USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

UNIVERSIDADE DA CORUÑA

Universidade Vigo

Master Thesis

Bivariate copula regression models in diabetes research

Óscar Lado Baleato

Máster en Técnicas Estadísticas

Curso 2016-2017

Propuesta de Trabajo Fin de Máster

Título en galego: Modelos de regresión para respostas bivariantes baseados en cópulas aplicación ao estudo da diabetes.

Título en español: Modelos de regresión para respuestas bivariantes basados en cópulas aplicación al estudio de la diabetes.

English title: Bivariate copula regression models in diabetes research

Modalidad: Modalidad B

Autor/a: Óscar Lado Baleato, Universidad de Santiago de Compostela

Director/a: Carmen María Cadarso Suárez, Universidad de Santiago de Compostela ; ,

Tutor/a: Francisco Gude Sampedro, SERGAS,

Breve resumen del trabajo: Las funciones cópula se han convertido en una potente herramienta para el modelado multivariante al recoger la estructura de dependencia de dos variables, su uso en regresión permite modelizar el efecto de un conjunto de covariables sobre una variable respuesta bivariada. Durante la realización de este trabajo se usarán modelos generalizados aditivos generalizados de localización,escala y forma GAMLSS de respuesta multivariante con cópulas a un caso de biomedicina usando datos disponibles en la Unidad de Epidemiología Clínica del Hospital Clínico Universitario de Santiago de Compostela.

Recomendaciones:

Otras observaciones:

Doña Carmen María Cadarso Suárez Catedrática de Estatística de la Universidad de Santiago de Compostela   don Francisco Gude Sampedro Director de la Unidad de Epidemiología Clínica (CHUS) informan que el Trabajo Fin de Máster titulado:

Bivariate copula regression models in diabetes research

fue realizado bajo su dirección por don Óscar Lado Baleato para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago, a 4 de septiembre de 2017.

La directora:

El tutor:

Doña Carmen María Cadarso Suárez

Don Francisco Gude Sampedro

El autor:

Don Óscar Lado Baleato

Agradecementos/Acknowledgements

A profesora Carmen Cadarso propúxome realizar este traballo o 30 de Xuño as 11 da mañá na Facultade de Medicina, recordo o día, porque estábase a celebrar o congreso encontro galaico-portugés de biometría en Santiago, e ese mesmo día as 5 da tarde tiven a oportunidade de asistir a unha conferencia de Thomas Kneib sobre o tema do que versa este traballo, a regresión para respostas bivariadas baseada en cópulas. Nunca tan motivado estiven como aquel día para seguir unha conferencia, posto que tiña que escribir a proposta de TFM para entregar catro días despois e pensei que esa conferencia me sería de axuda, tampouco nunca tanto medo lle ganei a un traballo cando me empecei a dar conta do complexo que era todo aquilo.

Teño que agradecer a profesora Carmen Cadarso por propoñerme facer este traballo e dirixirme no proceso no que se perde o medo, e sobre todo por tratarme como alguén máis do seu equipo, confiando en min dende o primeiro día.

Ao doutor Francisco Gude, primeiramente pola súa paciencia, durante o proceso da construción do modelo chequeou tantos resultados como modelos axustei eu, intentando buscarlle xeito a cousas que non o tiñan debido a miña inexperiencia. Ademáis como proxecto de científico que son debo agradecerlle que sempre reciba cos brazos abertos a todo o que queira traballar sobre calquera problema biomédico, brindando a todos os estudantes do máster ou da carreira de matemáticas os datos que tanto traballo costa recoller.

A doutoranda Jenifer Espasandín por confiar en mín e contar conmigo como colaborador en moitos traballos. Por botarlle un ollo a este TFM e ademais por darme sempre consellos acertados, a próxima vez fareiche máis caso.

To the Professor Giampiero Marra: I appreciate the sharing of preliminary versions of your work and the different variations developed during these months. In addition, I show you my gratitude for your prompt and enthusiastic answers to all my questions. Grazie tante.

A Alesandra por leer esto que non lle gusta nada, para axudarme a revisar o inglés deste traballo. E por pasar comigo esta última tarde de nervios. E sobre todo por animarme a facer este máster a pesar do medo que lle tiña.

A miña Nai por todas as becas, interships, e bolsas que me deu ata agora, e terme animado dende pequeno a estudiar.

O meu colega Antón por estes dous anos nos que descubrimos a estadística xuntos.

Finalmente a todo o profesorado do Mestrado de Técnicas Estadísticas, os máximos responsables de que eu, dous anos despois de non entender porque o número de sementes dispersada por unha formiga é un proceso de Poisson, agora estea presentando este traballo sobre modelos de regresión para respostas bivariadas baseados en cópulas, no contexto da regresión distribucional.

Xa para rematar a todos os gandeiros que fixeron folga e encheron o camiño de Tordoia a Santiago de tractores a mediados de setembro do ano 15 impedíndome chegar a UXA antes das dúas da tarde, se non fose por vós teríame cambiado de máster ese día.

Contents

Summary **xi**

1 Introduction **1**

2 A Case Study in Diabetes Research **3**

2.1 Glucose homeostasis and diabetes . . . 3

2.2 Glycated proteins and glycemic control . . . 5

2.3 AEGIS project . . . 6

3 Generalized Additive Models of Location, Scale, and Shape (GAMLSS) **9**

3.1 Generalized Additive Models (GAM) . . . 9

3.2 GAM for Location, scale and, shape (GAMLSS) . . . 11

3.3 Univariate modelling of glycemic markers . . . 12

4 Copula functions **19**

4.1 Sklar theorem . . . 19

4.2 Parametric families of the copula functions . . . 20

4.2.1 Archimidean copulas . . . 21

4.2.2 Elliptical copulas . . . 25

4.2.3 Other copulas . . . 26

4.2.4 Survival and other rotated copulas . . . 27

4.3 Concordance measurements and copulas . . . 28

5 Copula GAMLSS for bivariate responses **31**

5.1 Model formulation . . . 31

5.2 Estimation process . . . 33

5.3 Model building process . . . 37

6 Simulation study **39**

6.1 Performance of the model at different sample sizes . . . 39

6.2 Different strengths of dependence between margins . . . 45

6.3 Copula missespecification . . . 47

7 Bivariate modelling of glycemic markers **51**

7.1 Model Building . . . 51

7.2 Results . . . 54

8 Discussion **59**

A Marginal distributions **61**

65

Summary

Resumo en galego

A diabetes é unha enfermidade crónica caracterízada pola presenza dunha concentración elevada de glicosa no sangue. Esta enfermidade ten un gran número de complicacións derivadas dun diagnóstico tardío ou mal control da glicemia (glicosa en sangue).

A determinación da hemoglobina glicada (HbA1c) considerase a proba “gold standard” no control da diabetes. Sen embargo os seus niveis poden variar considerablemente, incluso entre individuos que presenten os mesmos niveis de glicemia. Ademáis da hemoglobina, outras proteínas presentes no sangue poden ser glicadas e usadas como marcadores da glicemia, sendo a albumina glicada a principal alternativa a HbA1c. Sen emabargo, as discorndancias entre os niveis de ambos marcadores son comúns e os investigadores mdicos están interesados en coñecer cal é a causa.

Para investigar de forma conxunta os factores que poidan influir nas discordancias atopadas entre os niveis de ambas medidas de control glicémico, proponse o uso de modelos flexibles para respostas bivariadas baseados en funcións cópula (CGAMLSS) (Marra e Radice, 2017). Estes modelos extenden o uso do GAMLSS (Rigby e Stasinopoulos, 2005), a respostas bivariadas, modelizando de forma flexible todos os parámetros que definen unha resposta bivariada construida por medio dunha función cópula.

Neste traballo de fin de mestrado o uso desta metodoloxía deu lugar a resultados non descritos previamente do efecto de distintas variables sobre a asociación entre os niveis de ambas proteínas glicadas.

English abstract

Diabetes is a serious chronic disease in which the blood glucose become chronically elevated, a delayed diagnosis or poor glycemic control lead in the long term, to serious complications.

Glycated haemoglobin (HbA1c) is considered the gold standard test in the glycemic control. Nevertheless, it has been seen that HbA1c levels can vary considerably, even among individuals with similar mean blood glucose levels. In addition to HbA1c, other circulating proteins can also become glycated and can therefore also be used as a marker of blood sugar levels, being the glycated albumin (GA) the main alternative to he HbA1c. However, discordances between the levels of HbA1c and GA are often encountered and clinicians are aware of the conditions that might explain them.

To simultaneously investigate factors that may influence on the discordance between measures of glycemic control, the bivariate additive conditional copula regression models (CGAMLSS) are proposed. This novel approach extends the use of GAMLSS (Rigby and Stasinopoulos, 2005) to situations in which each parameter of a bivariate response built using a copula function are modeled using additive predictors in a flexible way.

In this master thesis, the use of these models have shown hitherto unreported effects, in the association between both glycemic markers.

Chapter 1

Introduction

Regression modeling strategies have been widely applied in biomedical studies since they enable the identification and characterization of the relationship between a variable of interest y and a covariate x or a set of k covariates $x_1, \ldots, x_k$ identifying statistical associations between them that could be useful for a better understanding of different disase mechanisms. However, regression models are generally based on some assumptions that must be true to get valid conclusions from them, and which rarely holds in real data applications.

In order to get more flexible approaches to apply in real problems, different models have been proposed overcoming the less realistic assumptions of the Linear Models, where a Gaussian distribution was assumed for the response, with a constant variance and linear relationship between the predictor and the response variable. The development of the Generalized Linear Models (GLM; McCullagh and Nelder, 1989) relaxed the distributional assumption from gaussianity to any distribution from the exponential family and the proposal of the Generalized Additive Models (GAM; Hastie and Tibshirani, 1990; Wood, 2006) allowed to model the relation between covariates and the response variable in a non-linear way using additive predictors. However, the exponential distributional assumption and variance of the response independent from the covariates were maintained until the development of the Generalized Additive Models for Location, Scale and Shape (GAMLSS).

The GAMLSS regression models were proposed by Rigby and Stasinopoulos (2005) relaxing the exponential family assumption for the response variable by a very general distribution family and modelling each parameter of the distribution, not just the mean, in function to the explanatory variables in a non-linear way. GAMLSS let us model, virtually any variable found in biomedical studies, including quite skewed ones. Moreover, the modelization of the entire distribution of the response variable in function to a set of explanatory variables relaxed the assumption of fixed variance in the response that is quite useful in the clinical applications, where is common to find response variables whose variability changes with other covariates like age or gender.

Therefore most of the biomedical problems can be studied using the univariate regression framework introduced above, modeling the whole distribution of the response conditioned to a set of covariates in a non-linear way and with no restriction in the distribution of the response variable. However, in some situations is necessary study two responses jointly, regressing a bivariate response on a set of covariates in a non-linear way, in order to assess which factors modify the relationship between the responses, in addition to the changes in mean and variability. These can be achieved extending the flexible GAMLSS models to bivariate responses, using copula functions (Sklar, 1953). Such copula functions enable the construction of bivariate response from two given and arbitrary margins from the GAMLSS class. Furthermore, the parametric copulas are defined by one parameter measuring the strength of dependence between the marginals and that can be modeled as a function to a set of covariates as well. Also, the copula allows us to consider non symmetric structures of dependence between the responses.

From the different regression approaches for bivariate responses based on the use of copula functions and considering non-exponential responses in the margins, proposed in the literature, either in the bayesian (Klein and Kneib, 2016) or frequentist framework (Vatter and Cahez-Demoulin 2015; Yee 2015) we have chosen the recent proposal Bivariate copula additive models for location, scale and shape (CGAMLSS) (Marra and Radice, 2017a). In this type of models, the covariates effect on a bivariate response, built using a copula function, are estimated at the same time using a trust region algorithm and with integrated automatic multiple smoothing parameter selection for the smooth terms, overcoming the limitations of the models proposed in Vatter and Chaved-Demoulin (2015) based on a less efficient two-step estimation, and Yee (2015) models without an automatic way of chose the smoothing parameter. Furthermore, the only alternative that presents a simultaneous estimation and a automatic way of chosing the smoothing parameter for the smooth terms, is the Bayesian model proposed by Klein and Kneib (2016). However, these models present a high computational cost that is a limitation in the model building process.

Thereby the CGAMLSS approach is currently the most flexible framework to model bivariate responses, with more marginal distributions and copulas available, presenting a software implementation in the R package `GJRM` incorporating the utilities and syntax of the well known `mgcv` package, supposing a huge methodological advance, necessary to fill a gap in the study of some of the open problems in biomedicine. In the case of this master thesis the Clinical Epidemiological Unit from the Clinical Hospital of Santiago de Compostela, required the CGAMLSS approach to solve a problem derived from the AEGIS project (Clinical Trials: NCT01796184) to study which factors modify the concordances between the levels of two glycated proteins, the glycated hemoglobin (HbA1c) and glycated albumin (GA), used to monitor the Diabetes mellitus disease.

Diabetes mellitus is a disease characterized by the disability of human body to regulate blood glucose concentration, with a global incidence of 422 million four times more than in 1980 (WHO, 2016). An early diagnosis and strict glycemic control is essential to prevent diabetes complications such as cardiovascular disease, nephropathy or retinopathy. This control is based on the fasting levels of glucose and in the concentration of different glycated proteins, being the most used the HbA1c and the GA. However, discordances between the HbA1c and GA determinations are common in clinical practice and remain unexplained, given the number of HbA1c test performed in the world, studying which factors modify the association between the levels of both proteins will improve the diagnosis and control of diabetic patients.

To analyze which factors are involved in the discordances encountered between the levels of HbA1c and GA we propose the use of the CGAMLSS regression models, using the data collected in the AEGIS project. This is a cross-sectional population based study conducted between the years 2012 to 2015 in order to study the differential glycation of proteins. As pointed above the CGAMLSS approach will let us model the levels, variability and discordances between both glycemic markers without a restrictive distributional assumption for the responses and with the possibility of consider non-standard structures of dependence between them.

This master thesis is structured as follows: in Chapter 2 the medical problem that has motivated studying jointly the concentration of two glycemic markers is presented. In the Chapter 3 the GAMLSS regression is introduced to better understand their bivariate extension and will be applied to study the main cofounders that modify the glycation of proteins. In Chapter 4, copula functions are presented with the main parametric families used in CGAMLSS models. In Chapter 5, the CGAMLSS models are introduced. Chapter 6, presents a simulation study, in order to analyze the general performance of the CGAMLSS models. Chapter 7 show application of the CGAMLSS regression models to a medical problem. Finally the master thesis ends with a final discussion.

Chapter 2

A Case Study in Diabetes Research

Diabetes is a metabolic disease that occurs either when the pancreas does not produce enough insulin (a hormone that regulates blood glucose concentration), or when the body cannot effectively use the insulin it produces (WHO, 2016), resulting in an abnormal blood glucose control that becomes chronically elevated. Raised blood glucose, may, over time, lead to serious damage to the heart, blood vessels, eyes, kidneys and nerves. In order to avoid such complications is necessary an early diagnose and a frequent evaluation of the glycemic levels.

This evaluation of the glycemic status take advantage of a chemical reaction, known as glycation, that is a spontaneous attachment of a sugar molecule to proteins. In the case of the blood circulating proteins, are glycated according to the blood glucose levels deserve as circulating glycemic markers that offer us a summary measure of the mean blood sugar levels to which the blood proteins were exposed during their circulating life-span, that can be minutes, days or even months.

From the different circulating glycated proteins the most studied and used in the clinical practice are the glycated hemoglobin (HbA1c) and glycated albumin (GA), however discordances between their estimations in the assessment of glycemia is often encountered and clinicians are aware of the conditions that might explain it.

In this Chapter the main ideas of the glucose homeostasis are presented. Then the use of glycated proteins as glycemic markers is explained. Finally an outline of the AEGIS project is given.

2.1 Glucose homeostasis and diabetes

Glucose is an essential metabolic substrate of all mammalian cells. D-glucose is the major carbohydrate presented to the cell for energy production, supplying its catabolic and anabolic requeriments (Szablewski, 2011) and for some organs like the brain is the obligate metabolic fuel.

In fact, the brain is dependent on a continuous supply of glucose from blood plasma, because of the low circulating concentrations of other possible alternative substrates (e.g., ketone bodies) and due to transport limitations across the blood-brain barriers (e.g., free fatty acids) which separate the circulating blood from the brain extracellular fluid in the central nervous system. Moreover, the brain cannot synthesize glucose or store it as glycogen more than for a few minutes supply (Shrayyef and Gerich, 2010). A low concentration of glucose in plasma (hipoglycemia) impairs cerebral function and prolonged hypoglycemia causes convulsions, permanent brain damage and even death.

On the other hand, even midly elevated plasma glucose concentrations, can produce celular damages in the blood vessels and multiple organs because of the toxic effects of glucose, in the short term due to its osmotic activity and in the long term to its binding to different blood proteins (Serrano-Rios and Gutierrez-Fuentes, 2009) and contribution to the oxidative stress related to cancer disease. Thus the maintainance of normal plasma glucose levels, glucose homeostasis, is crucial to mantain the normal function of metabolism and physiology.

Plasma glucose levels are determined by the relative rates at which glucose enters and leaves the bloodstream (Shrayyef and Gerich, 2010) the mantaining of relatively stable plasma glucose levels requires precise matching of glucose utilization by the tissues, the glucose derived from the diet and in a minor proportion the endogenous glucose produced in the liver and kidneys through Glucongenolysis and Gluconeogenesis (Giugliano et al, 2008).

This regulation of the glycemia is conducted mainly by two hormones produced in the pancreas, the insulin and glucagon. The pancreas is a secretory organ located behind the stomach within the left upper abdominal region. Most part of this organ consists of exocrine cells that secrete different digestive enzimes in the duodenum and only 1.2% of the entire organ presents an endocrine action, releasing hormones into the blood stream. This endocrine cells are clustered together assembling the so-called islets of Langerhans, which are small, island-like structures that contain five different cell types releasing several hormones from the endocrine system, but the main ones are glucagon-producing $\alpha-$cells and insulin-producing $\beta-$cells.

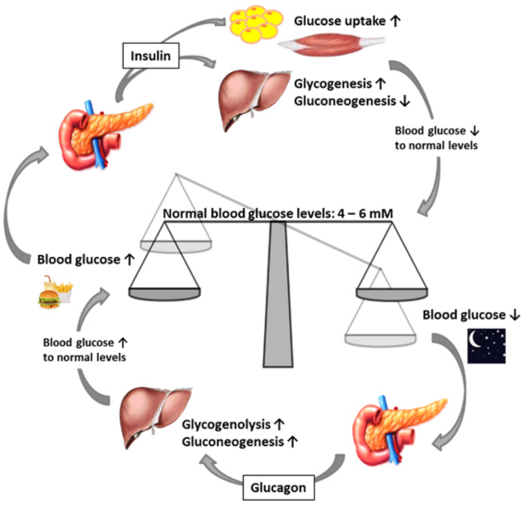


Figure 2.1: Regulation of the blood glucose concentration (Roder et al, 2016)

The pancreatic glycemic control is based on the action insulin and glucagon that work in an antagonic way. As it is shown in Figure 2.1 during sleep or in between meals, when blood glucose levels are low, glucagon is released from $\alpha-$cells to promote hepatic glycogenolysis. In addition, glucagon leads hepatic and renal gluconeogenesis to increase endogenous blood glucose levels during prolonged fasting. In contrast, insulin secretion from $\beta-$cells is stimulated by elevated exogenous glucose levels, such as those occurring after a meal. After docking to its receptor on muscle and adipose tissue, insulin enables the uptake of glucose into these tissues and hence, reduces blood glucose levels by removing the exogenous glucose from the blood stream. Furthermore, insulin promotes glycogenesis, lipogenesis and the incorporation of amino acids into proteins; thus, it is an anabolic hormone, in contrast to the catabolic activity of glucagon (Roder et al, 2016).

The well known Diabetes mellitus disease is a malfunction in the glucose homeostasis caused by a failure in the endocrine action of the insulin, being the intake of sugars by the tissues reduce and hence the glucose remains in the circulating blood. This failure in the endocrine action of the insuline, can be caused mainly by two mechanism that define two sub-types of the disease:

- Type 1: the less common diabetes, caused by the destruction of the $\beta-$cells within the islets of Lagerhans as a result of autoinmune response with multiple genetic predispositions and is also related to enviromental factors that are still poorly defined.
- Type 2: most common common and with an increasing incidence. It is highly related to the life style of western society. This type of diabetes, curse with peripheral insulin resistance and a compensatory hypersecretion of insulin from the pancreatic islets, that leads to a progressive decline in the islet secretory function.

The chronic hyperglycemia, causes different complications in the diabetic patients derived mainly from the toxicity of the glucose molecule, that in short term produces an increase in the inflamation processes (Rekeneire et al, 2006) and raises the blood osmotic pressure (Liamis et al, 2014), that in worst cases lead to a osmotic coma and dead and in the long term causes "microvascular disease" (damages in tiny blood vessels) associated with kidney disease and retinopathy as well as "macrovascular disease" (damages in the arteries) associated to Cardiovascular disease (Forbes and Cooper, 2013) being the cause of death of more than the 70% of diabetic patients (Laakso, 2010).

The monitoring of the diabetic patients in order to maintain the blood glucose levels back in the normal range is then a quite important task since it is the only way to prevent the complications of the disease. This control has been clasiccaly conducted using the plasma glucose levels in the fasting condition but currently is mainly based on the circulating levels of different glycated proteins, since the last option offers a global measurement of the mean blood glucose levels of the patient in the recent past.

2.2 Glycated proteins and glycemic control

The reducing sugars as glucose react with the biological amines present in proteins and both molecules become covalent binding forming a glycated protein (Ulrich and Cerami, 2001; Morais et al, 2013). For the circulating proteins present in blood this reaction occurs constantly in a spontaneous way and with a glycation rate of the 0.01% per day being higher at higher glucose levels (Zang et al, 2009). Despite of the slow rate of this reaction, for the proteins that show a high life-span in the bloodstream (slow turnover) the glycated products are gradually accumulated and serve as circulating glycemic biomarkers that offer an indirect measurement of the patients glycemia in the recent past.

From the circulating glycated proteins, the glycated hemoglobin discovered in 1968 (Rahbar 1968, Rahbar 2005) and proposed as diagnostic test in 2009 (ADA, 2009) has become the gold standard for the diabetes control. The hemoglobin is located inside of the Red Blood Cells (RBCs) that circulate in the bloodstream between three to four months before being broken down and replaced. During that time, the hemoglobin can bond, irreversibly, to glucose present in blood. The levels of glycated hemoglobin within the RBCs therefore reflects the average level of glucose to which the cell has been exposed during its life cycle. Thus, HbA1C readings higher than about 6.5% indicate higher than normal amounts of glucose roaming the blood stream in the past 120 days.

However, falsely elevated HbA1c in relation to mean blood glucose concentration can be obtained when RBC turnover is decreased, exposing the cell to glucose for a longer period of time, resulting in higher HbA1c levels (Radin, 2013). Similarly, any condition that shortens the life of the erythrocyte or is associated with increased red cell turnover that abbreviates the exposure time of the cell to glucose, resulting in lower HbA1c levels (Radin, 2013). Conditions such as acute and chronic blood loss, hemolytic anemia, and splenomegaly can all cause falsely lowered HbA1c results (Natin, 2010) . Moreover, due to the long life of the RBC the glycated hemoglobin is not adequate for study short-term changes in the concentration of blood glucose.

