TESI DI DOTTORATO

Dipartimento di Scienze Economiche Aziendali e Statistiche

The use of diagnostic tools in GAMLSS for liver fibrosis detection

Andrea Marletta

Tutor: **Prof.ssa Mariangela Sciandra**

Coordinatore Dottorato: **Prof. Marcello Chiodi**

Dottorato di Ricerca in "Statistica, Statistica Applicata e Finanza Quantitativa", XXVI Ciclo
Settore Scientifico Disciplinare: SECS/S01 - Statistica
Anno conseguimento titolo: 2016

Università degli Studi di Palermo

DSEAS

Dipartimento di Scienze Economiche,
Aziendali e Statistiche

Acknowledgements

I want to thank my three supervisors during my PhD course, in particular Prof. Massimo Attanasio and Prof. Marcello Chiodi for introducing me to the research project and guaranteeing the scholarship during these three years; ”my tutor” Mariangela Sciandra because she was the one that more believed in this project. A special thanks goes to Prof. Mikis Stasinopoulos and Prof. Robert Rigby because they helped me a lot in applying GAMLSS and they allowed me to spend a beautiful experience during my visiting period at London Metropolitan University.

Many thanks to Dr. Leonardo Rizzo, Ultrasuoni Srl, for collecting data and giving me the possibility to analyze them and all the Department staff, in particular, Prof. Gianfranco Lovison for the interest showed in the topic.

Thanks to all the colleagues and students that visited our PhD room in these years for having a nice time during our working hours, in particular Federico and Salvo because we started our academic career in 2004 and after a long time we are already here together. Thanks to Antonio, Gianluca and Marianna for the *R* support, Renato for the wise advices, Davide for the computer support and Luiz for helping me to live in London.

Last but not least, a very special thanks goes to Roberta and my family because even if they don’t know anything about GAMLSS or liver fibrosis, every time I needed, they were there for supporting me.

Contents

1 Introduction **1**

2 Liver fibrosis detection **5**

2.1 Background 5

2.2 New biomedical technology 8

3 GAMLSS **11**

3.1 Definition 11

3.2 Inference in GAMLSS 13

3.3 Worm plot 15

3.4 Open problems in GAMLSS 18

4 Dataset description **19**

4.1 Introduction 19

4.1.1 Exam variables 20

4.2 Data hierarchical structure 22

4.3 Some descriptive statistics 24

4.4 Changes in measuring from 2010 to 2013 26

5 Analysis **29**

5.1 Statistical models for high variability data 29
5.2 Focusing the attention on 2013 31
5.3 Selection of family distribution in GAMLSS 32
5.4 The BCPE Distribution 33
5.5 Model selection procedure 35
5.6 Use of other criteria for model selection 37
5.7 Inclusion of a random effect component 38
5.8 First solutions for M-shape worm plots 44
5.9 Comparisons with other models 46

6 Mixture models in GAMLSS 47

6.1 Overdispersion and mixture models 47
6.2 Simulation studies 55

7 ROC curve in GAMLSS 59

7.1 The ROC Curve 59
7.2 ROC Curve in GAMLSS 62

8 Conclusions 69

8.1 Summary and conclusions 69
8.2 Future work 71

Bibliography 77

Chapter 1

Introduction

The study of liver fibrosis involved several times, over the last years, the Department of Scienze Economiche, Aziendali e Statistiche of the University of Palermo thanks to the great amount of data collected in collaboration with the Ultrasuoni Srl company. A lot of studies were carried out focusing attention on the development of newer tools to detect liver fibrosis and a grant has been assigned to a research project involving ”New applications in biomedical industry”.
These previous works represent a good starting point for study because we have a lot of information about how to manage these particular kind of data. We already know that ARFI (Acoustic Radio Force Impulse) is a not-invasive tool to detect and classifies liver fibrosis measuring the stiffness of the liver tissue. This disease has a clear feature, it affects the liver tissue patchly, that is it can damage just some liver segments. For this reason, the easiest diagnoses are related to extreme cases in which all parts of liver are affected or healthy. Unfortunately, most of cases are intermediate providing different ARFI measures at different liver parts making more difficult

to derive a right diagnosis. From a statistical point of view this could be translated in asking more than one measurements and using some location measures, like for example the mean or the median.

We could overcome this problem by analyzing variability of the stiffness in terms of variance or other statistical parameters. In order to achieve this goal, we need a statistical tool, in particular a class of statistical models able to estimate a relationship between the stiffness and some predictors taking heterogeneity of the data into account. Besides, this class of models should be able to implement model with random effects since measurements are repeated during exam for the same subject.

We identified in GAMLSS (Generalized Additive Models for Location Scale and Shape) the best candidate among several class of statistical models to manage our data. Indeed, they are able to estimate jointly mean, variance skewness and kurtosis for a response variable as sum of linear and non-linear functions of some explanatory variables.

Starting from their standard definition two further extensions will be proposed in this work. Firstly, the use of a mixture model approach will follow from the analysis of residuals. Using this method a direct relationship between the components of the mixture and the health condition of the patient will be identified. Secondly, we will implement ROC (Receiver Operating Characteristic) curve in GAMLSS in order to show how ARFI can be also considered as a good tool for predicting liver fibrosis or cirrhosis.

The thesis is organized as follows. In Chapter 2 liver fibrosis will be introduced and a brief review of the tools used until now to detect liver diseases will be described. In Chapter 3 Generalized Additive Models for Location Scale and Shape will be presented. In Chapter 4 the framework of the full dataset and some preliminary descriptive statistics will be shown. Chapter 5 will contain a deeper analysis on the relationship between response variable and predictors applying a GAMLSS to a reduced dataset. Chapter 6 will propose the use of a mixture model approach in GAMLSS to deal with the problem of bimodality in the response distribution. In Chapter 7 the proposal to implement the ROC Curve in GAMLSS will be introduced in order to make predictions for liver fibrosis, while Chapter 8 will be devoted to discussion and future work. All the analysis in this thesis are implemented using the *R* statistical environment.

Chapter 2

Liver fibrosis detection

2.1 Background

Fibrosis is a disease which can affect the liver and culminates in cirrhosis, representing one of the ten most frequent causes of death in the world. It consists in the massive presence of connective tissue around portal areas and central veins causing non-functioning of the liver. Liver fibrosis is an asymptomatic and degenerative disease, it can be classified in 5 stages through the Metavir scoring system from F0 (normal liver) to F4 (cirrhosis) as in figure 2.1.

Scoring system Metavir is obtained as follows:

- F0 → Normal liver → No fibrosis surrounds the portal triads
- F1 → Portal fibrosis → Fibrous connective tissue is present but in limited areas
- F2 → Moderate fibrosis → Fibers begin to extend without connecting portal areas

- F3 → Severe fibrosis → Fibrous connective tissue links neighboring portal triads
- F4 → Cirrhosis → Most portal areas are connected by fibrous tissue also linking portal areas and central veins.

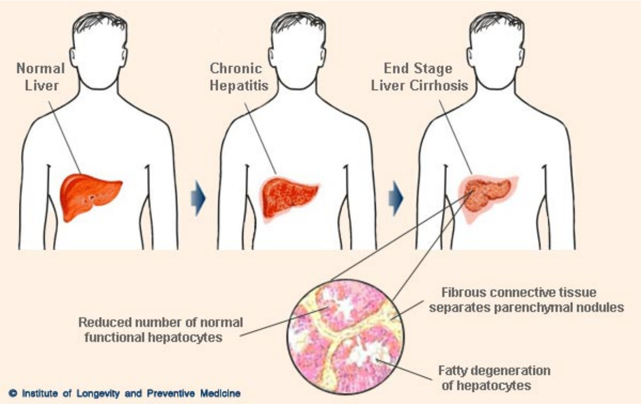
Normal Liver
Chronic Hepatitis
End Stage Liver Cirrhosis
Fibrous connective tissue separates parenchymal nodules
Reduced number of normal functional hepatocytes
Fatty degeneration of hepatocytes
© Institute of Longevity and Preventive Medicine

Figure 2.1: From fibrosis to cirrhosis

In medicine, biopsy is the most used exam to detect presence of liver diseases but it has both positive and negative aspects. Liver biopsy represents the gold standard test for staging liver disease. It is very useful in situations of uncertainty in diagnosis and it represents the best way to assess the possibility of rejection after liver transplant. Despite these positive aspects, biopsy has also some negative aspects, infact it is an invasive test which rarely presents possibility of complications for patients; in many situations it could be not predictive because fibrosis is a disease which affects the liver patchly and since only a very small part of the liver is involved in the biopsy, it could not able to find the sick part. Besides, it does not provide stable results in patients with nonalcholic fatty liver disease (Ratziu *et al.*, 2005). For these reasons biopsy was described as a tool ”far from an ideal test and liver diseases can be diagnosed precisely with laboratory tests and imaging studies” (Carey and Carey, 2010).

In order to overcome these problems it is necessary to take into account some alternative ways to detect liver fibrosis. During last years a lot of tools were proposed to substitute liver biopsy, but not satisfactory results were produced.

Even if liver fibrosis is an asymptomatic disease, most of the times in late stages it displays some clinical manifestations such as ascites and splenomegaly. Ascites consists in an accumulation of fluid in the peritoneal cavity leading to abdominal distension and in severe cases removed by paracentesis. Splenomegaly is an enlargement of the spleen caused by the reduction in the number of circulating blood cells affecting granulocytes or platelets. Since these symptoms are very related to cirrhosis, they appear only in the late stages of fibrosis and so they have high positive predictive value but low negative predictive value, making them not useful to diagnose or stage liver fibrosis.

Laboratory tests can help in detecting liver diseases, in fact anomalous values of ALT, AST, GGT and platelets are potential markers of hepatitis. Besides, construction of newer serologic markers have been proposed as aids in determining the degree of fibrosis in the liver. Most common indexes are build up as ratio index i.e. AST:ALT ratio or APRI (AST Platelets Ratio Index). As previously seen for clinical manifestations, laboratory tests are not sufficient to classify liver fibrosis, since anomalous values of these indexes could be symptoms related to other liver diseases.

Imaging studies as ultrasonography, tomography or magnetic resonance could be useful but only in late stages of liver fibrosis. They are actually used as more accurate exams for cirrhotic subjects. Other negative aspects of imaging studies are the high cost and the exposure of patients to radiations.

2.2 New biomedical technology

Ultrasound techniques named as hepatic elastography represent the future for detecting and staging liver disease, in fact contrary to biopsy they are not invasive or dangerous for the patient and consequently they could be repeated more times. Besides, these techniques are very rapid even in patients at the bedside and results are immediately displayed.

The two most famous ultrasound techniques are Fibroscan (produced by EchoSens) and ARFI (produced by Siemens). Both tools use a basic principle of physics: a wave is propagated more quickly in a stiffer tissue, where the wave is produced by a probe and the tissue is the liver. During recent years, a lot of studies have been considered a comparison between ARFI and Fibroscan (Friedrich-Rust *et al.*, 2009; Attanasio *et al.*, 2010). In particular, in Rizzo *et al.* (2011), ARFI imaging has been found to be a more accurate tool than Fibroscan for the non-invasive staging of both significant and severe classes of liver fibrosis. For this reason the attention will be focused on ARFI.

ARFI (Acoustic Radiation Force Impulse) measures the liver stiffness through mechanical excitation of tissue using acoustic pulses producing shear waves propagation. The shear wave speed is measured in m/s on a Region of Interest (RoI), a small box 1 x 0.5 (cm). The stiffer the liver, the faster the shear waves propagate. This speed could range from 0 m/s for patients with a normal liver to 5 m/s for cirrhotic patients. Since ARFI is a non-invasive test, it is possible to obtain measurements also in different segments or depths. This is a very important feature because in this way data are more reliable and it is possible to solve the problem of sparseness of fibrosis affecting the liver patchly. Moreover, ARFI is not only able to detect hepatic fibrosis but it is also important in staging the disease. In fact, it provides a corresponding scale compared with the Metavir scoring system for some thresholds of speed (Attanasio *et al.*, 2010). This correspondence is shown in Table 2.1.

Stage	ARFI (m/s)
F0-F1	< 1.3
F2	1.31 – 1.7
F3	1.71 – 1.99
F4	≥ 2

Table 2.1: ARFI vs METAVIR

Chapter 3

GAMLSS

3.1 Definition

Generalized Additive Models for Location Scale and Shape (GAMLSS) were introduced by Rigby and Stasinopoulos (2001). GAMLSS are defined as semi-parametric models. Actually, besides requiring definition of a parametric distribution for the response variable, it is possible to add non-parametric smoothing functions for each parameter considered in the model specification. The authors presented GAMLSS as a way to overcome some limitations of GLM (Generalized Linear Models) and GAM (Generalized Additive Models). GLM were introduced by Nelder and Wedderburn (1972) and represent a generalization of the linear regression model in which it is possible to use as response variable probability distribution different from the normal. A further generalization is represented by GAM introduced by Hastie and Tibshirani (1990) as an extension of GLM where a non-parametric smoothing component is considered. In comparison with GLM and GAM, the basic features of GAMLSS are two. Firstly, the Expo-

nential Family distribution assumption for the response variable is replaced by a more general distributions family. Moreover, GAMLSS allow to expand the modelling to scale and shape parameters as skewness and kurtosis too. For these reasons they are particularly flexible and suitable to model data in which the response variable shows some of these features.
GAMLSS assume independent observations y_i for $i = 1, 2, \ldots, n$ with probability density function $f(y_i|\theta^i)$ conditional on $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$ a vector of four distribution parameters. Each parameter can be a function of the explanatory variables. The first two parameters μ_i and σ_i represent location and scale parameters, while the remaining ν_i and τ_i refer to the shape parameters (skewness and kurtosis).

