment. Through this acquisitions, according to the *Protocol Total3Dgait*, it is possible to rebuild the trajectories of the markers applied to identified anatomical landmarks of the different patients, as well as the angles of rotation of the main joints of the lower limbs. Data used in this research work are just the angles of rotation of the main joints of the lower limbs collected during three gait cycles for all the 91 subjects and referred to three anatomical planes: sagittal, frontal, transverse.

The main objective of this research is the identification of *indicators* to be employed for the discrimination of the four groups. To this purpose, two methods are proposed. The first one is essentially based on the extraction of *static indicators* from the available functional variables.

In particular, taking into account both the clinical suggestions and the main results available in the literature, three indicators have been extracted from the functional data: Range Of Motion (ROM), Root Mean-Square (RMS) and Crest Factor (CF). The first provides information on the maximum angular excursion occurring in a gait cycle, the second on the variation of the time dependent waveform with respect to a constant value and the third on any impulsive phenomena characterizing the pendular movement typical of almost all forms of *diplegia*. A great number (72) of potential predictors have been employed for the construction of a linear discriminant model with stepwise procedure. Since all data have been used both for the construction of the model and its validation, in order to contain an over estimation of the hit ratio (HR), the *leave-one-out cross-validation* method has been used. Wilks' lambda criterion has been used to select significant predictors and then to measure the *goodness-of fit* of the model.

Although the use of *static indicators* is attractive for its simplicity, the selection of such a standard indicators from data is always and inevitably affected by the limits of the subjective evaluations. In contrast, multivariate statistical analysis has proved to be a powerful tool to eliminate collinearity and to facilitate the analysis, considering exclusively the essential structure hidden in the data. Principal Component Analysis (PCA) has showed to be extremely effective in the study of the human motion. This thesis proposes the use of functional PCA (FPCA) for the available data in the attempt of identifying a limited number of components that can explain most of the data variability and features to be used for discriminating purposes. Actually, differences among the *PC scores* (*PCs*) related to the subjects belonging to the four different groups have been tested and some of 48 available variables have been selected as predictors in a discriminant linear model. Correct classification rates between the two proposed methods are compared. For the selected indicators the clinical evaluation is supplied. The clinical interpretation of the statistical results is intended to make intelligible information for specialists.

The thesis is organized as follows. Chapter 1 introduces the term *diplegia* and its clinical classifications. Gait Analysis and gait cycle are presented in Chapter 2. Chapter 3 deals with data collection and the employed statistical methods. In Chapter 4 data analyses and results are presented. Chapter 5 presents the clinical feedback for the results discussed in Chapter 4 and some ideas for future developments. Finally, in the Conclusions, the main results are synthesized.

Chapter 1

Diplegia and clinical classifications

1.1 Introduction

Cerebral palsy (CP) is the most frequent cause of chronic disability in children. The estimated incidence varies between 2 and 3 cases per 1000 children (Stanley) [1]. According to one of the latest definitions, *CP describes a heterogeneous group of permanent disorders of movement and posture, causing limitations to activities, attributable to non-progressive disorders that occur in fetal or child development of the brain.* (Bax)[2]. CP is often accompanied by sensory, perceptual disorders or by cognitive, communicational difficulties and sometimes by secondary musculoskeletal problems.

Depending on the etiology or on some functional features or on the *topographic* distribution of the paralysis, medical literature is full of different definitions and classifications. However, it is very important to have a system of classification of the different forms of cerebral palsy, repeatable and usable by all. This would facilitate the prognosis and the rehabilitative intervention and would allow the definition of general criteria to be used in order to judge the effectiveness of the different adopted rehabilitative intervention protocols.

The working group of professor Ferrari (University of Modena-Reggio Emilia) has developed and validated [3][4][5] a method of classification of such paralysis based on the observational analysis of kinematics and kinesiology of the movement.

Despite the great number of CP definitions, the SCPE (Surveillance Cerebral Palsy Europe) suggests a classification system that minimizes the clinical categories, taking into account exclusively a distinction between bilateral and unilateral forms [6].

There are different opinions on the topic. Morris et al. [7] argue that the term should be completely abandoned as well as Colver AF et Al. [8] propose to completely eliminate the distinction between different classes of CP and suggest to keep only the generic term of "cerebral palsy" and to differentiate patients exclusively from a functional point of view.

The Ferrari group believes that the latter approach mix patients with very different degrees of CP severity, from children unable to perform any activity autonomously to others who exhibit only minor difficulties in motor skills. This complicates both the prognostic evaluation and the measurement of the effectiveness of the therapeutic solutions that can be eventually applied. In the light of these considerations, the Ferrari group supports the need of keeping the original definitions of *diplegia, hemiplegia, tetraplegia*. In order to delimit the *diplegia*, differential criteria should be adopted. In the next paragraph both the clinical signs used to make such a classification and the proposed differentiation into four groups are presented.

1.2 Proposed classifications

Among the clinical signs which may be useful for classification purposes, the following are taken into account.

Firstly, once the *motor sequence* has been activated, it is very difficult for a diplegic patient to stop and to decompose it in order to reverse the direction.

Speed reduction is another visible difficulty for a diplegic subject. Actually, patients encounter fewer difficulties to walk faster rather than slowly and to maintain the state of motion rather than standing still in place.

Coordination of the four limbs is also crucial. For diplegic subjects it is very difficult to coordinate, while riding, the upper limb movements with those of the lower limbs.

Stability and fixation: for diplegic patients maintaining a proper alignment of the body in space and preserving the overall balance is very difficult. Other clinical signs can be referred to the sensory functions, to the cortical functions and to the handling function. Although *sensory functions* are not generally affected, orientation in space as well as maintaining trajectories can become a problem. Almost all children affected by diplegia reach a quantitatively acceptable language, even if sometimes they can produce semantic errors.

Generally a fair expertise in handling is reached, but sometimes the difficulty of controlling the pulse can introduce uncertainty in more complex activities such as the use of utensils or writing and drawing.

The proposed classification is mainly based on the classic analysis of the *topographic distribution* of the injury (*tetraplegia* involving the four limbs, *diplegia* the lower ones, *hemiplegia* only one side) but takes into account specific *functions* as features to be used in order to classify the patients. In particular, *tetraplegia* is classified with relation to the so called *antigravity function*, *diplegia* with relation to the function "*walk*" and finally, *hemiplegia* with relation to the function "*manipulation*". Since the motor skills and the posture of a CP subject represent the best achievable result by a diseased central nervous system, it seems logical to use the movement (kinematics) to classify the degree of severity of the disease. *Diplegia*, as underlined before, is classified with relation to the function walk. Therefore, in the following paragraph the proposed classification for *diplegia* will be shown.

1.2.1 Four forms of *diplegia*

To distinguish the clinical forms of diplegia, the following elements involved in the *motor function* are considered [9]: 1) the use of the upper limbs and of aids for the walk, 2) the pendular movements of the trunk in the sagittal and in the frontal planes, 3) the movements of the pelvis (horizontal movement and antero-posterior tilt), 4) the progression mechanisms, 5) the movements of the foot, 6) the choice of the fulcrum during the walk.

According to these elements, four main forms of *diplegia* have been distinguished:

- Form I (*forward leaning propulsion*) including subjects that use aids for the upper limbs ("*quadripodi*" for defence) and subjects that do not use any device for the upper limbs.

- Form II (*tight skirt)* including subjects that use aids for the upper limbs ("*quadripodi*" for direction) and subjects that do not use any device for the upper limbs.
- Form III (*tight rope walkers)* including subjects that use aids for the upper limbs ("*quadripodi*" as barbell) and subjects that do not use any device for the upper limbs.
- Form IV (*dare devils*) including the generalized form, the distal form and the asymmetric form (double emiplegia).

The four forms are in a decreasing degree of severity. Patients belonging to the first group present severe pathological deficiency, while those of the fourth group are less compromised. Other pathological aspects common to all the identified forms, and not exclusively related to the motor function, are referred to the involvement of the upper and lower limbs with the lower ones more compromised, to the presence of possible perceptual and visuals disorders, especially for the I and III forms, to possible phenomena of epilepsy and mental retardation (infrequent) and to functional handling problems.

Perceptual problems may represent a significant component in the I and III form, while the II and the IV ones are mainly characterized by motor type problems.

The kinematic-based classification proposed by the Ferrari group can provide useful indications for the purposes of re-educational treatments. Actually, it does not aim at achieving a hypothetical normality, but aims at reaching the best adaptive functions for the patient. In addition, it is consistent with the generally accepted definition of CP as disorders of movement and posture. For more details about the clinical walking patterns employed by the medical staff to classify the four *diplegia* forms, see [3].

Chapter 2

Gait Analysis

2.1 Gait Analysis

The analysis of movement is a scientific discipline that deals with the evaluation of the human movement including the acquisition of experimental data, their processing and interpretation [10]. In the years '70 and '80, the first experimental procedures have been defined for the determination and analysis of the movement, through the use of computerized techniques. The collection of these techniques and of the used protocols[11] gave rise to the Gait Analysis (GA) [12][13]. By studying the kinematics, the dynamics and the patterns of the muscular activation, the clinicians can describe some features of the movement and the mechanisms of the motor disabilities with a reasonable accuracy. Then, they can identify the most appropriate therapeutic protocols and evaluate their effectiveness.

The tools used by Gait Analysis techniques allow a quantitative description with graphical representations in a virtual environment of the state of health of the musculoskeletal system, including the active function carried out by the muscles and the passive action suffered by the soft and the hard tissues. GA in the clinic is used to identify and evaluate accurately the severity of diseases disabling the musculoskeletal apparatus. The technique allows to quantify the motor skills of the patients when performing everyday movements, such as walking or climbing stairs, assessing the adequacy of the generated *motor performance*, called "*motor pattern*".

The usefulness of its clinical use as a tool to gain an in-depth knowledge of the joints function, in normal and/or pathological situations, has been amply demonstrated in the medical literature [14-23].

2.2 Gait Cycle

Gait cycle is defined as the *functional unit* in the analysis of movement. It is the time between the initial contact of one foot (stride) and its subsequent contact. Gait cycle represents the temporal unit reference that describes all the biomechanical events and the muscular activity.

It is divided into two distinct phases, the first one known as the *stance phase*, from 0% to 60% of the cycle (during this phase the reference limb is on the ground) and the *swing phase*, from 60% to 100% of the cycle (during this phase the reference limb makes his *swing* until its subsequent contact with the ground).

In turn, these two stages are divided into further stages as shown in Figure 1 :

1) IC (*Initial Contact*), representing the contact of the foot with the ground;

2) LR (*Loading Response*), representing the phase in which the limb that started the contact with the ground begins gradually to receive the load from the other limb (which simultaneously is preparing to break away from the ground);

3) MS (*Mid Stance*), is the phase in which there is the advancement of the limb in contact with the ground;

4) TS (*Terminal Stance*): this stage includes the boost of the limb in contact with the ground, while the other side is concluding the oscillation phase and is preparing the subsequent contact with the ground;

5) PSW (*Pre Swing*), at this stage there is the gradual transition of the load on the other limb;

6) IS (*Initial Swing*), first oscillation phase;

7) MS (*Mid Swing*), intermediate oscillation phase;

8) TS (*Terminal Swing*), final oscillation phase, the oscillation limb prepares for the next contact with the ground.

In healthy subjects the described phases as well as the biological phenomena related to the motion are extremely precise and similar (within the limits of the individual variability) among different subjects[24][25]. This rhythm is often lost in a pathological state, in particular when dealing with diseases of neurological nature.

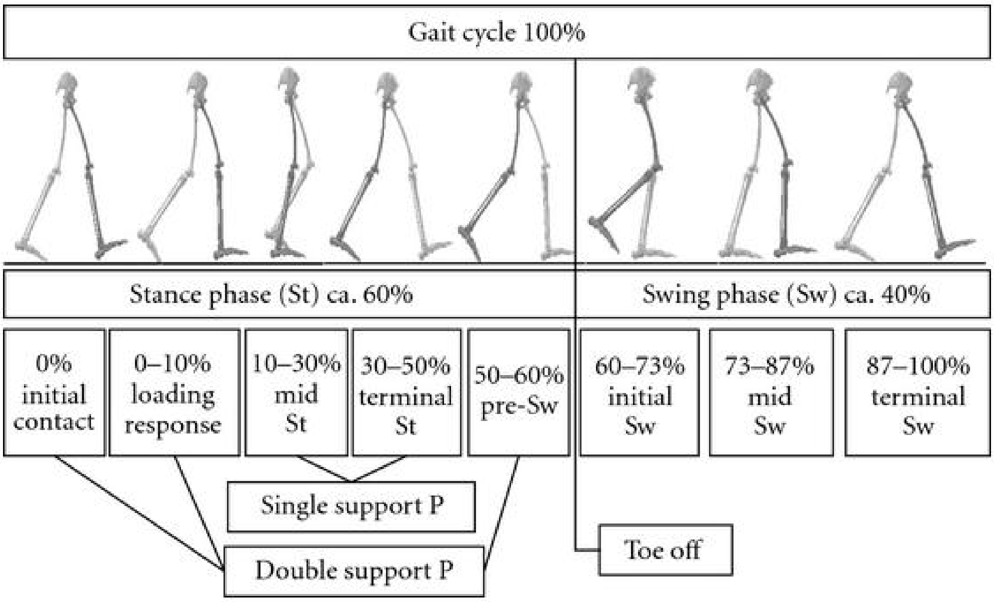


Figure 1: description of a gait cycle

Chapter 3

Data description and statistical methods

3.1 Subjects

The sample used for this research work is constituted of 91 subjects belonging to the four groups introduced in Chapter 1. Sample is not numerically balanced. Actually, 8 subjects belong to the first group, 28 to the second group, 16 to the third one and, finally, 39 to the fourth group.

The great numerical difference between, for example, the first and the fourth group occurs because the diffusion of CP is not homogeneous. Actually, the fourth form is less severe and more spread than the first one. Another practical reason is because patients belonging to the first group cannot face easily a long series of trials in the laboratory and then they may be sometimes stopped and not included in the sample.

3.2 Laboratory LAMBDA, instruments, tests

LAMBDA laboratory (i.e. Laboratory for the Analysis of the Movement of the Disabled Child) is located in the Spallanzani Hospital and belongs to the ASL Arcispedale of Santa Maria Nuova in Reggio Emilia. It is one of the largest research centres in Italy in the field of the cerebral palsy. The main elements constituting the equipment are described below.

3.2.1 Main instruments and Protocol *T3Dg*

LAMBDA laboratory is equipped with an *optoelectronic system* consisting of cameras with a CCD sensor sensitive to infrared light radiation. A LEDs strobe light, synchronized with the speed of image detection, allows a "freezing" effect of frames. Markers, placed on the subjects, are covered with an aluminum powder reflective material. The reflected image, properly processed, provides the three-dimensional coordinates for each point. By combining such information with the spatial position of the cameras, the three-dimensional position of markers can be obtained through a stereoscopic data elaboration. By recording 3D position at every instant, trajectories can be built.

Nexus software is used for the acquisition of signals from the optoelectronic system. It develops the 3D trajectories of markers, making possible the reconstruction of the body structure, through segments and points. Data are then saved in C3D files.