As a result of the problems present by the HbA1c test other glycated proteins, like the glycated albumin (GA), have been introduced in the glycemic control (Lee, 2015). The GA is present in the

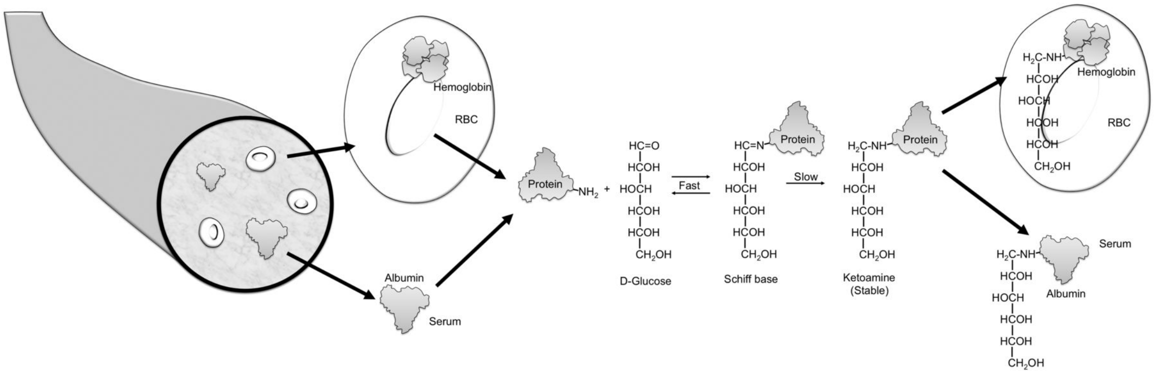


Figure 2.2: Formation of the HbA1c and GA, note that the first is in the celular space and the second in the intravascular space, (Welsh et al, 2016)

blood plasma and hence is not affected by the RBC life-span being proposed like a best glycemic control measurement in patients with anemia or acute renal failure (Inaba et al, 2007; Kobayashi et al, 2016; Wu et al, 2016). As the HbA1c the levels of these glycated protein are modified by non-glycemic factors like the Albumin turnover, a process enhanced by thyroid hormones and different drugs (Denese et al, 2015).

However, the levels of both glycemic markers are not concordant in all cases. These discordances encountered in the clinical practice can be partially explained by the factors that modify the life-span in blood of both proteins in a different way. However, other factor such as the use of tobacco and alcohol consumption have been reported to change the rate of glycation with independence of the glucose levels or time of exposition to it.

This problem of a differential glycation of proteins have been received much attention since the work of Cohen et al (2003). In this work the author proposed the difference between the levels of glycated hemoglobin predicted using other glycemic marker and the actual HbA1c levels, under the term of Glycation Gap (GG), as other marker to follow the diabetic patients, finding a relationship between the GG and the kidney disease.

This term of Glycation Gap and the necesity of developing statistical alternatives to analyze it, have motivated the AEGIS project in 2012.

2.3 AEGIS project

The A Estrada Glycation and Inflammation Study (AEGIS) is a cross-sectional study of a general population from A Estrada Municipality, selected by random sample, estratified by age 3500 persons, from these 428 were excluded due to the finalization of the study, 84 were dead, 211 did not response, 134 have changed their homes, 19 do not have medical asistance and 394 do not have the inclusion criteria because they suffered demencia, cerebrovascular disease, cancer, a terminal disease or inability of comunication. From these 1516 persons (55% Women, 45% Men) accepted to participate.

Between November 2012 and March 2015, all the participants went to the primary care centre from A Estrada in order to realize a clinic interview and determinations that included:

1. Demographic and Antropometric questionaire

2. Life-styles register, physical activity, diet, use of tobaco and alcohol consumption

3. Psicological tests

4. Periodontal analysis
5. Alergic tests
6. Blood sample
7. Continuous glucose monitoring (600 patients)

	Non-diabetic (1329)	Diabetic (187)
Male, n (%)	579 (45%)	99 (52%)
Age, years	50.61 (17.32)	66.56 (11.91)
BMI, kg/m^2	27.82 (4.93)	31.20 (5.00)
Alcohol		
Abstainers	478 (36%)	68 (37%)
Ligth Drinkers	541 (41%)	57 (30%)
Moderate Drinkers	201 (15%)	40 (21%)
Heavy Drinkers	109 (8%)	22 (12%)
Smoking		
Non-smokers	716 (54%)	109 (58%)
Ex-smokers	338 (25%)	57 (30%)
Smokers	275 (21%)	21 (12%)
Physical Activity		
Inactive	506 (38%)	90 (48%)
Minimally Active	482 (36%)	70 (37%)
HEPA-active	341 (26%)	27 (15%)
Glucose, mg/dL	88.91 (11.98)	134.24 (37.38)
Albumin, mg/dL	4.40 (0.23)	4.40 (0.23)
MCV, fL	89.75 (4.70)	90.16 (5.15)
Tyrosine (T3)	3.40 (0.41)	3.23 (0.41)
GA, %	13.67 (1.66)	17.53 (4.25)
HbA1c, %	5.41 (0.36)	6.96 (1.16)
Cor (HbA1c, GA)	0.12	0.71

Table 2.1: Characteristics of the population of study. The correlation between both proteins is presented in terms of Spearman ρ.

Regarding the variables considered in this study all laboratory analyses were performed on the day of sample collection in the Clinical Biochemistry Laboratory of the Hospital Clnico Universitario de Santiago de Compostela, Spain. The HbA1c was determined by high-performance liquid chromatography and the values obtained converted to Diabetes Control and Complications Trial-aligned units (Hoelze et al, 2004). The glycated albumin was estimated using the results of a glycated Serum Proteins Assay (Dyazyme Laboratories, 2017), that determines the concentration of all the glycated proteins in plasma and the levels of plasma albumin.

The Fasting Plasma Glucose was determined using the glucose oxidase peroxidase method. The Tyroid hormone T3, was determined using a competitive inmunoassay through direct chemiluminescent technology using the Advia Centaur XP, and the Mean Corpuscular Volume using a hematological autoanalizer ADVIA 2120.

Alcohol consumption was assessed using a questionaire where the patients were asked about the amount of alcoholic drinks consumed every week, using the standard unit system which consider 10g of alcohol for every glass of wine or bottle of beer and 20g for evey glass of a destilated liquor. Based on the grames of alcohol consumed per week the individuals were clasified in 1) Abstainers 0-9g, 2) Low drinkers 10-139g, 3) Moderate drinkers 140-279g and 4) Heavy drinkers 140-279g. Smoking habit was register as the number of cigarettes regularly consumed per day. Consumers of at least 1 cigarette per day were considered smokers, individuals who had quit smoking before the preceding year were considered ex-smokers and non-smokers the individuals who have never consumed tobacco. Moreover, the patients completed the International Physical Activity Questionnaire (short form) (IPAQ-Group, 2012), and based on the method described in Craig et al (2003) were clasified in 1) Inactive, 2) Minimally Active and 3) "HEPA-active", healt-enhancing physical activiy.

Chapter 3

Generalized Additive Models of Location, Scale, and Shape (GAMLSS)

The univariate smoothing regression techniques introduced by Hastie and Tibshirani (1990) have supossed a huge advance in the modelization of the relationship between a response variable and a set of covariates, letting us consider data-driven, non-linear relations between them. These models have been extended in Wood (2006) with contributions in the estimation method, automatic choosing of the smoothing parameter and estimation of the confidence intervals based on a bayesian approximation.

The extension of the smooth regression idea of the GAM models was made by the introduction of the GAMLSS models (Rigby and Stasinopoulos, 2005) letting us considerer any parameteric distribution in the response and model each of the parameters that define the response variable distribution.

In this Chapter the Generalized Additive Models (GAM) is introduced. Then, the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) is presented and applied to study which clinical variables or life-styles are involved in the mean levels and variability of both glycemic markers.

3.1 Generalized Additive Models (GAM)

These models were proposed by Hastie and Tibshirani (1990) to model the relationship between the continuous covariates and the response in a non-linear way being extended and popularized by Wood (2006). The general model specification is:

$$g(\mu_i) = X_i^* + \beta + s_1(x_{1i}) + s_2(x_{2i}) + \ldots + s_j(x_{ji}) \tag{3.1}$$

where

$$\mu_i \equiv E(Y_i)$$

Y_i is the response variable that must be a distribution form the exponential family, $g()$ is a known monotonic, twice differentiable link function that makes the response lie in the space of the parameters of the response, X_i^* is a row of the model matrix for any strictly parametric model components, β is the vector of parameters for the parametric model, and s_j are smooth functions of the covariates, x_j (Wood, 2006).

The smooth terms s_j are approximated by choosing a basis, defining the space of functions of which s is an element (Wood, 2006), representing the smooth term $s_j(x_j)$ as:

$$s_j(x_j) = \sum_{i=1}^{q_j} b_{ji}(x_j)\beta_{ji} \tag{3.2}$$

The model given in (2.1) is then can be represented as a linear model,

$$g(\mu_i) = X_i^*\beta + b_{1i}(x_1)\beta_{1i} + b_{2i}(x_2)\beta_{2i} + \ldots + b_{ji}(x_1)\beta_{ji} \tag{3.3}$$

In the original proposal of the GAM regression models (Hastie and Tibshirani, 1990) a backfitting algorithm was used to estimate the model parameters, then (Wood, 2006) proposed a better method defining the GAM model as a penalized GLM, in this case the coefficient estimates $\hat{\beta}$ of the model are obtained by minimizing the penalized least square objective:

$$||y - X\beta||^2 + \lambda_j \beta^T S_j \beta \tag{3.4}$$

In (3.4) the second term is a penalization of the smooth terms using their second derivative, we can see in Wood (2006) that the second derivarite of the basis in (3.2) is $\boldsymbol{\lambda\beta^T S\beta}$, where the λ parameter control the smoothness of the estimate. The automatic selection of these parameter is solved with different methods like UBRE or Cross-validation.

The uncertainty of the estimations is made in form of confidence intervals based on a bayesian approximation.

Despite of the flexibility that this approach has supposed to model the effect of continuous covariates in a response, the GAM models are just a penalized GLM (Wood, 2006) and thus the response must follow a distribution from the exponential family. A distribution is considered from the exponential family if can be written in the form (Fahremeir et al, 2013):

$$f(y|\theta, \phi) = exp(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \tag{3.5}$$

where a, b and c are arbitrary functions, ϕ is called dispersion parameter and represent the scale while θ is the "canonical parameter" that represents the location and which is made dependent on the covariates. The most commonly used distributions are:

Distribution		$(\theta(\mu)$	$b(\theta)$	ϕ
Normal	$N(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2
Bernoulli	$B(1, \pi)$	$\log(\pi/(1-\pi))$	$\log(1 + exp(\theta))$	1
Poisson	$P(\lambda)$	$log(\lambda)$	$\exp(\theta)$	1
Gamma	$G(\mu, \nu)$	$-1/\mu$	$-\log(-\theta)$	ν^{-1}
Inverse Gaussian	$IG(\mu, \sigma^2)$	$-1/(2\mu^2)$	$-(-2\theta)^{1/2}$	σ^2

Table 3.1: Some of the distributions belong to the exponential family, (Fahrmeir et al 2013).

These distributions let us handle with count and binary data but not with some continuous variables that appear in real data problems. Moreover, the ϕ parameter is not made dependent from the covariates, that is not so realistic in some cases.

3.2 GAM for Location, scale and, shape (GAMLSS)

The Generalized Additive Models for Location, Scale and Shape (GAMLSS) relaxed the exponential family assumption of the GLM and GAM models, and replace it by a very general distribution family. In addition the GAMLSS models given a response variable y of a distribution defined by p parameters $\theta^T = (\theta_1, \theta_2, \ldots, \theta_p)$, each parameter, and not just the mean, are related to covariates using additive non-parametric (smooth) functions (Rigby and Stasinopoulos, 2005), modeling the parameter of location (the mean), scale (variance) and shape (skewness and kurtosis) of the reponse variable y. A general formulation of the model is (Rigby and Stasinopoulos, 2005):

$$g_k(\boldsymbol{\theta}_k) = \eta_k = \boldsymbol{X}_k\beta_k + \sum_{j=1}^{J_k} s_{jk}(\boldsymbol{x}_{jk}), \quad ,k = 1, \ldots, 4 \tag{3.6}$$

The first two population parameters θ_1 and θ_2 in model (2.2) are usually location (μ) and scale parameter (σ), whereas the remaining parameters, if any, are characterized as shape parameters (skewness ν (θ_3) and kurtosis τ (θ_4)), we obtain then the model (Rigby and Stasinopoulos, 2005):

$$\begin{cases} g_1(\boldsymbol{\mu}) = \eta_1 = \boldsymbol{X}_1\beta_1 + \sum_{j=1}^{J_1} s_{j1}(\boldsymbol{x}_{j1}) \\[2ex] g_2(\boldsymbol{\sigma}) = \eta_2 = \boldsymbol{X}_2\beta_2 + \sum_{j=1}^{J_2} s_{j2}(\boldsymbol{x}_{j2}) \\[2ex] g_3(\boldsymbol{\nu}) = \eta_3 = \boldsymbol{X}_3\beta_3 + \sum_{j=1}^{J_3} s_{j3}(\boldsymbol{x}_{j3}) \\[2ex] g_4(\boldsymbol{\tau}) = \eta_4 = \boldsymbol{X}_4\beta_4 + \sum_{j=1}^{J_4} s_{j4}(\boldsymbol{x}_{j4}) \end{cases}$$

where $g()$ is an known monotonic, twice differentiable link function, $\boldsymbol{X}_k$ for $k = 1, \ldots, 4$ are the design matrices incorporating the linear additive terms in the model, β_k are the linear coefficient parameters and $s_{kj}(\boldsymbol{x}_{kj})$ represent smoothing functions for the explanatory variables (Rigby and Stasinopoulos, 2005; Stasinopoulos et. al, 2017). The smooth functions used within GAMLSS can be written in the form $s(x) = \boldsymbol{Z}\gamma$ where $\boldsymbol{Z}$ is the basis matrix which depends on the explanatory variable $\boldsymbol{X}$ and $\boldsymbol{\gamma}$ is a parameter vector to be estimated, subject to a quadratic penalty of the form $\lambda\boldsymbol{\gamma}^T\boldsymbol{G}\boldsymbol{\gamma}$, for a known matrix $\boldsymbol{G} = \boldsymbol{D}^T\boldsymbol{D}$ and where λ regulates the amount of smoothing needed for the fit (Stasinopoulos et. al 2017).

The estimation of the GAMLSS regression model is achieved by penalized maximum likelihood estimation for a fixed λ, maximazing the expression:

$$l = \sum_{i=1}^{n} \log f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$$

$$l_p = l - \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{kj}^T\boldsymbol{G}_{kj}(\lambda_{kj})\boldsymbol{\gamma}_{kj} \tag{3.7}$$

The maximization of the penalized log-likelihood function is based on the RS algorithm using a backffiting method.

The GAMLSS approach is of interest in the statistical field and has been developed in the Bayesian framework, first as a way of handle with zero inflated data (Klein et al, 2015a) and then extending it to any parametric response (Klein et al, 2015c). In the frequentist framework (Wood et al, 2016) have suggested an alternative way of chosing the λ parameter in the line of the GAM using a Laplace integration and (Marra and Radice, 2017b) have proposed a GAMLSS models using a penalized log-likelihood estimation based on a Trust-region algorithm with an automatic selection of smoothing parameter.

3.3 Univariate modelling of glycemic markers

Both the glycated hemoglobin (HbA1c) and the glycated albumin (GA) are indirect measurements of glycemia, that depend on the concentration of sugar in the bloodstream but also in the glycation rate of proteins and the blood glucose exposition time of the hemoglobin and Albumin. Hence any biological factor that modifies the glycation rate or the circulating proteins life-span, can modify the levels of both glycated markers with independence of glycemic levels.

The GAMLSS models will be applied to study such cofounders, using the estimation process of Rigby and Stasinopoulos (2005) implemented in the `gamlss` package and the more recent approach of Marra and Radice (2017b) present in the `GJRM` package. This study will let us assess what are the main biological factors modifying the levels and variability of the HbA1c and GA with independence of the glucose enviroment, evaluated using fasting plasma glucose, and to compare both approaches in a real data application.

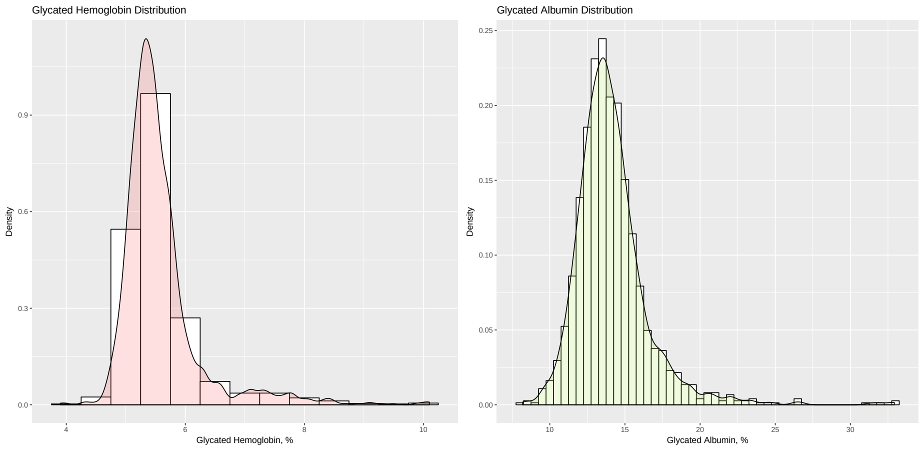


Figure 3.1: Histograms of Glycated hemoglobin (HbA1c) and Glycated Albumin (GA).

The use of the GAMLSS methodology is justified because if we want to study the mean levels of the HbA1c or GA in function of any covariate in a flexible way using a GAM model, we see that the levels of both proteins show a non-normal distribution (Figure 3.1). While if we considerer a non-exponential distribution like the reverse Gumbel we get a better fit (Figure 3.2). The choice of the reverse Gumbel distribution was based on terms of AIC, its interpretability and because is implemented in both R packages.

Furthermore, the use of GAM models will not be proper, because as we can see in the (Figure 3.3) the variability of both glycated proteins are not independent from the glucose levels and patient age.

In Figure 3.3 we can see that the variability of both glycated proteins increase with the glucose levels and patient age, this can be caused mainly because the diabetic patients show higher levels of sugar and most of them are up to 40 years old, showing their glucose levels a higher variability in such patients because they do not have the necessary mechanism to control their glycemia.

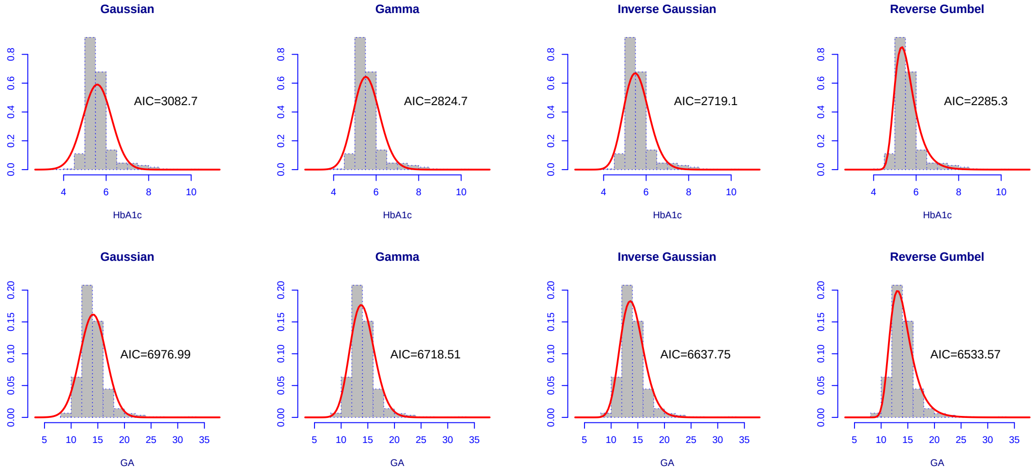


Figure 3.2: Density of the glycated hemoglobin (HbA1c) and glycated albumin (GA) for the continuous distributions of the exponential family and the reverse Gumbel. The red line represents different distributional assumptions.

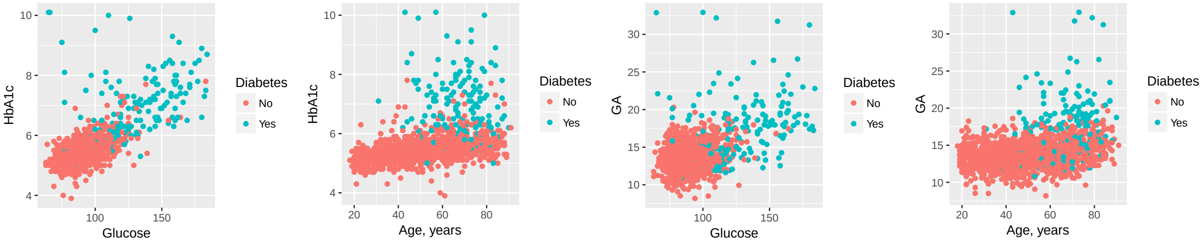


Figure 3.3: Scatter plot of the levels of Glycated hemoglobin (HbA1c) and Glycated Albumin (GA), regarding to Glucose levels and Age.

In the first model, to study the effects of different covariates in the mean levels and variability of the HbA1c levels, we propose:

$$\begin{cases} \eta_i^{\mu} = \beta_{0i}^{\mu} + Gender_i\beta_{1i}^{\mu} + Smoke\beta_{2i}^{\mu} + Alcohol\beta_{3i}^{\mu} + Exercise\beta_{4i}^{\mu} + s_i^{\mu}(Glucose) + s_i^{\mu}(Age) + \\ + s_i^{\mu}(BMI) + s_i^{\mu}(MCV)_i, \\ \eta_i^{\sigma} = \beta_{0i}^{\sigma} + Gender_i\beta_{1i}^{\sigma} + Smoke\beta_{2i}^{\sigma} + Alcohol\beta_{3i}^{\sigma} + Exercise\beta_{4i}^{\sigma} + s_i^{\sigma}(Glucose) + s_i^{\sigma}(Age) \end{cases} \tag{3.8}$$

The predictors (η_1 and η_2) are an additive composition of an intercept β_0 being the overall level of the predictor, linear effects (Gender, Smoke, Alcohol and Exercise), and the functions $s_i(\boldsymbol{z})$ represent non-linear effects of the continuous covariates (Age, Glucose, BMI, and MCV) with penalized splines representation for both estimations. The first equation refers to the location parameter μ of the reverse Gumbel distribution and the second to the scale parameter σ.

Table 3.2 shows the parametric effects obtained with both models. The levels of HbA1c are equal for men and women. Smoking increases its levels while the alcohol consumption reduces it with independence of the glycemic status. Its variability is higher in men, lower in ex-smokers and decreases with the alcohol consumption.

Both methods show a difference in the significance of the effect of ligth alcohol consumption and the minimal activity in the HbA1c variability.

Glycated hemoglobin μ				Glycated hemoglobin σ			
	Coefficients	SE	P value		Coefficients	SE	P value
μ^{Gender}	-0.01 -0.01	0.02 0.01	0.362 0.459	σ^{Gender}	0.19 0.37	0.05 0.09	< 0.01 < 0.01
$\mu^{ex-smoker}$	0.01 -0.008	0.02 0.02	0.416 0.648	$\sigma^{ex-smoker}$	-0.21 -0.10	0.05 0.09	< 0.01 0.305
μ^{smoker}	0.10 0.07	0.02 0.02	< 0.01 < 0.01	σ^{smoker}	-0.09 0.02	0.06 0.11	0.120 0.847
$\mu^{oh_{10-139}}$	-0.03 -0.03	0.20 0.01	0.07 0.08	$\sigma^{oh_{10-139}}$	-0.18 -0.31	0.05 0.09	< 0.01 < 0.01
$\mu^{oh_{140-279}}$	-0.12 -0.10	0.03 0.02	< 0.01 < 0.01	$\sigma^{oh_{140-279}}$	-0.13 -0.14	0.06 0.13	0.04 0.271
$\mu^{oh_{280+}}$	-0.12 -0.10	0.03 0.03	< 0.01 < 0.01	$\sigma^{oh_{280+}}$	-0.26 -0.53	0.08 0.16	< 0.01 < 0.01
$\mu^{Low-activity}$	-0.04 -0.02	0.02 0.01	0.057 0.264	$\sigma^{Low-activity}$	0.25 0.08	0.04 0.09	< 0.01 0.352
$\mu^{HEPA-activity}$	-0.03 -0.01	0.03 0.02	0.140 0.439	$\sigma^{HEPA-activity}$	0.06 0.11	0.05 0.10	0.212 0.290

Table 3.2: Parametric effects in the mean and variability of the HbA1c levels. In black the `gamlss` fit and in red the `GJRM` fit.