- μ_i = Location parameter or mean
- σ_i = Scale parameter or variance
- ν_i = Shape parameter 1 or skewness
- τ_i = Shape parameter 2 or kurtosis

The original formulation of GAMLSS is given by

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} Z_{jk}\gamma_{jk}$$

where for $k = 1, 2, 3, 4$, $g_k(.)$ are monotonic link functions relating the distribution parameters to explanatory variables, X_k is a known design matrix of order $n \times J'_k$, $\beta'_k = (\beta_1, \ldots, \beta_{J'_k})$ is a parametric vector of length J'_k and $Z_{jk}\gamma_{jk}$ the non-parametric additive terms.

Expanded formulation of GAMLSS is:

$$\begin{cases} g_1(\mu) = \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} Z_{j1}\gamma_{j1} \\ g_2(\sigma) = \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} Z_{j2}\gamma_{j2} \\ g_3(\nu) = \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} Z_{j3}\gamma_{j3} \\ g_4(\tau) = \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} Z_{j4}\gamma_{j4} \end{cases}$$

In this way each distribution parameter can be modelled as a linear function of explanatory variables and/or as linear functions of random variables. Other alternative formulations of GAMLSS could be considered.

The population probability (density) function $f(y|\theta)$ is left general with no explicit conditional distribution form for y. The only restriction that the R implementation of GAMLSS has for specifying the distribution of y is that function $f(y|\theta)$ and its first derivatives with respect to each of the parameters of θ must be computable. We shall use the notation:

$$y \sim D\{g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, \ldots, g_p(\theta_p) = t_p$$

to identify uniquely a GAMLSS, where D is the response variable distribution, $(g_1, \ldots, g_p)$ the link functions, $(t_1, \ldots, t_p)$ the model formulae for the explanatory terms in the predictors $(\eta_1, \ldots, \eta_p)$.

3.2 Inference in GAMLSS

There are two basic algorithms used for maximizing the penalized likelihood in GAMLSS. The first, the CG algorithm, is a generalization of the Cole and Green algorithm (Cole and Green, 1992) and it uses the first derivatives and the expected values of the second and cross derivatives of the likelihood function with respect to $\theta = (\mu, \sigma, \nu, \tau)$ for a four parameter distribution. However, for many probability distribution functions $f(y|\theta)$ the parameters θ are orthogonal. In this case the second, the RS algorithm is more suited. The RS is a generalization of the algorithm for fitting MADAM (Mean and Dispersion Additive Models). Essentially the RS algorithm has an outer cycle which maximizes the penalized likelihood with respect to the fixed and random effects in the model for each θ_k. At each iteration the current updated values of all the quantities are used. This algorithm is not a special case of the CG algorithm because in the RS the diagonal weight matrix W_{kk} is computed within the fitting of each parameter θ_k, whereas in the CG all weight matrices W_{ks} are evaluated after fitting all θ_k.

The aim of both algorithms is maximizing a penalized likelihood function l_p given by

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_k \gamma'_{jk} G_{jk} \gamma_{jk}$$

where $l = \sum_{i=1}^{n} logf(y_i|\theta^i)$.

This is achieved in two steps: firstly, the first and second derivatives of the aforementioned equation are obtained to give a Newton-Raphson step for maximizing it with respect to β_k and γ_{jk}; moreover each step of the Newton-Raphson algorithm is implemented by using a backfitting procedure cycling through the parameters and through the additive terms of the k linear predictors.

Each GAMLSS parametric model can be assessed by using its fitted global deviance GD given by $GD = -2l(\hat{\theta})$ where $l(\hat{\theta}) = \sum_{i=1}^{n} l(\hat{\theta^i})$. Two nested models M_0 and M_1 may be compared by using the test statistic $\Lambda = GD_0 - GD_1$ which has an asymptotic χ^2-distribution under M_0 with degrees of freedom $d = df_{M_0} - df_{M_1}$. For comparing non-nested GAMLSS the GAIC (Generalized Akaike Information Criterion) (Akaike, 1974) can be used. GAIC is obtained by adding a fixed penalty term for each effective degree of freedom used in the model. Model with the smallest value of GAIC will be selected.

For each model M the normalized randomized quantile residuals of Dunn and Smith (Dunn and Smyth, 1996) are used to check its global adequacy of M and the distribution component D. These residuals are given by $\hat{r}_i = \Phi^{-1}(u_i)$ where Φ^{-1} is the inverse CDF of a standard normal variate with $u_i = F(y_i|\hat{\theta^i})$ if y_i is an observation from a continuous response, whereas u_i is a random value from the uniform distribution on the interval $[F(y_i - 1|\hat{\theta^i}), F(y_i|\hat{\theta^i})]$ if y_i is an observation from a discrete integer response variable, where $F(y|\theta)$ is the CDF. The true residuals r_i have a standard normal distribution if the model is correct.

3.3 Worm plot

Diagnostics in GAMLSS is carried out through the use of worm plots. These graphs were introduced by Van Buuren and Fredriks (2001) for the LMS model (Cole and Green, 1992). Worm plots are widely used as diagnostic tool in growth curves studies and they are very similar to Q-Q plots. Quantile-quantile plots can be applied to compare the quantiles of a theoretical distribution of residuals scores (on the horizontal axis) against those of the empirical one (on the vertical axis). A worm plot consists in a detrended Q-Q plot where on the vertical axis the difference between location in the theoretical and empirical distribution is represented. The worm plot contains the 95 % confidence interval of the unit normal quantiles. For a given quantile z with associated probability p and a sample size n, the confidence interval is computed as $\pm 1.96 \times \varphi(z)^{-1} \sqrt{(p(1-p)/n)}$, where $\varphi(z)$ is the Normal density function. The interval becomes larger towards the extremes, so in the tails broader differences between theoretical and empirical quantiles are allowed.

This plot is called ”worm plot” because data points form a worm-like string. If the worm is flat, then the data follow the assumed distribution. Different patterns of the worm lead to underline some problems in the global fit. Worm plots are also useful to check assumptions on some particular intervals of the explanatory variables. For instance, in Van Buuren and Fredriks (2001), worm plots are displayed for models conditioning on specific class intervals in age.

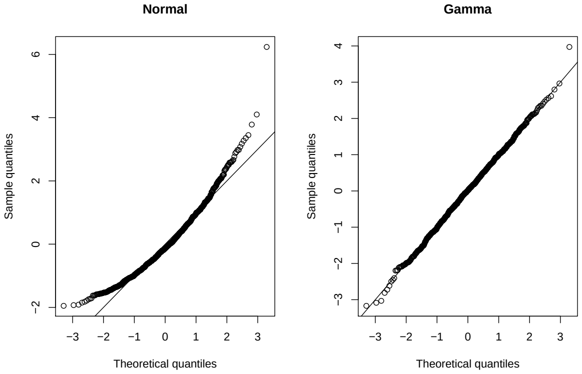

Figure 3.1: Example of Q-Q plots

In order to show the difference between standard Q-Q plot and worm plot we simulated $n = 1000$ observations from a Gamma distribution with shape parameter $\alpha = 5$ and we show Q-Q plot and worm plot for the null model considering Normal and Gamma distributions. As we expect, in Figure 3.1 Q-Q plot on the left (Normal distribution) is not aligned with the main diagonal while, of course, on the right (Gamma distribution) the fitting is definitely better. For same data Figure 3.2 represents the two worm plots for the same models. On the left representation of $z-$scores for worm plot is failing. The pattern is not a worm but a U-shape framework, moreover most of the points are out of the confidence bands. The graph on the right shows a worm-like string and all the points are in the confidence interval. This justifies the use of the Gamma distribution. This trivial example shows as the worm plot could be used to verify the correct specification of the model and, in particular, the choice of the response distribution.

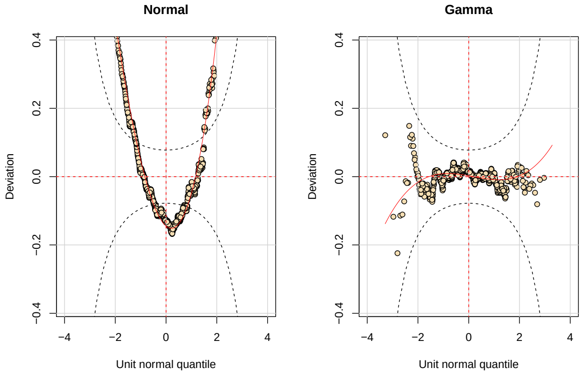

Figure 3.2: Example of worm plots

3.4 Open problems in GAMLSS

A lot of problems are still open in studying GAMLSS. So far, more than 80 distributions are implemented in GAMLSS, but every month new distributions are added by GAMLSS developers.

In particular, they are interested in looking for some developments in analyzing particular datasets exploiting the flexibility of GAMLSS method. Extension of the GAMLSS family of distributions it is possible not only through the definition of new theoretical distributions but also by adapting the existing one in a context of censored or truncated data.

Another developing issue is the implementation in GAMLSS of the additive terms. Up to now a lot of non-parametric functions could be applied using GAMLSS: splines, varying coefficients, fractional polynomials, and so on. There are other extra additive terms that, at the moment, do not lead to stable results and very often they return problem in convergence. So, some developments could concern the introduction of functions to fit break points within GAMLSS, GAM (General Additive Models) outside the exponential family, neural networks, penalized lag regression functions and regression trees.

Chapter 4

Dataset description

4.1 Introduction

Originally, ARFI measurements were collected by Ultrasuoni Srl from 2010 to 2013 and they were only available in DICOM (Digital Imaging and COmmunications in Medicine) format. To obtain our dataset it was necessary a pre-processing phase in which DICOM files were transformed into alpha-numeric strings by R-package **oro.dicom**. Strings of interest were extracted and the raw dataset was obtained. Three steps of data scrubbing were applied for the final dataset: some spelling errors were corrected, a control about coherence of values was applied and finally few duplicate rows were deleted.

In the final dataset each elastography includes also the total number of measurements; only a small number of patients repeated the exam more times in the four years. For this reason the dataset has a three-levels hierarchical structure: i) a macro-level patient; ii) exams by patient; iii) a final-level measurements in the exams. The hierarchy framework will be examined in depth in next section.

Response variable is liver stiffness stated as speed of the wave produced by ARFI measurements (measured in m/s). Explanatory variables are divided into two groups: risk factors concerning the patient and predictors about the exam. The first group is composed by sex, age, size and weight and they are obviously equal when different exams or measurements on the same patient are considered. Variables observed at the exam level include information about depth (in cm), liver segment and patient position during measurement. In Figure 4.1 below a subset of the whole dataset is displayed.

```
        id            patient variables   exam variables  response
        ||                     ||                 ||          ||
name      exam   measure |sex age  size weight|depth  seg posit| speed
Subject1  10.01    1     | M  73  1.65   75   | 5.4    7   ant |  2.64
Subject1  10.01    2     | M  73  1.65   75   | 5.5    7   ant |  2.58
.                               .                               .
.                               .                               .
Subject2  10.02    1     | F  62  1.60   55   | 5.4    6   ant |  1.68
Subject2  10.02    2     | F  62  1.60   55   | 4.0    6   pos |  1.88
.                               .                               .
.                               .                               .
Subject1  13.24    1     | M  75  1.65   75   | 4.3    8   lat |  2.12
Subject1  13.24    2     | M  75  1.65   75   | 5.5    7   ant |  1.31
```

Figure 4.1: Example of a subset from full dataset

4.1.1 Exam variables

In many studies about liver diseases antropometric data about subjects are available and the issue of these works is to investigate the relationship between these features and the presence of the disease. The possibility to observe exam variables represents an innovation. This is the reason why one of the aims of this study will be to analyze how the disease could be related to both antropometric data and exam variables. In particular, exam variables concern different ways of measuring elastography: depth in cm, liver segment and position of patient.

Since hepatic fibrosis affects the liver patchly, the advantage to obtain repeated measurements in different parts of liver becomes fundamental. The possibility to have data about depth represents the most important innovation of this dataset since for the first time data on depths are available. ARFI allows to measure the liver stiffness at different depths starting from 1.5 cm to a maximum of 8 cm. Understanding how speed changes its value when stiffness is measured at different values of depths represents a very important purpose for this study.

Liver is divided in 8 segments, in our dataset only 4 of 8 segments are considered. These segments are commonly numbered 5, 6, 7, 8 and positioned in left part of the liver. Segments 1, 2, 3, 4 are not considered because of their complexity in obtaining measurements.

Other studies show how position of patient during the examination affects the value of speed measured by ARFI (Attanasio *et al.*, 2010; Goertz *et al.*, 2012). This is why information about three kinds of assumed positions are added in the dataset. The three positions are: supine (ant), lateral (lat) and prone (pos) and they are displayed in Figure 4.2.