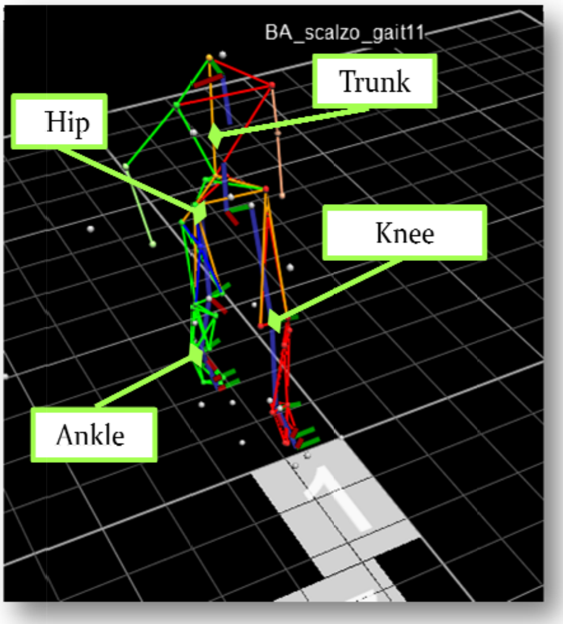


Figure 2: an example of Nexus software image

Protocols for the analysis of the movement are generally designed to meet two main objectives: 1) calculate the joint kinematics with the maximum accuracy and precision, 2) ensure the shortest test session to respect the needs of the involved patients. Actually, patients affected by cerebral palsy can walk only with the help of aids. Because of severe arthritis, they generally cannot stand for a long time. Therefore, the second objective of whatever standard protocol is practical and ethic. Protocol used in the LAMBDA laboratory is the *Total laboratory 3D Gait* (*T3Dg*) designed by Leardini in 2007 [11] and developed at the "Istituti Ortopedici Rizzoli"

(Bologna). By applying this protocol, a marker-set of 22 reflective and 2 identifiers technical markers are used.

In order to obtain consistent and comparable results from the Protocol, it is necessary to establish anatomical reference systems and joints rotation centres. The first ones were defined by Capozzo et al (1995) while the latter by Wu and Cavanagh (1995). The hip joint center is based on the geometric approach of Bell.

3.2.2 Walking tests

Several tests of at least 10m are performed. Data collected during these walks are then processed to obtain the necessary information, in particular on kinematics (the angular movement of the joints in the three anatomical planes: sagittal, frontal, transverse).

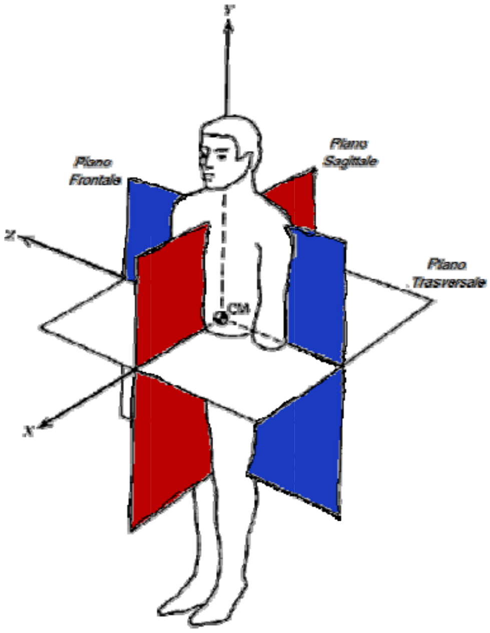


Figure 3: sagittal and frontal planes

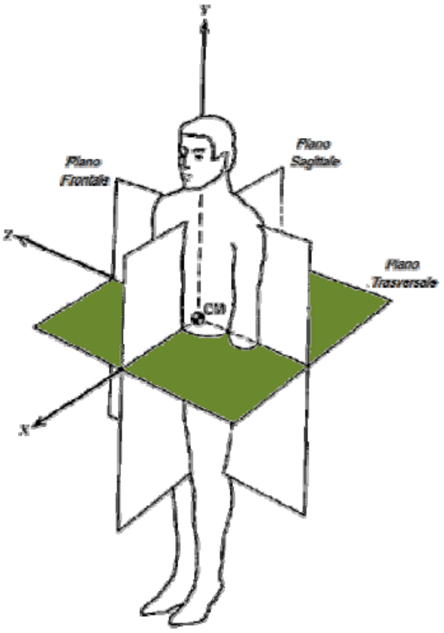


Figure 4 :transverse plane

Through the kinematic analysis, patients' joint movements can be reconstructed in the three planes. This reconstruction is achieved through the application of reflective spherical passive markers (visible to the infrared) on special landmarks. These markers, applied according to the *T3Dg Protocol*, are detected by the optoelectronic system, Vicon MX + System, (Vicon Motion System), consisting of 8 infrared light emission cameras with high resolution (100 Hz).

Single joint movements can then be analyzed separately on the three planes (frontal, sagittal, transverse) leading to the generation of kinematic graphs representing joints angular movements (expressed in degrees) referred (after normalization and sampling, as it will be detailed in Chapter 4) to the gait cycle percentages.

3.3 Available data structure

The choice of the variables to be monitored is mainly based on some clinical considerations. Actually, according to Ferrari classification, the standardized distinctive characteristics of the four *diplegia* forms (used at an observational level by specialists to produce such a classification) are referred to the joints rotation. For this reason, articulations of trunk, knees, ankles and hips have been taken into account.

In particular, **trunk rotation** in the sagittal plane may inform about the antepulsion that is typical of the patients belonging to the first group. Actually, they tend to face the entire walk bent forward, with constant support on four point canes which are placed in front and laterally to the trunk. Trunk rotation also indicates the presence of a sagittal pendulum, typical of the patients of the fourth group whose walk is quick and clocked. However, frontal pendulum highlights the lateral oscillations of the trunk used by patients of the third group to balance their weight.

The measurement of the **angular knee extension** on the sagittal plane should make possible the distinction of the first *diplegia* form from the second one. In fact, the first one is characterized by a joint block, basically due to musculoskeletal problems, the latter one by a knee flexion during the gait midstance phase.

Hip rotation in the transverse plane allows the identification of possible intra-rotations when the foot is in contact with the ground. This variable is also important for patients belonging to the first and the fourth group. Actually, for them the position of the lower limbs plays an adaptive equilibrium function. Finally, **ankle joint rotations** in the sagittal plane may indicate a possible block of the foot in flexion-extension, called club-foot. This factor is indicative for patients of the first form. In fact, they tend to balance their entire walk on tiptoe.

Then, functional variables, representative of the angular rotations of the previously described articulations in the three anatomical planes and recorded over three gait cycles, are coded with the following names:

- for the **knees:** RKANGLE_S, RKANGLE_F, RKANGLE_T
  LKANGLE_S, LKANGLE_F, LKANGLE_T
- for the **trunk:** RTRKANGLE_S, RTRKANGLE_F, RTRKANGLE_T
  LTRKANGLE_S, LTRKANGLE_F, LTRKANGLE_T
- for the **hip:** RHANGLE_S, RHANGLE_F, RHANGLE_T
  LHANGLE_S, LHANGLE_F, LHANGLE_T
- for the **ankles:** RAANGLE_S, RAANGLE_F, RAANGLE_T
  LAANGLE_S, LAANGLE_F, LAANGLE_T

where L and R means respectively left and right.

3.4 Methods

This research work aims at identifying suitable indicators to employ for the discrimination of four groups. Medical literature is full of examples of discrimination problems between two groups, healthy and pathological ones. Here the discrimination problem is extended to four groups. This complicates the identification of suitable indicators to employ as predictors. As it will be detailed in Chapter 4, two methods are proposed to achieve this purpose. They differ both for the theoretical approach and for the use of statistical methodologies. In particular, the first proposed method tries to identify suitable *static indicators* from the available functional data, while the second one preserves the functional nature of the data by investigating their hidden variability. However, despite the use of different predictors, both the methodologies propose a Linear Discriminant Analysis in order to identify a discriminant function. In the following paragraphs the most relevant theoretical aspects of the Linear Discriminant Analysis (LDA) and some theoretical considerations relating to the Principal Components Analysis for Functional data (FPCA), used in the second proposed approach, will be presented. For further theoretical deepening on FPCA, refer to [30][35][37].

3.4.1 Generalities on Linear Discriminant Analysis (LDA)

A discriminant model [26][27] among groups aims at predicting which group a new case belongs to. In most common applications of discriminant function analysis, many variables or predictors are considered in order to determine the ones with a high discrimination power. The linear discriminant function can be expressed by the following equation:

$$Z = a + W_1 X_1 + W_2 X_2 + \cdots + W_k X_k \qquad (3.1)$$

where *Z, the discriminant score,* is used to predict group membership, a is the discriminant constant and X_k are the explicative variables or predictors. In discriminant analysis the Total Sum of Squares (TSS) is partitioned into the Between Group SS (BSS) and the Within Group SS (WSS):

$$\boldsymbol{BSS} = (\overline{\boldsymbol{Z}}_{\boldsymbol{0}} - \overline{\boldsymbol{Z}})^2 + (\overline{\boldsymbol{Z}}_{\boldsymbol{1}} - \overline{\boldsymbol{Z}})^2 = \sum(\overline{\boldsymbol{Z}}_{\boldsymbol{i}} - \overline{\boldsymbol{Z}})^2 \qquad (3.2)$$

$$\boldsymbol{WSS} = (\boldsymbol{Z}_{\boldsymbol{i0}} - \overline{\boldsymbol{Z}}_{\boldsymbol{0}})^2 + (\boldsymbol{Z}_{\boldsymbol{i1}} - \overline{\boldsymbol{Z}}_{\boldsymbol{1}})^2 = \sum\left(\boldsymbol{Z}_{\boldsymbol{ij}} - \overline{\boldsymbol{Z}}_{\boldsymbol{j}}\right)^{\boldsymbol{2}} \qquad (3.3)$$

where i represents an individual case, j the group, Z_i an individual discriminant score, $\bar{Z}_j$ the mean discriminant score for group *j* (called *centroids*) and $\bar{Z}$ the grand mean of the discriminant scores. Discriminant analysis uses OLS to estimate the values of the parameters a and W_k that minimize the Within Group SS, WSS.

In this study, a stepwise discriminant function analysis is applied. Then, the discrimination model is built step-by-step. At each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination among groups. That variable will then be included in the model, and the process starts again. Wilks' lambda criterion is used to select significant predictors. The latter indicates whether or not there is a significant relationship between the predictors and the dependent variable.

To measure the *goodness-of-fit*, Wilk's lambda operates as follows. In case of two groups, the discriminant function can be extracted from data and the associated eigenvalue is:

$$\lambda = \frac{BSS}{WSS} \tag{3.4}$$

It turns out that if $\lambda = 0$ ($BSS = 0$), the model has no discriminatory power. The larger the value of λ, the greater the discriminatory power of the model. The Wilks' Λ for the discriminatory model is

$$\Lambda = \frac{1}{1+\lambda} = \frac{WSS}{TSS} \tag{3.5}$$

Λ is chi-square distributed with $df = (k - 1)$, where k is equal to the number of estimated parameters. Therefore, in terms of Λ, the more the parameter is close to 1 the less the discriminant power of the model is. For this reason Wilks' Λ is such an inverse quality criterion.

The stepwise introduction of predictors terminates when all the significant variables are considered and, of course, the discriminant power of the model is satisfying. Then an estimation of the *hit ratio* (HR) is needed. The latter gives the correctly classified observation units divided by the total number of observation units. If, for example, a classification matrix is considered for a two groups discriminant model,

TRUE CLASS MEMBERSHIP	PREDICTED CLASS MEMBERSHIP	
	GROUP I	GROUP II
GROUP I	$C11$	$C12$
GROUP II	$C21$	$C22$

Table 1: determination of class membership

the following relation holds:

$$HR = \frac{c11+c22}{c11+c12+c21+c22} \qquad (3.6)$$

When the same data set is used both for estimating the DA model and the classification, an over estimation of the HR is expected. To avoid this, the *leave-one-out cross-validation* method can be employed. This technique works by omitting each observation one at a time, recalculating the classification function using the remaining data, and then classifying the omitted observation.

The computation time is obviously longer, but an optimistic error rate is compensated. Assumptions for DA model are the same of those for the multivariate analysis of variance (MANOVA): a) data (for the variables/predictors) represent a sample from a multivariate normal (or quite) distribution inside each group; b) the variance/covariance matrices of variables are homogeneous across groups. Minor deviations are not so important; c) groups defined by dependent variables exist a priori; d) variables used to discriminate between groups need to be not completely redundant. If any one of the variables is completely redundant with the other variables used, then the matrix is said to be *ill-conditioned*, and it cannot be inverted. Therefore, it is necessary to evaluate if such assumptions hold for the chosen variables.

For multivariate analysis, data need to follow a multivariate normal distribution or, if not exactly, at least approximately. To assess multivariate normality, several visual procedures have been suggested in the literature. Here, it is proposed the Chi-square plot of squared Mahalanobis distance[28]. A plot of the ordered squared distances d_i^2 and $100\left(\frac{i-0.5}{n}\right)$ quantiles of the Chi-squared distribution with p degrees of freedom is called a Chi-square plot.

Distance d_i^2 is $d_i^2 = (x_i - \bar{x})'\Sigma^{-1}(x_i - \bar{x})$ with i=1, 2,….n and X_1, X_2,….X_n sample observations each measured on the p variables.

Before performing such an analysis, it is reasonable evaluating the univariate normality of the predictors. Actually, if predictors have a normal distribution, it does not imply the multivariate one but, if univariate predictors are not normal, it is sure that a multivariate distribution is far from normality. The assumptions of univariate normality are simply investigated with a Normal Probability Plot (NPP).

3.4.2 Principal Component Analysis (PCA)

So far, Ramsay and Dalzell (1991)[29] outlined the advantages of applying functional data analysis in practice. In particular, they highlighted the advantages provided by smoothing and interpolation procedures that can yield a functional representation of a finite set of observations. However, modeling problems are more natural to be considered functionally because functional pre-processing (i.e. derivatives) can provide insights into functional data display and functional linear regression models. Ramsay and Silvermann (2005)[30] proposed functional data analysis as an effective methodology to represent data in ways that aid further analyses and especially to study important sources of pattern and variation. Due to these practical advantages, in the last 10 years, functional data analysis has received a great attention in different scientific fields, from the analysis of handwritten in Chinese (Ramsay 2000)[31] to the analysis of price dynamics in online auctions (Wang, Jank, Shmueli &Smith2008)[32], to climatology (Meiring 2007)[33], to medical research (Erbas et al.2007)[34], and many other more. In the book *Applied Functional Data Analysis*, Ramsay & Silverman (2002)[35] gave a number of very interesting applications with continuous functional variables. Even additional non-parametric features have been incorporated into functional data analysis by Ferraty, F. & Vieu, P. (2006)[36]. Generally, the continuous functional variable is time, even though functional data may be observed over age, space, wavelength, molecular weight and so on.

As it will be shown in the next Chapter, the second proposed approach is based on the Principal Component Analysis applied to functional data. Therefore, in this section some useful details are provided with a continuous attention to the description of peculiarities that will be adopted (see Chapter 4) to analyze the data.

PCA finds the most informative or explanatory features hidden in the data, without needing an a priori-knowledge or hypotheses on their structure. It accomplishes this by computing a new smaller set of uncorrelated variables, (PCs) that represent the original data set. Each new variable is a linear combination of the original ones. In particular, the first principal component (PC1) is the linear combination of the original variables which accounts for the maximum amount of variance in a single direction. It is the line of best fit through the data, therefore its residual variance is a minimum for the complete data set. The second principal component (PC2) is orthogonal to the first one and accounts for the maximum amount of the remaining variance in the data. All the principal components are orthogonal to each other. Then there is no redundant information. Therefore, the first two components represent the plane of best fit through the data. All the remaining principal components can be defined in the same way, so that the lowest order components normally account for very little variance and can usually be ignored with respect to some components criteria selection[35]. It is very interesting to interpret PCA from a geometrical point of view. Actually, it can be considered as a rotation of the axes of the original variable coordinate system to new orthogonal axes, called *principal axes*. These new axes coincide with the directions of maximum variation of the original data.

From a mathematical point of view, PCA performs an orthogonal transformation that converts *p* variables, $X_1, X_{2,\ldots,} X_p,$ into *p* new uncorrelated principal components, $Z_1,$ $Z_{2, \ldots,}$ $Z_{p.}$ The PC model is then $Z=U^TX$ where the columns of *U* (U_1, $U_{2,\ldots,}$ U_p) are called *principal component loading vectors* and are the eigenvectors of the covariance matrix of *X*.