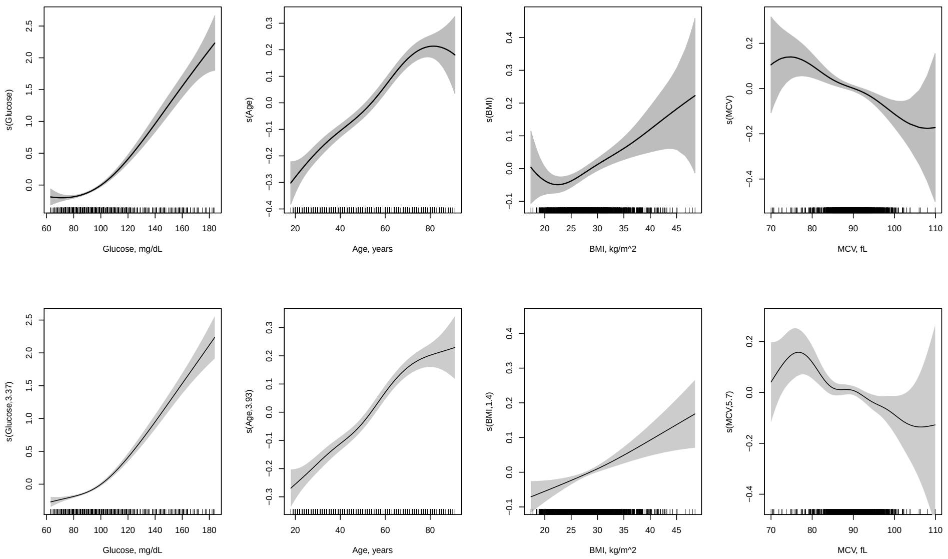


Figure 3.4: Estimated smooth centered effects of Glucose, Age, BMI and Mean Corpuscular Volume (MCV) in the levels of HbA1c with the associated 95% point-wise intervals. Using the `gamlss` R package with a RS estimation (up) and the using the `GJRM` package (down). The jittered rug plot, at the bottom of each graph, shows the covariate values.

Regarding the smooth effects (Figure 3.4), HbA1c increases with the Glucose from the $100mg/dL$, raises with age until 70 years and also with the BMI. As expected the levels of HbA1c decrease at higher levels of Mean Corpuscular Volume. In the `GJRM` estimate, lower levels of glycated hemoglobin are also associated with low values of MCV, as a consequence of the shorter life-span of the red blood cells in patients with microcytic anemia, when the hemoglobin exposure time to blood glucose is shorten.

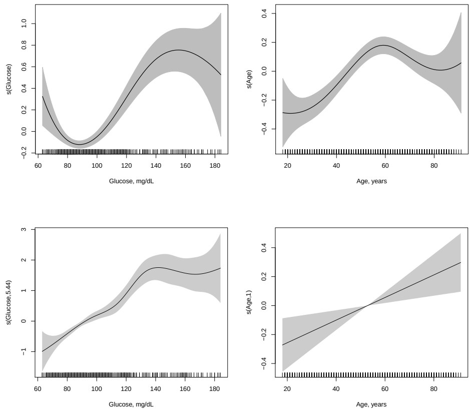


Figure 3.5: Estimated smooth centered effects of Glucose and Age in the variability of th HbA1c levels with the associated 95% point-wise intervals. Using the `gamlss` R package with a RS estimation (up) and using the `GJRM` package (down). The jittered rug plot, at the bottom of each graph, shows the covariate values.

The HbA1c variability increases with glucose and age for both methods (Figure 3.5). In the case of the `gamlss` fit, the HbA1c levels show higher variability at lower glucose levels as well, due to the presence of the diabetic patients taking anti-glycemic drugs with low fasting plasma glucose when measure, while the `GJRM` do not get this feature from the data. Regarding to the effect of the age the `GJRM` fit show a linear increase while the `gamlss` estimation raise only until 60 years.

In the second model, to study the effects of different covariates in the mean levels and variability of the GA the following predictors are proposed:

$$\begin{cases} \eta_i^{\mu} = \beta_{0i}^{\mu} + Gender_i\beta_{1i}^{\mu} + Smoke\beta_{2i}^{\mu} + Alcohol\beta_{3i}^{\mu} + Exercise\beta_{4i}^{\mu} + s_i^{\mu}(Glucose) + s_i^{\mu}(Age) + \\ + s_i^{\mu}(BMI) + s_i^{\mu}(T3)_i, \\ \eta_i^{\sigma} = \beta_{0i}^{\sigma} + Gender_i\beta_{1i}^{\sigma} + Smoke\beta_{2i}^{\sigma} + Alcohol\beta_{3i}^{\sigma} + Exercise\beta_{4i}^{\sigma} + s_i^{\sigma}(Glucose) + s_i^{\sigma}(Age) \end{cases} \tag{3.9}$$

As in the case of the HbA1c the predictors (η_1 and η_2) are an additive composition of an intercept β_0 being the overall level of the predictor, linear effects (Gender, Smoke, Alcohol and Exercise), and the functions $s_i(\boldsymbol{z})$ represent non-linear effects of the continuous covariates (Age, Glucose, BMI and T3) with penalized splines representation for both estimations. The first ecuation refers to the location parameter μ of the reverse Gumbel distribution and the second to the scale parameter σ.

In the Table 3.3 we can see the estimation of the parametric effects by both methods. The levels of the GA are lower for smokers and people with moderate and high alcohol consumption. Its variability is higher for smokers and decrease with the alcohol consumption either in the `gamlss` and `GJRM` packages.

Regarding the smooth effects (Figure 3.6) the levels of GA raise for levels of Glucose higher than $100mg/dL$ and increase for ages higher than 60 for both model estimations. Furthermore, higher BMI and Thyroid hormones levels reduce the GA concentration.

The variability of the GA levels increases with the Glucose from 100-120 mg/dL and raises for ages beyond 60 years (Figure 3.7), with similar smooth estimates for both methods. .

Glycated Albumin μ				Glycated Albumin σ			
	Coefficients	SE	P valor		Coefficients	SE	P valor
μ^{Gender}	-0.08	0.10	0.379	σ^{Gender}	-0.02	0.05	0.615
	-0.08	0.10	0.421		-0.07	0.10	0.496
$\mu^{ex-smoker}$	-0.02	0.10	0.846	$\sigma^{ex-smoker}$	-0.03	0.05	0.533
	-0.04	0.10	0.686		-0.06	0.11	0.600
μ^{smoker}	-0.78	0.12	< 0.01	σ^{smoker}	0.16	0.05	< 0.01
	-0.78	0.12	< 0.01		0.36	0.12	< 0.01
$\mu^{oh_{10-139}}$	-0.08	0.10	0.41	$\sigma^{oh_{10-139}}$	-0.20	0.05	< 0.01
	-0.09	0.10	0.359		-0.41	0.10	< 0.01
$\mu^{oh_{140-279}}$	-0.61	0.14	< 0.01	$\sigma^{oh_{140-279}}$	-0.16	0.07	0.010
	-0.62	0.14	< 0.01		-0.41	0.10	0.019
$\mu^{oh_{280+}}$	-0.71	0.18	< 0.01	$\sigma^{oh_{280+}}$	-0.19	0.08	0.02
	-0.76	0.18	< 0.01		-0.37	0.17	0.03
$\mu^{Low-activity}$	-0.09	0.10	0.356	$\sigma^{Low-activity}$	0.06	0.04	0.156
	-0.08	0.10	0.398		0.11	0.09	0.226
$\mu^{HEPA-activity}$	-0.14	0.10	0.183	$\sigma^{HEPA-activity}$	-0.01	0.05	0.718
	-0.13	0.10	0.210		-0.03	0.11	0.790

Table 3.3: Parametric effects in the mean and variability of the GA levels. In black the `gamlss` fit and in red the `GJRM` fit.

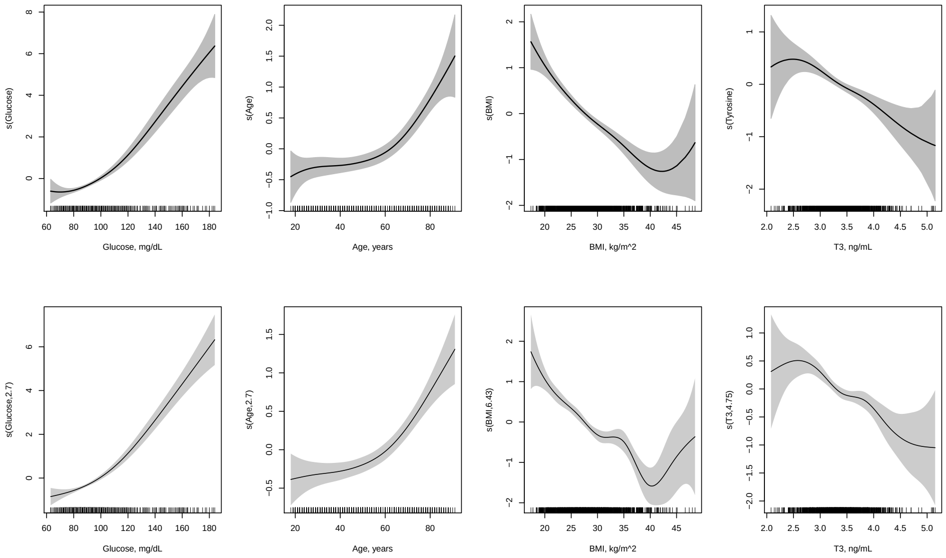


Figure 3.6: Estimated smooth centered effects of Glucose, Age, BMI and T3 (a thyroid hormone) in the levels of GA with the associated 95% point-wise intervals. Using the `gamlss` R package with a RS estimation (up) and the `GJRM` package (down). The jittered rug plot, at the bottom of each graph, shows the covariate values.

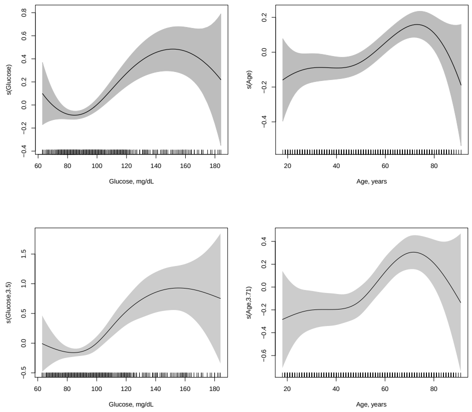


Figure 3.7: Estimated smooth centered effects of Glucose and Age in the variability of th GA levels with the associated 95% point-wise intervals. Using the `gamlss` R package with a RS estimation (up) and the `GJRM` package (down). The jittered rug plot, at the bottom of each graph, shows the covariate values.

Note about the interpretability

For a clinical interpretation of the model results, we have to take into account, that the parameters modeled in some of the distributions used in GAMLSS do not represent the expectation and variance of the response itself.

In the particular case of the reverse-Gumbel μ is a location parameter of the distribution but represents the mode instead of the mean. And the variance of the response can be obtained from σ using the transformation $V(Y) = (\frac{\pi^2}{6})\sigma^2$ while the mean can be obtained from μ using $E(Y) = \mu - 0.5777722\sigma$.

To ilustrates this transformation we present the effect of the moderate alcohol consumption and age in the mean levels of GA. In the case of the moderate alcohol consumption (see Table 3.3, `gamlss` fit), the effect in the expectation is:

$$(-0.606 + C * (-exp(0.166)) = -1.28), \quad C = 0.5777722$$

given a fitted model m1, in the case of a smooth effect like age the effect on the expectation can be obtained as:

```
Age_order <- order(with(dat,Age))
Variable=dat[Age_order,]$Age

prediction=predict(m1,what="mu",type="terms",se=T)

Intercept=13.40

mu_effect=as.numeric(Intercept+prediction$fit[,6])
mu_effect_low=as.numeric(Intercept+prediction$fit[,6]-1.96*prediction$se.fit[,6])
mu_effect_high=as.numeric(Intercept+prediction$fit[,6]+1.96*prediction$se.fit[,6])


prediction=predict(g11,what="sigma",type="terms",se=T)

Intercept=0.40

```

```
sigma_effect=exp(as.numeric(Intercept+prediction$fit[,6]))
sigma_effect_low=exp(as.numeric(Intercept+prediction$fit[,6]-1.96*prediction$se.fit[,6]))
sigma_effect_high=exp(as.numeric(Intercept+prediction$fit[,6]+1.96*prediction$se.fit[,6]))

plot(mu_effect[Age_order]+0.577722*sigma_effect[Age_order]~Variable,ylim=c(13,16.5),
     xlab="Age, years",ylab="Mean GA, %",type="l")
lines(mu_effect_low[Age_order]+0.577722*sigma_effect_low[Age_order]~Variable,col="grey")
lines(mu_effect_high[Age_order]+0.577722*sigma_effect_high[Age_order]~Variable,col="grey")
```

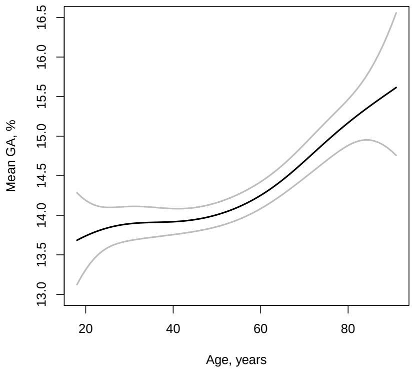


Figure 3.8: Effect of the Age in the expectation of the Glycated Albumin levels with the 95% confidence interval.

Chapter 4

Copula functions

Most of the approaches found in the literature for modeling two responses jointly conditioned to a set of covariates, assume a specific bivariate distribution like a bivariate gaussian (Klein and Kneib, 2015c). However, this assumption implies that both responses must be gaussian distributed and show a symmetric structure of dependence, known as elliptical (Durante and Sempi, 2016).

However, in our case of study, both assumptions are not realistic. In this case we have two non-gaussian responses that show higher correlation at higher vaules. A good alternative to study such kind of situations is to consider a bivariate distribution constructed by means of a copula function.

A copula function lets us build a bivariate distribution from two given marginals, that can belong to any parametric family, be equal or not and even consider two discrete variables or a mixed joint distribution of a continuous and discrete variable. Furthermore, for two given marginals, using the parametric copulas proposed in the literature we can consider different structures of dependence between them.

In this Chapter, the concept of copula function will be introduced. Then, a description of the main parametric families of copulas is given, as well as, the correspondence between copula functions and other correlation measurements.

4.1 Sklar theorem

Copulas are joint distributions generated from given marginals. Therefore, properties of copulas are analogous to properties of joint distributions for two or more random variables. In the bivariate case for two random variables (X, Y) the joint distribution H is defined as:

$$H(x, y) = P[X \leq x, Y \leq y]$$

The properties of this bivariate cdf are:

1. $lim_{x,y \to -\infty} \quad H(x, y) = 0,$
2. $lim_{x,y \to \infty \forall_j} \quad H(x, y) = 1$
3. By the rectangle inequality, for all (a_1, a_2) and (b_1, b_2) with $a_1 \leq b_1$, $a_2 \leq b_2$,

$$H(b_1, b_2) - H(a_1, b_2) - H(b_1, a_2) + H(a_1, a_2) \geq 0$$

Conditions 1 and 2 imply $0 \leq H \leq 1$. Condition 3 is referred to as the property that H is 2-increasing.

Given the bivariate cdf $H(x, y)$, univariate margins $F(x) = P(X < x)$ and $G(y) = P(Y < y)$ are obtained by letting $y \to \infty$ and $x \to \infty$, respectively.

The concept of copula function was introduced by Sklar (1953) under the Sklar theorem which states that the joint distribution function (c.d.f) $H(x, y)$ of any pair (X, Y) of continuous random variables can be written in the form:

$$H(x, y) = C(F(x), G(y)), \quad x, y \in \boldsymbol{R} \tag{4.1}$$

This theorem implies that copulas can be used to express a bivariate distribution in terms of its maringal distribution and a function that bind them together. Hence given a copula function C with margins F and G we can build a bivariate distribution of our chosen:

- By varying F and G in the previous expression, joint distributions with arbitrary marginals can be built.
- On the other hand, by varying C, one alters the nature of dependence between X and Y.

4.2 Parametric families of the copula functions

To proper consider the dependence structure between the margins of a bivariate random variable we have different parametric copulas available, most of them defined by one single parameter θ that measure the strength of dependence between the margins. The main differences between the parametric copula functions proposed in the literature are given by their ability to model assymetrical structures of dependence and the range of their θ parameter. To better understand what a assymetric structure of dependence is, I will introduce the concept of tail dependence.

The tail dependence concept is designed to show how large (or small) values of one random variable appear with large (or small) values of the other (Balakrishnan and Lai, 2009), a common situation in bivariate variables, including our biomedical problem (see Figure 4.1).

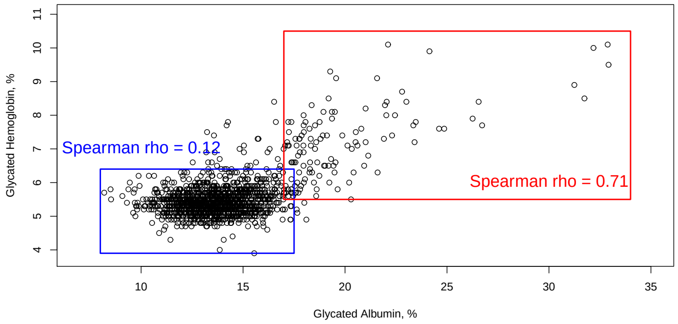


Figure 4.1: Relationship between the levels of glycated hemoglobin (HbA1c) and glycated albumin (GA), note the assymetric structure of dependence with higher relation at higher levels of both proteins.

In (Balakrishnan and Lai, 2009) the upper tail dependence is defined in terms of a coefficient λ_U as the limit (if it exist) of the conditional probability that Y is greater than the $100\alpha th$ percentile of G given that X is greater than the $100\alpha th$ percentile F as α approaches 1:

$$\lambda_U = \lim_{\alpha \uparrow 1} P[Y > G^{-1}(\alpha) | F^{-1}(\alpha)] \tag{4.2}$$

If $\lambda_U > 0$, then X and Y are upper tail dependent and asymptotically independent otherwise.

Similarly, the lower tail dependence coefficient is defined as:

$$\lambda_U = \lim_{\alpha \downarrow 1} P[Y < G^{-1}(\alpha) | F^{-1}(\alpha)] \tag{4.3}$$

Let C be the copula of X and Y. It can be shown that:

$$\lambda_U = \lim_{u \uparrow 1} \frac{\overline{C}(u,v)}{1-u}, \quad \lambda_L = \lim_{u \downarrow 1} \frac{C(u,v)}{u}$$

Several parametric copula functions have been proposed for both situations of upper and lower tail dependence as well as different strengths of association between the margins. Most of the proposed copulas can be classified in two families, Archimidean or Elliptical.

	Copula	Range of θ	λ_L	λ_U
Archimedean				
	AMH	$\theta \in (-1,1)$	0	0
	Clayton	$\theta \in (0,\infty)$	$2^{1/\theta}$	
	Frank	$0 \in R$	0	0
	Gumbel	$0 \in (1,\infty)$	0	$2-2^{1/\theta}$
	Joe	$0 \in (1,\infty)$	0	$2-2^{1/\theta}$
Elliptical				
	Gaussian	$\theta \in (-1,1)$	0	0
	t-Copula	$\theta \in (-1,1)$,$\zeta \in (2,120)$	$2 * t_{\zeta+1}(-\sqrt{\frac{(\zeta+1)(1-\theta)}{1+\theta}})$	

Table 4.1: Summary of the main copula functions, with the range of the θ parameter, the λ_L and λ_U coefficients.

4.2.1 Archimidean copulas

All the parametric copulas of the Archimidean family can be expressed as a sum of functions of marginals F and G of the form (Nelsen, 2006):

$$\varphi(C(u,v)) = \varphi(u) + \varphi(v) \tag{4.4}$$

Since we are interested in expressions that we can use for the construction of copulas, we want to solve the relation $\varphi(C(u,v)) = \varphi(u) + \varphi(v)$. We thus need to find an appropriately defined "inverse" $\varphi^{[-1]}$ so that:

$$C(u,v) = \varphi^{[-1]}(\varphi(u) + \varphi(v)) \tag{4.5}$$

In (Nelsen, 2006) φ is defined as a continuous, strictly decreasing function from $[0,1]$ to $[0,\infty]$ such that $\varphi(1)=0$. The pseudoinverse of φ is the function $\varphi^{[-1]}$, with domain $[0,\infty]$ and range $[0,1]$, given by:

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{[-1]}(t), & 0 \le t \le \varphi(0) \\ 0, & \varphi(0) \le t \le \infty \end{cases} \tag{4.6}$$

Note that if $\varphi(0)=\infty$, then $\varphi^{[-1]}(t)=\varphi^{-1}(t)$ and:

$$C(u,v) = \varphi^{-1}(\varphi(u), \varphi(v)) \tag{4.7}$$

The function φ is called the generator of the copula.

Clayton copula

The Clayton copula proposed by Clayton (1978), has the generation function $\varphi(t)=(1+t)^{-1/\theta}$ taking the form:

$$C(u,v;\theta) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta} \tag{4.8}$$

With the dependence parameter θ restricted on the region $(0,\infty)$. As θ approaches to zero, the marginals become independent and ∞ corresponds to the Frechet upper bound [1] (Triverdi and Zimmer, 2006).

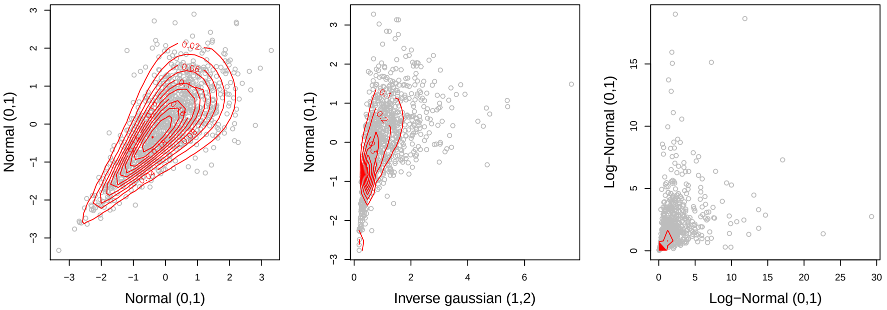


Figure 4.2: Contour plot of the bivariate density function of the Clayton copula with different marginals.

It exhibits strong left tail dependence ($\lambda_l = 2^{-1/\theta}$) and relatively weak right tail dependence, this is an appropiate copula for two random variables with a strong relation at low values but less correlated at high (Triverdi and Zimmer, 2006).

Joe copula

The copula function, proposed by Joe and Hu (1996) has the generator function $\varphi(t) = 1-(1-\exp(-t))^{1/\theta}$ and the expression:

$$C(u,v;\theta) = 1 - \{(1-u)^{\theta} + (1-v)^{\theta} - (1-u)^{\theta}(1-v)^{\theta}\}^{1/\theta} \tag{4.9}$$

With the dependence parameter θ restricted on the region $(1,\infty)$, being 1 the independence and ∞ the upper Frechet bound.