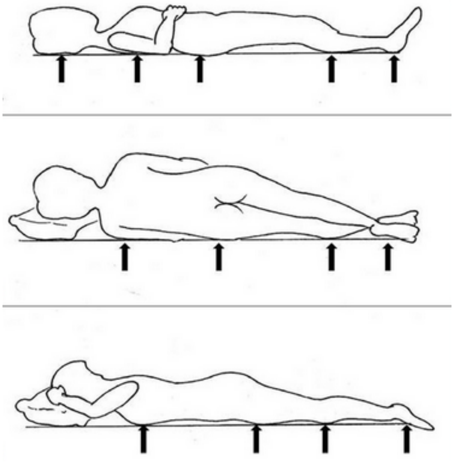

Figure 4.2: Three positions during elastography

4.2 Data hierarchical structure

The dataset presents a peculiar framework, as data show a three-levels structure. Liver fibrosis can be measured for three different kinds of statistical units. It is expected that observation is represented by speed for a patient, but as previously said, ARFI is not an invasive test, so having repeated measurements does not represent a problem. A lot of measurements are repeated at a single exam since it is possible to use ARFI for detecting liver fibrosis in different liver segments and at different depths. For this reason patient could be considered the statistical unit only at higher level. An intermediate level is represented by the exam; since our dataset includes data observed over 4 years (from 2010 to 2013), it is possible that a patient had more than one exam in this period. The lower level is the single measurement of the exam for a patient. This hierarchical structure is shown in the frame below.

$$\begin{tikzcd}
\text{Level 1} \arrow[d] & \text{Level 2} \arrow[d] & \text{Level 3} \arrow[d] \\
\text{Subject 1} \arrow[r] & \text{Exam } 1, \ldots, k_1 \arrow[r] & \text{Measurement } j_{1,1}, \ldots, j_{k_1,q_1} \\
\text{Subject 2} \arrow[r] & \text{Exam } 1, \ldots, k_2 \arrow[r] & \text{Measurement } j_{1,1}, \ldots, j_{k_2,q_2} \\
\text{Subject } n \arrow[r] \arrow[d] & \text{Exam } 1, \ldots, k_n \arrow[r] \arrow[d] & \text{Measurement } j_{1,1}, \ldots, j_{k_n,q_n} \arrow[d] \\
681 & 967 & 37.659
\end{tikzcd}$$

Data were collected by Ultrasuoni from 2010 to 2013 and the Tables in 4.1 present on the left the subjects grouped by number of exams and on the right the exams grouped by number of measurements.

The total number of the patients is 681 for 967 exams and 37.659 measurements. Most of patients (74%) did not repeat the exam during the four years, more than one hundred of subjects repeated twice the exam and only 9% did the exam 3 or more times. The distribution of exams grouped by number of measurements has a big variability, average for measurements

Level 2 → Level 1

Exams	Subjects
1	501
2	116
3	37
4	16
5	8
6	2
7	1
Total	681

Level 3 → Level 2

Measurements	Exams
0-10	92
11-20	170
21-30	293
31-40	141
41-50	106
51-100	73
101-200	92
Total	967

Table 4.1: Hierarchical structure of the dataset

is 39, but 30% of the exams contains between 21 and 30 measures. More than 250 exams include a number of measures less than 20 and more than 150 exams contains more than 50 measurements, this means that there is not a general rule about an optimal number of measurements done during a single elastography.

4.3 Some descriptive statistics

Some descriptive analysis were first conducted on the response variable. Speed is measured in m/s and it is ranged in (0.5, 9). The faster the value for speed is, the stiffer the liver. Distribution for speed is highly positively skewed, with a value for skewness equal to 1.7. Kurtosis is also present, value for kurtosis is 6.8, so distribution is leptokurtic. The graph on the right shows histogram and density for speed variable for our dataset.

Speed
Min. :0.500
1st Qu.:0.940
Median :1.190
Mean :1.484
3rd Qu.:1.780
Max. :9.390
Variance:0.666
Skewness:1.762
Kurtosis:6.894

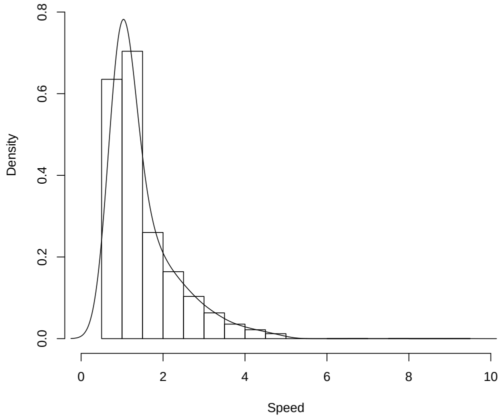

Table 4.2: Descriptive statistics and distribution for speed

In Table 4.3 descriptive statistics are shown for variables at patient level, differently from Table 4.4, where statistical units are measurements and not patients. Subjects are equally balanced between males and females (338 vs 343). Age of patients ranges from 17 to 88 years, subjects under 17 years old have been deleted because their liver could be not completely developed. Size ranges from 1.345 to 1.90 cm, while weight from 39 to 120 Kg. The average age of subjects is 54 years old, the average height is 1.63 cm tall and the average weight is 72 Kg.

Sex	Age	Size	Weight
M:343 (50,4%)	Min. :17.00	Min. :1.345	Min. : 39.00
F:338 (49,6%)	1st Qu.:44.00	1st Qu.:1.555	1st Qu.: 60.00
	Median :55.00	Median :1.630	Median : 69.50
	Mean :54.00	Mean :1.631	Mean : 70.50
	3rd Qu.:64.00	3rd Qu.:1.710	3rd Qu.: 79.00
	Max. :88.00	Max. :1.900	Max. :160.00

Table 4.3: Summary statistics for patient variables

In Table 4.4 some descriptive statistics of exam variables are presented. Measurements are collected in a range of depth from 1.5 to 8 cm; changes in depths will be shown later. The distribution of liver segments shows how all parts have a similar number of measurements. About position, the anterior one, the supine position, is the most frequent with more than 60% of the cases.

In Chapter 2 we explained that one of the advantages of using ARFI is the existence of a scale correspondence between speed and the METAVIR scoring system of liver biopsy. Hence, it is possible to derive a classification of statistical units a priori just on the basis of an average value of speed. Actually, a double classification is obtained considering differently statistical

Depth	Segment	Position
Min. :1.500	5:10526 (28%)	ant:23185 (61%)
1st Qu.:4.000	6: 8921 (24%)	lat: 8505 (23%)
Median :5.200	7: 9769 (26%)	pos: 5969 (16%)
Mean :5.194	8: 8443 (22%)	
3rd Qu.:6.200		
Max. :8.000		

Table 4.4: Descripitve statistics for exam variables

units at level 1 (Subject) or 2 (Exam). For each classification mean or median have been evaluated. About level 1 (Table 4.5), most of patients are classified in F0-F1 stages, so their liver presents no fibrosis and only 8% are cirrhotic. In terms of the median, the number of healthy patients increases.

Stage	ARFI Cutoff	Mean	Median
F0-F1	≤ 1.3	418	557
F2	1.3 – 1.7	186	55
F3	1.7 – 2	21	13
F4	≥ 2	56	56
Total		681	681

Table 4.5: ARFI vs METAVIR at level 1

Similar results are obtained at intermediate level, most of the patients are again in F0-F1 and 15% (13% using median) of exams belongs to cirrhotic patients. This percentage increased, in comparison with level 1, probably because cirrhotic patients repeated more times the exam during four years.

4.4 Changes in measuring from 2010 to 2013

Having a dataset about data observed in several years, it is possible to take into account some possible changes occurring during the period. First

Stage	ARFI Cutoff	Mean	Median
F0-F1	≤ 1.3	557	674
F2	1.3 – 1.7	201	114
F3	1.7 – 2	65	57
F4	≥ 2	144	122
Total		967	967

Table 4.6: ARFI vs METAVIR at level 2

change is in the average of measurements for exam during years. Until 2012 there is an increase to a maximum of 60 measures for exam, but this average decreases down to 29 in 2013 (see Table 4.7).

Year	2010	2011	2012	2013	Total
Exams	344	200	280	143	967
Measurements	9029	7772	16729	4129	37659
Average	26	39	60	29	39

Table 4.7: Number of exams and measurements for year

Secondly, during the exams it was preferable to obtain measures in deeper parts but an important update of the software used in ARFI was released in Febraury 2011 allowing measures until 8 cm, while the previous limit was 5.5 cm. As we can see in Table 4.8, for this reason the most frequent depth analyzed is 8 cm from 2011 to 2013.

Thirdly, other changes concern the liver segment analysed and the position of patient during the exam. Figure 4.3 shows on the left panel, the number of measurements grouped by liver segment per years. Most observed liver segment changed from the seventh segment in 2010 to the fifth one in 2013.

	2010	2011	2012	2013
1	5.5cm: 2461	8cm: 455	8cm: 1202	8cm: 283
2	5.4cm: 508	5.5cm: 307	4.6cm: 382	7.9cm: 129
3	3.8cm: 281	4.8cm: 203	5cm: 382	5.1cm: 106
4	3.9cm: 278	4.6cm: 200	5.4cm: 369	5.3cm: 101
5	4.7cm: 277	4.3cm: 197	4.7cm: 363	5.7cm: 97
6	5.1cm: 274	5.4cm: 182	5.8cm: 354	6.1cm: 97
Others	4950	6228	13677	3316
Total	9029	7772	16729	4129

Table 4.8: Number of annual measurements for depth

On the right panel, measures are grouped by patient position; the anterior one or supine is the most frequent for all the years but there is a significant increasing interest for lateral position in 2013.

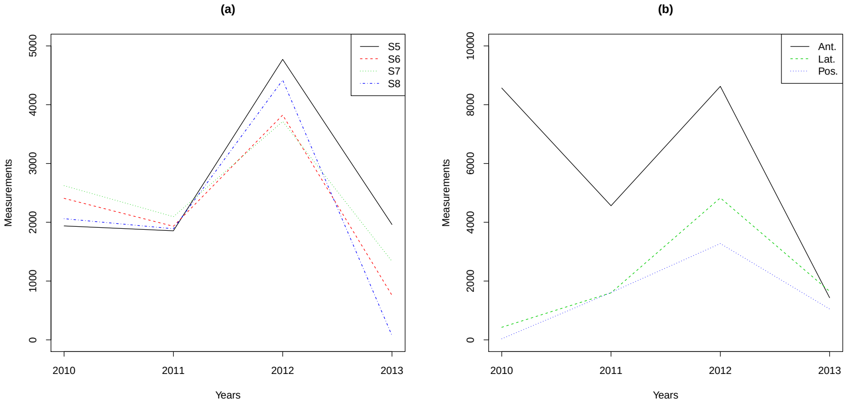

Figure 4.3: Number of annual measurements for segment (a) and position (b)

Chapter 5

Analysis

5.1 Statistical models for high variability data

In the previous chapter descriptive statistics have emphasized the large variability of ARFI observations in measuring liver stiffness.
When data show a very large variability or overdispersion, a possible solution to data modelling could be represented by the inclusion of parameters, not only for the mean effect.
In literature three extensions of classical linear model have been proposed as possible solution to this problem:

- ARCH (AutoRegressive Conditional Heteroskedasticity) (Engle, 1982);
- DGLM (Double Generalized Linear Models) (Smyth, 1989);
- GAMLSS (Generalized Additive Models for Location Scale and Shape) (Rigby and Stasinopoulos, 2005).

ARCH regression models were introduced in 1982 by Robert Engle. In his definition, two model equations are considered: a first one to estimate mean as a linear combination of lagged variables and a second variance equation that can include current and lagged explanatory variables. These models are mostly used for time series analysis and they are particularly useful to model financial volatility.

The class of DGLMs was proposed by Smyth (1989) derived some case deletion diagnostics for linear heteroscedastic models under maximum likelihood (ML) and restricted maximum likelihood (REML) estimation. In his paper, Smyth (1989) provides MLE for all the parameters when the population distribution is Normal Inverse Gaussian or Gamma. The method can be generalized using quasi-likelihoods. DGLMs are mainly used in presence of data with heteroscedasticity because they allow to estimate jointly mean and dispersion.

As we said in Chapter 3, the possibility to model also skewness and kurtosis with a distribution not necessarily belonging to the exponential family led us to choose the GAMLSS family.

Once GAMLSS family has been chosen, some fundamental issues will have to be decided to fit the best model to detect liver fibrosis. First of all, the choice of distribution for the response variable will be carried out among more than 80 distributions implemented in GAMLSS. Secondly, the choice about which distribution parameters (μ, σ, ν, τ) have to be included in the model. Thirdly, parameter model selection has to be considered on the basis of the GAIC (Generalized Akaike Information Criterion). Finally, since a hierarchical structure is present in dataset with repeated measurements for each subject, the possibility of including random effects has to be considered.

5.2 Focusing the attention on 2013

Data were collected for 4 years from 2010 to 2013 for a total of 37.659 measurements. When the number of observations is so big, applying GAMLSS requires a strong computational cost. To reduce this stress and get simpler structure for the dataset, we decided to reduce the total number of observations by focusing the attention on a specific year. The year 2013 has been selected, since it can be considered a stable year in terms of measurements. Indeed, during other years some changes occurred or the number of measures for patients was too variable. Moreover, the choice to take into account just a single year represents also a solution to the double hierarchy problem of the dataset structure. Actually, it is very uncommon that a subject repeated the exam during the year, so the exam effect is negligible and random effects have to be included only for subjects. Number of measurements is now reduced from 37659 to 4129. Some descriptive statistics for the 2013 dataset are presented in Tables 5.1 and 5.2.