In a functional context, each principal component is specified by a *principal component weight function* defined over the same range of time as the original functional data. Then the individual principal component scores z_i are given by a combination of the weight function and the original data. In real situations, when the plotted function representing the raw data collected over time is not smooth, a preliminary data treatment needs to be performed in order to smooth the curve. This can be accomplished with different approaches [35][37]. For the case under study, since data are collected at a frequency of 100Hz over an time interval (*gait cycle* time interval) that in the best situation (subjects belonging to the fourth group) lasts in mean for about 1.05 sec, the plotted function is reasonably smooth. Therefore, no previous treatment is necessary.

Chapter 4

Analyses, Results and Discussion

4.1 Introduction

In this chapter results of the analyses and a critical discussion about them are reported. As underlined in the previous chapter, in order to discriminate the four groups, two methodologies have been proposed and compared. The first method proposes some "*static" indicators*, derived from functional data, as predictors for the discriminant model, while the second one uses directly functional data reduced by PCA and proposes the individual PC scores as discriminating predictors.

4.2 The *"static" indicators* method

This first method is based on the extrapolation of synthetic indicators from the original functional variables to be employed for the identification of the proposed *diplegia* forms. This is particularly useful for the implementation of automatic classification systems as neural networks or Bayesian networks [38-42] that are easier to implement with discrete input variables. Commonly, extracted indicators include peak values or magnitudes of signals at specific gait cycle events. The choice of such indicators is subjective and strictly linked to the specific application context. Obviously no choice is the best choice, but only one possible among many others. However, if the choice is not accurate, it is possible that the considered indicators are highly correlated or not so representative of the curves. Therefore, in order to reduce possible sources of errors, the choice has been based on some clinical suggestions. Actually, since dependently from its degree of severity *diplegia* affects in a different way the angles representative of the articulations motion, the following indicators are considered and extracted from the gait cycle curves:

- the Range of Motion (ROM), i.e. the maximum variation of the rotation angle of the considered articulation. It is defined as the difference between the maximum and the minimum angles measured during a gait cycle, i.e.

$ROM = \max(x) - \min(x)$, where x represents the vector of measured angle values in a cycle;

- the Root Mean Square (RMS), to be interpreted as an indicator of variation with respect to a constant value of a time dependent signal, i.e. $RMS = \sqrt{\frac{\sum_{i=1}^{n}(x_i)^2}{n}}$, where x_i represent the measured angle values in a gait cycle;
- the Crest Factor (CF), defined as the ratio between the absolute value of the maximum in a set of data and the RMS, i.e. $CF = \frac{\max(x)}{RMS(x)}$, where x represents the vector of measured angle values in a gait cycle. It is representative of some impulsive phenomena characterizing the pendular movements that are typical in almost all the *diplegia* forms.

Considered the number of available functional variables and the three different detection planes, the following considerations hold. The 8 functional variables have been presented in the previous Chapter. Each of them is measured in three different planes: sagittal, frontal and transverse. Therefore, there are 24 available functional variables. Since from each variable 3 discrete indicators are extracted (ROM, RMS, CF), there 72 potential predictors to be used for the discrimination analysis. For each subjects three gait cycles are available. Then ROM, RMS, CF used in the discriminant analysis are the mean values related to the three gait cycles.

In order to assess the differences for these indicators among the four independent groups, the non-parametric Kruskal Wallis (KW) test has been chosen. This test offers a non-parametric alternative to the one-way analysis of variance when variables are non-normally distributed. When significant, KW test has been followed by the Mann Whitney (MW) post-hoc test. Since the level of significance of the test is chosen to be α=0.05 and 6 comparisons for four groups have been performed, Bonferroni correction has been considered. Actually, by choosing α = 0.00833 (i.e. 0.05/6), the probability to have at least one significant result is

$\mathrm{P(at\ least\ one\ significant)} = 1 - \mathrm{P(no\ significant\ results)} = 1 - (0.00833)^6 \approx 0.04895$, very closed to 0.05.

4.3 Results of the KW and the MW test

In the following Table 2, the corresponding p-values related to some of these tests with the corresponding variables are reported.

Kruskal Wallis test	
PREDICTORS	***p*-value**
ROM_LAANGLE_S	0,0000
ROM_RAANGLE_S	0,0000
ROM_LHANGLE_S	0,0000
CF_LAANGLE_S	0,0030
CF_LKANGLE_S	0,0000
RMS_RTRKANGLE_S	0,0000
ROM_RHANGLE_S	0,0000
ROM_RKANGLE_S	0,0000
CF_RKANGLE_S	0,0000
ROM_RTRKANGLE_F	0,0000
RMS_RAANGLE_F	0,0000
ROM_LAANGLE_F	0,0015
RMS_LAANGLE_F	0,0000
CF_LAANGLE_T	0,0000

Table2: KW test for some variables

To know which differences are significant, *post-hoc* Mann Whitney test has been performed. Results of such tests are reported below for some variables related to the sagittal and the frontal planes. Figures 5, 6 and 7 show the boxplots of RMS_RAANGLE_S, CF_RAANLE_S and ROM_RAANGLE_S

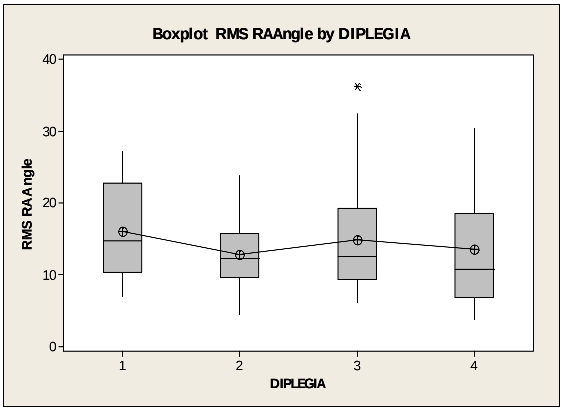


Figure 5: boxplot of RMS_RAANGLE_S

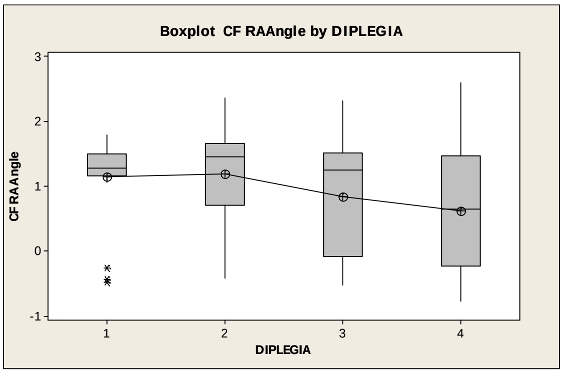


Figure 6: boxplot of CF_RAANGLE_S

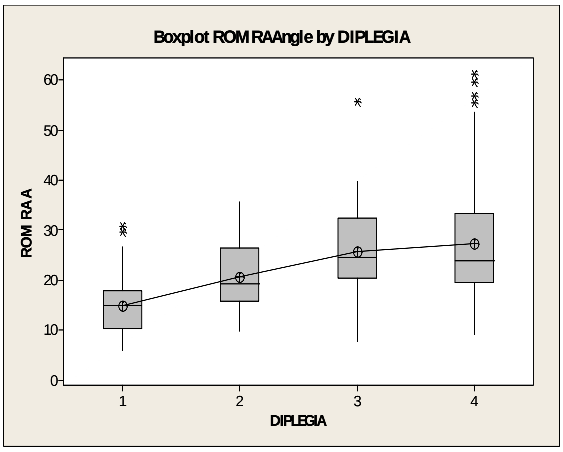


Figure 7: boxplot of ROM_RAANGLE_S

p-values for the variables indicated above are synthesized in the following Tables 3, 4 and 5.

***p*-value of Mann-Whitney test (Alpha Bonferroni=0,00833)**				
RMS RAAngle_S	**Diplegia 1**	**Diplegia 2**	**Diplegia 3**	**Diplegia 4**
Diplegia 1				
Diplegia 2	0,057			
Diplegia 3	0,393	0,375		
Diplegia 4	0,031	0,456	0,091	

Table3: MW results for RMS_RAANGLE_S

***p*-value of Mann-Whitney test (Alpha Bonferroni=0,00833)**				
CF RAAngle_S	**Diplegia 1**	**Diplegia 2**	**Diplegia 3**	**Diplegia 4**
Diplegia 1				
Diplegia 2	0,1108			
Diplegia 3	0,3484	0,0072		
Diplegia 4	0,0072	0,0000	0,1922	

Table4: MW results for CF_RAANGLE_S

***p*-value of Mann-Whitney test (Alpha Bonferroni=0,00833)**				
ROM RAAngle_S	**Diplegia 1**	**Diplegia 2**	**Diplegia 3**	**Diplegia 4**
Diplegia 1				
Diplegia 2	0,0003			
Diplegia 3	0,0000	0,0000		
Diplegia 4	0,0000	0,0000	0,7077	

Table5: MW results for ROM_RAANGLE_S

Values of RMS_RAANGLE_S are not significantly different for the four groups. On the contrary, for CF_RAANGLE_S differences between I/IV (diplegia forms), II/III and II/IV are significant. For ROM_RAANGLE_S, all the tests are significant, except that related to III and IV.

In the following Table 6 are reported the results of the MW tests for some of the previous significant variables with the correspondent *p*-values. Since most of them are significant, they will be considered for the discriminant analysis.

Mann-Whitney test (Alpha Bonferroni=0,00833)		
PREDICTORS	**DIPLEGIA FORM**	***p*-value**
ROM_LAANGLE_S	II/IV	0,0022
ROM_LAANGLE_S	I/IV	0,0012
ROM_RAANGLE_S	I/IV	0,0018
ROM_RKANGLE_S _MID	I/IV	0,0020
ROM_RKANGLE_S _MID	II/IV	0,0001
ROM_LKANGLE_S _MID	II/IV	0,0070
ROM_LHANGLE_S	II/IV	0,0000
ROM_LHANGLE_S	III/IV	0,0009
CF_LAANGLE_S	I/IV	0,0036
CF_LKANGLE_S	II/IV	0,0001
RMS_LKANGLE_S	II/IV	0,0012
RMS_RTRKANGLE_S	I/III	0,0014
RMS_RTRKANGLE_S	I/IV	0,0001
RMS_RTRKANGLE_S	II/IV	0,0004
ROM_RHANGLE_S	II/IV	0,0012
ROM_RTRKANGLE_F	I/IV	0,0047
RMS_RAANGLE_F	II/IV	0,0017
ROM_LAANGLE_F	I/II	0,0018
ROM_LAANGLE_F	I/III	0,0074
ROM_LAANGLE_F	I/IV	0,0067
RMS_LAANGLE_F	III/IV	0,0075
CF_LAANGLE_T	I/IV	0,0056
ROM_RKANGLE_S	II/IV	0,0000
CF_RKANGLE_S	II/IV	0,0000

Table 6: MW post-hoc test for some variables

4.4 The *gait cycle meantime* (*gcm*) variable

As underlined in Chapter 3, variables employed in this study are functional variables representative of the angular rotations related to some joints during a gait cycle. For each subject, three independent gait cycles have been detected. Gait cycle can be defined as a repetitive pattern involving steps and strides[10]. A step is one single step and a stride is a whole gait cycle. Step time is defined as the time from one foot hitting the floor to the other foot hitting the floor. Gait speed determines the contribution of each body segment. Normal walking speed primarily involves the lower extremities. Actually, arms and trunk provide stability and balance. The faster the speed, the more the body depends on the upper extremities and trunk for propulsion as well as for balance and stability [12]. In subjects affected by CP, balance and stability are generally compromised. This has an evident impact on the speed and, as a consequence, on the duration of the gait cycle. In order to understand if differences in cycle duration hold, the following considerations have been made.

The three gait cycles available for each subject can be considered as replicates. In Figure 8 the three gait cycles related to subject 18 are reported (variable RKANGLE_S).

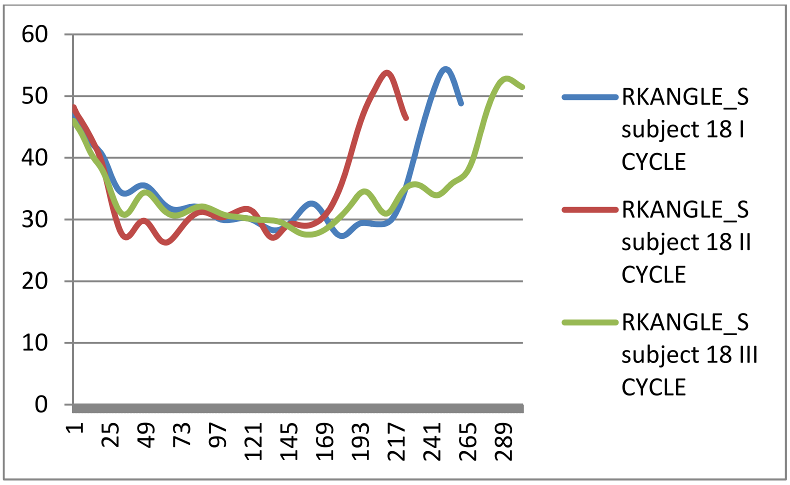


Figure 8: the three gait cycles for RKANGLE_S subject 18

The first cycle is constituted of 263 angular values, the second one of 226 and the third one of 303 values. Since the system frequency is 100Hz, it is possible to determine the duration of the three gait cycles and, for each subject, to consider the mean cycle duration. For this case, first cycle lasts for 2.63 sec, second for 2.26 sec and the third for 3.03 sec. Therefore, subject 18 needs in mean 2.64 sec. to complete a gait cycle. Performing the same considerations for all the 91 subjects involved in the study, the following mean cycle durations per group result: I group 3.04 sec, II group 1.48 sec, III group 1.14 sec. and IV group 1.05 sec. In order to test if such differences are significant, Kruskal Wallis test has been performed. Since it was significant (*p*-value=0.0000), MW post-hoc test has been used to detect which groups are significantly different in terms of the *gcm* variable. In the following Table 7 results are reported.

***p*-value**	**Mann-Whitney test (Alpha Bonferroni=0,00833)**			
	DIPLEGIA 1	**DIPLEGIA 2**	**DIPLEGIA 3**	**DIPLEGIA 4**
DIPLEGIA 1				
DIPLEGIA 2	0,0004			
DIPLEGIA 3	0,0003	0,0063		
DIPLEGIA 4	0,0000	0,0012	0,2209	

Table 7: MW test for *gcm* variable

Differences are significant for all the groups except for groups III and IV. This result is coherent with some clinical considerations. Actually, the four groups are in a decreasing order of degree of severity. The first group is the most compromised, then the difficulty in maintaining stability and balance during the walk has an evident impact on the duration of the gait cycle. Therefore, in consideration of the proved discrimination feature of the mean gait cycle duration, variable *gcm* will be considered for the further discriminant analysis.

4.5 Univariate normality tests.

Multivariate normality within each group is required for the variables involved in the discriminant analysis. Therefore, in order to understand if univariate distributions of each considered variable are normal, normal probability plots (NPP) have been previously built. This condition is not sufficient to assess multivariate normality, but for sure, if univariate distributions are far from normality, the multivariate one won’t be normal. Variables whose univariate distribution was found far from normality, have been transformed by Box-Cox transformations. For instance, in the following Figure 9, 11 the NPPs for variables CF_RKANGLE_S, RMS_RKANGLE_S, ROM_RKANGLE_S are reported, while in Figure 10, 12 their related NPPs after log-transformation are shown.