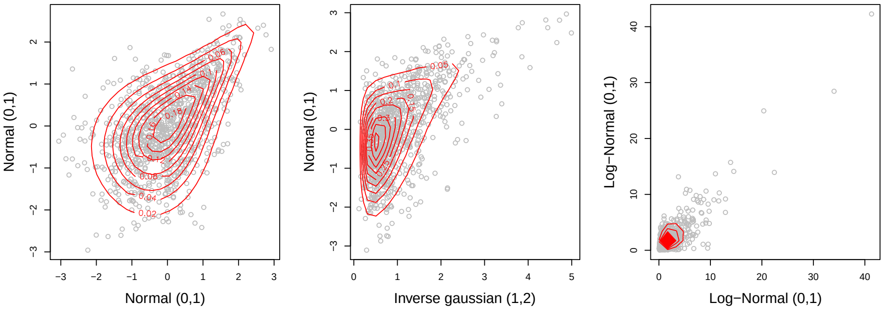


Figure 4.3: Contour plot of the bivariate density function of the Joe copula with different marginals.

The Joe copula exhibits strong right tail dependence ($\lambda_u = 2 - 2^{-1/\theta}$) and relatively weak left tail dependence and is appropiate when two random variables are strongly correlated at high values and less correlated at low (Triverdi and Zimmer, 2006).

Gumbel copula

The Gumbel copula has been proposed by Gumbel (1960) with the generator function $\varphi(t) = exp(-t^{1/\theta})$, and takes the form:

$$C(u, v; \theta) = exp(-(-logu)^{\theta} + (-logv)^{\theta})^{1/\theta} \tag{4.10}$$

The dependence parameter θ is restricted to the interval $[1, \infty)$. Values 1 and ∞ correspond to independence and the upper Frechet bound (Triverdi and Zimmer, 2006). As the Joe copula exhibits strong rigth upper tail dependence and relatively weak left tail dependence.

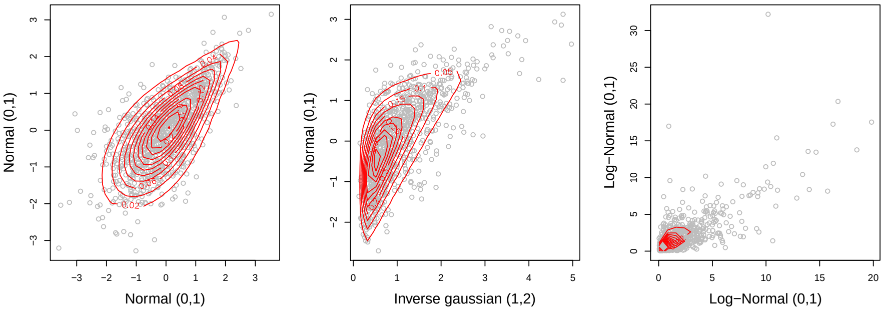


Figure 4.4: Contour plot of the bivariate density function of the Gumbel copula with different marginals.

Frank copula

The generator of the Frank copula is, $\varphi(t) = log(\frac{1-\theta}{1-\theta^{t}}), 0 < \theta < 1$, and:

[1] we encountered the Fréchet-Hoeffding bounds as universal bounds for copulas, fo any copula C: $W(u, v) = max(u + v - 1.0) \leq C(u, v) \leq min(u, v) = M(u, v)$ being W and M the upper and lower bound respectively (Nelsen, 2006)

$$C(u, v) = log_\theta(1 + \frac{(\theta^u - 1)(\theta^v - 1)}{(\theta - 1)}) \tag{4.11}$$

The dependence parameter may assume any real value $(-\infty, \infty)$. Values of $-\infty$, 0, ∞ corresponding to the Frechet lower bound, independence and Frechet upper bound, respctively (Triverdi and Zimmer, 2006). This is a comprehensive copula function given that both Frecht bounds are included in the range of permissible dependence (Devroye, 1986).

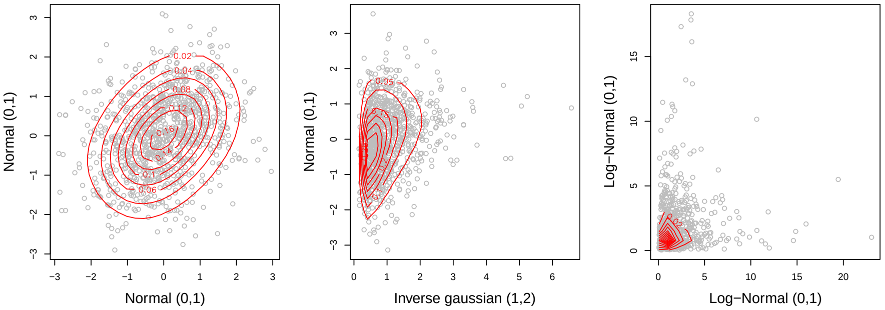


Figure 4.5: Contour plot of the bivariate density function of the Frank copula with different marginals.

Dependence in the tails of the Frank copula tends to be relatively weak and the strongest dependence is centered in the middle of the distribution.

Ali-Mikhail-Haq copula

The Ali-Mikhail-Haq (AMH) copula proposed by Ali et al (1978) has the generator function $\varphi(t) = \frac{1-\theta}{\exp(t)-\theta}$ and the expression:

$$C(u, v) = \frac{uv}{1 - \theta(1 - u)(1 - v)} \tag{4.12}$$

This copula function is appropiate for modeling margins showing strong correlations and the dependence in the tails tends to be weak.

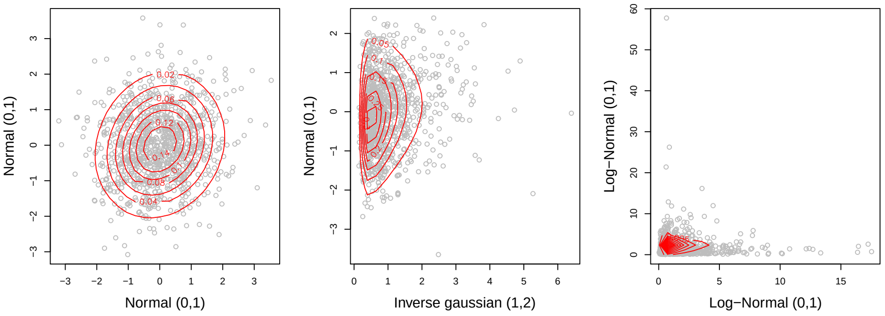


Figure 4.6: Contour plot of the bivariate density function of the Ali-Mikhail-Haq copula with different marginals.

4.2.2 Elliptical copulas

Elliptical copulas are obtained from the multivariate elliptical distributions by applying the inverse transformation related to Sklars theorem (Durante and Sempi, 2016). A random vector has a multivariate elliptical distribution if can be expressed in the form:

$$\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + \boldsymbol{RAU} \tag{4.13}$$

where $\boldsymbol{\mu} \in R^d$, $\boldsymbol{A} \in \boldsymbol{R}^{d*k}$ with $\boldsymbol{\Sigma} := \boldsymbol{AA}^T \in \boldsymbol{R}^{d*d}$ and $rank(\boldsymbol{\Sigma}) = k \leq d$, $\boldsymbol{U}$ is a d-dimensional random vector uniformly distributed on the sphere $\boldsymbol{S}^{d-1} = \{u \in \boldsymbol{R}^d : u_1^2 + \ldots + u_d^2 = 1\}$ and R is a positive random variable independent of $\boldsymbol{U}$ (Durante and Sempi, 2016).

Because of the transformation matrix $\boldsymbol{A}$, the uniform random variable $\boldsymbol{U}$ produces elliptically contoured density level surfaces, whereas the generating random variable R gives the distribution shape; in particular determines the tails of the distribution (Durante and Sempi, 2016). This radial symmetry causes that the lower and upper bivariate tail dependence coefficients coincide.

Gaussian copula

The Gaussian copula is given by the expression:

$$\boldsymbol{X} \overset{d}{=} \boldsymbol{AZ} \tag{4.14}$$

where $\boldsymbol{A} \in \boldsymbol{R}^{d*k}$, $\Sigma := \boldsymbol{AA}^T \in \boldsymbol{R}^{d*d}$ is the covariance matrix, $rank(\boldsymbol{\Sigma}) = k \leq d$ and $\boldsymbol{Z}$ is a d-dimensional random vector whose independent components have the univariate standard gaussian distribution, given the expression (Durante and Sempi, 2016):

$$\Phi_2\left(\Phi^{-1}(u), \Phi^{-1}(v); \theta\right) \tag{4.15}$$

where Φ is the cdf of the standard normal distribution and Φ_2 the standard bivariate normal distribution with the correlation parameter θ restricted to the interval $(-1, 1)$, attained the lower and upper Frechet bound (Nelsen, 2006).

The gaussian copula has both and lower tail dependence coefficients equal to 0.

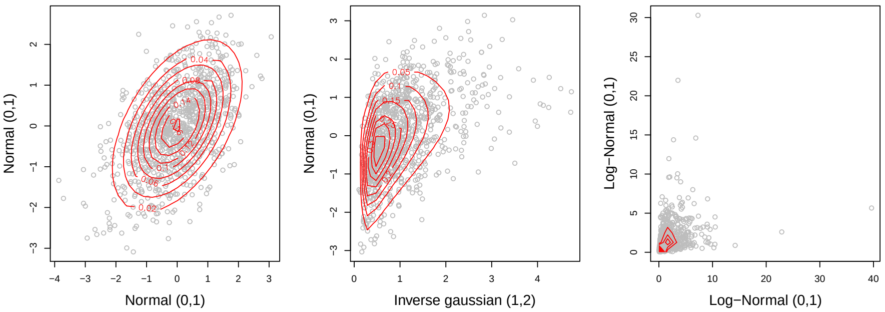


Figure 4.7: Contour plot of the bivariate density function of the Gaussian copula with different marginals.

Student-t copula

The Students t-copula is given by:

$$\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\sqrt{W}\boldsymbol{Z} \tag{4.16}$$

where $\boldsymbol{Z} \in \boldsymbol{N}_d(0, \boldsymbol{I}_d)$ is a Gaussian distribution, $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$ is positive definite. Moreover, W and $\boldsymbol{Z}$ are independent, and W follows a inverse Gamma distribution with parameters $(\nu/2, \nu/2)$ (Durante and Sempi, 2016). The bivariate Students t-copula is given by:

$$C(u, v) = t_{2,\zeta}\left(t_\zeta^{-1}(u), t_\zeta^{-1}(v); \zeta, \theta\right) \tag{4.17}$$

where θ is in the range $(-1, 1)$, t_2 denotes the bivariate of a Student t-distribution, and the t_ζ^{-1} denotes the inverse of the standard t-distribution, ζ is the degrees of freedom, t-copula becomes a Gaussian copula in the limit $\zeta \to \infty$. As a consequence of the radial symmetry, the lower and upper tail dependence coefficients are identical and given by Durante and Sempi (2016):

$$\lambda_L(C_{\theta,\zeta}) = \lambda_U(C_{\theta,\zeta}) = 2 * t_{\zeta+1}(-\sqrt{\frac{(\zeta+1)(1-\theta)}{1+\theta}}) \tag{4.18}$$

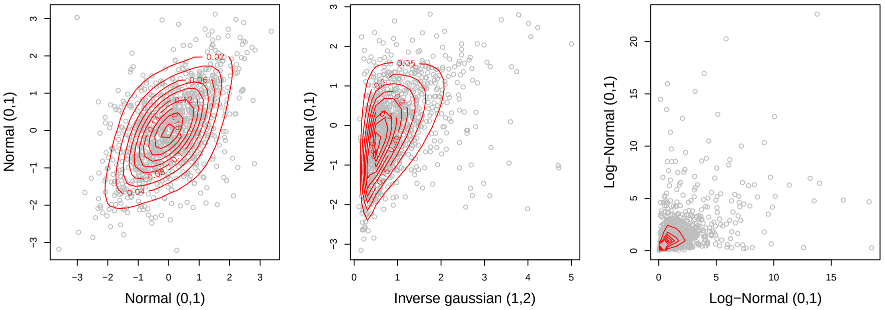


Figure 4.8: Contour plot of the bivariate density function of the t-Student copula with different marginals.

4.2.3 Other copulas

Eyraud-Farlie-Gumbel-Morgesten copula

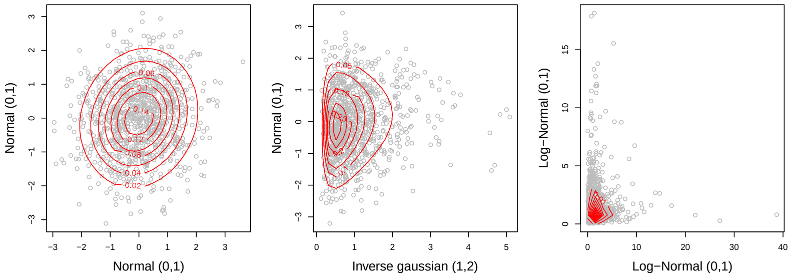


Figure 4.9: Contour plot of the bivariate density function of the EFGM copula with different marginals.

The Eyraud-Farlie-Gumbel-Morgesten (EFGM) copula is not a member of the Archimidean or Elliptical families, Nelsen (2006) describes it like a copula with a quadratic section and a perturbation of the product copula $C(u, v) = uv$. This copula takes the form:

$$C(u, v) = uv(1 + \theta(1 + u)(1 - v)) \tag{4.19}$$

Do not allow for upper or lower tail dependence and like the AMH copula is appropiate for two random variables with weak dependences (Triverdi and Zimmer, 2006).

Plackett copula

As the EFGM this copula function defined by Plackett (1965) do not belong to the Archimidean or elliptical family and has the expression:

$$C = \frac{[1 + (1 + \theta)(u + v)] - \sqrt{[1 + (\theta - 1)(u + v)]^2 - 4\theta(\theta - 1)uv}}{2(\theta - 1)} \tag{4.20}$$

For $\theta > 0$, and do not present upper or tail dependence.

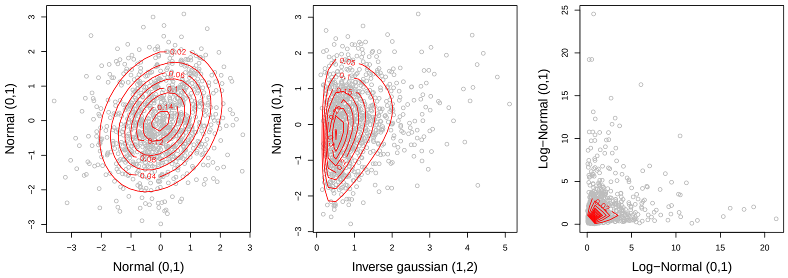


Figure 4.10: Contour plot of the bivariate density function of the Plackett copula with different marginals.

4.2.4 Survival and other rotated copulas

For any pair (X, Y) of random variables with joint distribution function H, the joint survival function is given by $\bar{H}(x, y) = P(X > x, Y > y)$. The marginals of $\bar{H}$ are the functions $\bar{H}(x, -\infty)$ and $\bar{H}(-\infty, y)$, which are the univariate survival functions $\bar{F}(x) = P(X > x)$ and $\bar{G}(y) = (Y > y)$, respectively (Nelsen, 2006) if we apply the Sklar theorem to this survival functions instead of considering the distribution of the form $F(x) = P(x < X)$ and $G(y) = P(y < Y)$, we obtain a survival copula also known as 180^o rotated copulas:

$$\begin{aligned}\bar{H}(x, y) = 1 - F(x) - G(y) + H(x, y) &= \bar{F}(x) + \bar{G}(y) - 1 + C(F(x), G(y)) \\ &= \bar{F}(x) + \bar{G}(y) - 1 + C(1 - \bar{F}(x), 1 - \bar{G}(y))\end{aligned} \tag{4.21}$$

So we have:

$$\bar{H}(x, y) = \hat{C}(\bar{F}(x), \bar{G}(y)) \tag{4.22}$$

Where $\hat{C}$ is the 180^o rotated version of the original copula function. This rotation only make sense for assymetric copula functions as the Joe, Gumbel and Clayton. Similarly to the survival copula we can obtain other rotated versions following the definition given in Brechmann and Schepsmeier (2013) as:

$$C_{90}(u_1, u_2) = u_2 - C(1 - u_1, u_2) \quad 90^o \quad rotation \tag{4.23}$$

$$C_{270}(u_1, u_2) = u_1 - C(u_1, 1 - u_2) \quad 270^o \quad rotation \tag{4.24}$$

An interesting property of the 90^o and 270^o rotated copulas is that present the opposite ranges of the θ parameter than the non-rotated versions and 180^o, that let us model non-symmetrical structures of depdence not possible with the non-rotated versions that present a $\theta > 0$ for the Calyton and $\theta > 1$ for the Gumbel and Joe.

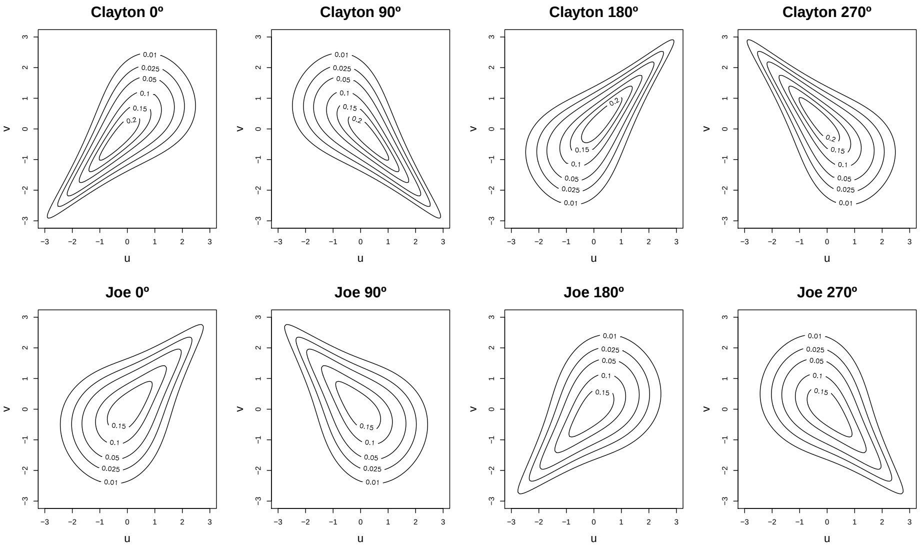


Figure 4.11: Contour plots of the bivariate density function of the rotated Joe and Clayton copulas that let us model positive and negative correlation of two random variables.

Furthermore, this rotated versions let us model a bivariate variable showing an assymetric structure of dependence with both negative and positive correlations. This can be achieved by mixing any of the Clayton, Joe and Gumbel with any of the rotated versions of the same family. For instance, mixing the Clayton copula with its 90 rotation allows one to model positive and negative tail dependence. (Marra and Radice, 2017a).

4.3 Concordance measurements and copulas

Although the θ parameter measures the dependence between two random variables, for some families the range of θ has no upper or lower limit. Given that, it is not easy realize how strong is the correlation between the margins, and in the regression context, how much a set of covariates modify the relationship between the responses. In order to have a correlation measurement in a more interpretable scale we can express the copula parameter in terms of concordance measures that fall in the $(-1, 1)$ interval.

Kendall τ and Copula functions

Let (x_i, y_i) and (x_j, y_j) be two observations from (X, Y) of continuous random variables. The two pairs (x_i, y_i) and (x_j, y_j) are concordant if $(x_i - x_j)(yi - y_j) > 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$ (Balakrishnan and Lai 2009).

There are $\binom{n}{2}$ distinct pairs (x_i, y_i) and (x_j, y_j) of observations in the sample, and each pair is either concordant or discordant. Let **c** denote the number of concordant pairs and **d** the number of discordant pairs. Then Kendall's tau for the sample is defined as (Nelsen 2006):

$$\tau = \frac{\mathbf{c} - d}{\mathbf{c} + d} \tag{4.25}$$

Just as H can be expressed as a function of copula C, Kendalls τ can be expressed in terms of the copula (Balakrishnan and Lai 2009) as:

$$\tau = 4\int_0^1 \int_0^1 C(u,v)\mathbf{c}(u,v) \quad du \quad dv - 1 = 4E(C(U,V)) - 1 \tag{4.26}$$

Spearman ρ and Copula functions

Like Kendalls τ, the Spearmans ρ is based on concordance and discordance. Let $(X_1, Y_1), (X_2, Y_2)$ and (X_3, Y_3) be three independent pairs of random variables with a common distribution function H. Then, ρ_s is defined to be proportional to the probability of concordance minus the probability of discordance for the two pairs (X_1, Y_1) and (X_2, Y_3) (Balakrishnan and Lai 2009):

$$\rho_s = 3(P[(X_1 - X_2)(Y1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]) \tag{4.27}$$

Equation 4.27 can be expressed in terms of the copula as:

$$\rho_s = 12\int_0^1 \int_0^1 C(u,v) \quad du \quad dv - 3 = 12E(UV) - 3 \tag{4.28}$$

Chapter 5

Copula GAMLSS for bivariate responses

The univarite regression models GAMLSS allow us to study how a set of covariates is related to distributional aspects of one single response. However, the objective of this work is study jointly the levels of the glycated hemoglobin and glycated albumin to determine which factors could modify the association between them.

To study both proteins jointly, flexible copula regression models for bivariate responses are proposed. The reason to consider such kind of modeling is that both proteins show non-standard distributions with variability dependent on the covariates and the relationship between both is not symmetric, as seen in the previous chapter the use of copula functions let us address both situations.

From the alternatives found in the literature we have considered Copula Regression Models for Location, Scale and Shape (CGAMLSS) proposed by Marra and Radice (2017a). These models consider a copula function to build a bivariate response of the GAMLSS-class, estimating the copula dependence and marginal distribution parameters simultaneously. CGAMLSS model each parameter using an additive predictors casting different types of covariate effects (linear, non-linear, spatial or random effects). The simultaneous parameter estimation is achieved within a penalized likelihood framework using a trust region algorithm with integrated automatic multiple smoothing parameter estimation.

These models are implemented in the R package `GJRM` (Marra and Radice, 2017b) and are easily applicable using the `copulaReg()` function with a similar syntax of the well known `mgcv` R package (Wood, 2006).

This Chapter introduces the model formulation of the CGAMLSS. Then, the estimation process is explained. Finally, the key points of the model building process are given.

5.1 Model formulation

In these models the joint cumulative distribution function (cdf) of two continuous random variables, Y_1 and Y_2 is expressed by means of a copula function and two marginals distributions as:

$$H(y_1, y_2|\boldsymbol{\vartheta}) = C\left(F_1(y_1|\mu_1, \sigma_1, \nu_1), F_2(y_2|\mu_2, \sigma_2, \nu_2); \zeta, \theta\right), \tag{5.1}$$

where $\boldsymbol{\vartheta} = (\mu_1, \sigma_1, \nu_1, \mu_2, \sigma_2, \nu_2, \theta, \zeta)$ is a vector containing all the parameters that define the bivariate response, $F_1(y_1|\mu_1, \sigma_1, \nu_1)$ and $F_2(y_2|\mu_2, \sigma_2, \nu_2)$ are the marginal cdfs of Y_1 and Y_2 taking values in $(0, 1)$, μ_m, σ_m and ν_m, for $m = 1, 2$ are marginal distribution parameters, C is a uniquely defined two-place copula function with the association parameter θ and ζ represents the number of degrees of freedom of the Student-t copula, only relevant when this copula is used.

The distribution available in the `GJMR` to consider in the marginals $F_1(y_1|\mu_1, \sigma_1, \nu_1)$ and $F_2(y_2|\mu_2, \sigma_2, \nu_2)$ are included in the appendix 1. These distributions have two or three parameters, however the CGAMLSS approach can be extended to any twice differentiable parametric distribution. Regarding the copula functions, CGAMLSS have implemented all the copulas presented in the Table 5.1, in addition to them the `GJRM` displays the rotated versions of the Clayton, Joe and Gumbel and the mixed copulas that can be constructed using it.