Speed
Min. : 0.500
1st Qu.: 0.890
Median : 1.170
Mean : 1.511
3rd Qu.: 1.900
Max. : 4.900
Variance: 0.786
Skewness: 1.432
Kurtosis: 4.529

Table 5.1: Descriptive statistics for 2013 speed

Sex	Age	Size	Weight
M:75 (50,4%)	Min. :19.00	Min. :1.370	Min. : 41.00
F:66 (49,6%)	1st Qu.:44.00	1st Qu.:1.540	1st Qu.: 59.00
	Median :55.00	Median :1.630	Median : 68.00
	Mean :54.00	Mean :1.617	Mean : 70.25
	3rd Qu.:64.00	3rd Qu.:1.700	3rd Qu.: 76.00
	Max. :88.00	Max. :1.855	Max. :160.00

Table 5.2: Descriptive statistics for 2013 subjects

5.3 Selection of family distribution in GAMLSS

Focusing on 2013, the density of speed is displayed in figure 5.1. Also in this case, the distribution for response variable is positively skewed and lepotkurtic. This means that, among the more than 80 distributions implemented in GAMLSS, we have to search for a continuous and positive skewed distribution in R^+ taking kurtosis also into account. We have selected 6 probability distributions and we fit a null model for the reduced dataset. The criterion used for the distributions comparison was the GAIC (Generalized Akaike Information Criterion). In particular we fitted the following probability distributions: IG (Inverse Gaussian) (Johnson *et al.*, 1994); BCCG (Box-Cox Cole and Green) (Cole and Green, 1992); BCPE (Box-Cox Power Exponential) (Rigby and Stasinopoulos, 2004); BCT (Box-Cox generalized t) (Rigby and Stasinopoulos, 2006); GB2 (Generalized Beta 2) (McDonald and Xu, 1995); ex-GAUS (exponentially modified Gaussian (EMG) distribution) (Grushka, 1972). The number of parameters involved (p) and AIC are presented in Table 5.3. Null model with Box-Cox Power Exponential distribution has the smaller AIC, hence BCPE distribution has been selected in order to model speed in our dataset.

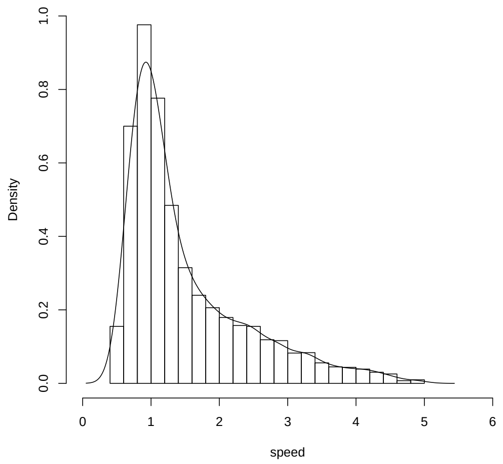

Figure 5.1: Distribution for speed in reduced dataset

Distr.	*p*	*AIC*
IG	2	8402.869
BCCG	3	8215.707
ex-GAUS	3	7980.840
BCPE	**4**	**7968.587**
BCT	4	8261.270
GB2	4	8219.455

Table 5.3: AIC for 6 distributions in GAMLSS (Null model)

5.4 The BCPE Distribution

The BCPE (Box-Cox Power Exponential) distribution was introduced by Rigby and Stasinopoulos (2004) to estimate smooth centile curves for skewed and kurtotic data. This distribution was developed to model both skewness and kurtosis in the distribution of a continuous response variable Y. The

BCPE could be considered as a generalization of Box-Cox Normal distribution. The standard Power Exponential family includes Normal, Uniform, Laplace but it does not consider skewness. On the other hand, the Box-Cox Normal distribution is able to model skewness but not kurtosis. Matching Box-Cox Normal with Power Exponential the result is a continue four parameters distribution denoted BCPE (μ, σ, ν, τ). This distribution provides a flexible model for a positive Y in presence of skewness and kurtosis. Unlike the BCT (Box-Cox t) distribution (Rigby and Stasinopoulos, 2006), this distribution fits well also platykurtic data and not just leptokurtic. The parameters of the model may be interpreted as related to location, scale, skewness and kurtosis and since we are using GAMLSS each can be modelled as a linear parametric or smooth nonparametric function of explanatory variables. A positive random variable having a Box-Cox Power Exponential distribution, denoted by BCPE (μ, σ, ν, τ), is defined through the transformed random variable Z given by:

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\tau} - 1\right] & \text{if } \nu \neq 0 \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right) & \text{if } \nu = 0 \end{cases}$$

for $0 < Y < \infty$ where $\mu, \sigma > 0$ and where the random variable Z is assumed to follow a standard Power Exponential distribution with power parameter, $\tau > 0$, treated as a continuous parameter.

The probability density function of Y is given by:

$$f_Y(y) = f_z(z)\left|\frac{dz}{dy}\right| = \frac{y^{\nu-1}}{\mu^{\nu}\sigma} f_z(z)$$

with standard link functions

$g_1(\mu) = 1; \qquad g_2(\sigma) = log; \qquad g_3(\nu) = 1; \qquad g_4(\tau) = log$

5.5 Model selection procedure

For a given distribution for the response variable (BCPE), the selection of the terms for all the parameters of the distribution uses a stepwise GAIC procedure. We now describe the steps of the procedure employed in our analysis.

1. From the null model build a model for μ using a forward approach.
2. given the model for μ build a model for σ (forward)
3. given the models for μ and σ build a model for ν (forward)
4. given the models for μ, σ and ν build a model for τ (forward)
5. given the models for μ, σ, ν and τ check whether the terms for ν are needed using backward elimination.
6. given the models for μ, σ, ν and τ check whether the terms for σ are needed (backward).
7. given the models for μ, σ, ν and τ check whether the terms for μ are needed (backward).

At each step we are conditioning further steps to the previous choices. Application of the default stepwise procedure leads to an unfeasible GAMLSS model. Indeed, several problems occur in the interpretation of a so complex statistical model. The biggest problem is represented by the high number of variables involved in a linear system of 4 equations. Our aim is to find a

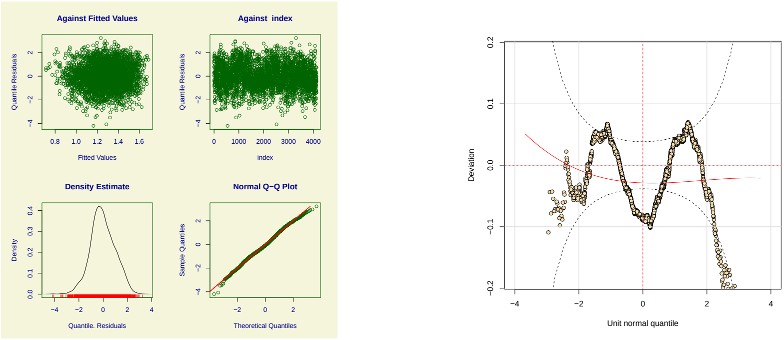

Figure 5.2: Diagnostic tools for model A

specific set of predictors for single equations; in this way we could detect variables having an effect on each specific parameter of the speed distribution.

Besides, we discussed in previous chapter the hierarchical structure of the dataset with repeated measurements. A proper specified GAMLSS model should include a random effect for this framework. It seems clear that adding a patient effect gives back a more complicated model. For this reason fixed effects have to be reduced substantially.

Finally, a consideration about diagnostics of the model, plot of the residuals is not so bad, no evidence in pattern of residuals against fitted values and indexes and some typical problems in the tails for the Q-Q plot. But this is not enough, indeed the worm of the plot of the quantile residuals is not so flat, but it has a very strange M-shape with a lot of points out of the boundaries representing confidence bands as shown in Figure 5.2.

These are several reasons to consider the model not appropriate and we need some enhancements to make it simpler and easier to interpret from a medical point of view.

5.6 Use of other criteria for model selection

During stepwise procedure for model selection, different criteria could be used to select variables to include in the linear predictor of the final model. So far, a penalty $k = 2$ was applied obtaining AIC. Increasing the penalty is a way to simplify the model structure. A grid of values from 2 (AIC) (Akaike, 1974) to $\log n$ (SBC) (Schwarz *et al.*, 1978) was given to k and the relative number of explanatory variables included in the model for each parameter (μ, σ, ν, τ) is displayed in Table 5.4. Choice of best k has to be based on a satisfactory trade off between simplicity of the model and loss of information.

k	μ	σ	ν	τ
2 (AIC)	7	6	4	5
3 - 4	4	6	2	3
5 - 6	3	5	2	2
7	3	5	1	2
log(n) = 8.32	3	4	1	1

Table 5.4: Number of selected explanatory variables for different k

Worm plots for $k = 4$ and $k = log(n) = 8.32$ are displayed in Figure 5.3. M-shape pattern is maintained in the graph but now models have a reasonable number of parameters. So problems of complexity of the model and difficulty in the interpretation of the estimate from a clinical point of view were reduced.

Once the value k for the penalty has been selected, another possibility to improve the model fit consists in choosing proper link functions. We try to use a modified version of BCPE distribution with log link for μ, this version is called BCPE-original (BCPEo). The structure of the model does not change, the only difference is in the use of the logarithm as link function

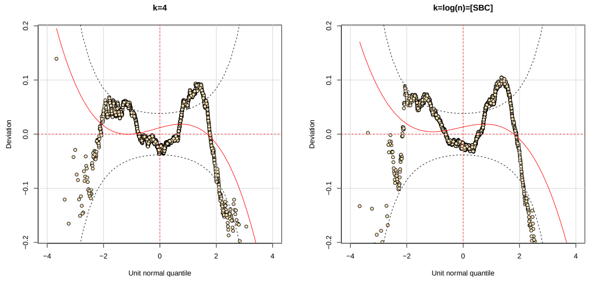

Figure 5.3: Worm plots for model selected with different penalty k

for μ. As it is possible to see from Table 5.5, the BCPEo models have a smaller Global Deviance. This means that whatever value of penalty k we choose models with log link for μ are preferable.

k	$Dev.BCPE$	$Dev.BCPEo$	df
2 (AIC)	7221.774	7199.201	33
3 - 4	7263.652	7241.626	26
5 - 6	7278.113	7257.81	21
7	7300.611	7285.142	18
log(n) = 8.32	7316.697	7300.487	16

Table 5.5: Comparison in terms of Global Deviance between GAMLSS with BCPE and BCPEo

5.7 Inclusion of a random effect component

Finally, in order to take into account correlation between observations from the same patient, a random effect component for the patient variable has to be considered. The previous model was estimated without taking into account hierarchical structure of data. Focusing on 2013, patient effect cannot be evaluated but we have to consider an exam effect. There are two functions for fitting random effects in GAMLSS, `random`() and `re`().

The function `random`() is based on the original function of Trevor Hastie in the package **gam**. Following this approach it is possible to find a ”local” maximum likelihood estimation of the smoothing parameter λ. This method is equivalent to the PQL method of Breslow and Clayton (1993) applied at the local iterations of the algorithm. Venables and Ripley (2002) claimed that this iterative method was first introduced by Schall (1991).

The function `re`() is similar to the `lme`() function of the package **lme** (Laird and Ware, 1982). Using this function it is possible to fit complex random effect models where the assumption of the normal distribution for the response variable is relaxed. The theoretical justification comes again from the fact that this is a PQL method (Breslow and Clayton, 1993). So we add a random effect for exam using alternatively the function `random`() and `re`() in **gamlss**. The two models are compared using AIC and worm plot as diagnostic tool for goodness of fit. Worm plot for two alternative models are shown in Figure 5.4. On the basis of these criteria the model using `re` was chosen.

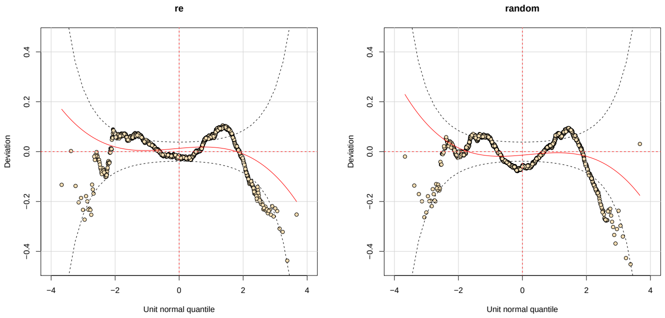

Figure 5.4: Worm plots for model selected with different random effects

Final specification of the model and table of coefficients are displayed in Table 5.6.

$$speed = \begin{cases} \log\mu = \alpha_1 + age + depth + segment + r.e.(subject) \\ \log\sigma = \alpha_2 + age + weight + size + position \\ \quad \nu = \alpha_3 + age \\ \log\tau = \alpha_4 + age \end{cases}$$

$\log\mu$	Estimate	P-value	
α_1	0.049	0.0491	*
depth	-0.048	< 2e-16	***
age	0.008	< 2e-16	***
seg6	-0.017	0.3677	
seg7	-0.128	1.25e-11	***
seg8	0.076	0.0566	.
$\log\sigma$	Estimate	P-value	
α_2	-0.460	0.0672	.
age	0.007	1.26e-09	***
weight	0.008	5.24e-16	***
size	-0.913	1.22e-09	***
positlat	-0.021	0.4395	
positpos	0.109	0.0004	***
ν	Estimate	P-value	
α_3	-2.539	< 2e-16	***
age	0.032	< 2e-16	***
$\log\tau$	Estimate	P-value	
α_4	-0.736	3.35e-07	***
age	0.024	< 2e-16	***

Table 5.6: Coefficients for selected GAMLSS model

Some considerations can be derived from this output. In order to make the interpretation of the model coefficients simpler, a graphical representation of the regression terms against predictors is plotted for each parameter of the model.