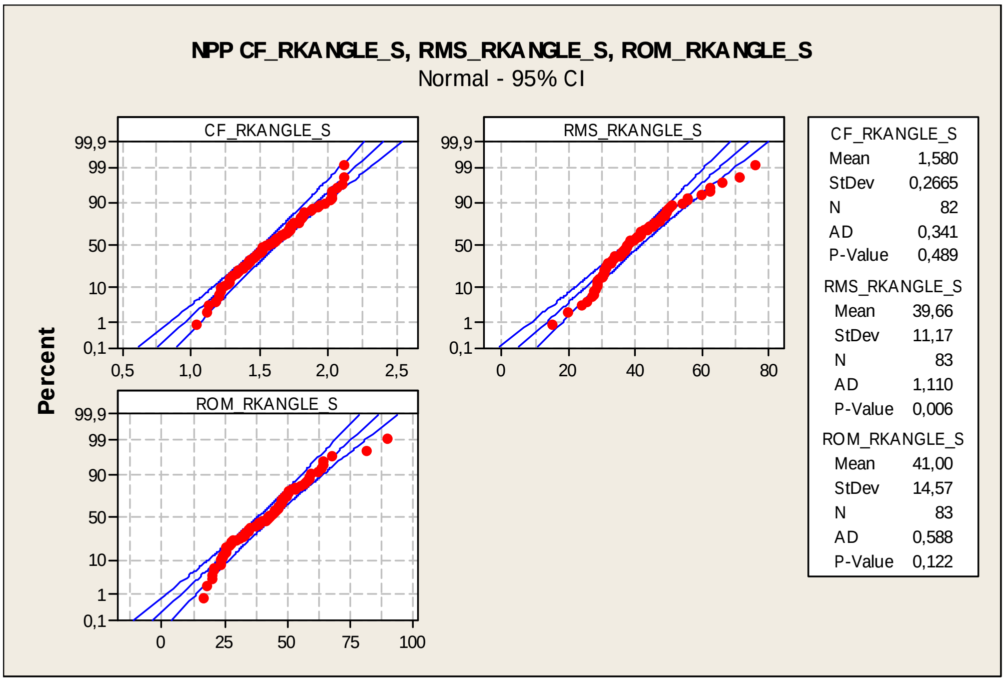


Figure 9: NPP for three variables in the sagittal plane

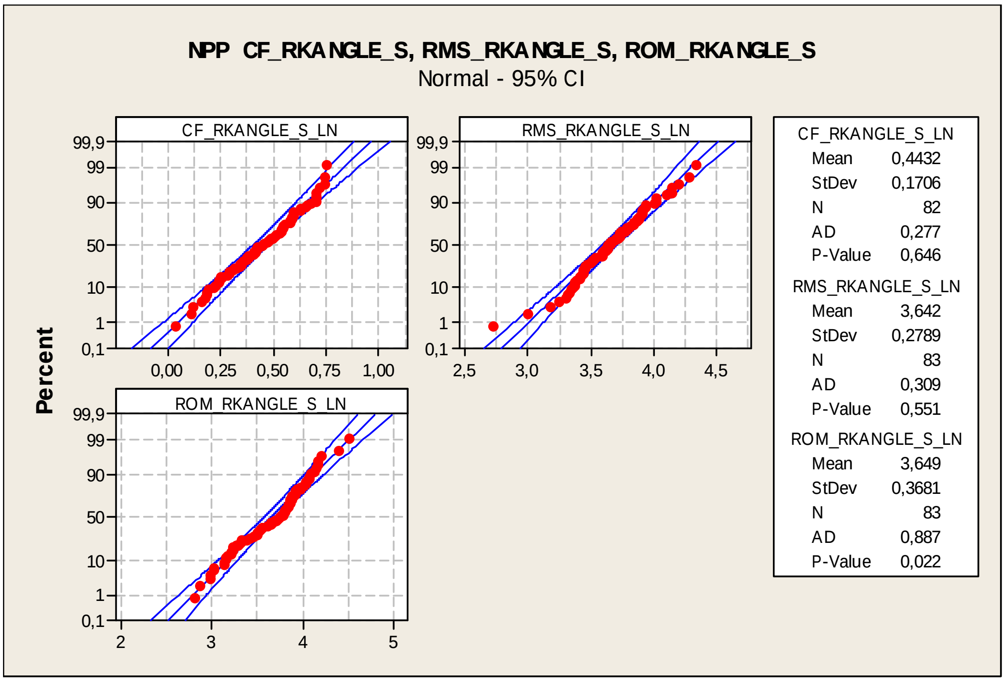


Figure 10: NPP three variables after log-transformation in the sagittal plane

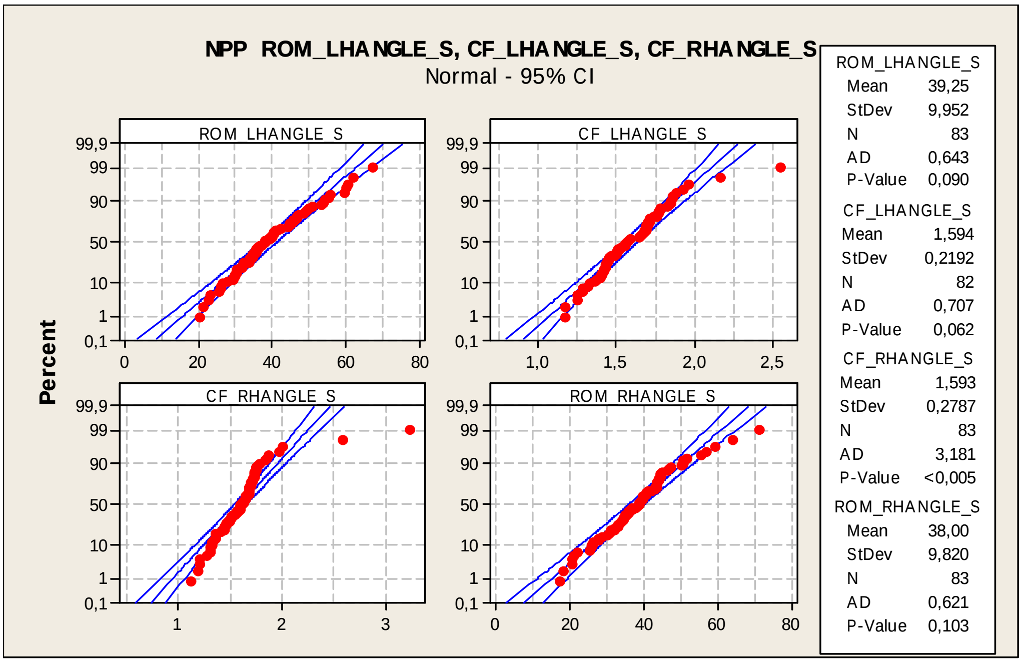


Figure 11: NPP of four variables in the sagittal plane

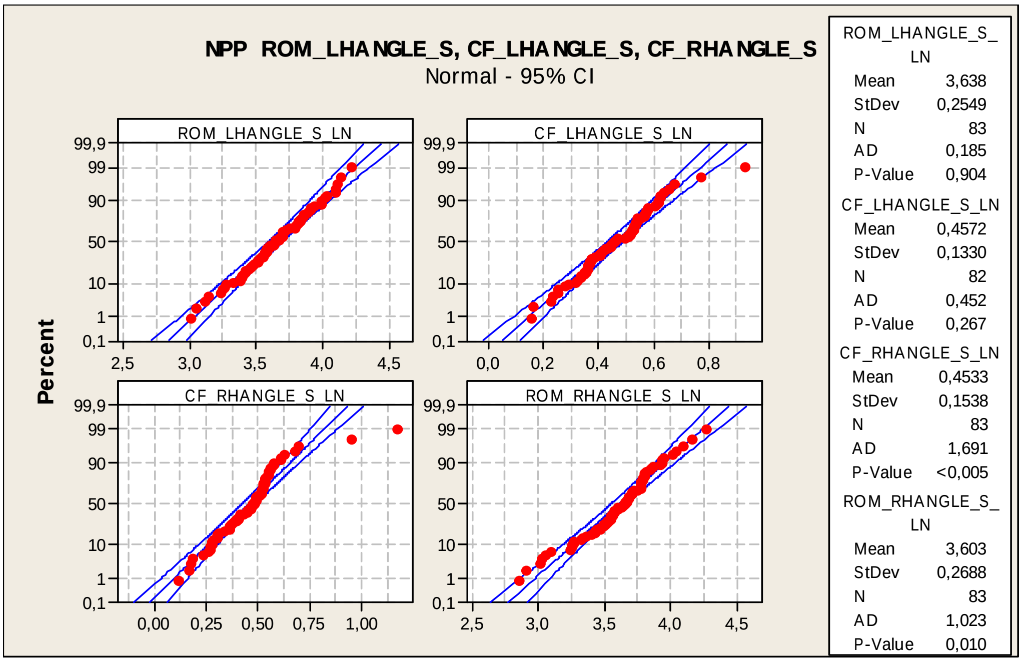


Figure 12: NPP of four variables in the sagittal plane after log-transformation

It can be observed from Figure 12 that CF_RHANGLE_S is not completely normalized after log-transformation.

4.6 The *midstance segments* of the waveforms

In Chapter 2 gait cycle has been described in detail. As underlined, it involves two main phases, the *stance* phase and the *swing* phase. The *stance* phase occupies the 60% of the whole cycle. In particular, the midstance phase (from 10% to 30% of the stance phase) is characterized by the settlement of the foot at the lateral border. *Midstance segments* have been focused and analyzed because of some clinical considerations. Actually, the peculiarity of subjects belonging to the II group is just the knee flexion during midstance. Considering that gait cycles have a different duration, in order to detect the midstance phase, the original functional variables involving knee joints have been normalized and sampled at each 1% from 1% to 100%. Therefore, after normalization, segments related to the percentage of gait cycle going from 10% to 30% have been extracted and ROM, CF and RMS computed. Figures 13 and 14 represent respectively the normalized gait cycles of the variable RKANGLE_S referred to two subjects belonging to the II group and the extracted midstance segments.

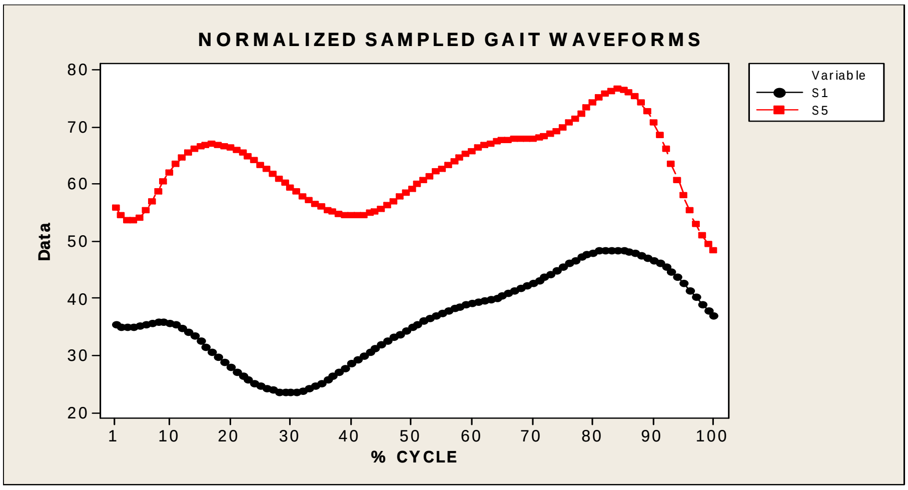


Figure 13: normalized RKANGLE_S variables for two subjects belonging to the II group

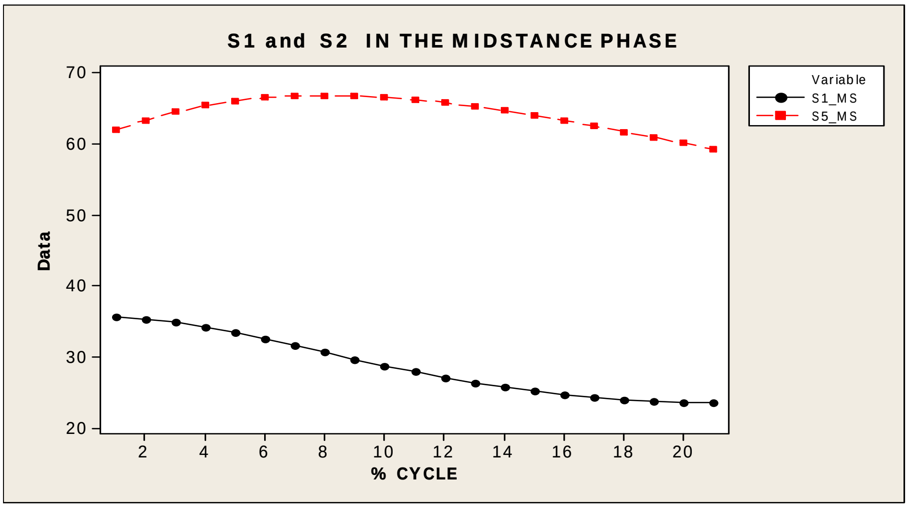


Figure 14: *midstance segments* for the normalized RKANGLE_S variables

Variables involved in this analysis are RKANGLE and LKANGLE related to the three detection planes. The correspondent *midstance variables* have been renamed as RKANGLE_MID and LKANGLE_MID. ROM, CF and RMS indicators extracted from these waveforms have proved to discriminate better among groups than those related to the whole cycle. Therefore, they have been considered as possible predictors for the discriminant model.

4.7 Discriminant model

As described in Chapter 3, a stepwise discriminant procedure has been used to determine the discriminant model. Considering that three discrete indicators have been extracted from the 24 functional variables, the potential predictors for the discriminant analysis are 72. In particular, 24 referred to the sagittal plane, 24 to the frontal plane and 24 to the transverse one. The discriminant procedure, performed as described in Chapter 3, indicates that only 11 of the 72 available variables are significant as predictors for the discrimination of the four forms of *diplegia*.

The selected predictors are reported in the following Table 8.

SELECTED PREDICTORES		CODED NAMES
CF_LKANGLE_S		E
RMS_LKANGLE_S		F
ROM _RTRKANGLE_S		G
ROM_LAANGLE_S		P
ROM_LHANGLE_S		X
MEANTIME		A2
ROM_LAANGLE_F		A20
RMS_LAANGLE_F		A22
CF_LAANGLE_T		A28
CF_RKANGLE_T		A40
ROM_RKANGLE_S _MID		RM

Table 8: selected predictors

A great number of the original variables have been found to be highly correlated and then they have been discarded at the beginning. Results of the discrimination analysis with the 11 predictors show that the proportion of correctly classified observations intra groups is 77.3%. In particular, the proportion of correctly classified for the first group is 100%, for the second group is 66.7%, for the third group 73.3% and, finally, for the fourth is 81.6%. However, since all the available data have been used both for the model and the validation, the accuracy of the discriminant model has been assessed by cross-validation technique. The proportion of correctly classified among groups after cross-validation is 70.5%, while that one related to the individual groups are: 87.5% (I group), 59.3% (II group), 60% (III group), 78.9% (IV group). Changes are particularly evident for groups II and III. The misclassification rate is very high and, as a consequence, the global result is not very satisfying. As previously underlined, Wilks' criterion was used to assess the *goodness-of-fit* for the model. Wilks'Λ is found to be 0.20836.

The number of misclassified observations is high. In particular, observations 5, 17, 22, 42, 44, 69 are misclassified after cross-validation (see Table 9), while those reported in Table 10 are not predicted in the correct group.

Misclassified observations			
OBSERVATION	**TRUE GROUP**	**PREDICTED GROUP**	**CROSS VALIDATION**
5	1	1	2
17	2	2	1
22	2	2	3
42	3	3	2
44	3	3	4
69	4	4	3

Table 9: misclassified observations after cross-validation

Misclassified observations			
OBSERVATION	**TRUE GROUP**	**PREDICTED GROUP**	**CROSS VALIDATION**
9	2	4	4
11	2	4	4
12	2	3	3
13	2	1	1
14	2	1	1

17	2	2	1
28	2	3	3
31	2	4	4
34	2	3	3
35	2	3	3
43	3	4	4
45	3	4	4
47	3	2	2
50	3	2	2
52	4	3	3
60	4	3	3
68	4	1	1
71	4	2	2
74	4	3	3
76	4	3	3
87	4	2	2

Table 10: misclassified observations

To identify a new observation, the linear discriminant functions (see Table 11) associated with the four groups have been computed. The new observation belongs to the group whose discriminant function value is higher.