Copula	$C(u, v; \zeta, \theta)$	Ranges of θ and ζ	Link function	Kendall's τ
AMH	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$-\frac{2}{3\theta^2}\left\{\theta + (1-\theta)^2\right.$ $\left.\log(1-\theta)\right\} + 1$
Clayton	$\left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}$	$\theta \in (0, \infty)$	$\log(\theta - \epsilon)$	$\frac{\theta}{\theta+2}$
FGM	$uv\left\{1 + \theta(1-u)(1-v)\right\}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$\frac{2}{9}\theta$
Frank	$-\theta^{-1}\log\left\{1 + (e^{-\theta u} - 1)\right.$ $\left.(e^{-\theta v} - 1)/(e^{-\theta} - 1)\right\}$	$\theta \in \boldsymbol{R} \setminus \{0\}$	$-$	$1 - \frac{4}{\theta}\left[1 - D_1(\theta)\right]$
Gaussian	$\Phi_2\left(\Phi^{-1}(u), \Phi^{-1}(v); \theta\right)$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$\frac{2}{\pi}\arcsin(\theta)$
Gumbel	$\exp\left[-\left\{(-\log u)^{\theta}\right.\right.$ $\left.\left.+(-\log v)^{\theta}\right\}^{1/\theta}\right]$	$\theta \in [1, \infty)$	$\log(\theta - 1)$	$1 - \frac{1}{\theta}$
Joe	$1 - \left\{(1-u)^{\theta} + (1-v)^{\theta}\right.$ $\left.-(1-u)^{\theta}(1-v)^{\theta}\right\}^{1/\theta}$	$\theta \in (1, \infty)$	$\log(\theta - 1 - \epsilon)$	$1 + \frac{4}{\theta^2}D_2(\theta)$
Student-t	$t_{2,\zeta}\left(t_{\zeta}^{-1}(u), t_{\zeta}^{-1}(v); \zeta, \theta\right)$	$\theta \in [-1, 1]$, $\zeta \in (2, \infty)$	$\tanh^{-1}(\theta)$, $\log(\zeta - 2 - \epsilon)$	$\frac{2}{\pi}\arcsin(\theta)$
Plackett	$(Q - \sqrt{R})/\left\{2(\theta - 1)\right\}$	$\theta \in (0, \infty)$	$\log(\theta)$	-

Table 5.1: Definition of copulas implemented in `GJRM`, with corresponding parameter ranges of association parameter θ, transformation/link function of θ and ζ, and relation between Kendall's τ and θ. $\Phi_2(\cdot, \cdot; \theta)$ denotes the cdf of a standard bivariate normal distribution with correlation coefficient θ, and $\Phi(\cdot)$ the cdf of a univariate standard normal distribution. $t_{2,\zeta}(\cdot, \cdot; \zeta, \theta)$ indicates the cdf of a standard bivariate Student-t distribution with correlation θ and ζ degrees of freedom, and $t_{\zeta}(\cdot)$ denotes the cdf of a univariate Student-t distribution with ζ degrees of freedom. $D_1(\theta) = \frac{1}{\theta}\int_0^{\theta}\frac{t}{\exp(t)-1}dt$ is the Debye function and $D_2(\theta) = \int_0^1 t\log(t)(1-t)^{\frac{2(1-\theta)}{\theta}}dt$. Quantity ϵ is set to 1e-07 and is used to ensure that the restrictions on the space of θ are maintained. Finally $Q = 1 + (\theta - 1)(u + v)$ and $R = Q^2 - 4\theta(\theta - 1)uv$ (Marra and Radice 2017a).

In these models all the parameters in the vector $\boldsymbol{\vartheta} = (\mu_1, \sigma_1, \nu_1, \mu_2, \sigma_2, \nu_2, \zeta, \theta)$, are related to covariates and regression coefficients using additive predictors and in the case of the distribution parameters with restricted ranges, known monotonic link functions which ensure that the restrictions on the parameter spaces are maintained.

The additive predictors employed to model all the parameters that characterize the bivariate response can be defined as a generic predictor η_i function of an intercept and smooth functions of sub-vectors of a generic covariate vector $\mathbf{z}_i$. That is,

$$\eta_i = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}), \ i = 1, \ldots, n \tag{5.2}$$

where $\beta_0 \in \boldsymbol{R}$ is an overall intercept, $\boldsymbol{z}_{ki}$ denotes the k^{th} sub-vector of the complete covariate vector $\boldsymbol{z}_i$ (containing, for example, binary, categorical, continuous, and spatial variables) and the K functions

$s_k(\boldsymbol{z}_{ki})$ represent generic effects which are chosen according to the type of covariate(s) considered. We obtain then the model:

$$\begin{cases}\eta_{\mu_1} = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}) & \eta_{\mu_2} = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}) \\ \eta_{\sigma_1} = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}) & \eta_{\sigma_2} = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}) \\ \eta_{\nu_1} = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}) & \eta_{\nu_2} = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}) \\ \qquad\qquad \eta_{\theta} = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}) & \end{cases}$$

Each $s_k(\boldsymbol{z}_{ki})$ of the additive predictor can be approximated as a linear combination of J_k basis functions $b_{kj_k}(\boldsymbol{z}_{ki})$ and regression coefficients $\beta_{kj_k} \in \boldsymbol{R}$, i.e.

$$\sum_{j_k=1}^{J_k} \boldsymbol{\beta}_{kj_k} b_{kj_k}(\mathbf{z}_{ki}). \tag{5.3}$$

The generic effects $s_k(\boldsymbol{z}_{ki})$ can represent then linear effects, for which the equation 5.3 becomes $\boldsymbol{z}_{ki}^T\boldsymbol{\beta}_k$. For continuous variables the smooth functions are represented using a regression spline approach. For each continuous variable z_{ki}, $s_k(z_{ki})$ is approximated by $\sum_{jk=1}^{J_k} \boldsymbol{\beta}_{kj_k} b_{kj_k} z_{ki}$ where b_{kj_k} are known basis functions. To enforce some properties in the estimation, such as smoothness, a conventional integrated square second derivative spline penalty is typically employed, that is $\boldsymbol{D}_k = \int \boldsymbol{d}_k(z_k)\boldsymbol{d}_k(z_k)^T \boldsymbol{d}_{z_k}$, where the j_k^{th} element of $\boldsymbol{d}_k(z_k)$ is given by $\partial^2 b_{kj_j}(z_k)/\partial z_k^2$ and the integration over the range of z_k. As in the GAM models (Wood, 2006) each $\boldsymbol{\beta}_k$ has an associated quadratic penalty $\lambda_k \boldsymbol{\beta}_k \mathbf{D}_k \boldsymbol{\beta}_k$ to enforce the smoothness. The smoothing parameter $\lambda_k \in [0, \infty)$ controls the trade-off between fit and smoothness, and plays a crucial role in determining the shape of $\hat{s}_k(\boldsymbol{z}_{ki})$. The overall penalty can be defined as $\boldsymbol{\beta}\mathbf{D}\boldsymbol{\beta}^T$, where $\mathbf{D} = \text{diag}(0, \lambda_1 \mathbf{D}_1, \ldots, \lambda_K \mathbf{D}_K)$.

Marra and Radice (2017a) suggest the use of low rank thin regression splines (Wood, 2003; Wood, 2006) as basis function b_{kj_k} , however any of the splines implemented in the `mgcv` R package can be employed instead. Finally, the smooth functions are subject to centering (identifiability) constraints (Wood, 2006).

Furthermore, this models can also handle with spatial data information, if we have an area split up into discrete contiguous geographic units we can employ a Markov random field approach. In this case the equation (5.3) becomes $\boldsymbol{z}_{ki}^T\boldsymbol{\beta}_k$ where $\boldsymbol{\beta}$ represents the vector of spatial effects and $\boldsymbol{z}_{ki}$ is made up of a set of area labels. Moreover, a smoothing penalty based on the neighborhood strucuture is applied, being $\boldsymbol{D}_k$ an adjacency matrix, a recent application of the CGAMLSS for spatial data can be see in Duarte et al (2017).

5.2 Estimation process

The inference is based on a penalized maximum likelihood estimation. First, it is considered the log-likelihood function for a copula model with two continuous margins (Kolev and Pavia, 2009):

$$l(\boldsymbol{\delta}) = \sum_{i=1}^{n} log\left\{C(F_{1i}(y_{1i}|\mu_{1i}, \sigma_{1i}, \nu_{1i}), F_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i}); \zeta_i, \theta_i)\right\} +$$
$$\sum_{i=1}^{n}\sum_{m=1}^{2} log\left\{f_m(y_{mi} \mid \mu_{mi}, \sigma_{mi}, \nu_{mi}))\right\}$$

where parameter $\boldsymbol{\delta}$ is defined as $(\beta_{\mu1}^T, \beta_{\mu2}^T, \beta_{\sigma1}^T, \beta_{\sigma2}^T, \beta_{\nu1}^T, \beta_{\nu2}^T, \beta_{\theta}^T, \beta_{\zeta}^T)^T$.

Given the additive predictors structure casting non-linear relationships considered in these models, the use of a classic unpenalized optimization algorithm could excessively smooth the effects and therefore would not reflect the true trends found in the data. For that reason, Marra and Radice (2017a) propose the use of a penalized maximum likelihood, maximizing the expression:

$$l_p(\boldsymbol{\delta}) = l(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{S_\lambda} \boldsymbol{\delta} \tag{5.4}$$

where $\boldsymbol{S_\lambda} = \text{diag}(\boldsymbol{\lambda}_{\mu1}\boldsymbol{D}_{\mu1}, \boldsymbol{\lambda}_{\mu2}\boldsymbol{D}_{\mu2}, \boldsymbol{\lambda}_{\sigma1}\boldsymbol{D}_{\sigma1}, \boldsymbol{\lambda}_{\sigma2}\boldsymbol{D}_{\sigma2}, \boldsymbol{\lambda}_{\nu1}\boldsymbol{D}_{\nu1}, \boldsymbol{\lambda}_{\nu2}\boldsymbol{D}_{\nu2}, \boldsymbol{\lambda}_{\theta}\boldsymbol{D}_{\theta}, \boldsymbol{\lambda}_{\zeta}\boldsymbol{D}_{\zeta})$ with each smoothing parameters related to the corresponding $\boldsymbol{D}$ component and the overall $\boldsymbol{\lambda}$ is defined as $(\lambda_1, ..., \lambda_K)^T$.

To maximize expression (5.4) with an automatic selection of the smoothing parameter, $\boldsymbol{\lambda}$, CGAMLSS use a two step algorithm where the estimation of the vector $\boldsymbol{\delta}$ is achieved using a Trust Region algorithm in the first step and in the second, the smoothing parameter is selected using the methodology proposed in Wood (2004):

- Step 1: at a iteration index a, holding $\boldsymbol{\lambda}$ at a vector of values and with a starting vector $\boldsymbol{\delta}^{[a]}$ (obtained for the marginals parameters using the `gamlss` function from the `GJRM` package and for θ from the kendall's τ between the responses) the equation (5) is then maximized using a Trust Region algorithm, as follows:

  $$\min_{\boldsymbol{p}} \breve{l}_p \overset{\text{def}}{=} -\{l_p(\boldsymbol{\delta}^{[a]}) + \boldsymbol{p}^T\boldsymbol{g}_p^{[a]} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{H}_p^{[a]}\boldsymbol{p}\} \quad \text{so} \quad \text{that} \quad \|\boldsymbol{p}\| \leq \Delta^{[a]}$$

  $$\boldsymbol{\delta}^{[a]} = \arg\min_{\boldsymbol{p}} \quad \breve{l}_p(\boldsymbol{\delta}^{[a]}) + \boldsymbol{\delta}^{[a]} \tag{5.5}$$

  where $\boldsymbol{g}_p^{[a]}$ and the $\boldsymbol{H}_p^{[a]}$ are the gradient vector and Hessian matrix of the log-likelihood function penalized by the $\boldsymbol{S}_{\hat{\lambda}}$ matrix, $\| \cdot \|$ the euclidean norm and $\Delta^{[a]}$ is the radius of the trust region algorithm of each iteration.

  At each iteration of the algorithm, $\hat{l}(\boldsymbol{\delta}^{[a]})$ is minimized using a quadratic approximation and subject to the the constraint that the solution falls within a trust region with radius $\Delta^{[a]}$. Based on the ratio between the improvement in the objective function when going from $\boldsymbol{\delta}^{[a]}$ to $\boldsymbol{\delta}^{[a+1]}$ the proposed solution is then rejected or accepted and the region $\Delta^{[a]}$ expanded or shrunked (Conn et al, 2000, Nocedal and Wright, 2006), that is, given the expression:

  $$\phi = \frac{l_p(\boldsymbol{\delta}^{[a+i]} - l_p(\boldsymbol{\delta}^{[a]})}{g_p^{T[a])}p + \frac{1}{2}p^T H_p^{[a]} p}$$

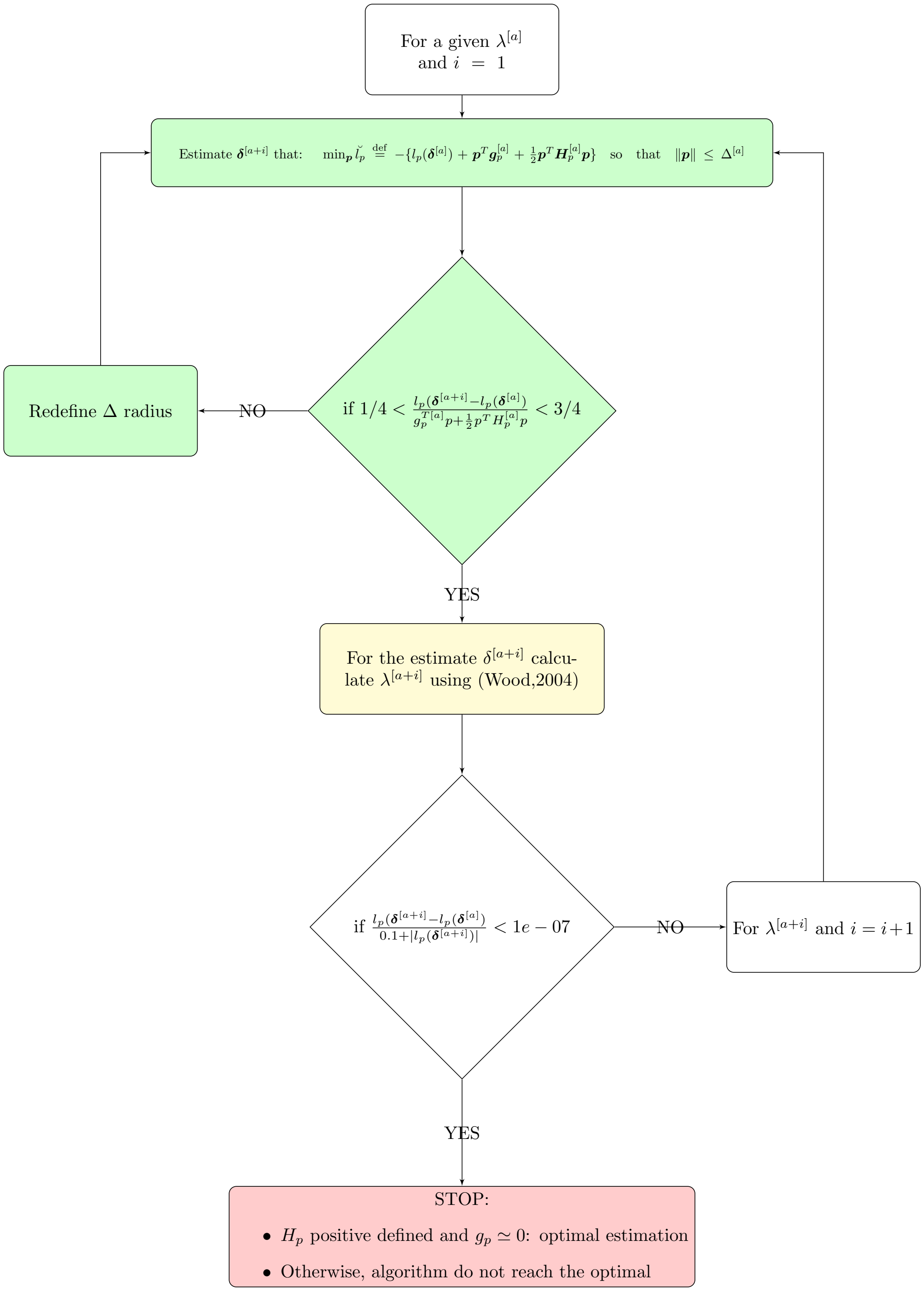

For a given $\lambda^{[a]}$ and $i = 1$
Estimate $\boldsymbol{\delta}^{[a+i]}$ that: $\min_{\boldsymbol{p}} \breve{l}_p \overset{\text{def}}{=} -\{l_p(\boldsymbol{\delta}^{[a]}) + \boldsymbol{p}^T \boldsymbol{g}_p^{[a]} + \frac{1}{2}\boldsymbol{p}^T \boldsymbol{H}_p^{[a]} \boldsymbol{p}\}$ so that $\|\boldsymbol{p}\| \leq \Delta^{[a]}$
if $1/4 < \frac{l_p(\boldsymbol{\delta}^{[a+i]} - l_p(\boldsymbol{\delta}^{[a]})}{g_p^{T[a]} p + \frac{1}{2} p^T H_p^{[a]} p} < 3/4$
NO
Redefine Δ radius
YES
For the estimate $\delta^{[a+i]}$ calculate $\lambda^{[a+i]}$ using (Wood,2004)
if $\frac{l_p(\boldsymbol{\delta}^{[a+i]} - l_p(\boldsymbol{\delta}^{[a]})}{0.1 + |l_p(\boldsymbol{\delta}^{[a+i]})|} < 1e - 07$
NO
For $\lambda^{[a+i]}$ and $i = i+1$
YES
STOP:
• H_p positive defined and $g_p \simeq 0$: optimal estimation
• Otherwise, algorithm do not reach the optimal

Figure 5.1: Scheme of the CGAMLSS estimation process. In green the first step and in yellow the second. Note that the algorithm stops when it is not able to get a better estimation of $\boldsymbol{\delta}$. If this happens and the Hessian matrix is not positive defined or/and the gradient vector is not close to zero, the algorithm was not able to reach an optimal estimation.

If $\phi \geq 1/4$, then we accept $[a + 1]$ as the next iterate, however if $\phi < 1/4$ then we do not accept

$[a+1]$ as the next iterate and decrese Δ to $1/4$ of its current value, while if $\phi > 3/4$, we do not accept $[a+1]$ as the next iterate either and increase Δ to 2 times its current value (Geyer, 2015; Nocedal and Wright, 2006).

The maximization of the equation (5.5) by means of a Trust Region algortihm instead of the use of a line search methods such as the Newton Raphson algorithm is justified becasue the first is generally more stable and faster, particularly for functions that are, for example, non-concave and/or exhibit regions that are close to flat (Marra and Radice, 2016; Nocedal and Wright 2006) common in this flexible copula regression models.

- Step 2: holding the models parameter vector value fixed at $\boldsymbol{\delta}^{[a+1]}$, solve the problem:

$$\boldsymbol{\lambda}^{[a+1]} = \underset{\boldsymbol{p}}{\arg\min} \quad V(\boldsymbol{\lambda}) \overset{\text{def}}{=} ||\boldsymbol{z}^{[a+1]} - \boldsymbol{A}_{\lambda^{[a]}}^{[a+1]} \boldsymbol{z}^{[a+1]}||^2 - \hat{n} + 2tr(\boldsymbol{A}_{\lambda^{[a]}}^{[a+1]}), \tag{5.6}$$

where, after defining $\boldsymbol{I}^{[a+1]} = -\boldsymbol{H}^{[a+1]}$, $\boldsymbol{z}^{[a+1]} = \sqrt{\boldsymbol{I}^{[a+1]}}\boldsymbol{\delta}^{[a+1]} + \sqrt{\boldsymbol{I}^{[a+1]}}^{-1} g^{[a+1]}$, $\boldsymbol{A}_{\lambda^{[a]}}^{[a+1]} = \sqrt{\boldsymbol{I}^{[a+1]}}(\boldsymbol{I}^{a+1} + \boldsymbol{S}_{\lambda^{[a]}})^{-1}\sqrt{\boldsymbol{I}^{[a+1]}}$, $tr(\boldsymbol{A}_{\lambda^{[a]}}^{[a+1]})$ represent the number of effective degrees of freedom (edf) of the penalized model and $\hat{n} = 7n$ (if we employ a three parameter distribution for both margins). Note that gradient vector and Hessian matrix needed in this expression is obtained as a byproduct of the previous step.

This expression is equivalent to the Un-Biased Risk Estimator (UBRE) (Wood 2006), solved with the methodology proposed in Wood (2004), the equivalence between the equation (5.6) and the UBRE estimator is given in Radice et al (2015).

These two steps are iterated until they are not able to improve the objective function $l_p(\boldsymbol{\delta})$, that is:

$$\frac{l_p(\boldsymbol{\delta}^{[a+i]} - l_p(\boldsymbol{\delta}^{[a]})}{0.1 + |l_p(\boldsymbol{\delta}^{[a+i]})|} < 1e - 07 \tag{5.7}$$

Note that (5.7) only depends in the $\boldsymbol{\delta}$ parameter vector estimation, and as pointed in Marra and Radice (2017a) proving the algorithm convergence when smoothing parameters are estimated in a performance iteration fashion is difficult and to the best of our knowledge this is still an open issue.

Some Inferential details

If the model reaches the convergence, reliable point-wise confidence intervals for linear and non-linear functions of the models coefficients are obtained using a Bayesian large sample approximation, as in GAM (Wood, 2006):

$$\boldsymbol{\delta} \quad \dot{\sim} \quad N(\hat{\boldsymbol{\delta}}, -\boldsymbol{H}_p(\hat{\boldsymbol{\delta}})^{-1}) \tag{5.8}$$

The justification to use this Bayesian approximation is given in Marra and Wood (2012) and it can be justified making the large sample assumption that $\boldsymbol{H}(\boldsymbol{\delta})$ can be treated as fixed, and making a Bayesian assumption on the prior of $\boldsymbol{\delta}$ for smooth models (Wood, 2006). Note that the equation (5.12) do not take into account the $\boldsymbol{\lambda}$, but based on Marra and Wood (2012) is not a problem provide that a heavy oversmoothing is avoided.

5.3 Model building process

This kind of modelling has three key points in the model building process that the researcher must solve, 1) choose suitable distributions for the marginals, 2) to chose the covariates for building the predictors for each parameter and 3) choose the copula function that best accommodate the type and/or strength of dependence between the margins. To solve these problems Marra and Radice (2017a) propose:

1. Choose suitable distributions for the marginals: use of the normalized quantile residuals for each marginal to choosing the marginal‘s distributions graphically. The normalized quantile residuals are defined as $\hat{r}_{mi} = \Phi^{-1}\{F_m(y_{mi}|\hat{\mu}_{mi}, \hat{\sigma}_{mi}, \hat{\nu}_{mi})\}$ for $i = 1, \ldots, n$ and $m = 1$, where Φ^{-1} is the quantile function of a standard normal distribution. If $F_m(y_{mi}|\hat{\mu}_{mi}, \hat{\sigma}_{mi}, \hat{\nu}_{mi})$ is close to the true distribution then the $\hat{r}_{mi}$ follow approximately a standard normal distribution, and using a Q-Q plot we can detect a lack of fit of the marginal distribution graphically. Moreover, we can base our selection in the AIC and BIC.

2. The selection of the covariates for the predictors can be based on AIC and BIC, but a good knowledge of the problem is required.

3. Given the marginals distributions and the predictors of the parameters, the choice of the copula function is based on the model AIC and BIC defined as $-2l(\hat{\delta}) + 2edf$ and $-2l(\hat{\delta}) + \log(n)edf$, respectively.