- Age seems to be the most important predictor for speed since it has to be included in each equation of the selected model. Age effect is always significant and its coefficient is positive for all the equations

- Depth has a negative effect on speed, since $\log \mu = -0.048$ then $\mu = exp(-0.048) = 0.953$

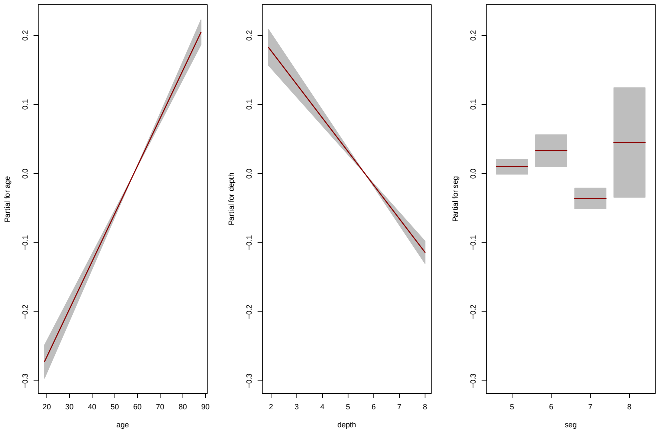

Figure 5.5: Term coeffcients for μ

- Segment is significant for μ with baseline Segment 5 not significantly different from Segment 6 and Segment 8. Coefficient for Segment 7 is negative $\log \mu = -0.128$ then $\mu = exp(-0.128) = 0.880$.

- For $\log \sigma$ we have 4 explanatory variables with size with a negative coefficient and lateral position not significant different from the baseline position (anterior) $\log \sigma = 0.109$ then $\sigma = exp(0.109) = 1.115$. (see Figure 5.6)

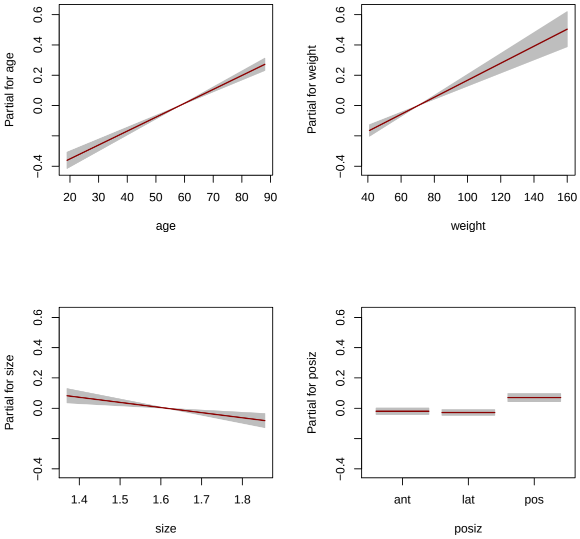

Figure 5.6: Term coeffcients for σ

- For skewness and kurtosis the only significant terms are just negative intercept and a positive coefficient for age. (see Figure 5.7)

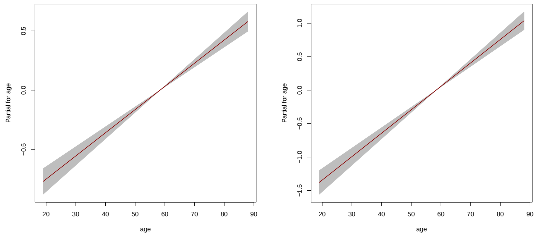

Figure 5.7: Term coeffcients for ν and τ

Residuals plot are shown in Figure 5.8. On the left, a four panels plot shows representation of residuals against fitted values and index variable, density estimate and Q-Q plot comparing theoretical and empirical distribution of the residuals. The density estimate seems to be not so far from a Normal distribution while the Q-Q plot shows a not good fitting in the tails. On the right the worm plot (see Chapter 3) of the model, most of the points are within confidence interval bands but they form a M-shape pattern more than a worm-like string. In the next section some solutions will be explored to fix this issue.

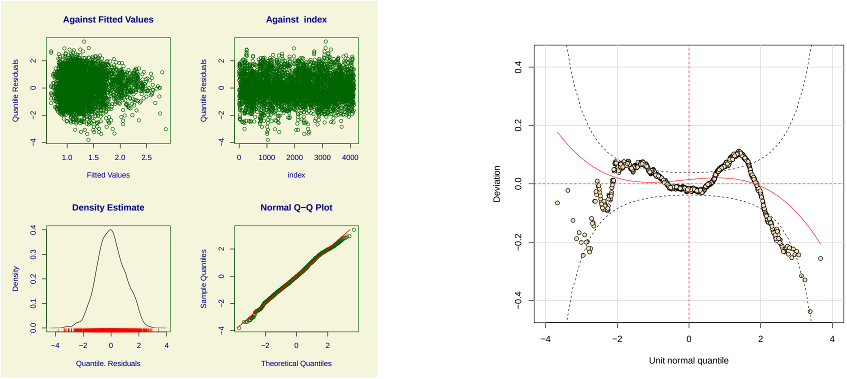

Figure 5.8: Diagnostic plots of the final model

5.8 First solutions for M-shape worm plots

The M-shape pattern in the worm plot could suggest the presence of bimodality in the data. A first solution could consist in the introduction of some significant interaction terms in the model. After choosing SBC as criterion to select the model, interaction terms were only introduced for μ with `addterm`() function. Using this function we try to fit all models that differ from the starting model by adding a single term from those supplied, maintaining marginality. The choice to include or not the interaction term is made on the basis of the AIC. The only significant interaction is the one involving Depth and Segment variables. In Figure 5.9 the worm plot from the model with interaction has been compared with that of the final model. As we can see from the two graphs there is no substantial difference between the plots, hence the model without interaction terms is chosen.

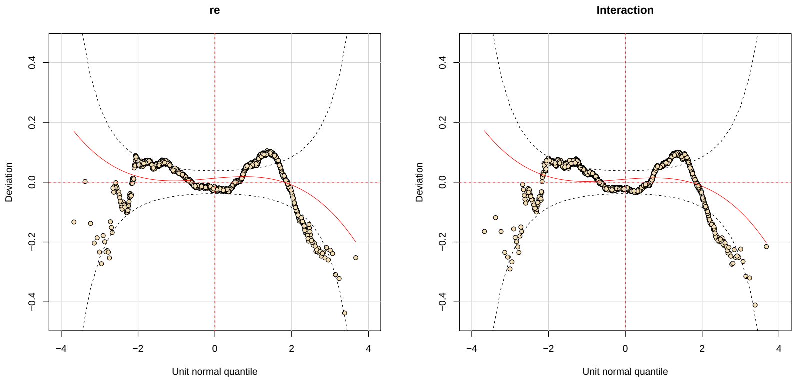

Figure 5.9: Simple model vs Interaction model

Since the idea to add interaction terms does not fix our M-shape pattern problem, we try to detect presence of two different groups of observations related to some features of the predictors to justify this bimodality in worm plots. For this reason we split up the dataset in two subsets for each predictor. The split has been obtained both conditioning on one of the factors of the qualitative explanatory variables and choosing an appropriate cut-off for quantitative predictors. Even using this solution the worm plots show again the same pattern. Just for example in Figure 5.10, conditioned worm plots for Sex are shown.

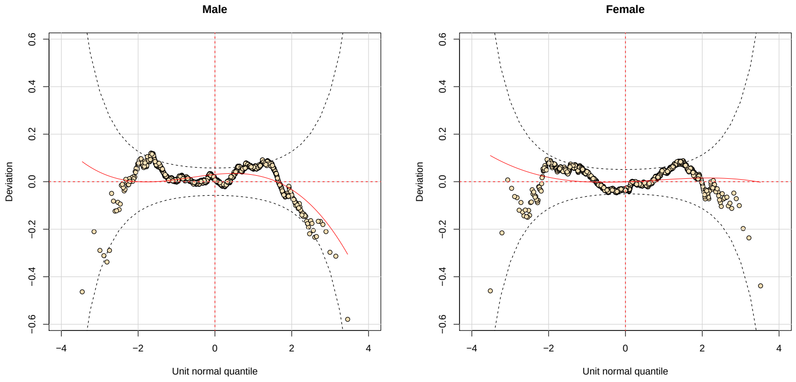

Figure 5.10: Males vs Females

Usually, an explanation of this pattern could be represented by skewness in residuals. Fitting a curve to the model residuals, the suggested distribution is plotted in Figure 5.11. This distribution does not seem to be so different from a Normal one. In next chapter we try to implement a mixture model approach for GAMLSS to deal with this M-shape in worm plots.

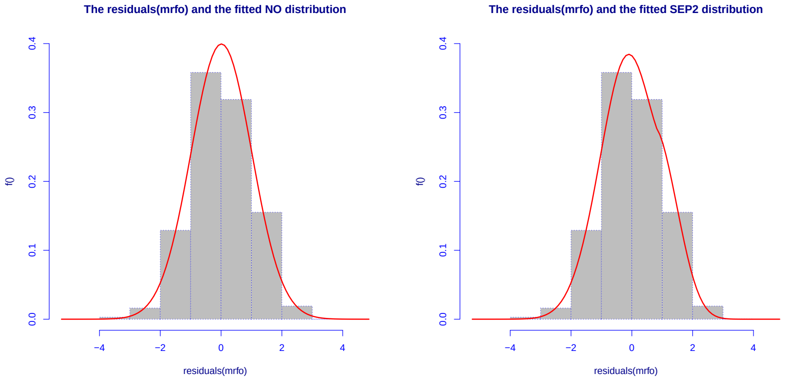

Figure 5.11: Histogram of residuals for final GAMLSS

5.9 Comparisons with other models

After the final model has been selected, a posterior comparison between the distribution in Table 5.3 is done. Using AIC as selection criterion in Table 5.3, BCPE was chosen as distribution function for stiffness response variable. A comparison using same probability distributions is shown in Table 5.7 where AIC is now computed on the final model. BCPE is again the best probability distribution since AIC for BCPE model is the lowest.

Distr.	*p*	*AIC*
IG	2	6916.590
BCCG	3	6676.640
ex-GAUS	3	6983.527
BCPE	**4**	**6616.631**
BCT	4	6673.272
GB2	4	6658.703

Table 5.7: AIC for 6 distributions in GAMLSS (Final model)

Chapter 6

Mixture models in GAMLSS

6.1 Overdispersion and mixture models

Overdispersion is the most common form of unexpected variation. Data are overdispersed when there is too much variation in comparison with the variation expected by the assumed distribution. There is much literature about overdispersion, since it is a frequently recurring situation when real data are analyzed. In this study we try to manage the problem of overdispersed data following the solution treated by Aitkin (1996). He proposed to use an EM algorithm (Dempster *et al.*, 1977) for maximum likelihood estimation in GLM with overdispersion. The algorithm is initially derived as a form of Gaussian quadrature assuming a normal mixing distribution. The approach we are going to use can be seen as an extension of this approach to probability distributions not belonging to the exponential family.

As we can see from Figure 6.1, our data are characterized by overdipersion. Taking into consideration the hierarchical structure of the data the two plots show the relationship among mean and variance (a) or mean and

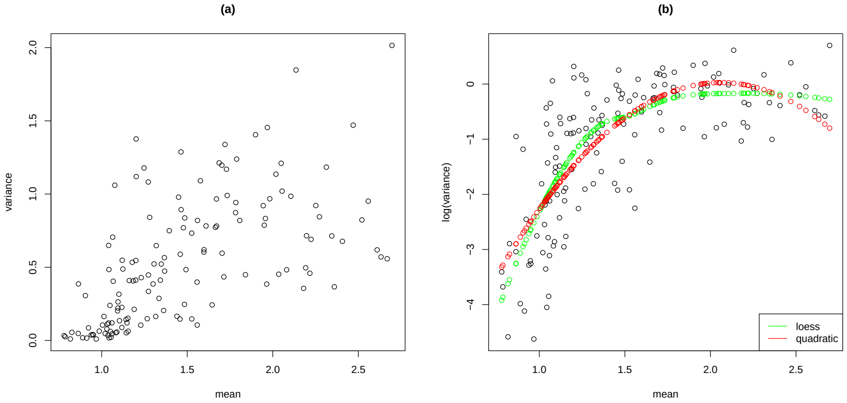

Figure 6.1: Overdispersed data

log variance (b) for each exam in the full dataset. On the left it is shown how variance increases when mean is increasing, on the right a quadratic curve (red) and a non parametric local polynomial curve (green) are plotted to describe this relationship. So probably we could use some technique for overdispersed data to avoid problems emphasized in the previous chapters. In particular, the mixture approach in GAMLSS could represent one of the possible solutions to fix the M-shape worm plots.

There is an extensive literature on mixture distributions and their use in modelling data. Everitt and Hand (1981), Titterington et al. (1985), and McLachlan and Peel (2000) are some of the books dedicated exclusively to mixture distributions.