LINEAR DISCRIMINANT FUNCTIONS FOR GROUPS				
	I GROUP	**II. GROUP**	**III GROUP**	**IV GROUP**
Constant	-267,88	-270,311	-262,127	-278,281
E	56,232	63,699	57,685	63,853
F	72,579	73,508	65,483	70,062
G	11,061	11,333	11,907	11,597
P	-18,332	-16,976	-15,761	-14,609
X	72,181	65,114	65,385	67,56
A2	17,664	29,541	36,566	34,657
A20	-3,238	0,834	1,727	-0,366
A22	3,991	4,331	6,924	4,106
A28	7,103	4,968	6,844	5,084
A40	2,639	2,148	2,22	2,579

RM	-2,907	-2,072	-1,743	-1,215

Table 11 : linear discriminant functions for groups

In Table 12, distances among the four groups for the validated model are shown.

	SQUARED DISTANCE AMONG GROUPS			
	1	**2**	**3**	**4**
1	0			
2	8,7794	0		
3	18,8569	4,2835	0	
4	16,7252	4,0714	3,5215	0

Table 12 : squared distances among groups

Variables selected for the discriminant model and shown in Table 8 exhibit a normal univariate distribution. In the following Figures 15 and 16, NPP are shown for some of them.

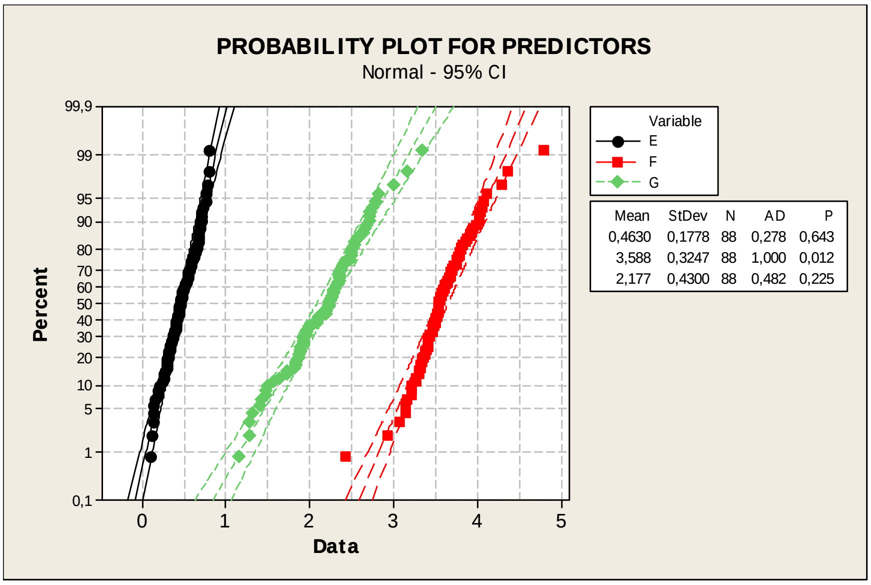
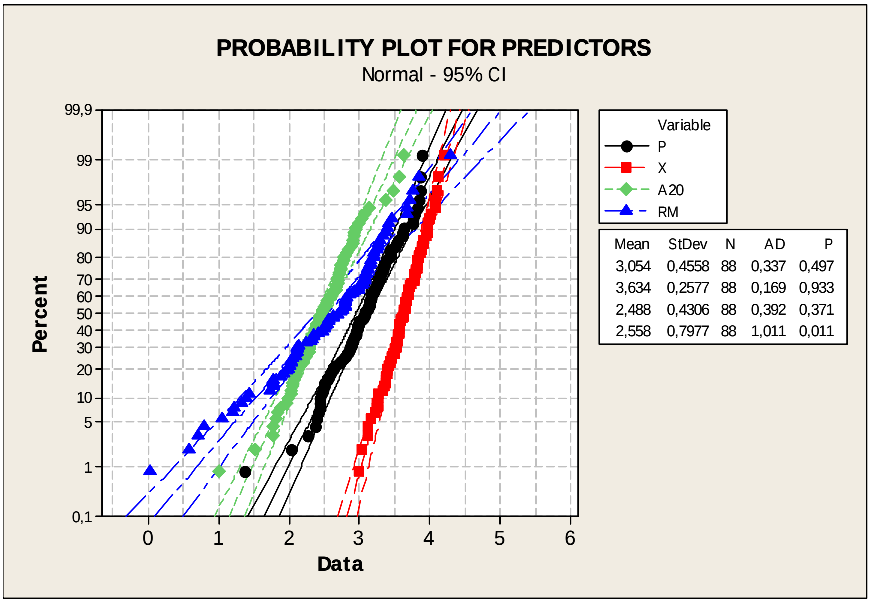


Figure 15: probability plots for variables E, F, G



Mean	StDev	N	AD	P
3,054	0,4558	88	0,337	0,497
3,634	0,2577	88	0,169	0,933
2,488	0,4306	88	0,392	0,371
2,558	0,7977	88	1,011	0,011

Figure 16: probability plots for variables P, X, A20, RM

In order to assess if data follow a multivariate normal distribution inside each group, a graphical procedure has been applied [28]. Even if this procedure is more rigorous when both the number *n* (sample observations each measured on the *p* variables) and *p* are greater than 30, results can suggest if data are very far from multivariate normality. This empirical rule holds because, only when *n* and *p* are great, squared distances behave like chi-square random variables. The following Figures 17 and 18 represent the Chi-square plots for group II and IV. Plots follow almost a linear pattern. No systematic curved pattern is visible, the assumption of multivariate normality can be excepted.

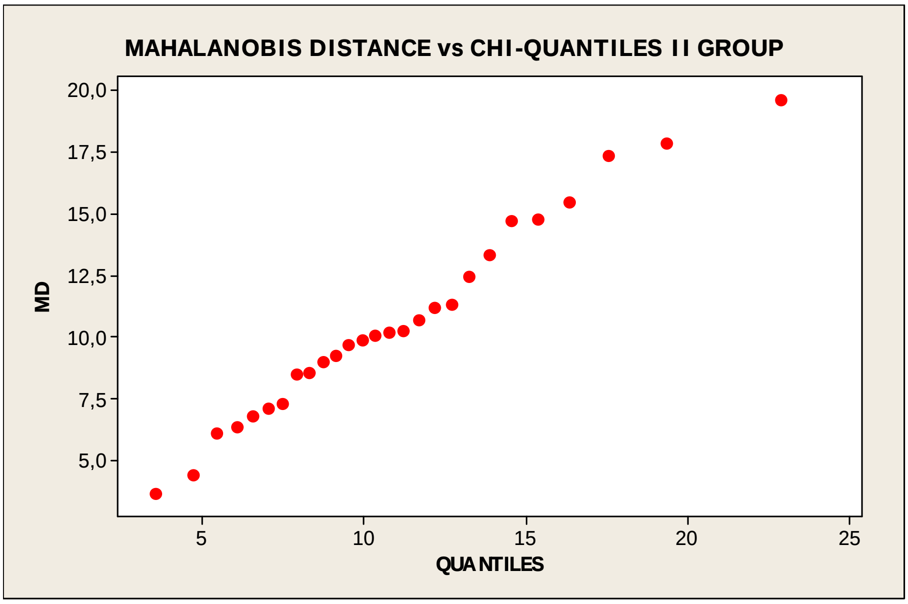


Figure 17: Mahalanobis distance vs Chi-quantiles II group

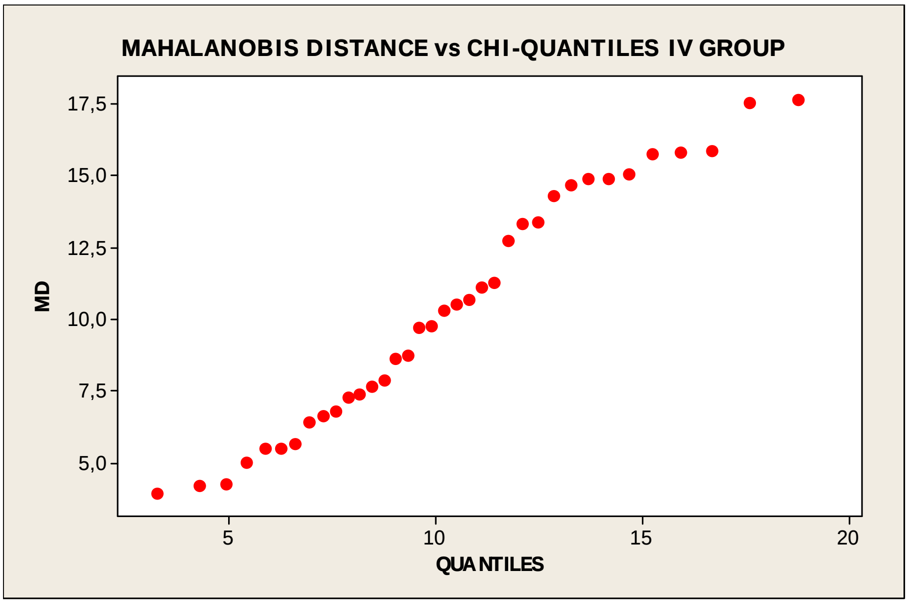


Figure 18: Mahalanobis distance vs Chi-quantiles IV group

4.8 Considerations on the *"static" indicators* model

Model classification rate is low. Only 70.5% of observations are correctly classified after cross-validation. Of the original potential 72 variables, only 11 have been selected and used for discrimination purposes. This because of the high collinearity and the lower discriminant power of a great number of involved variables. In order to improve the quality of the classification and then to reduce the error rate, principal component analysis has been applied to the functional data. As previously underlined, such a methodology has been generally used for medical purposes in order to compare two groups (healthy and pathological ones), while here the discrimination problem is extended to four groups. This increases the difficulty. Actually, if it is easy to distinguish a I-form diplegia (the most severe) from the other ones, features of the third and fourth forms are more closed. Then differences are more difficult to capture. PCA can better capture data variability[43-44] because the entire waveform associated to each gait cycle is taken into account. However, PCA provides few components that can explain the great part of data variance. In the next paragraph this analysis is detailed.

4.9 Preliminary considerations on the FPCA

As highlighted in Chapter 3, Principal Component Analysis is a standard approach for the exploration of variability in multivariate data. PCA uses an eigenvalue decomposition of the variance matrix of the data to find directions along which data have the highest variability.

Therefore, as a first exploratory step in the analysis of the available data, PCA is used for reduction purposes and for exploring their variability structure.

Actually, the data set is constituted of 24 functional variables, 8 for each detection plane (sagittal, frontal, transverse): RKANGLE, LKANGLE, RTRKANGLE, LTRKANGLE, RAANGLE, LAANGLE, RHANGLE, LHANGLE (for details about each variable, see Chapter 2).

For each variable, three complete gait cycles are available. For instance, considering RKANGLE_S (i.e. RKANGLE in the sagittal plane), Figures 19-22 represent the waveforms related to 4 different subjects belonging to the four groups (angles amplitude vs sample number). Each gait cycle has a different length. Actually, 459 detected angle values are represented in Figure 19 for subject 1 belonging to the first group, 225 for subject 8 belonging to group II, 114 for subject 43 belonging to group III and, finally, 111 for subject 91 belonging to group IV. Since the sampling frequency is 100Hz, the duration in the time domain is different for the four subjects.

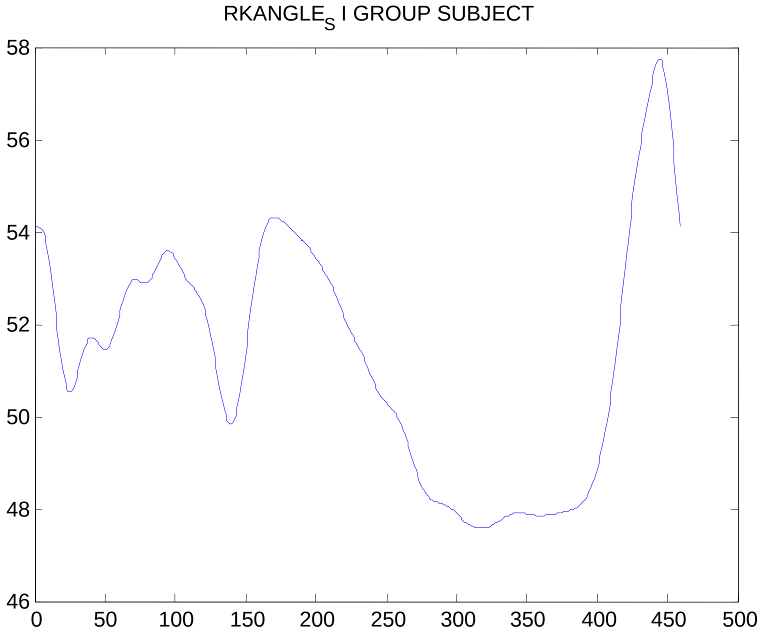


Figure 19: RKANGLE_S in a gait cycle for subject 1(I group)

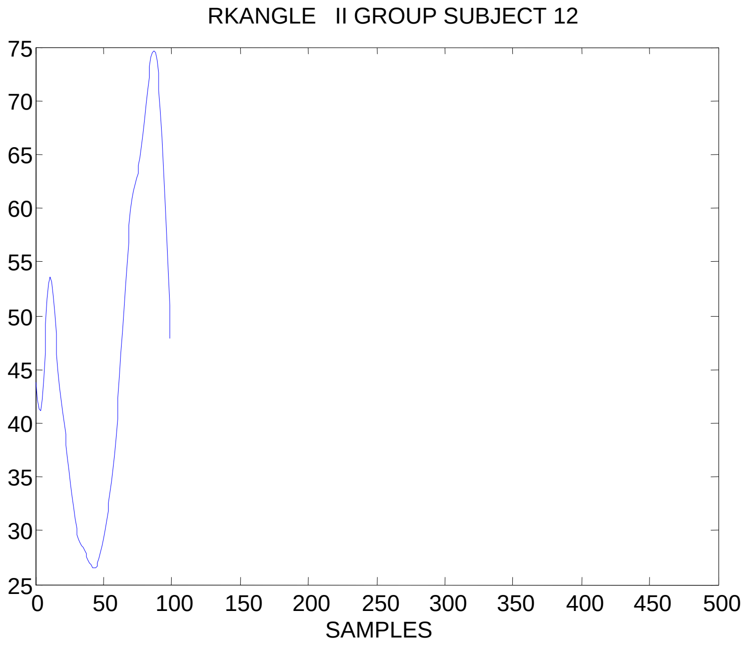


Figure 20: RKANGLE_S in a gait cycle for subject 12 (II group)

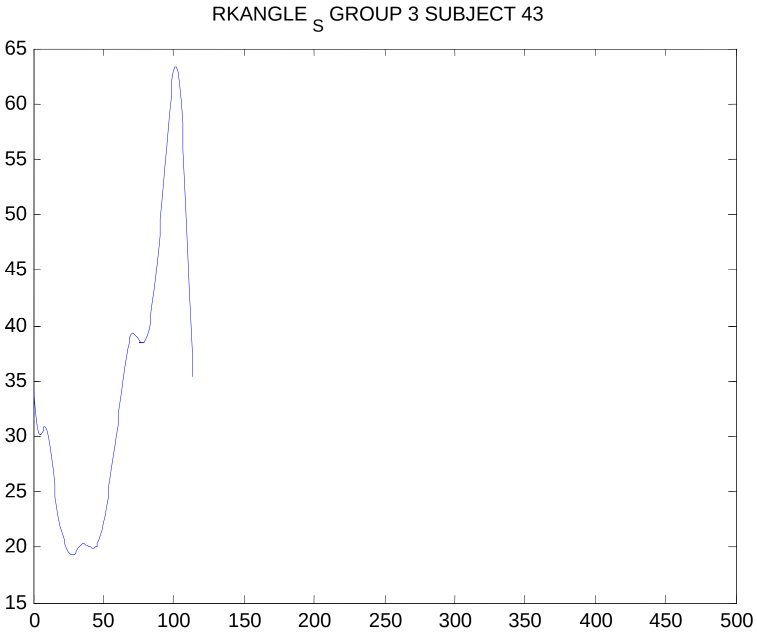


Figure 21: RKANGLE_S in a gait cycle for subject 43 (III group)

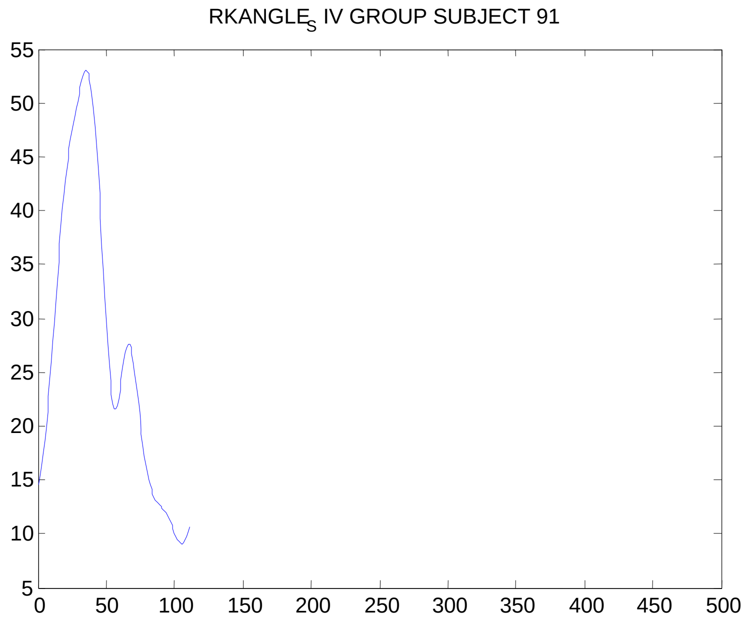


Figure 22: RKANGLE_S in a gait cycle for subject 91 (IV group)

In order to apply PCA to the functional variables introduced above, some preliminary precautions need to be taken.