Chapter 6

Simulation study

Due to the recent proposal of the CGAMLSS there is only a simulation study of the general perform of the model in Marra and Radice (2017a), where CGAMLSS outperform the models proposed by Vatter and Chavez-Demoulin (2015), at sample sizes $n = 500$ and $n = 5000$ and for fixed variances the estimation of the smooth and linear effects in the marginal means and θ parameter were quite accurate, as well as the coverage properties of the point-wise confidence intervals.

However, given our medical problem where both responses present higher variability at higher glucose levels and increase with age, a simulation scenario where the variances are dependent on the predictors will be displayed to study the general perform of the model at different sample sizes. Moreover, given the recent introduction of CGAMLSS different model features have not been studied yet, mainly regarding the association parameter. For that reason, in a second scenario, the effect of the dependence strength between both margins in the estimation of the association parameters effects will be explored, as well as the effect of a copula misspecification in the estimation of the smooth and linear effects in the association parameter.

In this Chapter, the performance of the model will be study for different sample sizes. Then the effect of the strength of dependence between the margins in the estimation of the covariates effects in the association parameter is given. Finally, the effect that a copula missespecification has in the estimation of parametric and smooth effects in the association parameter is presented.

6.1 Performance of the model at different sample sizes

Two continuous outcomes following both a reverse-Gumbel distribution were considered. We have created two continuous covariates i.i.d uniformly distributed and a third variable was dichotomized so that each value had a 50% chance of appearing. Both continuous outcomes were joined using a gaussian copula to obtain a bivariate response. Linear and non-linear effects of the regressors on the parameters of the bivariate response were considered, of the form:

$$\begin{cases} \eta_{\mu_1} = 1.25x_1 + s_1(x_2) \\ \eta_{\sigma_1} = 1.1x_3 + s_1(x_2) \\ \eta_{\mu_2} = 0.9x_1 + s_2(x_2) \\ \eta_{\sigma_2} = 0.8x_3 + s_2(x_2) \\ \eta_{\theta} = 0.5x_1 + s_3(x_2) \\ \end{cases}$$

where $x_1, x_2 \in U(0,1)$ and x_3 is a categorical variable that can take the value 0 or 1. s_1,s_2 and s_3 are smooth functions, $s_1(x_2) = x_2 \sin(3x_2)$, $s_2(x_2) = \sin(\pi x_2)$ and $s_3(x_2) = \sin(2\pi x_2)$:

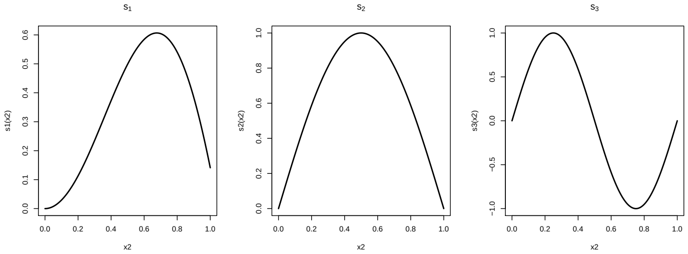


Figure 6.1: Simulated smooth effects in the simulation setup.

Sample sizes were set to $200, 500, 1000$ and 2000, and the number of replicates 1000 each. Models were fitted using the `copulaReg()` function of the `GJRM` package. In practice the R code used to simulate and fit the models is:

```r
library(copula)  # Let us simulate a bivariate response
library(gamlss)  # Include the distributions for the marginals
library(GJRM)    # Let us fit the model

# We define some non-linear effects

f1 <- function(x) x * sin(3 * x)
f2 <- function(x) sin(pi*x)
f3 <- function(x) sin(2*pi*x)

# The data.gen function (Marra and Radice, 2017a), let us generate a bivariate
#response conditioned to a set of covariates

data.gen <- function(f1,f2,f3){

# Define the covarites:

x1=runif(1,0,1)
x2=runif(1,0,1)
x3=round(runif(1,0,1),0)

# Define the predictors

eta_mu1 <- 1.25*x1 + f1(x2)
eta_mu2 <- 0.9*x1 + f2(x2)
eta_sigma21 <- 1.1*x3 + f1(x2)
eta_sigma22 <- 0.8*x3 + f2(x2)
eta_theta <- 0.5+x1+f3(x2)

theta.para <- tanh(eta_theta) # Theta parameter with the inverse of the link function

speclist1 <- list( mu = eta_mu1, sigma = exp(eta_sigma21)) #Marginal 1 Parameters
speclist2 <- list( mu = eta_mu2, sigma = exp(eta_sigma22)) #Marginal 2 Parameters

#Note that we use exp(eta_sigma21) because of the link function (log) for sigma

spec <- mvdc(copula = Cop, c("RG", "RG"),list(speclist1, speclist2))

#The mvcd build a bivariate response defined by the marginals and the Copula function

c(rMvdc(1, spec), x1, x2, x3)}

# Any time that we run the funcion we obtain one row with the bivariate
#response and the covariates

#Simulate data
```

```r

datos=t(replicate(n,data.gen(f1,f2,f3),simplify=T))
colnames(datos)=c("y1","y2","x1","x2","x3")
datos=as.data.frame(datos)
datos$x3=as.factor(datos$x3)

#Define the predictors for the model

mu1=y1~x1+s(x2)
mu2=y2~x1+s(x2)
sd1=~x3+s(x2)
sd2=~x3+s(x2)
theta=~x1+s(x2)

f=list(mu1,mu2,sd1,sd2,theta)

#Fit the model

m1=copulaReg(f,margins=c("rGU","rGU"),BivD ="N",data=datos)

#Evaluate the model estimation

m1$fit$converged
m1$iter.if
m1$iter.sp
m1$iter.inner


s1=summary(m1) #From the summary we obtain the parametric effects, std. deviation and p values

#Evaluate the Smooth Effects in 200 equally spaced fixed points:

newdata=data.frame(x1=0,x2=seq(0,1,len=200),x3=0)

p1=predict(m1,eq=1,newdata=newdata,type="terms",se=T)

p2=predict(m1,eq=2,newdata=newdata,type="terms",se=T)

p3=predict(m1,eq=3,newdata=newdata,type="terms",se=T)

p4=predict(m1,eq=4,newdata=newdata,type="terms",se=T)

p5=predict(m1,eq=5,newdata=newdata,type="terms",se=T)
```

The package `copula` (Hofert et al, 2017) contains the functions, `ellipCopula()`, `mvcd()` and `rMvcd()` which let us to simulate from any elliptical copula. The package `gamlss` contains the function required to simulate values from a reverse Gumbel distribution. The various **eta** refer to the marginal and copula parameter predictors. These are transformed in **speclist1**, **speclist2** and **theta.para** with the inverse of the link function to ensure that the restrictions on the parameters' spaces are maintained. Given the data simulated using the **data.gen** function (Marra and Radice, 2017a) we have fitted a model from which we stored the estimated linear effects with the standard error of the estimation (evaluated with the `summary()` function), and in the case of the smooth effects have been evaluated in 200 equally spaced fixed points in the $(0, 1)$ range along with the estimated standard error associated to each point, using the `predict()` function. In addition, some estimation details like the number of iterations needed by the algorithm to reach convergence were stored.

From these stored results we can analyze graphically the estimators of the smooth and parametric effects in the marginal parameters and given that the main breakthrough of CGAMLSS is the possibility of study the effect of different covariates in the association between two responses. The estimates of the predictor effects in the θ parameter will be studied deeply. The estimation accuracy and precission will be studied using Bias, given by $|\bar{\hat{\beta}} - \beta|$, for the linear effects being $\bar{\hat{\beta}} = \frac{1}{1000}\sum_1^{1000}\beta_i$ the mean of the estimates, and β the real effect.

The same formula will be applied for the smooth terms using the expression $n_s^{-1}\sum_{i=1}^{n_s}|\bar{\hat{s}}_i - s_i|$ being $\bar{\hat{s}}_i = \frac{1}{1000}\sum_{rep=1}^{1000}\hat{s}_{rep,i}$, n_s the 200 equally spaces points and s_i the equivalent point from the real effect. Moreover, we will use the RMSE given by $\sqrt{\frac{1}{1000}(\hat{\beta}_i - \beta_i)^2}$ for the linear terms and by $\sum_{i=1}^{n_s}\sqrt{\frac{1}{1000}\sum_{rep=1}^{1000}(\hat{s}_{rep,i} - s_i)^2}$ for the smooth ones.

The coverage properties of the CGAMLSS intervals will be studied for the linear terms, checking how many times the real β was inside of the confidence intervals constructed for the linear terms, while the coverage for the smooth terms will be evaluated in the 200 equally spaced points along the range of the covariate x_2.

Since the biggest advance of the CGAMLSS is the possibility of study which covariates are related to the association between two responses given by the copula association θ parameter, we will present graphically the results of the estimations of the parametric and smooth effects on the margins parameters and study more exhaustively the parametric and smooth effects on the association parameter θ.

Marginal predictors

Figures 6.3 and 6.4 show the estimation of the smooth effects $s_1(x_2)$ and $s_2(x_2)$ in the marginal parameters. The estimation of the smooth effect from the η_{σ_1} and η_{σ_2} predictors, seems satisfactory for all samples sizes, with the mean of the one thousand estimates closer to the true effect as the sample size increases. However, the smooth functions estimates of the location parameter, needs a bigger sample size to get closer to the real effect.

Figure 6.2 displays the estimates of the parametric effects for the marginal predictors. We can see that all mean estimates are very close to true values and their variability decrease as the sample size increases.

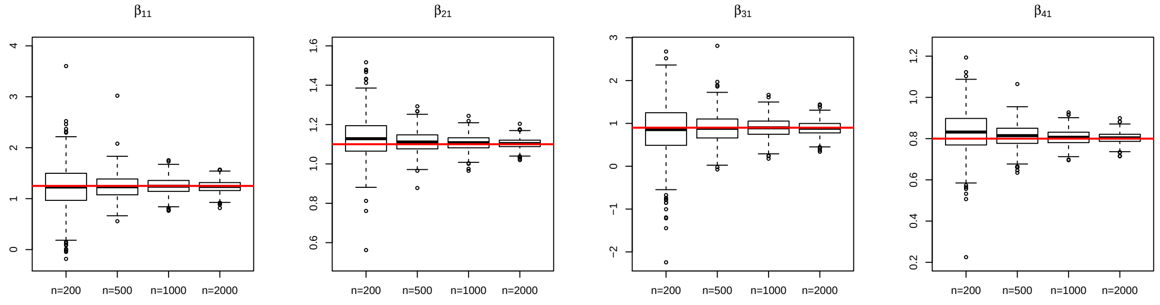


Figure 6.2: Estimation of the parametric effects for the margin parameters. β_{11} refers to the parametric effect of the η_{μ_1} predictor, β_{21} refers to the parametric effect of the η_{σ_1} predictor, β_{31} refers to the parametric effect of the η_{μ_2} predictor and β_{41} refers to the parametric effect of the η_{σ_2} predictor.

Association between the responses

Figure 6.5 depicts the parametric effect estimates of the θ parameter predictor, as well as the coverage rate of the confidence intervals constructed for each estimate. All mean estimates are very close to the real effect (Table 6.1) and the variability of the estimations decreases as the sample size increases

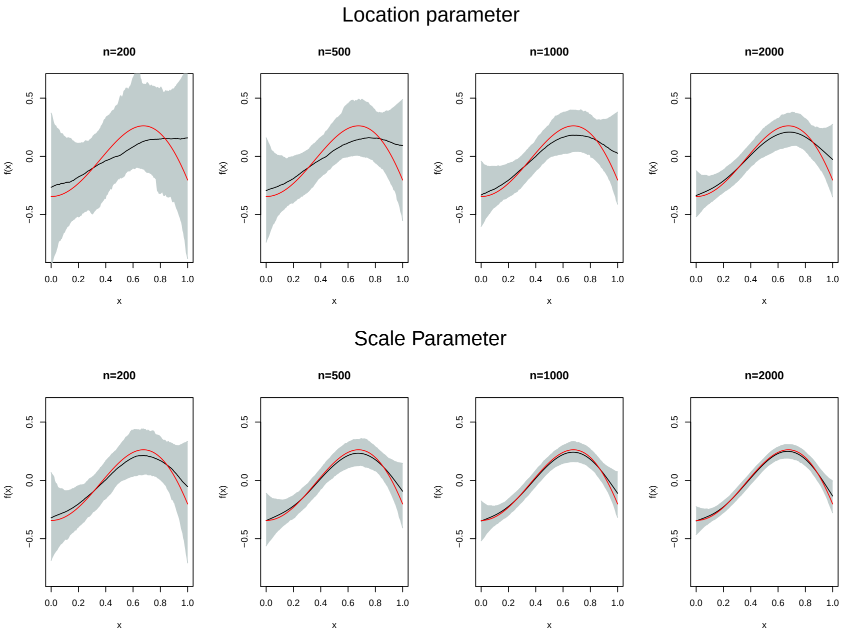


Figure 6.3: Estimates of the smooth effect on the parameters of the first margin at different sample sizes. In red the real function, and in black the mean of the 1000 estimates, the grey shade represent the 2.75 and the 97.5 quantiles.

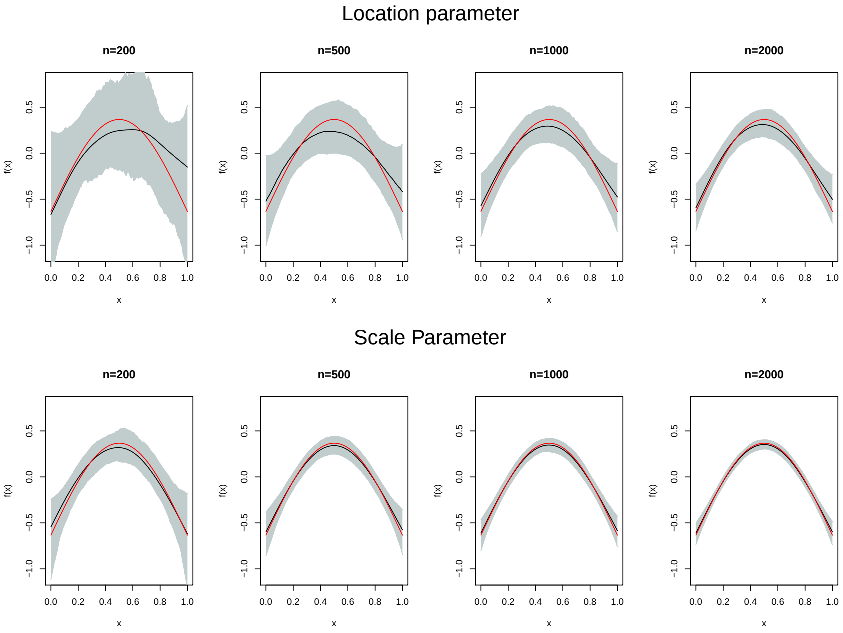
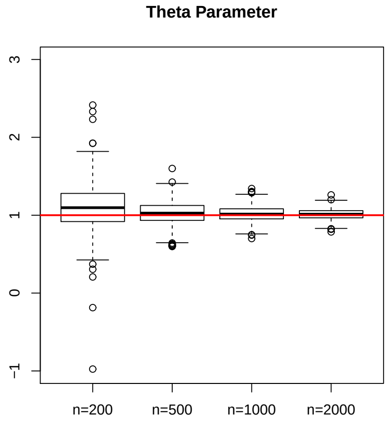


Figure 6.4: Estimates of the smooth effect on the parameters of the second margin at different sample sizes. In red the real function, and in black the mean of the 1000 estimates, the grey shade represent the 2.75 and the 97.5 quantiles.

(Figure 6.5). The coverage rates of the confidence intervals constructed for the parametric effect are close to the significance levels for all sample sizes and get closer as this increases (Figure 6.5).



	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
n=200	0.833	0.898	0.965
n=500	0.871	0.928	0.978
n=1000	0.879	0.934	0.986
n=2000	0.884	0.948	0.987

Figure 6.5: Parametric effects estimation of parameter θ (left) and the coverage of the confidence intervals for the linear effects at different samples sizes ($n = 200, 500, 1000, 2000$) and significance levels, α (0.1, 0.05, 0.01) (right).

In the Figure 6.6 the estimates of the non-linear effect $s_3(x_2)$ in the association parameter at different sample sizes are shown. For all sample sizes the estimation presents a good performance with the mean of the estimates close to the real effect, as expected, as the sample size increases, the variability of the estimates decreases and their mean becomes closer to the real effect. In the Table 6.1, we can also see that both the RMSE and bias drop with sample size. Moreover, the rates of coberture of the point-wise confidence intervals constructed for the smooth effect $s_3(x_2)$ are near to the significance level along the range of the predictor covariate x_2, and getting closer for higher sample sizes (Figure 6.7).

	Parametric Effect		Smooth Effect	
	Bias	RMSE	Bias	RMSE
n=200	0.100	0.24	0.05	1.28
n=500	0.03	0.12	0.02	0.82
n=1000	0.02	0.08	0.01	0.80
n=2000	0.01	0.06	0.008	0.80

Table 6.1: RMSE and Bias for the estimates for the parametric and smooth terms at different sample sizes.

The estimation process is quite similar for different sample sizes, showing failures of convergence in 4 every one thousand models at n=200, and 1 every one thousand at n=500. In terms of the iterations needed by the algorithm to reach the optimal estimation in the trust region and smoothing parameter

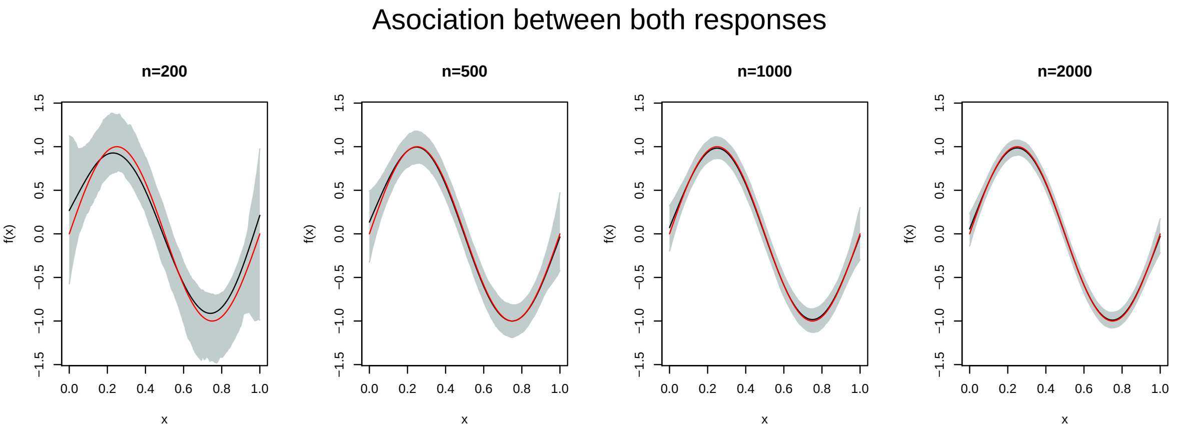


Figure 6.6: Estimates of the smooth effect on the association parameter at different sample sizes. In red the real function, and in black the mean of the 1000 estimates, the grey shade represents the 2.75 and the 97.5 quantiles.

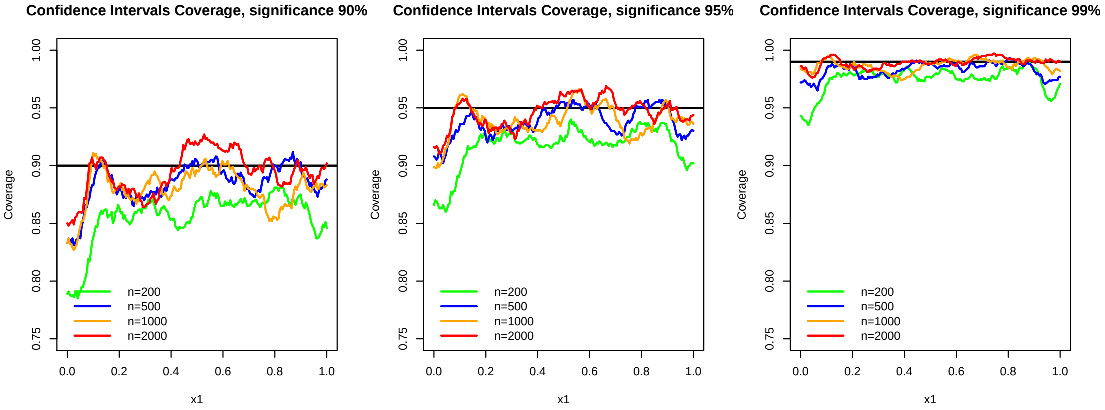


Figure 6.7: Coverage properties at different levels of significance of the CGAMLSS confidence intervals for the smooth effect on the θ parameter at different sample sizes.

loops there is no significant differences between the different sample sizes, however the trust region iterations within smoothing loops decrease for higher samples sizes, with a mean of 114 loops at n=200, 46 at n=500, 30 at n=1000 and 24 at n=2000.

6.2 Different strengths of dependence between margins

In this scenario following the work of Klein and Kneib (2016) we will study the effect of dependence strength between the margins in the estimation of the predictor effects, considering three scenarios that represent weak, constant and strong dependence.

Two gaussian responses were assumed, and joined under a gaussian copula. The copula association parameter was specified as follow:

1. Constant Dependence: $\theta = \log(3)$

2. Weak Dependence: $\theta = 0.50\cos(\pi x)$
3. Strong Dependence: $\theta = \log(4.5 - 1.7\sin(\pi x))$

Where $x \in U(0,1)$ and considering four sample sizes $n = 500, 1000, 2000, 5000$ with 1000 replicates each.

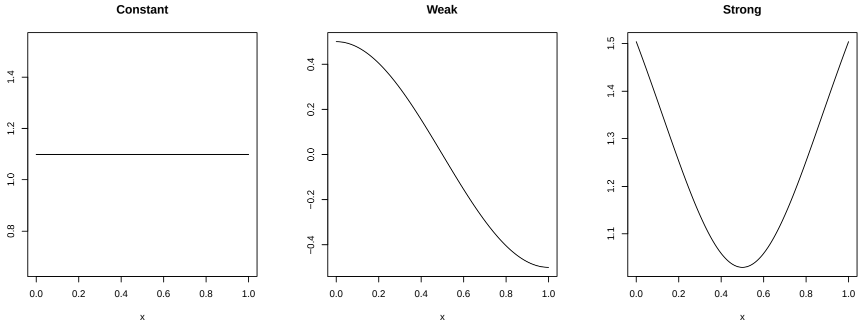


Figure 6.8: Representation of the different strengths of dependence simulated.

The code to simulate and fit the model is analogous to the applied in the previous scenario, and the evaluation of the effects was done using a 200 equally spaced points as well.

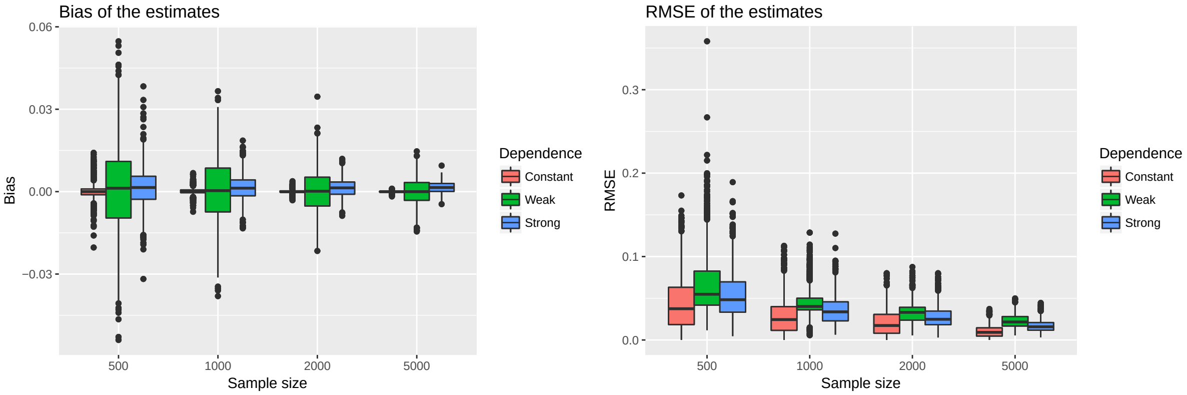


Figure 6.9: Bias and RMSE of the 1000 estimates for the different samples sizes and degrees of dependence. The formulas considered to represent this graphs were for the bias $n_s^{-1}(\hat{s_{i,j}} - s_i)$ and for the RMSE $\sum_1^{ns}\sqrt{(\hat{s_{i,j}} - s_i)^2}$, being n_s two hundred equally espaced fixed points, and $j = 1, \ldots, 1000$.