As for other statistical models, using mixtures we suppose that our random variable Y comes from k component.

Suppose that the random variable Y comes from component k represented GAMLSS models, having density function $f_k(y)$, with probability π_k for

$k = 1, 2, \ldots, K$ then the marginal density of Y is given by

$$f_Y(y) = \sum_{k=1}^{K} \pi_k f_k(y)$$

where $0 \leq \pi_k \leq 1$ is the prior probability of each k component.

In GAMLSS two ways to apply mixtures are defined. According to the first method, each k component has a proper structure and there is no need to have parameters in common. According to this approach, it is possible that the conditional distributions $f_k(y)$, $k = 1, 2, \ldots, K$ could have different GAMLSS family distributions. Using the second approach, the k components of the mixture may have parameters in common, i.e. the parameter sets $(\theta_1, \theta_2, \ldots, \theta_k)$ are not disjoint. The prior (or mixing) probabilities are either assumed to be constant or may depend on predictors x_0 and parameters α through a multinomial logistic model. Note that, since some of the parameters may be common to the k components, the distribution used must be the same for all components. Similarly the link functions of the distribution parameters must be the same for all the components. In both cases likelihood function is maximized iteratively using the EM algorithm, with respect to ψ, i.e. with respect to θ and π.

We model the mixing probabilities π_{ik} using a multinomial logistic model where δ_i is a single draw from a multinomial distribution with probability vector π, i.e. $\delta_i \sim M(1, \pi)$. Consequently the complete log likelihood is given by:

$$l_c = l_c(\psi, y, \delta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} log f_k(y_i) + \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \log \pi_{ik}$$

Summary of the $(r+1)$-th iteration of the EM algorithm

- E-step

  Replace δ_{ik} in previous equation by $\hat{w}_{ik}^{(r+1)}$, for $k = 1, 2, \dots, K$ $i = 1, 2, \dots, n$ to give:

  $$Q = \sum^{a} K_{k=1} \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \log f_k(y_i) + \sum_{K}^{k=1} \sum_{n}^{i=1} \hat{w}_{ik}^{(r+1)} \log \pi_{ik}$$

  where $\hat{w}_{ik}^{(r+1)} = \dfrac{\hat{\pi}_{ik}^{(r)} f_k(y_i|\hat{\theta_k}^{(r)})}{\sum_{k=1}^{K} \hat{\pi}_{ik}^{(r)} f_k(y_i|\hat{\theta_k}^{(r)})}$

- M-step

  1. Since components $f_k(y)$ for $k = 1, 2, \dots, K$ have parameters in common, Q cannot be maximized separately with respect to each θ_k. Obtain $\theta^{(r+1)}$ by fitting a single GAMLSS model to an expanded response variable y_e, with expanded explanatory variable design matrix X_e, using weights $\hat{w}^{(r+1)}$.
  2. Obtain $\hat{\alpha}^{(r+1)}$ by fitting a multinomial logistic model.
  3. $\hat{\psi}^{(r+1)} = [\hat{\theta}^{(r+1)}, \hat{\alpha}^{(r+1)}]$

Note that the M-step (1) is achieved by expanding the data set K times.

Using *R* software and package **gamlss.mx**, a mixture model has been fitted for liver fibrosis data with parameters in common through the function `gamlssNP()`. In *R* output, the column headed as MASS identifies the k mixture components. If this coefficient is significant then the use of a mixture with parameters in common is justified. This column is declared as a factor in the R implementation of the EM algorithm. If this factor MASS is included in the predictor for a distribution parameter μ, σ, ν, or τ, then the predictor intercepts differs between the k components. We choose $k = 2$ so the two components of mixture have a probability distribution BCPEo. No further actions were applied to perform model selection. The model in Chapter 5 was used as reference model to define parameters that has to be included in the model.

Some considerations are important on the results from this method. Firstly, it is possible to compare the coefficients tables of the two models: the one in Table 5.6 and the mixture model one in Table 6.1. Most of the conclusions derived from the first output are confirmed here. Among predictors, Age is again the most important since it is present in all the equations of the model. Depth has a significant negative effect on Speed. The only liver segment that differs from the baseline (segment 5) is segment 7. As for as $\log \sigma$ is concerned, even in this case there is no significant difference between anterior or lateral position, while the posterior one has a positive effect on the variance. All other coefficients are equally signed and similar in terms of absolute value. Moreover, the new coefficient named MASS is positive and statistically significant. The use of a mixture model is justified and the difference between the 2 components is positive.

Besides, comparisons between linear GAMLSS and GAMLSS mixture extension are carried out in two ways. Firstly, a comparison in terms of goodness of fit is achieved using Global Deviances GD. Secondly, the worm plot is used as diagnostic tool and since, according to our hypotheses, the use of a mixture model gets a more flat worm, the h number of points outside the confidence interval bands, is used as criterion for comparison. We wish for a lower GD and a lower h for the mixture model. Mixture model has a lower Global Deviance with $\Delta GD = 34.30$. As we can see from Figure 6.2, there is a difference between the two worm plots: compared to the one on the left (original model), the second one with the mixture approach is more flat and on the right tail more points are now within the boundaries with $\Delta h = 1355$.

$\log \mu$	Estimate	P-value	
α_1	-0.105	1.32e-06	***
depth	-0.046	$< 2e-16$	***
age	0.007	$< 2e-16$	***
seg6	0.012	0.2818	
seg7	-0.044	4.29e-07	***
seg8	0.058	0.0529	.
MASS	0.276	$< 2e-16$	***
$\log \sigma$	Estimate	P-value	
α_2	-1.444	$< 2e-16$	***
age	0.012	$< 2e-16$	***
weight	0.005	$< 2e-16$	***
size	-0.262	0.0002	***
positlat	-0.016	0.2114	
positpos	0.083	9.14e-09	***
ν	Estimate	P-value	
α_3	-2.246	$< 2e-16$	***
age	0.026	$< 2e-16$	***
$\log \tau$	Estimate	P-value	
α_4	-0.435	0.00345	***
age	0.040	$< 2e-16$	***

Table 6.1: Coefficients for mixture models in GAMLSS

Now, each statistical unit will have a posterior probability to belong to each of the two components. Let us consider these probabilities derived from the

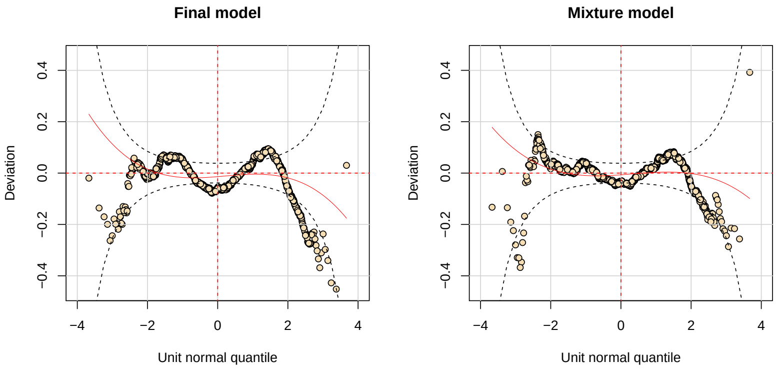

Figure 6.2: Final model vs mixture approach

estimated mixture model and let us connect them with the Metavir stage classification of liver fibrosis. As we have said in previous chapters, according to this classification a liver is divided into five stages from F0 to F4 on the basis of the presence of connective tissue in the liver. For simplicity, here we use a stage classification in three groups: F0-F1 (normal liver), F2-F3 (mild fibrosis) and F4 (cirrhosis). The probability to belong to one of the two identified components is divided in three groups too.

In Figure 6.3 we have on the x-axis the speed and on the y-axis the posterior probability to belong to component 1. We are surprised to see that there is a well-defined pattern in this plot. In fact, partitioning the two variables in 3 sectors, we obtain a grid of 9 sectors and measurements cluster only in some specific sectors. This situation could let us suppose the existence of a direct relationship between the posterior probability and the Metavir staging system. In particular, F0-F1 values of the speed are related to a percentage to belong to component 1 between 0% and 69%. Moreover, measurements belonging to cirrhotic patients seem to be linked to percentages over 90%.

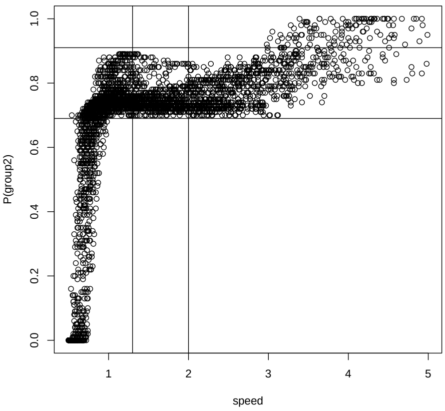

Figure 6.3: Partition of posterior probabilities related to speed

A 3-by-3 table could be derived from the graph to summarize the results of this method (see Table 6.2). Taking into account all $n = 4129$ measurements and using this approach, the direct relationship between Metavir and the three different groups involves 531 units in the first group and 121 in the third group. Actually, we know that these units represent just 15% of the entire number of observations and most of measurements are grouped in the middle classes, but it is however important to underline the presence of 4 zero-cells in the association table. For these reasons, the hypothesis that the two identified components of the mixture approach in GAMLSS could coincide with healthy and cirrhotic patients can not be rejected.

P \ D	F0-F1	F2-F3	F4
0-69%	531	0	0
69-91%	1730	807	827
91-100%	0	0	121

Table 6.2: Association matrix

6.2 Simulation studies

Some simulations have been run in order to evaluate the goodness of a mixture approach in GAMLSS when the response variable seems to be bimodal. The starting scenario is very similar to the one in liver fibrosis data. $M = 50$ datasets are simulated with $n = 1000$ observations. Each dataset includes a response variable Y and explanatory variables X_1, X_2, X_3. The Y variable is obtained as mixture of 2 BCPEo distributions and the weights for the mixture components are $\pi_1 = 0.75$ (in liver fibrosis data the proportion of observations belonging to non-cirrhotic patients is about this value) and $\pi_2 = 0.25$. Plot of densities for $Y = Y_1, Y_2$ for a simulated dataset is displayed in Figure 6.4.

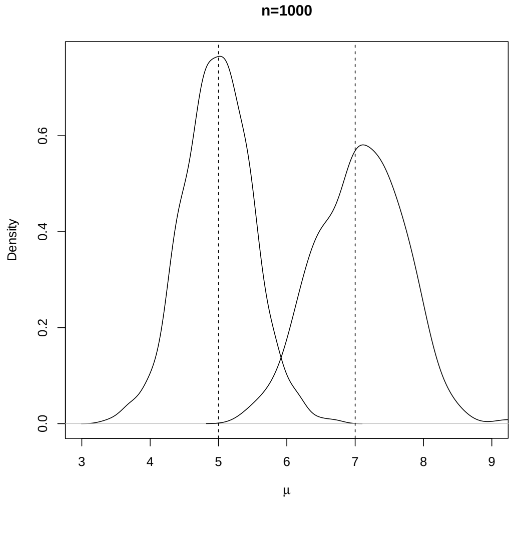

Figure 6.4: Plot of density for Y (Scenario 1)

Three predictors X_1, X_2, X_3 are simulated from a Normal distribution. A GAMLSS involving 4 parameters (μ, σ, ν, τ) has been estimated for these simulated data. The framework of the estimated GAMLSS is similar to the final model presented in Chapter 5 and it is shown below.

$$Y = Y_1, Y_2 = \begin{cases} \log \mu = \alpha_1 + X_1 + X_2 + X_3 \\ \log \sigma = \alpha_2 + X_1 + X_2 \\ \quad \nu = \alpha_3 + X_1 \\ \log \tau = \alpha_4 + X_1 \end{cases}$$

where $Y_1 \sim BCPEo(5, 0.1, 1, 2), Y_2 \sim BCPEo(7, 0.1, 1, 2)$ and $X_1, X_2, X_3 \sim N(5, 1)$. Using the same dataset, a GAMLSS mixture model approach is estimated. As seen for liver fibrosis data, linear GAMLSS and GAMLSS mixture extension can be compared using Global Deviances GD and the number of points outside the confidence interval bands h .

Besides the already described scenario, different scenarios have been implemented here, changing starting values for μ_2, in order to obtain different mixtures (scenario 2-5). Other scenarios have been obtained assuming different default values of random generalization of BCPEo distribution, values of $\sigma = 0.25, 0.5$ (scenario 6-7), $\nu = -1, 0$ (scenario 8-9), $\tau = 1, 3$ (scenario 10-11). Finally in scenario 12-13 different weights π_1 for the mixture components are considered: $\pi_1 = 0.5, 0.9$.