Firstly, in a functional PCA, variables X_i are referred to the individual samples of the waveform (in this context, time samples). Considering that waveform data have a different length, it is necessary to normalize them and to perform a sampling at each 1% from 1% to 100%. This would correspond to the generation of an *n* x 100 data matrix for each variable where *n* represents the number of subjects. When data are not smooth, it is previously necessary to interpolate them. This can be accomplished, for example, by using cubic splines and then sampling the resulting waveform. However, in the case under study, considering that the electronic acquisition system works at the frequency of 100Hz, data are quite smooth and the spline has been considered as unnecessary.

Secondly, the way of variation of each single curve needs to be considered. Generally, curves vary in two ways, vertically and horizontally. Vertical variations give information about the amplitude of the curve, while the horizontal ones represent the phase. In order to compare single gait cycles among different subjects, curves have to be in phase. For the employed registration system in LAMBDA Laboratory, curves related to different gait cycles are supplied in phase and then, no additional analysis is needed. In other circumstances, it would have been necessary to rescale each curve to a common standard interval by, for example, the employment of a *time warping function* [35]. The previous waveforms represented in Figures 19-22, are reported in Figures 23-26 after normalization.

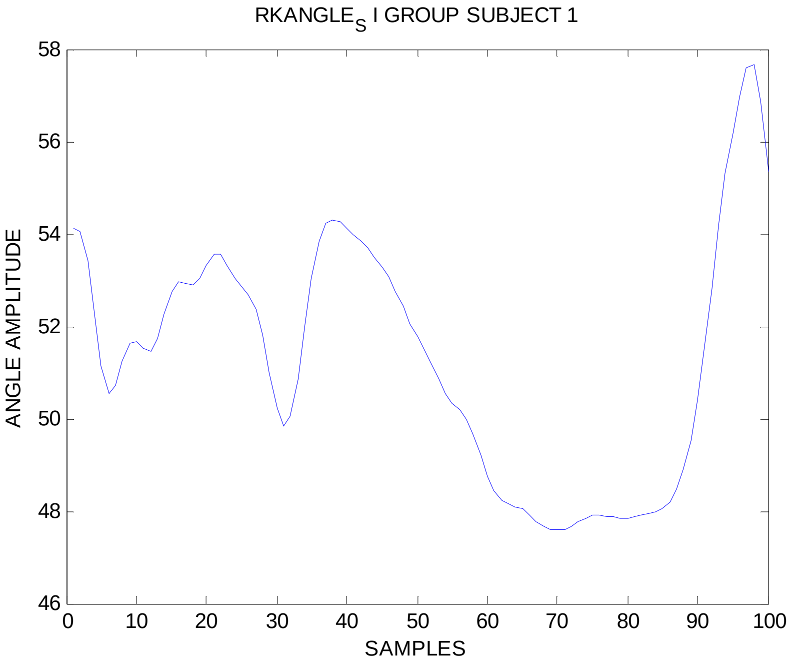


Figure 23: RKANGLE_S in a gait cycle for subject 1 after normalization (I group)

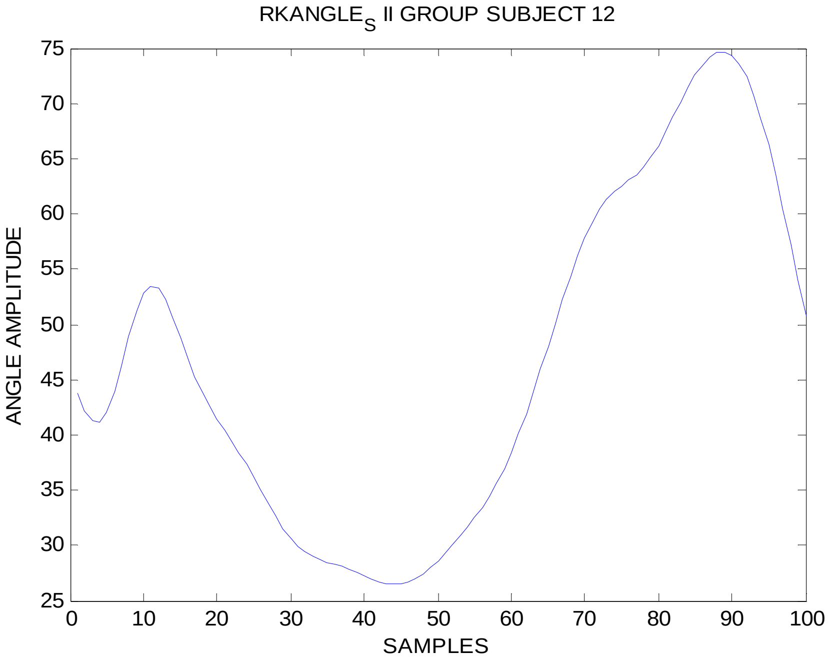


Figure 24: RKANGLE_S in a gait cycle for subject 12 after normalization (II group)

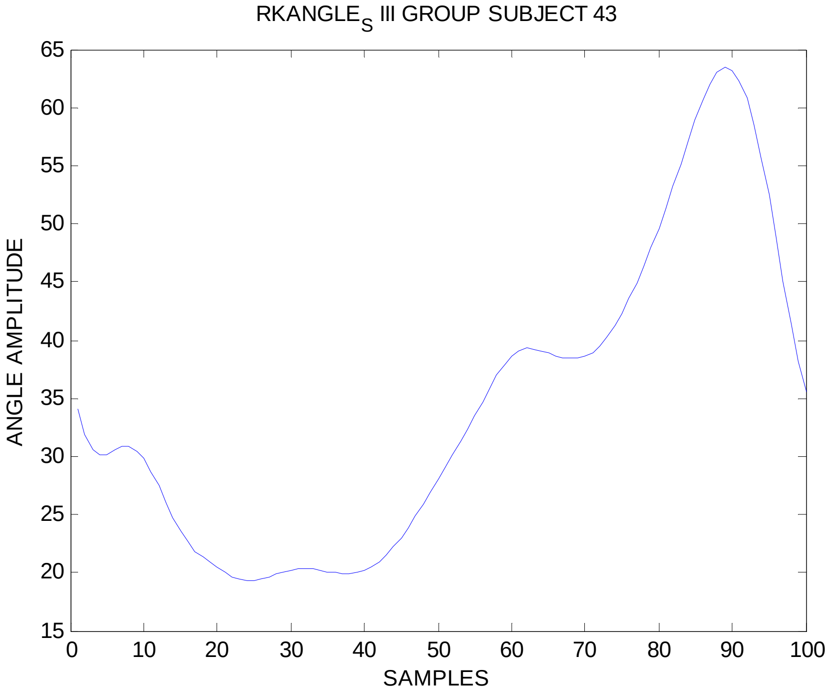


Figure 25: RKANGLE_S in a gait cycle for subject 43 after normalization (III group)

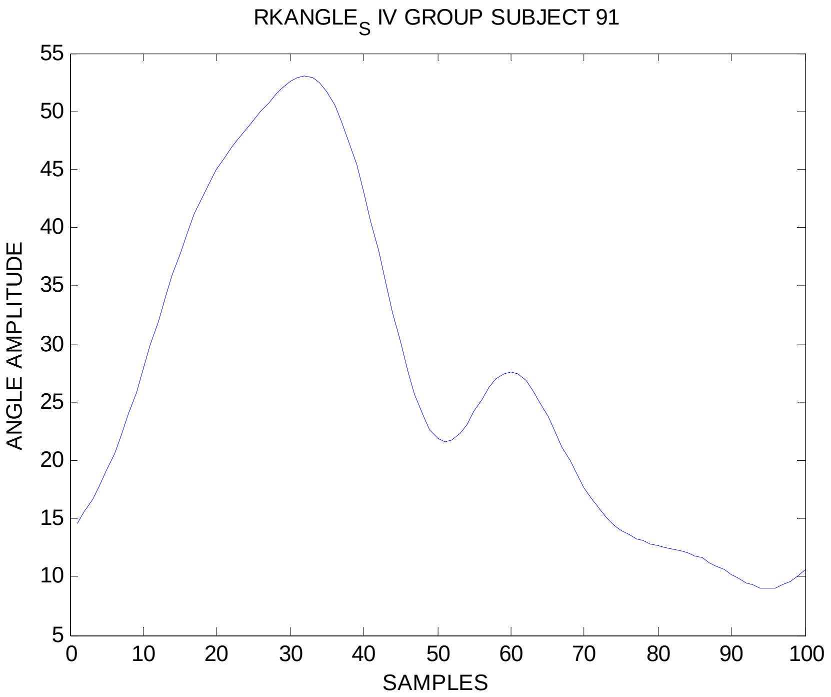


Figure 26: RKANGLE_S in a gait cycle for subject 91 after normalization (IV group)

4.10 Functional Principal Component Analysis

Now, for each of the 24 functional variables, 91x100 data matrices are considered to perform the FPCA. In the PC model $Z=U^TX$, elements of matrix U, i.e. the *principal component loading vectors*, represent the orthogonal basis set for the waveform data while, the *principal component score vectors* (PCscores, PCs) in Z are composed of the coefficients that measure the contribution of the principal component to each individual waveform. The original waveform data for each subject are then transformed into a set of PC scores that measure the degree to which the shape of their waveform corresponds to each feature. Since PC scores represent in synthesis the gait waveform data for each subject, it seems particularly attractive to employ them as discrimination features of the four different groups [45-50]. Actually, differently from the global predictors (ROM, RMS, CF) approach discussed in the previous sections, PC scores are more realistically representative of the waveform data.

4.10.1 *Selection of the number of principal components and analysis of group differences in the PC scores*

PCA is particularly attractive because it is able in capturing the greater part of data variation by few principal components. Different criteria can be used in order to select the number of components to be considered [43]. To analyze data involved in this study, only the first two principal components have been considered because they can generally capture more than the 90% of data variation. Other principal components could have been considered, but the smaller variance they could explain was hard to interpret. Therefore, they have been discarded.

4.10.2 Results

Can a combination of the first two principal component scores related to the 24 functional variables discriminate the four groups? To give an answer to this question, it is preliminary necessary to understand if the generated scores are statistically different for the four groups and then to build a discriminant model where they can be used as predictors. Therefore, in the next paragraphs the following results are presented: a) the normal univariate tests on the potential PC scores predictors, b) the group differences tests with respect to the generated PCs, c) the discriminant model.

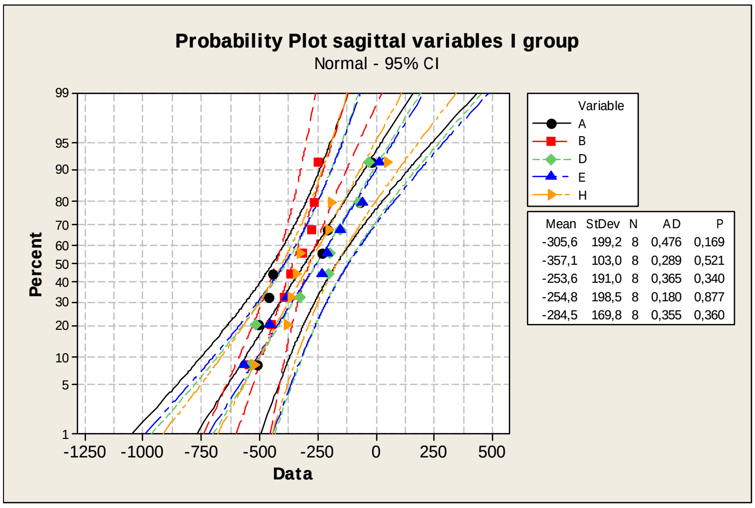
4.10.2.1 Univariate normality tests

Before verifying if inside each group data are multivariate normal distributed, normal probability plots (NPP) are built for all the 48 possible predictors. In the next Figures, NPPs are reported for some of those variables that will be selected (see next paragraph) for the discriminant model. Variables are coded as follows (Table 13):

SELECTED PREDICTORS	CODED NAMES
RKANGLE_S_PC1	A
RKANGLE_S_PC2	AA
RTRKANGLE_S_PC1	B

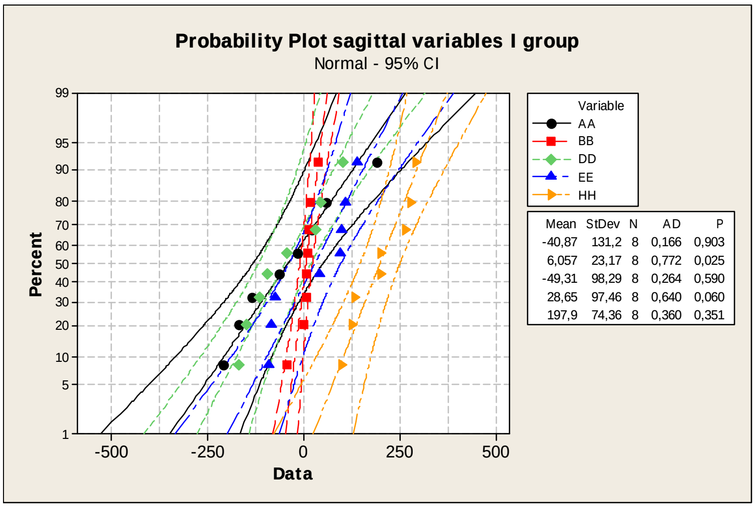
RTRKANGLE_S_PC2	BB
RHANGLE_S_PC1	D
RHANGLE_S_PC2	DD
LHANGLE_S_PC1	E
LHANGLE_S_PC2	EE
LKANGLE_S_PC1	H
LKANGLE_S_PC2	HH
RKANGLE_F_PC1	I
LKANGLE_F_PC1	J
RTRKANGLE_F_PC1	K
LAANGLE_F_PC1	N
RAANGLE_F_PC2	O
RHANGLE_T_PC1	MM