We can see that for the weak dependence model both the Bias and RMSE of the estimation is a slightly higher (Figure 6.9) and clearly the coverage of the point wise estimated intervals, is worst than for the constant and strong dependences, requiring a higher sample size to become closer to the significance level (Figure 6.10).

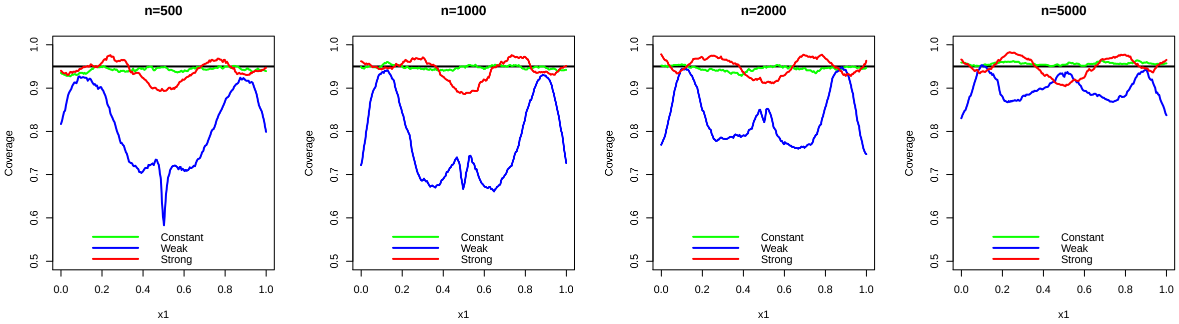


Figure 6.10: Coverage for the smooth terms at 95% for the different samples sizes and strengths of dependence.

6.3 Copula missespecification

In this simulation scenario two gaussian responses were considered and joined with a Clayton copula to produce a bivariate distribution. The bivariate distribution parameters were related to two continuous covariates and one binary covariate through predictors specified as in the simulation study presented by Marra and Radice (2017a):

$$\begin{cases} \eta_{\mu_1} = 0.5 - 1.25x_2 - 0.8x_3 \\ \eta_{\mu_2} = 0.1 - 0.8 * x_1 + s_1(x_2) \\ \eta_{\sigma_1} = 1.8 \\ \eta_{\sigma_2} = 0.1 \\ \eta_\theta = 0.2 + 2x_1 + s_4(x_2) \end{cases}$$

where $x_1, x_2 \in U(0,1)$ and x_3 a covariate dichotomized as in the first scenario, being $s_1(x_2) = x\sin(3x_2)$ the smooth term used in the first scenario as well, and $s_4(x_2) = \sin(3\pi x_2)$:

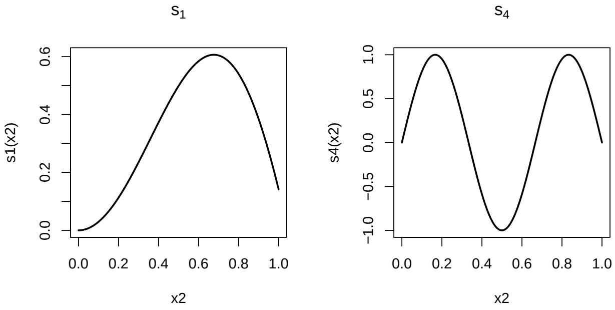


Figure 6.11: Smooth effects consider in this simulation study.

In this scenario two different models will be fitted, one under the right copula assumption using a Clayton copula, and the other under a copula misspecification considering a Clayton 180^o rotated instead. Both models will be fitted using the same dataset, `set.sedd(123)`, considering a big sample size of $n = 2000$ and 1000 replicates. The choice of these two copula functions is justified because for the same range of the θ parameter show show two antagonic structures of dependence. The clayton copula, from which the data were simulated, is appropiate lower tail dependences, while the $C180$ is appropiate the opposite situation of upper tail dependence (Figure 6.12).

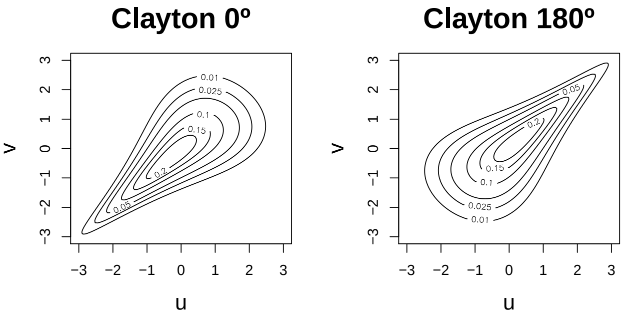


Figure 6.12: Contour plots of both copula functions with gaussian margins, the data will be generated with the structure of dependence like the Clayton 0 (left) and fitted under both copula assumptions.

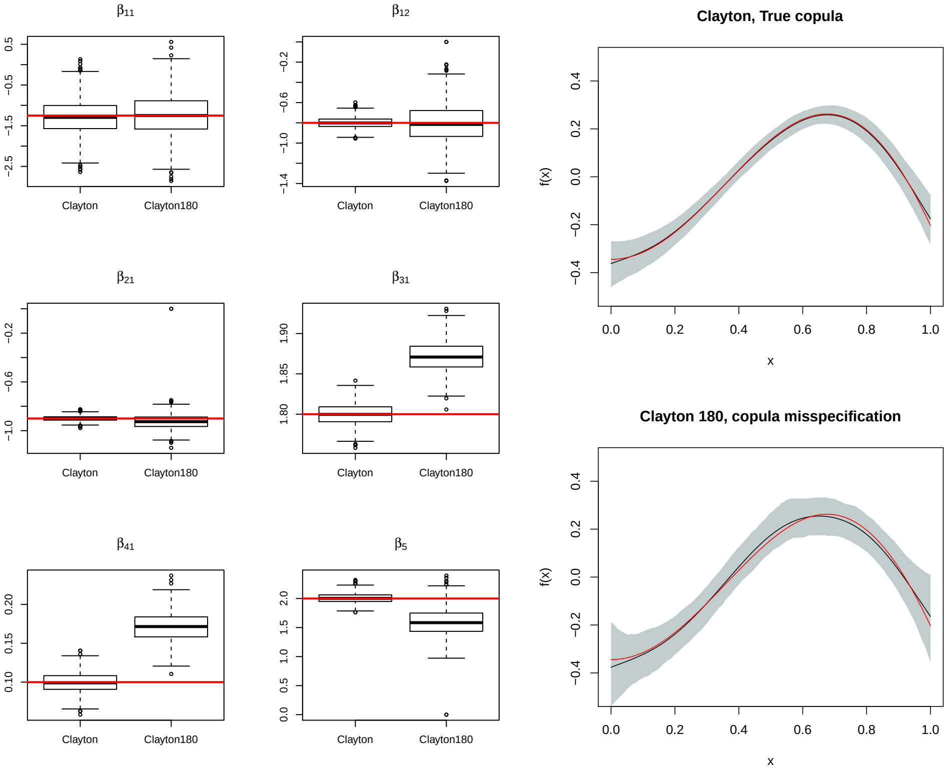


Figure 6.13: Estimation of the parametric and smooth effects in the θ parameter under both copula assumptions, for the smooth effec, in red the real effects, in black the mean of the estimations. In the parametric effects, β_{11} and β_{12} represent the parametric of the η_{μ_1} predictor, β_{21} the parametric effect of the η_{μ_2} predictor, β_{31} and β_{41} the intercept of the η_{σ_1} and η_{σ_2} respectively, and finally β_5 the parametric effect of the association parameter predictor, and the horizontal line represent the true effect.

In Figure 6.13 we can see the estimation of the parametric and smooth effects in the margins parameters. Concerning the effects on the means, the model under the right copula assumption delivery lower variability in the estimates, for both the parametric and smooth effects, but the mean of the 1000 estimates are very similar in both cases.

However, in the estimation of the effects of the scale parameters predictors we see a higher effect of the copula misspecification, displaying a clearly biased estimate of the variability levels for both margins, while in the model under the right copula assumption, the estimation of the variability levels for both responses are very close to the true effect.

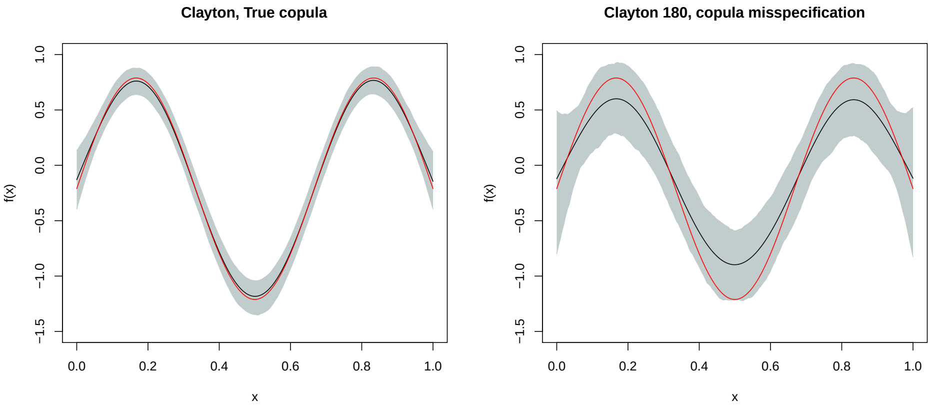


Figure 6.14: Estimation of the parametric effects for the marginal and association and the estimation of the smooth term for the μ_2 under both copula assumptions.

Focusing on the estimation of the θ smooth and linear effect, in Figure 6.13 we can see that the model under the copula misspecification show higher variability in the estimates and their mean is far from the true effect. Additionally, in Figure 6.14 similar results are found for the smooth term estimation.

Chapter 7

Bivariate modelling of glycemic markers

The Clinical Epidemiological Unit of the Clinical Hospital of Santiago de Compostela (CHUS), presented a biomedical problem derived from the AEGIS project. One of the aims of this project, launched in the year 2012, was to determine which factors were involved in the glycation rate of the glycated hemoglobin and the discordances encountered between it and other glycated proteins, such as the glycated albumin.

The introduction of the flexible copula regression models, by means of CGAMLSS, discussed in this work, offers a methodological solution letting us study properly the association between the responses, as well as their mean and variability conditioned to a set of clinical factors, all in a flexible framework where non-standard marginal can be considered and also different structures of dependence between them.

In this Chapter, the proposed model to study the concentration of both proteins jointly is presented. Finally the results obtained from the analysis with the CGAMLSS approach is given.

7.1 Model Building

The choice of the parametric distributions for both responses was based on the AIC criteria, being for the HbA1c and GA the Sigh-Maddala and DAGUM the distributions that better fit our data. Both distributions have been developed in the econometrical field and their interpretability requires that each of the three parameters must be modeled by the covariates in order to understand the effect on terms of the expectation and variability of the responses. However, for our data we have had convergence failures under both distributional assumptions. In the case of the GA, the FISK distribution also displays a good fit to the response, however the right interpretability of this distribution is also complex and only is possible if $\sigma > 2$.

For that reason, both proteins have been modeled under the assumption of a reverse Gumbel distribution (see Apendix A), this distribution is appropiate for skewed distribution with right heavy tails, with a higher interpretability in the medical field, since the variance is equal to the product of a constatn for σ^2 and μ is the mode of the distribution and can be easily transformed to get the expectation using σ.

The covariates selection was based on the AIC of the fitted models, as well as in the knowledge about the medical problem. In our experience the best way to build the model predictors in the CGAMLSS class is a previous characterization of the covariates effects in the marginals using first a GAM for study the effects in the means and GAMLSS for studying the variability, given that the covariates selection

Glycated Hemoglobin			Glycated Albumin		
Distribution	**AIC**	**BIC**	**Distribution**	**AIC**	**BIC**
Sigh-Maddala	2049.50	2065.45	Sigh-Maddala	6438.64	6454.58
DAGUM	2159.77	2175.72	DAGUM	6453.16	65469.10
Reverse Gumbel	2249.78	2260.41	FISK	6485.16	6495.80
FISK	2365.52	2376.48	Reverse Gumbel	6533.56	6544.20
Logistic	2568.52	2579.15	Inverse Gaussian	6619.41	6630.04
Log-normal	2713.5	2724.12	Log-normal	6621.44	6632.06
Inverse Gaussian	2716.44	2727.07	Logistic	6637.15	6647.78
Gamma	2807.08	2817.71	Gamma	6667.84	6678.47
Gausian	3011.92	3022.55	Gausian	6667.84	6824.11
Weibull	3671.40	3682.03	Weibull	7163.43	7174.062
Gumbel	4096.78	4107.41	Gumbel	7552.15	7562.78

Table 7.1: AIC and BIC for all the continuous distributions available in the GJMR package for the levels of Fructosamine and glycated hemoglobin.

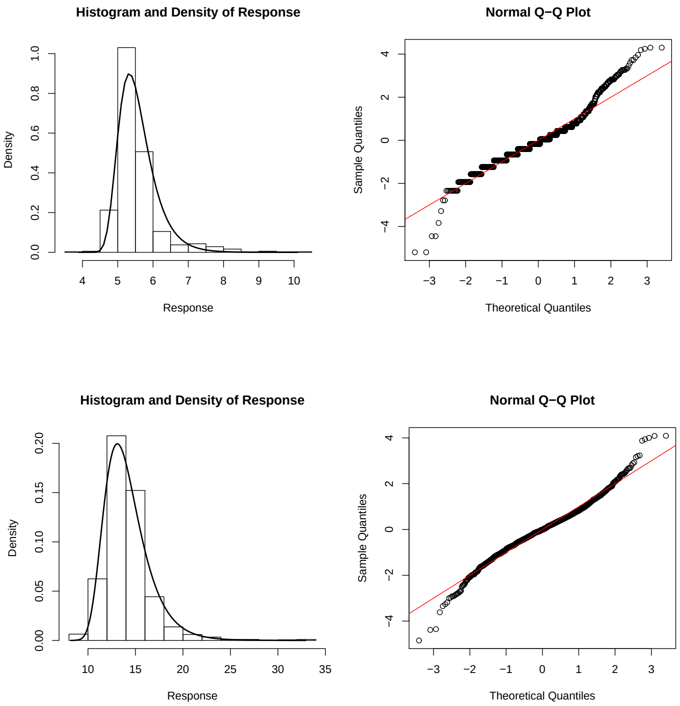


Figure 7.1: Distributions and QQ-plots of the normalized residuals for both responses. For the glycated hemoglobin (up) and glycated albumin (down).

based exclusively in the fitted models AIC, or computational cost, is a time consuming task due to the fact that the inclusion of a non-informative covariates in the first margin has a negative effect on the estimation of the other predictors effects, given wiggly estimates, even existing an association between the covariate and the reponse parameter. Based on the clinical knowledge of the problem and a previous characterization of the margins using a GAMLSS the proposed model is:

$$\begin{cases} \eta_i^{\mu_1} = \beta_{0i}^{\mu_1} + Gender_i\beta_{1i}^{\mu_1} + Smoke\beta_{2i}^{\mu_1} + Alcohol\beta_{3i}^{\mu_1} + Exercise\beta_{4i}^{\mu_1} + s_i^{\mu_1}(Glucose) + s_i^{\mu_1}(Age) + \\ + s_i^{\mu_1}(BMI) + s_i^{\mu_1}(MCV)_i, \\ \eta_i^{\sigma_1^2} = \beta_{0i}^{\sigma_1^2} + Gender_i\beta_{1i}^{\mu_1} + Smoke\beta_{2i}^{\mu_1} + Alcohol\beta_{3i}^{\mu_1} + Exercise\beta_{4i}^{\mu_1} + s_i^{\sigma_1^2}(Glucose) + s_i^{\sigma_1^2}(Age) \\ \eta_i^{\mu_2} = \beta_{0i}^{\mu_2} + Gender_i\beta_{1i}^{\mu_2} + Smoke\beta_{2i}^{\mu_2} + Alcohol\beta_{3i}^{\mu_2} + Exercise\beta_{4i}^{\mu_2} + s_i^{\mu_2}(Glucose) + s_i^{\mu_2}(Age) + \\ + s_i^{\mu_1}(BMI) + s_i^{\mu_2}(T3), \\ \eta_i^{\sigma_2^2} = \beta_{0i}^{\sigma_2^2} + Gender_i\beta_{1i}^{\sigma_2} + Smoke\beta_{2i}^{\sigma_2} + Alcohol\beta_{3i}^{\sigma_2} + Exercise\beta_{4i}^{\sigma_2} + s_i^{\sigma_2^2}(Glucose) + s_i^{\sigma_2^2}(Age), \\ \eta_i^{\theta} = \beta_{0i}^{\theta} + Gender_i\beta_{1i}^{\theta} + Smoke\beta_{2i}^{\theta} + Alcohol\beta_{3i}^{\theta} + Exercise\beta_{4i}^{\theta} + s_i^{\theta}(Glucose) + s_i^{\theta}(Age) + \\ + s_i^{\theta}(MCV), \end{cases} \tag{7.1}$$

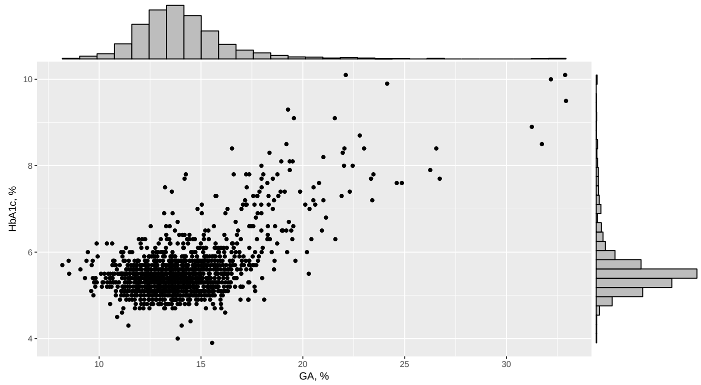


Figure 7.2: Relationship between the levels of Glycated Albumin and Glycated Hemoglobin. This is our bivariate response.

The copula was chosen based on the AIC and BIC of the models fitted under different copula assumptions. As expected in this case where at higher concentration of both proteins higher correlation, the model fitted under the Joe copula assumption presents the best AIC and BIC, this suggests that the glycated hemoglobina and the glycated albumin present an association with an upper tail dependence structure.

However, the differences between the AIC and BIC under the Gaussian and the Joe copula assumption are not big and the "scarcity" of the Hessian matrix for the Joe copula model is bigger, this suggest that the model under the Joe copula assumption could need more data to model the relation between both proteins properly. Moreover, the confidence intervals for the Kendalls τ and θ parameter under the assumption of Gaussian copula are not fully positive showing ($\theta = (-0.0946, 0.379), \tau = (-0.0618, 0.26)$).

Taking in account that the effect on the marginals is almost equal in both models, and the range of the Kendalls τ does not seem really negative, the parametric and smooth estimation for the θ parameter predictor will be presented under both copula assumptions, while the effects of the covariates for margins only for the model built under the Joe copula assumption, being pretty similar for the gaussian copula model.

Copula	AIC	BIC
Joe	6074.62	6654.23
Student-T	6130.31	6663.61
N	6150.02	6737.74
F	6207.19	6763.40
AMH	6234.68	6771.96
FGM	6294.20	6789.90
Plackett	6394.90	7022.66
Clayton	7192.50	7986.98
Clayton180	7497.39	8417.58
Gumbel	33241.51	33299.64
Gumbel180	40705.27	41645.00
Joe180	66877.99	66957.16

Table 7.2: AIC and BIC for different copula assumption for modeling the association between both proteins. In red, models that have shown failures of convergence.

7.2 Results

Regarding the marginal expectations (Figure 7.3 and Figure 7.4), fasting plasma glucose is the main covariate on predicting both glycated hemoglobin and glycated albumin concentrations and the functional form of the effect of glucose levels on both proteins is similar increasing for values higher than $100mg/dL$. Mean concentrations of glycated hemoglobin increase linearly with age while glycated albumin concentrations only do so for elderly people ($> 60years$). Body mass index, an index of obesity is a negative factor influencing serum glycated albumin. However, it showed a significant positive relation with glycated hemoglobin

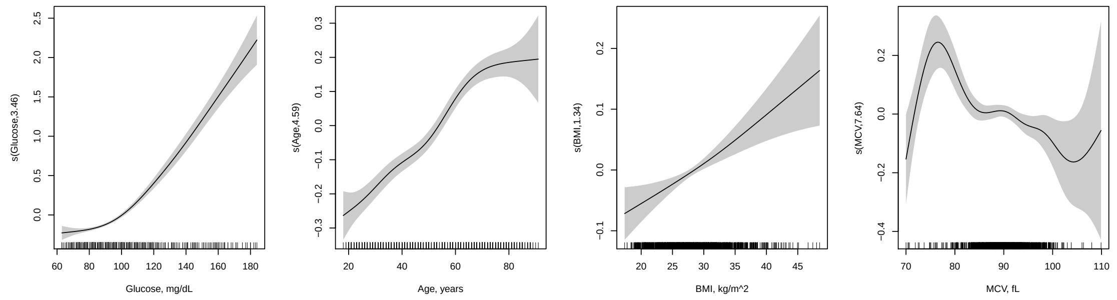


Figure 7.3: Smooth functions estimate and associated 95% and point-wise confidence intervals of the effect of Glucose, Age, Body mass index (BMI) and mean corpuscular volume (MCV) of the red blood cells (RBC) on the glycated hemoglobin levels.

As explained in Chapter 2, we have selected mean corpuscular volume as covariate in the first marginal and T3 hormone in the second. Mean corpuscular volume is an indirect measure of the red blood cells turnover (or life-span), being hemoglobin a protein that is inside the red cells. Glycated hemoglobin decrease in response to macrocytosis (higher mean corpuscular volume), a condition that is most commonly associated with vitamin deficiencies and alcohol consumption, and it is also decrease

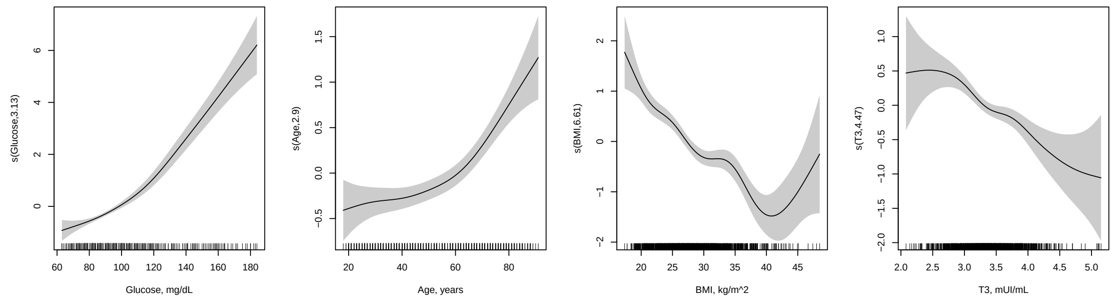


Figure 7.4: Smooth functions estimate and associated 95% and point-wise confidence intervals of the effect of Glucose, Age, Body mass index (BMI) and thyroid hormones (T3) on the glycated albumin levels.

in response to blood loss (microcytosis), a frequent condition in young women, in both circustances the red blood cells are destroying at a higher rate than in health subjects and hence there is a bigger proportion of young RBCs in the blood stream, assembled in response to this higher RBCs destruction.