As we can see from Table 6.3, good results are obtained with the starting scenario. Indeed, average difference Global Deviance among linear model and the mixture approach is $\overline{\Delta GD} = 19.4$ and the difference in terms of number of points outside the confidence interval bands is $\overline{\Delta h} = 52$. Looking at different scenarios, when distance between μ_1 and μ_2 decreases the

	μ_1	μ_2	π_1	$\sigma_1 = \sigma_2$	$\nu_1 = \nu_2$	$\tau_1 = \tau_2$	$\overline{\Delta GD}$	$\overline{\Delta h}$
1	5	**7**	0.75	0.1	1	2	19.4	52
2	5	**6**	0.75	0.1	1	2	-0.2	-2
3	5	**8**	0.75	0.1	1	2	62.9	47
4	5	**9**	0.75	0.1	1	2	90.4	63
5	5	**10**	0.75	0.1	1	2	107.9	43
6	5	7	0.75	**0.25**	1	2	2.4	1
7	5	7	0.75	**0.5**	1	2	0.2	-1
8	5	7	0.75	0.1	**-1**	2	25.3	111
9	5	7	0.75	0.1	**0**	2	27.5	104
10	5	7	0.75	0.1	1	**1**	0	0
11	5	7	0.75	0.1	1	**3**	15.4	49
12	5	7	**0.5**	0.1	1	2	0.1	-4
13	5	7	**0.9**	0.1	1	2	2.64	2

Table 6.3: Simulation scenarios for $n = 1000$

two approaches appear very similar. When this distance increases, the two comparing indicators increase too, except for some convergence problems. Fixing μ_1 and μ_2 and increasing σ, mixture components are more flat and similar results to scenario 2 are obtained. Negatively skewed or symmetrical scenarios give better results than the first one. Different values for kurtosis $\tau = 1, 3$ show that, when low values are selected, results for two models are similar, instead using higher values results are similar to scenario 1. Finally, to choose different weights for the π mixing components leads to slight differences between the two methods.

Similar results are obtained increasing number of observations $n = 2000$ for simulated datasets (Table 6.4).

In conclusion, we could state that the use of a mixture approach in GAMLSS leads to good results for dataset similar to the liver fibrosis data. Firstly, since our scenarios are similar to the applied one in liver fibrosis data, sim-

	μ_1	μ_2	π_1	$\sigma_1 = \sigma_2$	$\nu_1 = \nu_2$	$\tau_1 = \tau_2$	$\overline{\Delta GD}$	$\overline{\Delta h}$
1	5	**7**	0.75	0.1	1	2	41.5	152
2	5	**6**	0.75	0.1	1	2	-0.7	-19
3	5	**8**	0.75	0.1	1	2	125.2	172
4	5	**9**	0.75	0.1	1	2	176	97
5	5	**10**	0.75	0.1	1	2	217.9	97
6	5	7	0.75	**0.25**	1	2	1	0
7	5	7	0.75	**0.5**	1	2	0.1	-2
8	5	7	0.75	0.1	**-1**	2	47	251
9	5	7	0.75	0.1	**0**	2	51.3	232
10	5	7	0.75	0.1	1	**1**	-0.5	10
11	5	7	0.75	0.1	1	**3**	35.5	117
12	5	7	**0.5**	0.1	1	2	-0.4	-23
13	5	7	**0.9**	0.1	1	2	2	50

Table 6.4: Simulation scenarios for $n = 2000$

ulations are limited to the use of a BCPEo distribution. Moreover, the difference in terms of $\overline{\Delta GD}$ and $\overline{\Delta h}$ is related to the framework of the mixture components. When they are well defined and not overlapped there is a clear gain in the global goodness of fit. Finally, the choice of weights π influences the results.

Chapter 7

ROC curve in GAMLSS

7.1 The ROC Curve

Receiver operating characteristic (ROC) curve is one of the most used tool to measure the accuracy of a binary medical test. Let D be the dummy variable to indicate the presence of disease and Y the result of the diagnostic test ($Y = 1$ positive test for disease and $Y = 0$ negative test for disease). A binary medical test is informative if it is able to predict the disease better than randomly. For this reason, in the presence of a dichotomous outcome and a binary prediction, four different situations can appear:

- True Positive (TP) when you have disease and your prediction test is positive;

- True Negative (TN) when you have not disease and your prediction test is negative;

- False Positive (FP) when you have not disease and your prediction test is positive;

- False Negative (FN) when you have disease and your prediction test is negative.

Arranging the outcomes in a 2-by-2 table, if D is used for the disease and Y for the test result we will have the following table:

	D=0	D=1
Y=0	TN	FN
Y=1	FP	TP
Total	TN + FP	FN + TP

Table 7.1: Definition of 2-by-2 table for ROC Curve

The accuracy of a test could be computed as the sum of the main diagonal ($TP + TN$) over the n total number of subjects. Two important factors that characterize a binary test are sensitivity and specificity. Sensitivity measures the proportion of subjects that are correctly predicted when disease is present, so it is defined as TP / ($TP + FN$). On the other hand, specificity measures the proportion of subjects that are correctly predicted when the outcome is negative, defined as TN / ($TN + FP$). Sensitivity is also called True Positive Rate (TPR) or True Positive Fraction (TPF), while specificity is also named True Negative Rate (TNR) or True Negative Fraction (TNF). Most of the times TNF is expressed as the difference between 1 and the False Positive Fraction ($1 - FPF$). An ideal test supposes all patients correctly predicted with $TPF = 1$ and $TNF = 1$ and all observation in 2-by-table will be on the main diagonal.

For a binary test, ROC curve is a graphical plot of sensitivity vs (1 - specificity), i.e (TPF) vs (FPF), where each point of the curve represents a different value for the cutoff to classify a subject as diseased or non-diseased. Since specificity and sensitivity are ranged between 0 and 1, this curve is always included in a square of dimensions (0,1) x (0,1). The point (0,0) represents $TPF = 0$ and $FPF = 0$ which predicts all subjects to be negative, while the point (1,1) represents $TPF = 1$ and $FPF = 1$ which predicts all subjects to be positive. When all subjects are correctly classified for all cutoff points then ROC curve is just a broken line following the points (0,0), (0,1) and (1,1), where the first value is on the horizontal axis and the second value is on the vertical axis. On the contrary, a completely random test would give a diagonal line from the left bottom to the top right corner. So every test whose curve is above the diagonal line is an informative test. Consequently the closer to the upper left corner is the curve, the better is the test (see Figure 7.1).

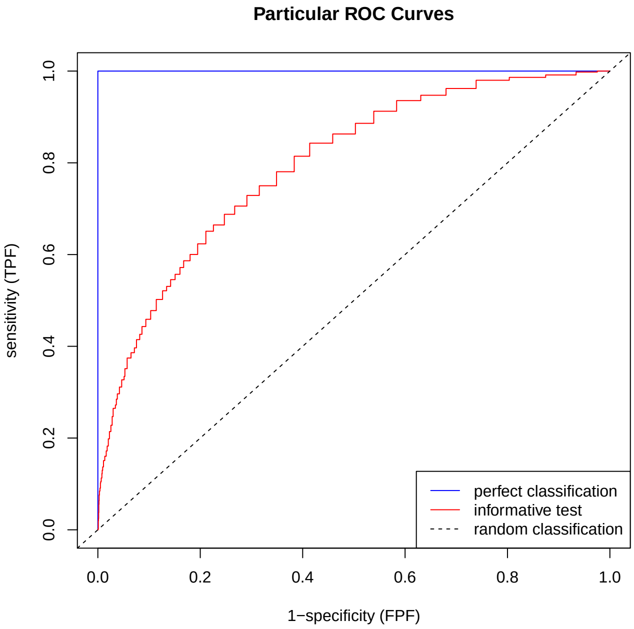

Figure 7.1: ROC Curve examples

Another way to check if a medical test is informative is to compute the Area Under a ROC Curve (AUC). This index is the most commonly used method for summarizing a diagnostic test's overall accuracy. It ranges from 0 to 1 (perfect classification) and takes value 0.5 for a random test. Hence the higher above 0.5 the AUC is, the more informative is the test.
ROC curves are also present in a binary regression framework (Pepe, 2003; Alonzo and Pepe, 2002), in fact it is possible to draw a ROC curve starting from a 2-by-2 table generated from the fitted model $\hat{p}$ and the true binary classification D. A cutoff or threshold value $0 < t < 1$ is chosen and set $Y = 1$ if $\hat{p} \geq t$, while $Y = 0$ if $\hat{p} < t$. The 2-by-2 table is a frequency cross tabulation of $Y = 0, 1$ against $D = 0, 1$. All the points of the curve are obtained as FPF and TPF corresponding to different values of t. For each t a 2-by-2 table is generated with resulting values for sensitivity and specificity of prediction. All these values are plotted in a square of dimensions (0,1) x (0,1) creating a binary regression ROC curve.

7.2 ROC Curve in GAMLSS

In the previous section ROC Curve has been shown as a tool used for binary test in prediction. Using our dataset the first issue is that our response variable is not binary but continuous, and hence the distribution to fit the data is continuous too. Actually, we could dichotomize the speed variable, as we have seen in Chapter 2, by using some established cutoffs to classify liver fibrosis in stages from F0 to F4. Since it needs just a binary classification, one solution could be to focus attention not on liver fibrosis but on liver cirrhosis. Moving to cirrhosis, it could be possible to split up observations in cirrhotic and non-cirrohtic choosing 2 m/s as a binary threshold (Table

7.2). Hence set $D = 0$ if $Y < 2$ and $D = 1$ if $Y \geq 2$.

Stage	ARFI (m/s)	Dataset	2013
F0-F1-F2-F3	$0 - 2$	30112 (80%)	3181 (77%)
F4	≥ 2	7547 (20%)	948 (23%)

Table 7.2: Dichotomization of ARFI values

In this way the response variable will be dichotomized, of course it will lead to a great loss of information. This categorization will led to a ROC curve for our dataset but following this approach it is necessary to fit statistical models just using a binary distribution for the response, coming back to a logistic regression model.

For this reason a new approach is proposed, in which it is possible to use the ROC curve starting from a model with continuous response variables.

ROC curves are suitable to binary data because in logistic regression FPF and TPF are computed starting by fitted values of $\hat{p} = P(Y = 1)$ in a range (0,1). The difference between logistic regression and GAMLSS is that, fitted values for data is not ranged in (0,1) but in $(0, \infty)$, so it is necessary to calculate $\hat{p} = P(Y > 2)$, where 2 is the threshold for diagnosing cirrhosis.

It is made possible by considering the density function of the chosen distribution for GAMLSS. As seen in the previous chapters, this distribution is BCPEo, the original Box-Cox Power Exponential with log link for μ function (Rigby and Stasinopoulos, 2004).

In the proposed approach, for values in the estimates (0,1), $\hat{p} = P(Y > 2) = 1 - P(Y \leq 2) = 1 - F(2|\mu = \hat{\mu}, \sigma = \hat{\sigma}, \nu = \hat{\nu}, \tau = \hat{\tau})$ are obtained using the difference between 1 and the density function of BCPEo distribution at an established cut-off (2 m/s), where parameters are the fitted values computed for GAMLSS model. Using this approach there exists a direct correspondence between each observation y and a probability $\hat{p}$ that lies in

(0,1). Then we can use these n probabilities to derive the ROC curve. Using a ROC curve in GAMLSS has a double aim: first, to justify the use of this approach compared with the standard logistic regression and secondly, to compare distributions by using the same method with other distributions for the response variable.

In order to obtain the ROC curve as a prediction tool using these data, it is necessary to split up the dataset in two subsets: the training and the validation set.

Definition of training and validation sets has to take into account that measurements belong to different subjects, therefore observations on the same subject have to fall in the same subset. This constrain can be easily reached by sampling not for measurements but for subjects. Considering the 2013 dataset, 141 subjects were analyzed. We decide to sample 100 individuals for training set and the remaining 41 for the validation. This is equivalent to split up the whole dataset in 70% and 30%.

The selected GAMLSS in Chapter 5 represents the starting point for estimating the ROC curve. This model was fitted on the complete dataset with different weights for training ($w = 1$) and validation ($w = 0$) individuals. Predicted values $\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}$ were extracted for this weighted model. Predictor values for each parameter are included in $1 - F(cutoff)$ where F is the density function for BCPEo and the selected cutoff is 2.

$$\hat{Y} \rightarrow \hat{p} \qquad\qquad \hat{p} = 1 - F(2|\mu = \hat{\mu}, \sigma = \hat{\sigma}, \nu = \hat{\nu}, \tau = \hat{\tau})$$

This is shown in detail in Figure 7.2 where an approximation of the density function for the speed variable is plotted and the coloured area is the probability $\hat{p}$ derived from the model.

Step procedure to implement ROC curve predictions in GAMLSS

- Sampling for individuals
- Fit a GAMLSS weighted model
- Extract predicted values $\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}$ for each y and evaluate $\hat{y}$
- Transform $\hat{y}$ to $\hat{p}$ only for observations in the validation set
- Compute specificity and sensitivity and draw ROC curve using the true indicator of cirrhosis D and the fitted probabilities $\hat{p}$ for each observation y in the validation set

A vector of probabilities $\hat{p}$ has been obtained and now it is possible to use the same procedure used in binary logistic regression to compute accuracy, sensitivity and specificity of the prediction for the validation set. To do this, it is enough to compare these $\hat{p}$ with the binary classification in Table 7.2. Now that classification is useful because, for validation set that represents the true value $D = 0$ (healthy) or $D = 1$ (cirrhotic), while $\hat{p}$ represents prediction probabilities to have or not cirrhosis.

Since for prediction training and validation sets are used, the sampling could affect results; for this reason, the sampling procedure has been repeated 50 times to make more accurate predictions and to obtain more robust results.