Table 13: coded names for some variables



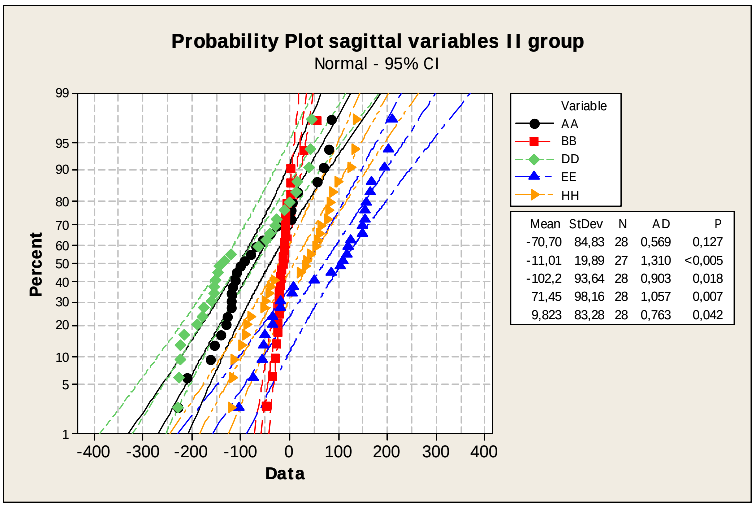
Mean	StDev	N	AD	P
-305,6	199,2	8	0,476	0,169
-357,1	103,0	8	0,289	0,521
-253,6	191,0	8	0,365	0,340
-254,8	198,5	8	0,180	0,877
-284,5	169,8	8	0,355	0,360

Figure 27: NPP for variables A,B,D,E,H related to group I



Mean	StDev	N	AD	P
-40,87	131,2	8	0,166	0,903
6,057	23,17	8	0,772	0,025
-49,31	98,29	8	0,264	0,590
28,65	97,46	8	0,640	0,060
197,9	74,36	8	0,360	0,351

Figure 28: NPP for variables AA, BB, DD, EE, HH related to group I



Mean	StDev	N	AD	P
-70,70	84,83	28	0,569	0,127
-11,01	19,89	27	1,310	<0,005
-102,2	93,64	28	0,903	0,018
71,45	98,16	28	1,057	0,007
9,823	83,28	28	0,763	0,042

Figure 29: NPP for variables AA, BB, DD, EE, HH related to group II

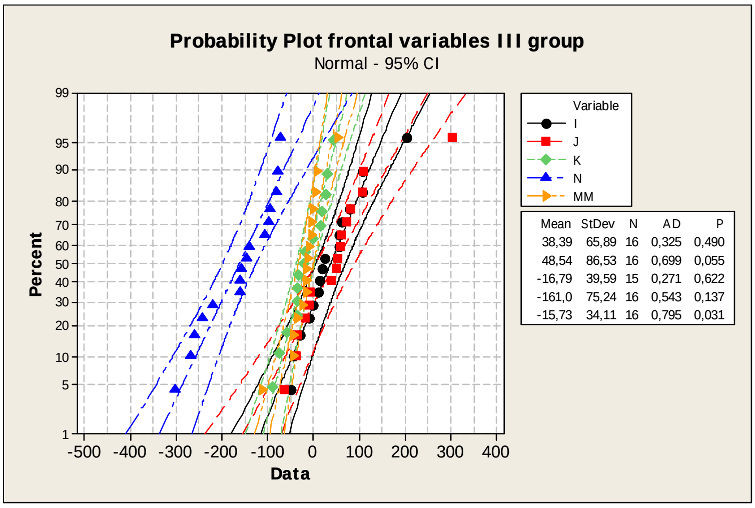


Figure 30: NPP for variables I, J K, N,MM related to group III

4.10.2.2 Group differences for the PC scores

Group differences with respect to the generated PC scores have been tested through the Kruskal Wallis test. Significant differences emerged. For instance, a *p*-value equal to 0.000 has been found for the variables RKANGLE_S_PC1,RKANGLE_S_PC2,RTRKANGLE_S_PC1,RTRKANGLE_S_PC2,RHANGLE_S_PC1,RHANGLE_S_PC2,LHANGLE_S_PC1,LHANGLE_S_PC2,LKANGLE_S_PC1, LKANGLE_S_PC2, LAANGLE_F_PC1 (p-value = 0.036). Then, the *post-hoc* Mann Whitney test with Bonferroni correction has been performed in order to understand which group differences have been found significant. In the following Table 14, *p*-values for some tests are reported. They are related to some of the 24 variables selected for the model during the stepwise discriminant procedure. Variables are reported in the first column, while the second one shows which groups are significantly different with respect to the PC scores.

Mann-Whitney test (Alpha Bonferroni=0,00833)		
PREDICTORS	**DIPLEGIA FORM**	***p*-value**
RKANGLE_S_PC1	I/IV	0,0031
RKANGLE_S_PC1	II/III	0,0001
RKANGLE_S_PC1	II/IV	0,0001

RKANGLE_S_PC2	I/III	0,0004
RKANGLE_S_PC2	I/IV	0,0000
RKANGLE_S_PC2	II/III	0,0000
RKANGLE_S_PC2	II/IV	0,0000
RTRKANGLE_S_PC1	I/III	0,0011
RTRKANGLE_S_PC1	I/IV	0,0000
RTRKANGLE_S_PC1	II/IV	0,0002
LAANGLE_F_PC1	III/IV	0,0037

Table 14 : MW post-hoc test for some variables

4.10.2.3 Results of the stepwise discriminant analysis

Considering that 8x3 functional variables were available and that only the first two principal components have been retained after performing PCA, 48 possible predictors have been included at the beginning in the discriminant model.

The stepwise discrimination procedure indicates that only 16 of the 48 available variables are significant for the discrimination of the four groups. Variables reported below have shown a significant discriminant power:

1. RKANGLE_S_PC1,
2. RKANGLE_S_PC2,
3. RTRKANGLE_S_PC1,
4. RTRKANGLE_S_PC2,
5. RHANGLE_S_PC1,
6. RHANGLE_S_PC2,
7. LHANGLE_S_PC1,
8. LHANGLE_S_PC2,
9. LKANGLE_S_PC1,
10. LKANGLE_S_PC2,
11. RKANGLE_F_PC1,
12. LKANGLE_F_PC1,
13. RTRKANGLE_F_PC1,
14. LAANGLE_F_PC1,
15. RAANGLE_F_PC2,

16. RHANGLE_T_PC1

Actually, employing the previous variables, the proportion of correctly classified observations into groups is 0.955. In particular, the proportion of correctly classified for group I is 100%, for group II is 96.3%, for group III is 80% and for group IV is 100%. However, since all the available data have been used both for the model and the validation, the accuracy of the discriminant model has been assessed by cross-validation technique. This has modified the results. For instance, after cross-validation the proportion of correctly classified observations into groups becomes 93.3%, while the proportion of correctly classified ones for each group is: 100% (I group), 92.6% (II group), 80% (III group), 97.4% (IV group). Cross-validation changes the proportions of correctly classified for the II and the IV group, while the III group is always 80%. The misclassification rate is 6.7% and represents a very satisfying result. The stepwise procedure gives the possibility to understand that PCs that can explain a large amount of variability are not necessarily important for group discrimination. For instance, RAANGLE_S _PC1 explains 80% of data variation, but is not relevant from a discrimination point of view. Then, it is not included in the model. In order to assess the *goodness-of-fit* of the model, Wilks'Λ criterion has been used. Wilks'Λ is here 0.01074. This result is very good and indicates that the final model represents a satisfying compromise between a good proportion of correctly classified observations (low error rate) and a Wilks'Λ as near as possible close to zero. Therefore, although the great number of detected variables to represent the gait cycle, the dimension of the gait is found much smaller and the discriminant power of the selected predictors very satisfying. Actually, with FPCA data reduction is based on features that are extracted from the entire gait waveform and not a-priori fixed by subjective considerations of the experts. This objectivity leads to more robust results.

In Table 15 the misclassified observations are shown. Two observations (17 and 54) are misclassified after cross-validation.

Misclassified observations			
OBSERVATION	TRUE GROUP	PREDICTED GROUP	CROSS VALIDATION
15	2	1	1
17	2	2	3
37	3	4	4
42	3	4	4
50	3	4	4
54	4	4	3

Table 15 : misclassified observations after cross-validation

To identify a new observation, the linear discriminant functions (see Table 16) associated with the four groups have been computed. The new observation belongs to the group whose discriminant function value is higher.

LINEAR DISCRIMINANT FUNCTIONS FOR GROUPS				
	I GROUP	II GROUP	III GROUP	IV GROUP
Constant	-21,3085	-15,2901	-30,7734	-19,3344
A	-0,0213	-0,0187	0,0077	-0,0043
AA	-0,0329	0,0034	-0,0781	-0,0693
B	-0,0693	-0,0276	-0,0494	-0,0455
BB	0,0636	-0,0182	0,0359	0,0099
D	-0,0065	-0,0021	-0,0241	0,0009
DD	0,0372	-0,0144	0,0312	0,0248
E	0,0085	-0,0111	0,0399	0,0118
EE	-0,0183	0,0183	-0,0706	-0,0574
H	0,0078	-0,0138	-0,0148	0,0115
HH	0,0501	-0,0038	0,0388	0,012
I	0,028	0,0093	0,0241	0,0204
J	-0,0042	0,002	-0,0006	-0,001
K	-0,0176	0,0005	-0,0376	-0,0147
N	-0,0404	-0,0199	-0,0565	-0,0358
O	-0,0029	0,0001	-0,0113	-0,0078
MM	0,0276	-0,0071	-0,001	0,0531

Table 16 : linear discriminant functions for groups

For each group the location of the points that represent the means for all the variables in the multivariate space is found. These points are called group *centroids*. Squared distance among groups are so determined as distances among *centroids*. In Table 17, distances among the four groups for the validated model are shown.

	SQUARED DISTANCE AMONG GROUPS			
	1	**2**	**3**	**4**
1	0,0000			
2	23,0249	0,0000		
3	45,5789	61,8867	0,0000	
4	45,9587	54,5877	15,7916	0,0000

Table 17: squared distance among groups

As underlined in the previous paragraphs, in order to assess if data follow a multivariate normal distribution inside each group, the same graphical procedure, used for the *static indicators* approach, has been applied. The following Figures 31 and 32 represent the Chi-square plots for group II and IV. Plots follow almost a linear pattern. Since no systematic curved pattern is visible, the assumption of multivariate normality can be accepted.

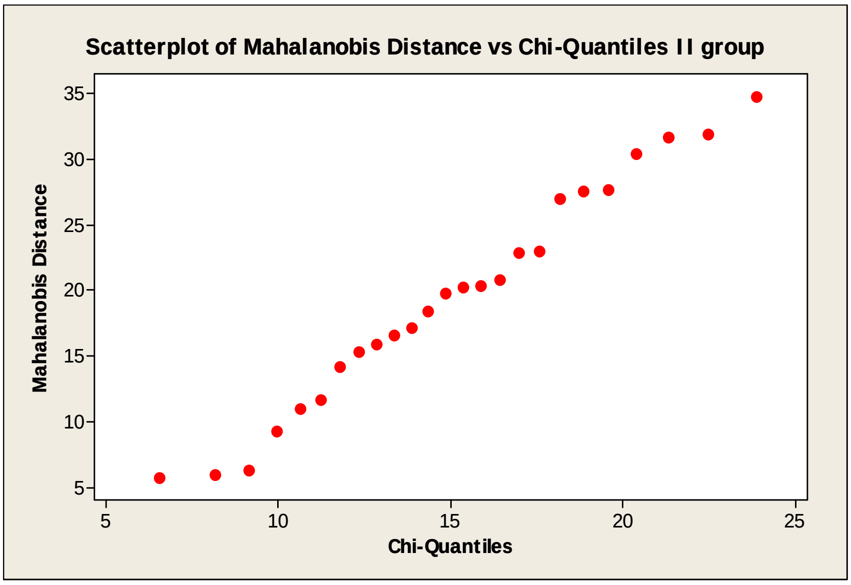


Figure 31: MD vs Chi-Quantiles group II

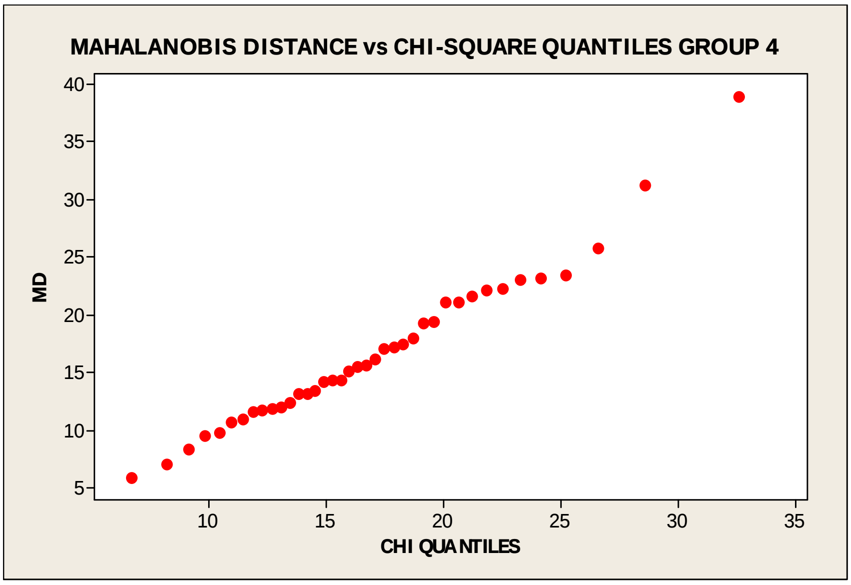


Figure 32: MD vs Chi-Quantiles group IV

Chapter 5

Clinical feedback and future developments

5.1 Clinical interpretation of the discriminant model

In the previous Chapter 4 two methodologies have been presented for the discrimination of four groups. *Static indicators* approach has proved to be less effective than *PCA approach* for functional data. Actually, the proportion of correctly classified observations, after *cross-validation* for both the methodologies, passes from 70.5% to 93.3%. With relation to the latter approach, it is particularly interesting to understand which is the clinical interpretation and significance of the variables selected as predictors for the linear discrimination model. These predictors are represented by the PC scores of the first two principal components of the functional variables measured in the LAMBDA laboratory. For example, the RKANGLE_S variable has been introduced into the model through the individual scores of its first two principal components (RKANGLE_S_PC1, RKANGLE_S_PC2). Without deepening the potential statistical significance of the coefficients of the first two principal components, which will be opportunely considered for future analysis, the variables used in the model have been examined by medical experts.

They were asked to give a score from 1 to 5 (Min = 1, Max = 5) to the selected variables in order to "weight" their clinical interest. These predictors have been found to match with the observed variables that the experts take into account to classify the diplegia form. Table 18 lists the variables and the relative scores attributed.

Model Variables	Score
RKANGLE_S	4
RTRKANGLE_S	5
RHANGLE_S	4
LHANGLE_S	4
LKANGLE_S	4
RKANGLE_F	3
LKANGLE_F	3
RTRKANGLE_F	5
LAANGLE_F	4
RAANGLE_F	4
RHANGLE_T	4

Table 18: score attributed to the selected variables by experts

It seems extremely relevant (score 5) from a clinical point of view the presence of RTRKANGLE_S and RTRKANGLE_F, both representing the pendulum of the trunk on the sagittal and the frontal planes during a gait cycle. In particular, the trunk pendulum on the sagittal plane is typical of subjects belonging to the first and the fourth groups, while the rotation (oscillation) on the frontal plane is the main characteristic of those subjects belonging to the third group. Score 4 is attributed to the remaining variables with the exception of two variables measuring the angular excursion of the right and left knees on the frontal plane, i.e. RKANGLE_F and LKANGLE_F (score 3). In particular, these two variables are representative of the main characteristic of subjects belonging to the second group, called “tight skirt” patients. Actually, during the gait cycle, they exhibit knees with a valgus deformity in the frontal plane.

A variable representative of the angular rotation of a monitored joint may be significant both for the left and right side. This happens for RKANGLE_F and LKANGLE_F or for RHANGLE_S and LHANGLE_S and is clinically relevant. In fact, the movement of patients is strongly asymmetrical. Therefore, the information content of the two variables is different and both can be significant for the discriminant model.

The three detection planes are not homogenous. Actually, variables related to the transverse plane do not exhibit a great interest for discriminating purposes. Only RHANGLE_T has been found to be significant for the classification of groups. This result is coherent with clinical considerations. Actually, the sagittal plane is the one in which most of the joint rotations occur. If joints are considered as hinges, the range of rotations between the segments foot-leg or leg-thigh or thigh-pelvis is maximum in this plane.

The presence of the variables RAANGLE_F and LAANGLE_F in the model can be partially explained because of an intra-rotation of the lower limbs typical of subjects belonging to the third and fourth groups.

Two other general questions have been proposed to the clinical experts. The first concerns the introduction of the *gait cycle meantime* (*gcm*) variable in the model related to the *static indicators* approach. Patients belonging to the first group are significantly slower than those belonging to the fourth one. This result is coherent with their clinical profile. Actually, each form of diplegia express not only a different organization of the motion, to be correlated to the different degrees of severity of the disease, but also to the different efficiency of the movements that result slower and more difficult for patients of the first group and faster and more similar to the motion of asymptomatic children for patients belonging to the fourth one.

The second general question is related to the correct classification rate for subjects belonging to the third group. With the FPCA approach this rate is 80% before and after *cross-validation*. On the contrary, the fourth group classification rate after *cross- validation* is 97.4%. Experts generally agree with the difficulty to discriminate subjects of the third group from those of the fourth one. From a statistical point of view, this might be of great interest for future developments. Some general considerations on these future analyses are presented in the next paragraph.

5.2 Future developments

Although the classification rate of the third group reaches the 80%, it seems appropriate to understand why it is far from the classification rate of the other groups, more closed to 100%. To distinguish a first or a second diplegia form from the fourth one is simpler than to distinguish the third from the fourth. The increased severity of the first form makes less difficult the identification of the first group by the experienced medical staff.

The identified model is able to give a greater objectivity to the classification. Actually, the selected predictors are related to the measurements of the angles of some joint rotations performed in the LAMBDA laboratory and processed by Nexus software. The classification rate of 93.3% is a good confirmation of the work done in the clinics by the medical staff on an observational basis. However, what would change if subjects of the third group, classified with the highest error rate, were included in the second or in the fouth group?

By following the FPCA approach, the following preliminary results have been obtained. Including subjects of the third group into the fourth one, the proportion of correctly classified observations after cross-validation becomes 98.9% with a Wilks' lambda equal to 0,03704. In particular, the classification rate for the first group is 100%, for the second one 96.3% and for the new third group, now constituted by 54 subjects, is 100% . Selected predictors are A, AA, B, BB, D, DD, E, EE, H, HH, J according to the coded names of Table 4.2 (see Chapter 4).

Even if such a result offers a lower error rate, it could be not meaningful from a medical point of view. Actually, the separation of the third group from the fourth one could be clinically necessary and the consequent adoption of different therapeutic protocols could be more effective from a therapeutic point of view. Nevertheless, considering the available data employed for this research work, the performance of a three group model is found better. Such a situation deserves to be adequately deepened and correctly interpreted.

Finally, the FPCA approach has taken into account the first two principal components of the available functional variables because they could explain more than 90% of the data variability. Since the coefficients of these principal components are available, it could be interesting to understand if they can be clinically interpreted

CONCLUSIONS

Cerebral Palsy (CP) is a lesion of the central nervous system that alters the motion functions. The main objective of this study was the identification of some *indicators* derived from functional data (obtained by 3D Gait Analysis) to be employed for the discrimination of four forms of *diplegia* according to the classification proposed by Ferrari et al. Two different approaches have been proposed.

In a first moment, *"static" indicators* have been extracted from the waveforms representing the angles of rotation of the main joints of the lower limbs, collected for 91 subjects during three gait cycles. Data are referred to three anatomical planes: sagittal, frontal, transverse. Range of motion (ROM), root mean square (RMS) and crest factor (CF) have been considered for their immediate clinical interpretability. Actually, ROM reflects the different maximum angular excursion of the main joints of the lower limbs depending on the degree of severity of diplegia as well as RMS the variation of the waveform with respect to a constant value and CF any impulsive phenomena characterizing the pendular movement typical of almost all the four forms of *diplegia*. The discrimination procedure indicates that only 11 of the 72 available variables can be effectively used for the discrimination purposes. Actually, most of the variables have been discarded because of their high correlation or because of their lower discriminant power. The classification rate of this model is not very satisfying. Actually, after *cross-correlation*, it is 70.5 %. The limit of this approach is probably due to the high subjectivity in the selection of the potential predictors that reflect more the clinical specialists' points of view than the intrinsic structure of the data.

In order to overcome these limits and to improve the classification rate, a functional approach has been proposed. Multivariate statistical analysis has proved to be a powerful tool to eliminate collinearity and to facilitate the analysis of data. In particular, Functional Principal Component Analysis (FPCA) has been proposed for the analysis of the available data in the attempt of identifying a limited number of components that can explain most of data variability and the *indicators* to be used for discrimination purposes.

PC scores related to the first two principal components (that can capture more than 90% of data variability) have been tested for differences among the four groups and employed as predictors in the discrimination model. The stepwise discrimination procedure indicates that only 16 of the 48 have a significant discrimination power. Actually, after *cross-validation*, the classification rate of the model is 93.3% . Wilks' criterion was used to assess the *goodness-of-fit* for the model. Wilks'Λ was 0.01074.

Although the great number of detected variables to represent the dynamic of gait cycle, the dimension of the gait is found to be much smaller and the selected variables are clinically meaningful. Then, principal component modeling of the gait waveforms is found to be a more effective technique with respect to the first proposed approach. PCA method is more objective and robust than the previous approach, because data reduction is based on features that are extracted from the entire gait cycle waveform. The subjective choice of indicators is simple but not necessarily able in reflecting adequately the variability of the curves. The proposed FPCA methodology could be used as a tool to help specialists in classifying the *diplegia* forms, especially when different opinions occur.

Bibliography

[1] F. Stanley, E Blair, E. Alberman, *Cerebral Palsies: epidemiology and causal pathways*, Books google.com

[2] Bax, et Al., *Proposed definition and classification of cerebral palsy*, Developmental Medicine & Child Neurology Volume / Issue 08 / August 2005, pp 571-576

[3] Ferrari, Alboresi, Pascale, Perazza, *The term Diplegia should be enhanced, Part I,* EURJ PHYS REHABIL 44:195-201, 2008

[4] Cioni, Lodesani, Pascale et Al., *The term Diplegia should be enhanced, Part II,* EURJ PHYS REHABIL 44:203-11, 2008

[5] Pascale, Perazza, Ferrari et Al., *The term Diplegia should be enhanced, Part III*, EURJ PHYS REHABIL MED 2008 44:213, 20, 2008

[6] Christine Cans, H.Dolk , Mj Platt, A. Colver, A. Prasauskiene, Rageloh-Mann, *Recommendations from the SCPE collaborative group for defining and classifying cerebral palsy*, Article first published online: 23 JUN 2009 DOI: 10.1111/j.1469-8749.2007

[7] C-Morris, *Definition and classification of Cerebral palsy,: a historical perspective*, Developmental Medicine & Child Neurology, 2007 - Wiley Online Library

[8] A.F. Colver et Al, , The term Diplegia should be abandoned, *Arch Dis Child 2003;88:286-290 doi:10.1136/adc.88.4.286*

[9] Ferrari A., et al., *Le forme spastiche della paralisi cerebrale infantile* , Springer (2005)

[10] Perry J., *Analisi del movimento,* Elsevier, 2005

[11] Leardini A., et al., *A new anatomically based protocol for gait analysis in children,* Gait Posture (2007), doi: 10.1016/j.gaitpost.2006.12.018

[12] Michael W.Whittle, *Gait analysis, an introduction*, Butterworth Heinemann Elsevier Fourth edition, 2007

[13] Giannini S,et al, *Gait Analysis: methodologies and clinical applications*, IOS press for BTS, 1994

[14] Andriacchi T. et al., *Knee Joint Kinematics during Walking Influences the Spatial Cartilage Thickness Distribution in the Knee*, J Biomech, Apr 29, 2011; 44(7):1405-1409

[15] Carrie Stackhouse, Patricia A. Shewokis, et Al., *Gait initiation in children with cerebral palsy*, Gait & Posture Volume 26, Issue 2, Pages 301–308, July 2007

[16] Chang-Soo Yang, Gye-San Lee, Bum-Kwon Choi, David O'Sullivan, Young-Hoo Kwon, and Bee-Oh Lim, *Gait analysis in children with autism using temporal-spatial and foot pressure variables*, 30th Annual Conference of Biomechanics in Sports – Melbourne 2012

[17] Kelly A. McKean, Scott C. Landry, Cheryl L. Hubley-Kozey, Michael J. Dunbar, William D. Stanish, Kevin J. Deluzio, *Gender differences exist in osteoarthritic gait*, Clinical Biomechanics 22 (2007) 400–409

[18] D.H. Sutherland, *The evolution of clinical gait analysis part III – kinetics and energy assessment*, Gait & Posture Volume 21, Issue 4, Pages 447–461, June 2005

[19] Maathuis, Karel G. B, van der Schans, Cees P, Van Iperen, Andries, Rietman, Hans S, Geertzen, Jan H., *Gait in Children With Cerebral Palsy: Observer Reliability of Physician Rating Scale and Edinburgh Visual Gait Analysis Interval Testing Scale*, Journal of Pediatric Orthopaedics: May/June 2005 Volume 25 - Issue 3 - pp 268-272

[20] Kirtley C., *Clincal Gait Analysis: Theory and Practice*, Elsevier Health Sciences, 2006

[21] Fiona Dobson, Meg E. Morris, Richard Baker, H. Kerr Graham, *Gait classification in children with cerebral palsy*: *A systematic review*, Gait & Posture (2006)

[22] TF Winters, JR Gage and R Hicks, *Gait patterns in spastic hemiplegia in children and young adults*, J. Bone Joint Surg. Am. 69:437-441, 1987.

[23] Sebastian I. Wolf, Frank Braatz, Dimitrios Metaxiotis, Petra Armbrust, Thomas Dreher, Leonhard Doderlein , Ralf Mikut, *Gait analysis may help to distinguish hereditary spastic paraplegia from cerebral palsy,* Gait & Posture 33 (2011) 556–561

[24] K Jordan, JH Challis, KM Newell, *Walking speed influences on gait cycle variability*, Gait & posture, vol 26 issue 1 June 2007, pp. 128-134, Elsevier

[25] Rolf Moe-Nilssen, Jorunn L. Helbostad, *Estimation of gait cycle characteristics by trunk accelerometry*, Journal of Biomechanics, vol. 37 issue 1, January 2004, Pages 121–126

[26] Geoffrey McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, 2004

[27] F Hair, RL Tatham, RE Anderson, W Black – 2006, *Multivariate data analysis*, Pearson Prentice Hall 2006

[28]Shahla Ramzan, Faisal Maqbool Zahid and Shumila Ramzan *Evaluating Multivariate Normality: A Graphical Approach* Middle-East Journal of Scientific Research 13 (2): 254-263, 2013

[29] Ramsay,J. O.& Dalzell, C. J., *Some tools for functional data analysis (with discussion),* Journal of the Royal Statistical Society: Series B53(3), 539-572,1991

[30] Ramsay, J. O. & Silverman, B. W., *Functional Data Analysis*, 2nd ed, Springer, New York, 2005

[31] Ramsay, J. O., *Functional components of variation in handwriting,* Journal of the American Statistical Association 95 (449), 9-15,2000

[32] Wang, S., Jank, W., Shmueli, G. & Smith, P., *Modeling price dynamics in eBay auctions using differential equations*, Journal of the American Statistical Association 103(483), 1100-1118,2008

[33] Meiring, W., *Oscillations and time trends in stratospheric ozone levels: a functional data analysis approach'*, Journal of the American Statistical Association 102 (479), 788-802,2007

[34] Erbas, B., Hyndman, R. J. & Gertig, D. M., *Forecasting age-specific breast cancer mortality using functional data models*, Statistics in Medicine 26(2), 458-470, 2007

[35] Ramsay, J. O. & Silverman, B. W., *Applied Functional Data Analysis: Methods and Case Studies*, Springer, New York, 2002

[36] Ferraty, F. & Vieu, P., *Nonparametric Functional Data Analysis: Theory and Practice*, Springer, New York, 2006

[37] Johnson, R.A. and Wichern, D. W. *Applied Multivariate Statistical Analysis*. Fifth edition, New Jersey: Prentice Hall, 2002

[38] Leen Van Gestel , Tinne De Laet, Enrico Di Lello, Herman Bruyninckx, Guy Molenaers, Anja Van Campenhout, Erwin Aertbelie, Mike Schwartz, Hans Wambacq, Paul De Cock, Kaat Desloovere, *Probabilistic gait classification in children with cerebral palsy: A Bayesian approach*, Research in Developmental Disabilities 32 (2011) 2542–2552

[39] Nilsson N. J., *Introduction to Machine Learning*, Artificial Intelligence Laboratory, Department of Computer Science, Stanford University, 2005, available on http://ai.stanford.edu/~nilsson/mlbook.html

[40] Bishop J C. M., *Neural networks for pattern recognition* , Oxford University Press, 2004

[41] Bai-ling Zhanga, Yanchun Zhang, Rezaul K.Begg, *Gait classification in children with cerebral palsy by Bayesian approach*, Pattern Recognition 42 (2009) 581-586

[42]Sebastian Wolf, Tobias Loose, Matthias Schablowski, Leonhard Doderlein, Rudiger Rupp, Hans Jurgen Gerner, Georg Bretthauer, Ralf Mikut, *Automated feature assessment in instrumented gait analysis* Gait & Posture 23 (2006) 331–338

[43] R. Berrendero, A. Justel , M. Svarc, *Principal components for multivariate functional data*, Computational Statistics and Data Analysis 55 (2011) 2619–2634

[44] Shahid Ullah and Caroline F Finch, *Applications of functional data analysis: A systematic review* Ullah and Finch BMC Medical Research Methodology 2013, 13-43 http://www.biomedcentral.com/1471-2288/13/43

[45] Alessandra Carriero, Amy Zavatsky, Julie Stebbins, Tim Theologis, Sandra J. Shefelbine, *Determination of gait patterns in children with spastic diplegic cerebral palsy using principal components,* Gait & Posture 29 (2009) 71–75

[46] K.J. Deluzio, J.L. Astephen, *Biomechanical features of gait waveform data associated with knee osteoarthritis. An application of principal component analysis*, Gait & Posture 25 (2007) 86–93

[47] Brigitte Toro, Christopher J. Nester, Pauline C. Farren, *Cluster analysis for the extraction of sagittal gait patterns in children with cerebral palsy*, Gait & Posture 25 (2007) 157–165

[48] D.J. Rutherford, C.L. Hubley-Kozey, W.D. Stanish, *The neuromuscular demands of altering foot progression angle during gait in asymptomatic individuals and those with knee osteoarthritis* Osteoarthritis and Cartilage 18 (2010) 654-661

[49] Kevin J. Deluzio, Urs P. Wyss, Benny Zee, Patrick A. Costigan,Charles Sorbie, *Principal component models of knee kinematics and kinetics: Normal vs. pathological gait patterns*, Human Movement Science 16 (1997) 201-217

[50] Sebastian I. Wolf, Frank Braatz, Dimitrios Metaxiotis, Petra Armbrust, Thomas Dreher, Leonhard Dederlein, Ralf Mikut, *Gait analysis may help to distinguish hereditary spastic paraplegia from cerebral palsy* Gait & Posture 33 (2011) 556–561