As seen in the previous Section, the use of this approach needs to be validated comparing it with standard ROC curve of binary logistic regression (LR). Secondly, it is possible to compare GAMLSS also with other statis-

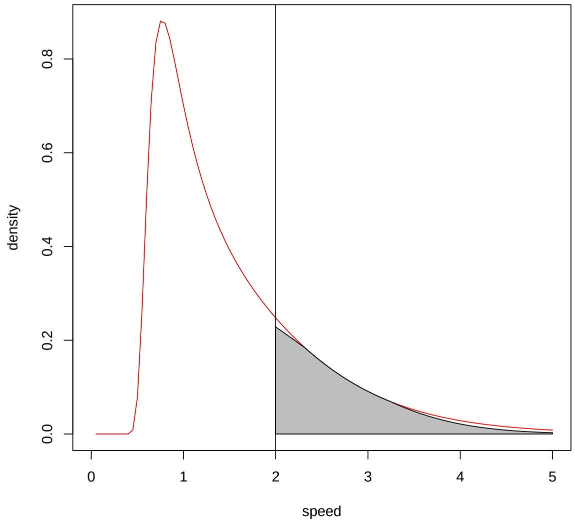

Figure 7.2: From GAMLSS fitted values to predicted probabilities

tical models. For the comparison we select LMM (Linear Mixed Model) and a Gamma response GLMM (Generalized Linear Mixed Model) since the response variable distribution is positive skewed.

Two possible ways of comparing different statistical models are possible using ROC curves. The first one is a graphical comparison, where different ROC curves are drawn in order to identify the higher curve. The higher the curve, the better the prediction. Secondly, the AUC index can be computed for all models; the model with a higher AUC index will be better.

In Figure 7.3 for a single training and validation sample, ROC curves are shown for different statistical models. As it is possible to observe, the blue one representing GAMLSS with BCPEo is slightly above all the other curves, even if for small values of FPF some ROC curves cross. However all the curves are above the diagonal line, this means that our prediction test is informative and better than a random guess. Graphical comparison is not enough because it could depend on the chosen sample. Using the AUC index it is possible to compute an average measure for each model for all samples.

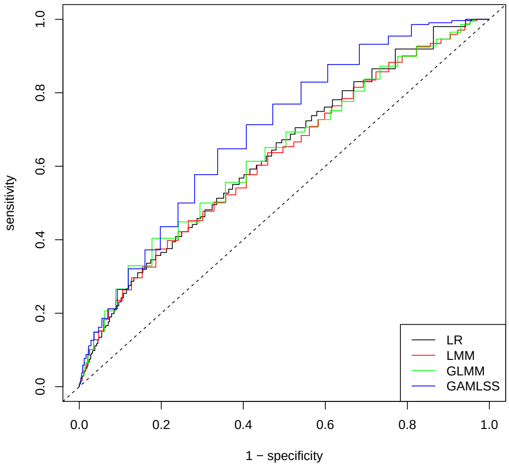

Figure 7.3: ROC curves for different statistical models

In Table 7.3, average values for AUC over 50 test and validation samples are presented. The average AUC for GAMLSS is the best among the statistical models and it is the only one above 0.70. Match pairs t-tests were conducted to test the hypothesis that GAMLSS average AUC is greater than others and all p-value are less than 0.01.

In Figure 7.4, on the left the AUC indices for 50 training and validation samples are shown on the same graph. On the x-axis we have the sampling index and each point represents a resulting value of AUC. It is possible to

LR	LMM	GLMM	**GAMLSS**
0.674	0.669	0.678	**0.701**

Table 7.3: Mean of AUC

note that blue points denoting GAMLSS are the highest point in most of the cases (44 over 50). On the right, boxplot for 50 AUC indexes are displayed. The yellow box-plot related to GAMLSS is the higher as we could expect from previous considerations. In terms of variance all statistical models have similar interquartile ranges for AUC so no problems of stability are detected.

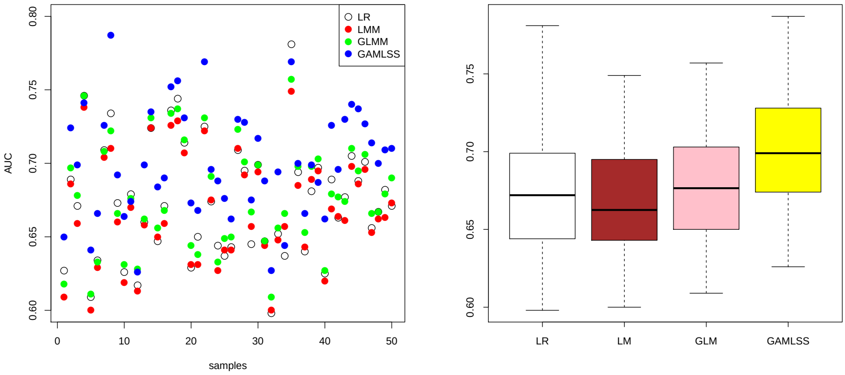

Figure 7.4: AUC indices for 50 test and validation samplings

Chapter 8

Conclusions

8.1 Summary and conclusions

When the research project has been submitted, the aim was to find a way to estimate and evaluate liver stiffness using a new biomedical technology detecting hepatic diseases.

The main problem in dealing with liver fibrosis is the heterogeneity of observations on the disease; so we need some tools to manage it. ARFI (Acoustic Radio Force Impulse) is proposed as a way to catch the heterogeneity replicating a substantial number of measurements in different part of the liver. GAMLSS have been proposed as a way to estimate the heterogeneity derived by ARFI measurements.

The main aim of the thesis is to use diagnostic tools in a double way. Firstly, from a medical point of view, ARFI is used as diagnostic tool to detect and classify liver fibrosis and cirrhosis. Secondly, from a statistical point of view, the worm plot is used as a diagnostic tool in GAMLSS for liver fibrosis data.

Some difficulties arose since it was the first time that such a large and complex dataset was collected using ARFI. The particular framework of the full dataset led us to consider a subset taking into consideration only the year 2013. This reduction was also suggested by the need to limit the computational cost.

A first simple linear GAMLSS was applied to the 2013 dataset considering as predictors both patient level and exam level variables. Since this model presents a too complicated structure, some enhancements have been introduced in order to get an easier model. After the inclusion of a random effect component due to the presence of repeated measurements for patient, the final model has been obtained. The most important predictors are age of the subject, depth of the measurements and the liver segment. In particular, depth has a negative effect and the segment numbered 7 is significantly different from others. Besides, the position of the patient during the exam has resulted as an important predictor of the variability of the liver stiffness. Finally the age variable has also a positive effect on the skewness and kurtosis of the speed response variable.

Two statistical extensions are provided to develop the use of GAMLSS in analyzing liver fibrosis data. The first extension concerns about the use of a mixture approach in GAMLSS, for the first time this method is used in GAMLSS to deal with bimodality in the response variable. Furthermore, we try to find a relationship between the identified mixture components and the state of a normal or cirrhotic liver. Some simulation studies have been carried out considering several scenarios; the use of mixture model in GAMLSS has been shown to produce good results in terms of goodness of fit and diagnostics.

The second extension refers to the implementation of ROC (Receiver Operating Characteristic) curve in GAMLSS as prediction tool. The idea is to dichotomize the response variable in two categories: cirrhotic and non-cirrhotic measurements. In the proposed approach, (0,1) ranged values $\hat{p}$ are obtained using the density function of BCPEo distribution at an established cut-off (2 m/s), where parameters $\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}$ are the fitted values computed for GAMLSS model. Splitting up the dataset in training and validation set, 50 samples have been extracted and ROC curve and AUC (Area Under the Curve) indexes have been computed. In order to make a comparison with other classes of statical models, AUC has been computed for LR, LMM and GLMM. In most of the cases AUC for GAMLSS assumed the higher value.

8.2 Future work

Several problems are still open in studying liver fibrosis. First of all, a comparison with datasets related to other years have to be conducted in order to have more stable results and to confirm interpretation of the estimated parameters. The possibility to collect data about laboratory tests (ALT, AST, GGT and platelets) and their inclusion in a GAMLSS could enhance the knowledge of the different stages of liver fibrosis and the relationships of these markers with the used predictors.

From a methodological point of view, the use of an approach for censored data is suggested since the liver stiffness seems apparently ranged only for positive values. Actually, since the speed has been measured as shear wave propagation on a tissue, for physical reasons the minimum possible value is around 0.5 m/s. Then, a possible development could concern the use of a GAMLSS for left-censored data.

About the mixture approach, the possibility to solve bimodality problems using this method should be verified through similar datasets and more simulation studies; this might cover, for example, simulated data from other probability distributions, different from the BCPE.
Finally, the implementation of ROC curve in GAMLSS should be improved since for high values of specificity the curves of the compared models cross each other and some constrains about crossing should are requested.

Bibliography

Aitkin, M. A., Francis, B. and Hinde, J. (2005). *Statistical modelling in GLIM 4*, volume 32. Oxford University Press.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19**(6), 716–723.

Alonzo, T. A. and Pepe, M. S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, **3**(3), 421–432.

Arena, U., Vizzutti, F., Corti, G., Ambu, S., Stasi, C., Bresci, S., Moscarella, S., Boddi, V., Petrarca, A., Laffi, G. *et al.* (2008). Acute viral hepatitis increases liver stiffness values measured by transient elastography. *Hepatology*, **47**(2), 380–384.

Attanasio, M., Enea, M. and Rizzo, L. (2010). Some issues concerning the statistical evaluation of a screening test: the ARFI ultrasound case. *Statistica*, **70**(3), 311–322.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **88**(421), 9–25.

Calvaruso, V., Cammà, C., Di Marco, V., Maimone, S., Bronte, F., Enea, M., Dardanoni, V., Manousou, P., Pleguezuelo, M., Xirouchakis, E. *et al.* (2010). Fibrosis staging in chronic hepatitis C: analysis of discordance between transient elastography and liver biopsy. *Journal of viral hepatitis*, **17**(7), 469–474.

Carey, E. and Carey, W. D. (2010). Noninvasive tests for liver disease, fibrosis, and cirrhosis: Is liver biopsy obsolete? *Cleveland Clinic journal of medicine*, **77**(8), 519–527.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in medicine*, **11**(10), 1305–1319.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**(3), 236–244.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007.

Everitt, B. S. and David, J. (1981). Finite mixture distributions. *Monographs on Applied Probability and Statistics. Chapman and Hall, London, New York.*

Friedrich-Rust, M., Wunder, K., Kriener, S., Sotoudeh, F., Richter, S., Bojunga, J., Herrmann, E., Poynard, T., Dietrich, C. F., Vermehren, J. *et al.*

(2009). Liver fibrosis in viral hepatitis: Noninvasive assessment with Acoustic Radiation Force Impulse imaging versus Transient Elastography 1. *Radiology*, **252**(2), 595–604.

Goertz, R., Egger, C., Neurath, M. and Strobel, D. (2012). Impact of food intake, ultrasound transducer, breathing maneuvers and body position on acoustic radiation force impulse (ARFI) elastometry of the liver. *Ultraschall in der Medizin (Stuttgart, Germany: 1980)*, **33**(4), 380–385.

Grushka, E. (1972). Characterization of exponentially modified gaussian peaks in chromatography. *Analytical Chemistry*, **44**(11), 1733–1738.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, volume 43. CRC Press.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). Continuous univariate distributions, vol. 1-2.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lee, Y. and Nelder, J. A. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 619–678.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37. CRC press.

McDonald, J. B. and Xu, Y. J. (1995). A generalization of the Beta distribution with applications. *Journal of Econometrics*, **66**(1), 133–152.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **135**(3), 370–384.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Ratziu, V., Charlotte, F., Heurtier, A., Gombert, S., Giral, P., Bruckert, E., Grimaldi, A., Capron, F., Poynard, T., Group, L. S. *et al.* (2005). Sampling variability of liver biopsy in nonalcoholic fatty liver disease. *Gastroenterology*, **128**(7), 1898–1906.

Rigby, R. and Stasinopoulos, D. (2001). The GAMLSS project: a flexible approach to statistical modelling. In *New trends in statistical modelling: Proceedings of the 16th international workshop on statistical modelling*, pages 337–345.

Rigby, R. A. and Stasinopoulos, D. M. (2004). Smooth centile curves for skew and kurtotic data modelled using the box–cox power exponential distribution. *Statistics in medicine*, **23**(19), 3053–3076.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(3), 507–554.

Rigby, R. A. and Stasinopoulos, D. M. (2006). Using the box-cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, **6**(3), 209–229.

Rizzo, L., Calvaruso, V., Cacopardo, B., Alessi, N., Attanasio, M., Petta, S., Fatuzzo, F., Montineri, A., Mazzola, A., L'abbate, L. *et al.* (2011). Comparison of transient elastography and Acoustic Radiation Force Impulse for non-invasive staging of liver fibrosis in patients with chronic hepatitis C. *The American journal of gastroenterology*, **106**(12), 2112–2120.

Schall, R. (1991). Estimation in Generalized Linear Models with random effects. *Biometrika*, **78**(4), 719–727.

Schwarz, G. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 47–60.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape GAMLSS in r. *Journal of Statistical Software*, **23**(7), 1–46.

Titterington, D. and Smith, A. (1985). Statistical analysis of finite mixture distributions.

Van Buuren, S. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, **20**(8), 1259–1277.

Venables, W. and Ripley, B. (2002). Modern applied statistics with S. 4th edition.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J. and Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, **39**(6), 1279–1293.