For the second margin we have considered the T3 hormone, that modifies albumin metabolism, being this protein the substrate of the glycation process. As expected, a negative relation between the levels of glycated albumin and thyroid hormones concentration is present, in consequence of the higher albumin turnover in response to the elevated levels of Thyroid hormones.

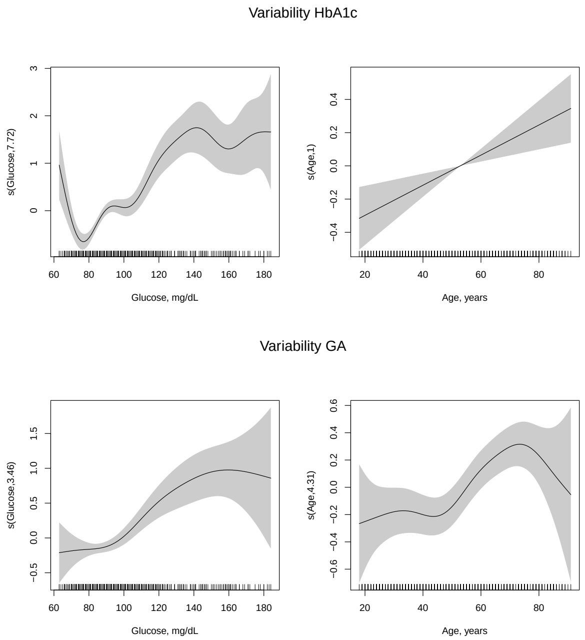


Figure 7.5: Smooth functions estimate and associated 95% and point-wise confidence intervals of the effect of Glucose and Age on the variability of the glycated hemoglobin and albumin levels.

Regarding the marginal variance (Figure 7.5), variabilities of glycated hemoglobin and glycated albumin are higher at higher glucose levels and increase with age. A larger variability is expected in both glycated proteins at higher levels of glucose, indicating those people with diabetes. The high variability in the glycated hemoglobin at lower levels of glucose could also indicate the presence of some people with diabetes being treated with anti-diabetic drugs and thus presenting low glucose levels when measured.

Finally, the association between glycated hemoglobin and glycated albumin, is presented under two copula assumptions.

Under the gaussian copula assumption, the correlation between both responses increase with the fasting plasma glucose concentration and is also higher at low glucose levels, as in the variance case, due to the presence of people with diabetes being treated with anti-diabetic drugs and thus presenting low fasting plasma glucose levels when measure. Aging has a clear effect on the correlation between both proteins, and the MCV shows a U-shape effect, reflecting that the correlation between both is higher when the RBC turnover is modified in case of anemia disease, either microcytic or macrocytic.

Under the Joe copula assumption the effect of the age is not statistical significant, and the fasting plasma glucose effect does not get the diabetic patients being treated with antidiabetic drugs, showing a linear increase until 130 mg/dL. However, the U-shaped MCV effect is very similar to the effect obtained under the gaussian copula assumption.

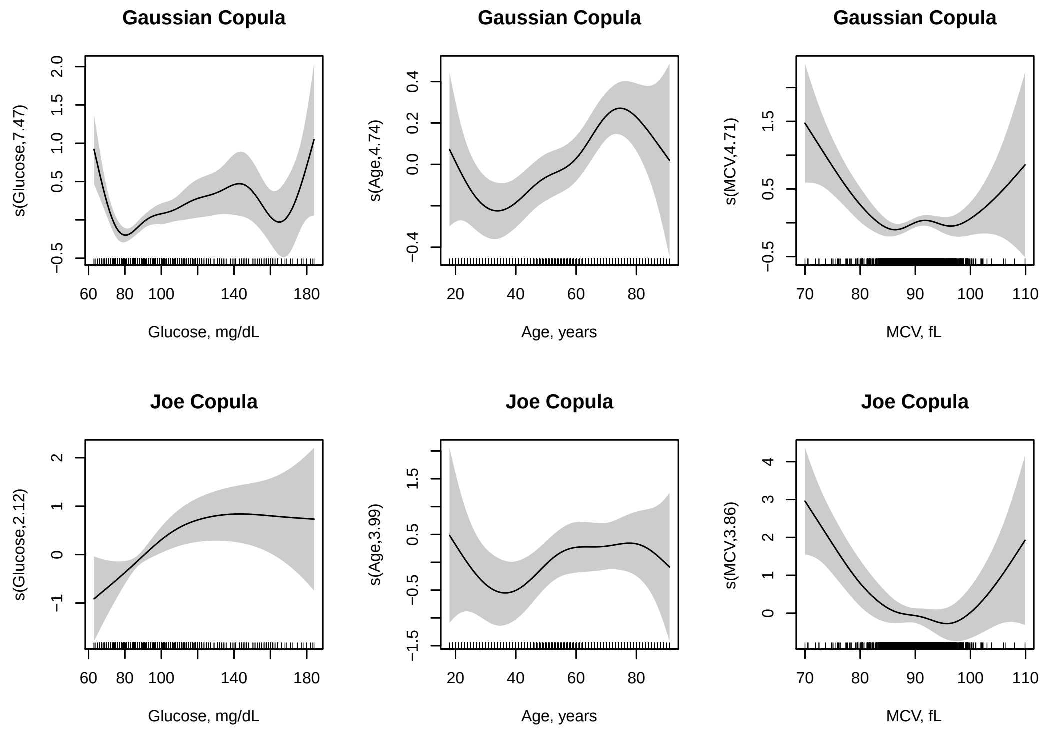


Figure 7.6: Smooth functions estimate and associated 95% and point-wise confidence intervals of the effect of Glucose, Age and the mean corpuscular volume of the red blood cells (MCV) on the association between the glycated albumin and glycated hemoglobin levels under a gaussian copula(up), and joe copula (down) assumption.

A better idea of the covariates effect on the mean levels, variability and association between both glycated protein levels, can be obtained looking at the contours plots depicted in Figures 7.7 and 7.8.

	HbA1c μ			HbA1c σ			GA μ			GA σ		
	Coefficient	SE	P value	Coefficient	SE	P value	Coefficient	SE	P value	Coefficient	SE	P value
Gender (ref: women)	-0.01	0.01	0.503	0.25	0.09	0.01	-0.01	0.10	0.310	-0.22	0.09	0.02
Ex-smoker (ref: non-smoker)	0.01	0.01	0.57	-0.09	0.100	0.370	-0.11	0.10	0.257	-0.05	0.10	0.626
Smoker	0.08	0.02	< 0.01	0.14	0.12	0.240	-0.92	0.11	< 0.01	0.21	0.11	0.06
oh_{10-139} (ref: abstainer)	-0.35	0.01	0.03	-0.26	0.09	< 0.01	-0.08	0.09	0.413	-0.35	0.09	< 0.01
$oh_{139-279}$	-0.10	0.02	0.01	-0.05	0.13	0.700	-0.57	0.13	< 0.01	-0.14	0.13	0.265
oh_{280+}	-0.09	0.03	< 0.01	-0.36	0.17	0.042	-0.68	0.18	< 0.01	-0.15	0.17	0.358
Low-activity (ref: inactive)	-0.01	0.01	0.464	0.06	0.09	0.516	-0.09	0.09	0.334	0.131	0.09	0.151
HEPA-activity	-0.02	0.01	0.362	0.14	0.11	0.180	-0.06	0.10	0.543	0.06	0.11	0.530

Table 7.3: Parametric effects on the location and scale parameters of both reponses, for the model under the Joe copula assumption, the results assuming a gaussian copula is quite similar.

	Joe Copula			Gaussian Copula		
	Coefficient	Std Error	P value	Coefficient	Std Error	P value
Gender (ref: women)	-0.18	0.33	0.577	0.12	0.07	0.128
Ex-smoker (ref=non-smoker)	0.10	0.35	0.768	-0.03	0.07	0.65
Smoker	0.10	0.35	0.705	-0.05	0.09	0.610
oh_{10-139} (ref: abstainer)	0.17	0.30	0.572	0.06	0.07	0.341
$oh_{140-279}$	0.29	0.43	0.498	-0.08	0.09	0.395
oh_{280+}	0.56	0.46	0.224	-0.07	0.13	0.593
Low-activity (ref: inactive)	0.26	0.30	0.393	-0.003	0.06	0.959
HEPA-activity	0.37	0.33	0.251	0.05	0.08	0.482

Table 7.4: Parametric effects in the association parameter θ under both copula assumptions. Note the different range of the θ parameter for both copulas.

These plots have been obtained using the `predict()` function from the `GJRM` package to both fitted models, considering different values for the principal predictive covariates. Using the `copula` package (Hofert et al, 2017), changes in the bivariate response with glycemia, age and anemia are represented.

Concerning the parametric effects (Table 7.3), smoking increases the mean levels of the glycated hemoglobin while decreases the mean levels of glycated albumin, despite this inverse effect on both responses smoking does not modify the correlation between the concentration between glycemic markers, either under the gaussian or Joe copula assumption (Table 7.4). Alcohol consumption decreases the levels and variability of both glycated proteins. Regarding differences between males and females, the glycated hemoglobin show a higher variability in men, while the glycated hemoglobin present a higer variability in women (Table 7.3). For all the models fitted in this work, the physical activity has not shown any effect on the levels of the glycated proteins, once adjusted by the glucose levels.

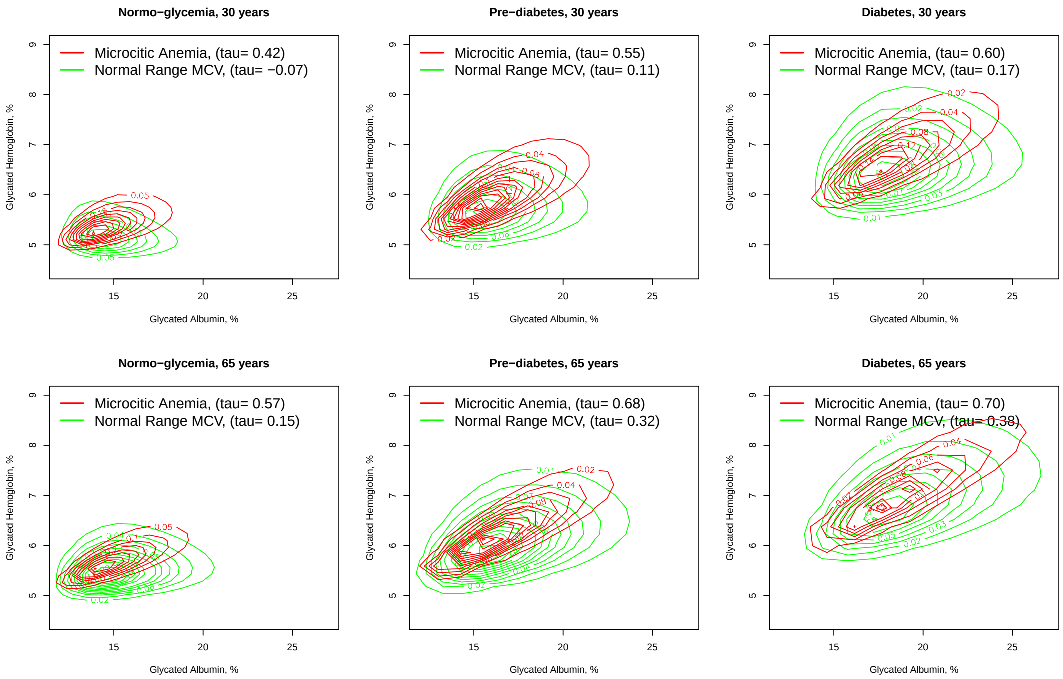


Figure 7.7: Controur plots representing the changes in the association between the levels of both proteins under the gaussian copula assumption for two different ages. The normoglycemia is defined by 90 mg/dL, the Pre-Diabetes = 120 mg/dL and the Diabetes = 150 mg/dL. The normal range of Mean Corpuscular Volume was fixed in 90 fL and the microcytic anemia in 75 fL.

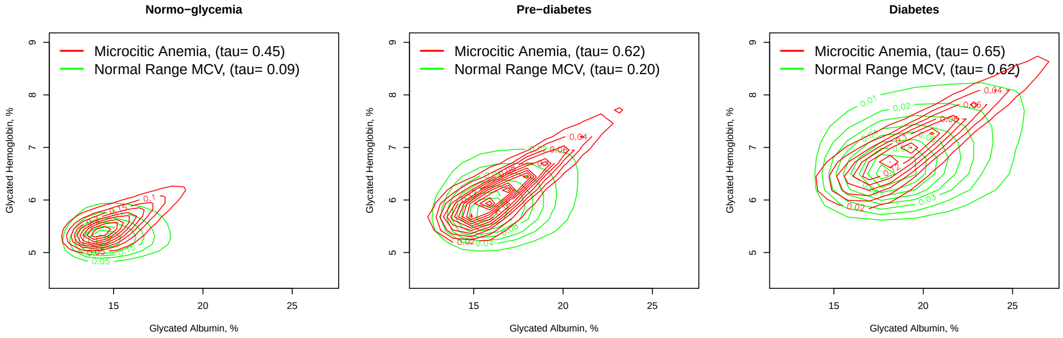


Figure 7.8: Controur plots representing the changes in the association between the levels of both proteins under the Joe copula assumption. The normoglycemia is defined by 90 mg/dL, the Pre-Diabetes = 120 mg/dL and the Diabetes = 150 mg/dL. The normal range of Mean Corpuscular Volume was fixed in 90 fL and the microcytic anemia in 75 fL. The Age was fixed to the mean Age of non-diabetic patients of our sample 50 years.

Chapter 8

Discussion

In this master thesis we have shown the usefulness of bivariate copula additive models (CGAMLSS) in diabetes research. The use of this approach revealed hitherto unreported effects concerning the mean, variability and the relationship between glycemic control markers.

Regarding the statistical framework, CGAMLSS models have proved to be fast and present unbiased estimates of the parametric and smooth effects in the association parameter of two response variables, with reliable point-wise confidence intervals. CGAMLSS regression offers us a flexible framework to use in biomedicine, that allows for a broad family of response variables joined by different copula functions, modelling each of the parameters that define the bivariate response, in a flexible way using additive predictors.

Concerning the clinical problem, the main variables explaining the correlation between our glycemic markers are glucose and mean corpuscular volume (MCV). The MCV presents a U-shaped effect on the correlation, because in conditions of anemia the blood circulating time is closer for both glycemic markers. The choice of copula leads to a different interpretation of the age effect on the association parameter. Under a gaussian copula assumption the correlation between the responses raises with age, a result that can be explained by the increment of the glycation rate with aging (Nakashima et al, 1933; Kilpatricks et al, 1996) while under the Joe copula assumption this effect is not statistically significant, changing the medical interpretation of the results.

Since concentrations of both proteins are used as glycemic markers for the Diabetes control these findings have clinical implications and the determination of both glycemic markers must be interpreted carefully for patients suffering anemia.

In this study we have focused on two continuous response variables. However, CGAMLSS models allow for binary bivariate responses, to study concordances between diagnostic tests. Two of these criteria, established by the American Diabetes Association (2014), are fasting glucose levels ($> 126mg/dL$) and glycated hemoglobin ($HbA1c > 6.5$). On this basis the author of this master thesis is currently working on the development of a bivariate binary model that allows us to investigate whether there is concordance between these two diagnostic criteria and determine if these threshold levels are the most appropriate. Moreover, given that the diabetes progression is preceding by a previous state of low hyperglycemia known as pre-diabetes, a bivariate model for multinomial responses will be also deserible in diabetes research, considering that a good identification of a pre-diabetes state let the physician stop the progression of the disease to a chronic hyperglycemia .

In addition to the glycated hemoglobin and fasting plasma glucose tests, there are also other types of diagnostic criteria needed by the ADA such as the measure of plasma glucose at 2 hours (200 mg/dL) in a oral glucose tolerance test. From a statistical point of view, work is required to develop regression models for trivariate responses, in the line of the work of Fillipou et al (2017), accounting for three binary responses.

Appendix A

Marginal distributions

Distribution	μ	σ	ν	$E(Y_m)$	$V(Y_m)$
Beta	$I(-)$	$I(-)$	-	μ	$\sigma^2\mu(1-\mu)$
Gamma	log	log	-	μ	$\sigma^2\mu^2$
Gumbel	1	log	-	$\mu - 0.577722\sigma$	$1.64493\sigma^2$
Normal	1	log	-	μ	σ^2
Inverse Gaussian	log	log	-	μ	$\mu^3\sigma^2$
Log-normal	log	log	-	$\sqrt{\exp(\sigma^2)}\exp^{\mu}$	$\exp(\sigma^2)(\exp(\sigma^2)-1)e^{2\mu}$
Logistic	1	log	-	μ	$\frac{\pi^2\sigma^2}{3}$
Reverse Gumbel	1	log	-	$\mu + 0.577722\sigma$	$1.64493\sigma^2$
Weibull	log	log	-	$\mu\Gamma(\frac{1}{\sigma}+1)$	$\mu^2[\Gamma(\frac{2}{\sigma}+1) - \{\Gamma(\frac{1}{\sigma}+1)\}^2]$
DAGUM	log	log	log	$-\frac{\mu}{\sigma}\frac{\Gamma(-\frac{1}{\sigma})+\Gamma(\frac{1}{\sigma}+\nu)}{\Gamma(\nu)}$ if $\sigma > 1$	$-(\frac{\mu}{\sigma})^2(2\sigma\frac{\Gamma(-\frac{2}{\sigma})\Gamma(\frac{2}{\sigma}+\nu)}{\Gamma(\nu)}$ $+\{\frac{\Gamma(-\frac{1}{\sigma}\Gamma(\frac{1}{\sigma}+\nu))}{\Gamma(\nu)}\}^2)$ if $\sigma > 2$
Sigh-Maddala	log	log	log	$\mu\frac{\Gamma(1+\frac{1}{\sigma})\Gamma(-\frac{1}{\sigma}+\nu)}{\Gamma(\nu)}$ if $\sigma\nu > 2$	$\mu^2\{\Gamma(1+\frac{2}{\sigma})\Gamma(\nu)\Gamma(-\frac{2}{\sigma}+\nu)$ $-\Gamma(1+\frac{1}{\sigma})^2\Gamma(-\frac{1}{\sigma}+\nu)^2\}$ if $\sigma\nu > 2$
FISK	log	log	-	$\frac{\mu\pi/\sigma}{\sin(\pi/\sigma)}$ if $\sigma > 1$	$\mu^2\{\frac{2\pi/\sigma}{sin(2\pi/\sigma)} - \frac{(\pi/\sigma)^2}{sin(\pi/\sigma)^2}\}$ if $\sigma > 2$

Table A.1: The distributions implemented in the GJRM package, with the link function for each parameter of the distribution and in the cases when the parameter are not location or scale parameters the transformation needed to get location and scale measurements of the distribution. In the table $I()$ represents the distribution function of a standardized logistic and Γ is the gamma function.

With exception of the Sigh-Maddala, DAGUM and FISK distribution, obtained from the book (Kleiber and Kotz, 2003), the remaining dsitribution was been consulted in the `gamlss` manual.

Beta distribution: appropiate when the response variable take values in a known restricted range, excludion the endpoints of the range. This distribution present two parameterizations, Marra and Radice (2017a) use:

$$f_Y = (y|\mu,\sigma) = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} \tag{A.1}$$

for $0 < y < 1$, where $\alpha = \mu(1-\sigma^2)/\sigma^2$, $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$ and B is the regularized beta function, $\sigma > 0$ and $\beta > 0$, hence $0 < \mu < 1$ and $0 < \sigma < 1$. In this parameterization, the mean of Y is $E(Y) = \mu$ and the variance is $Var(Y) = \sigma^2\mu(1-\mu)$.

Gamma: the gamma distribution is appropriate for positively skew data. Its distribution, is defined by:

$$f_Y(y|\mu,\sigma) = \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y^{1/\sigma^2-1}e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)} \tag{A.2}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. Here $E(Y) = \mu$ and $Var(Y) = \sigma^2\mu^2$.

Gumbel: appropiate for moderately negative skew data. Its distribution is defined by:

$$f_Y(y|\mu,\sigma) = \frac{1}{\sigma}\exp[(\frac{y-\mu}{\sigma}) - \exp(\frac{y-\mu}{\sigma})] \tag{A.3}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$ and $\sigma > 0$, with $E(Y) = \mu - \gamma \approx \mu - 0.577722\sigma$ and $Var(Y) = (\pi^2\sigma^2)/6 \approx 1.64493\sigma^2$.

Inverse Gaussian: this distribution is appropriate for highly positive data. Its distribution is defined as:

$$f_Y(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2y^3}}\exp[-\frac{1}{2\mu^2\sigma^2y}(y-\mu)^2] \tag{A.4}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$ with $E(Y) = \mu$ and $Var(Y) = \sigma^2\mu^3$.

Log-normal: appropriate for positively skew data. Its pdf is defined by:

$$f_Y(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{y}\exp\{-\frac{[log(y)-\mu]^2}{2\sigma^2 s}\} \tag{A.5}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. Here $E(Y) = \sqrt{\exp(\sigma^2)}e^{\mu}$ and $Vat(Y) = \exp(\sigma^2)(\exp(\sigma^2) - 1)e^{2\mu}$.

Logistic: is appropriate for moderately kurtotic data, this distribution is given by:

$$f_Y(y|\mu,\sigma) = \frac{1}{\sigma}\{\exp[-(\frac{y-\mu}{\sigma})]\}\{1+\exp[-(\frac{y-\mu}{\sigma})]\}^{-2} \tag{A.6}$$

Normal: the default option of the `GJRM` package, is defined by:

$$f_Y(y|\mu,\sigma) = \frac{1}{sqrt2\pi\sigma}\exp[-\frac{(y-\mu)^2}{2\sigma^2}] \tag{A.7}$$

where $-\infty < y < \infty$, where $-\infty < \mu < \infty$ and $\sigma > 0$. The mean os Y is give by $E(Y) = \mu$ and the variance of Y by $Var(Y) = \sigma^2$. So μ is the mean and σ is the standard deviation of Y.

Reverse Gumbel: appropriate for moderately positive skew data.

$$f_y(y|\mu,\sigma) = \frac{1}{\sigma}\exp\{-(\frac{y-\mu}{\sigma}) - \exp[-\frac{(y-\mu)}{\sigma}]\} \tag{A.8}$$

for $\infty < y < \infty$, where $-\infty < \mu < \infty$ and $\sigma > 0$, with $E(Y) = \mu + \gamma\sigma \approx \mu + 0.57722\sigma$ and $Var(Y) = (\pi^2\sigma^2)/6 \approx 1.64493\sigma^2$.

Weibull: the distributions is defined by:

$$f_Y(y|\mu,\sigma) = \frac{\sigma y^{\sigma-1}}{\mu^{\sigma}}\exp[-(\frac{y}{\mu})^{\sigma}] \tag{A.9}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$, the mean and variance of this distribution is given by $E(Y) = \mu\Gamma(\frac{1}{\sigma}+1)$ and $Var(Y) = \mu^2\{\Gamma(\frac{2}{\sigma}+1) - [\Gamma(\frac{1}{\sigma}+1)]^2\}$.

DAGUM: the distribution is defined by:

$$f_Y(y|\mu,\sigma,\nu) = \frac{\sigma\nu}{y}(\frac{(y/\mu)^{\sigma\nu}}{\{(y/\mu)^{\sigma}+1\}^{\nu+1}}) \tag{A.10}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$, the mean and variance of this distribution is given by $E(Y) = -\frac{\mu}{\sigma}\frac{\Gamma(-\frac{1}{\sigma})+\Gamma(\frac{1}{\sigma}+\nu)}{\Gamma(\nu)}$ if $\sigma > 1$ and $Var(Y) = -(\frac{\mu}{\sigma})^2(2\sigma\frac{\Gamma(-\frac{2}{\sigma})\Gamma(\frac{2}{\sigma}+\nu)}{\Gamma(\nu)} + \{\frac{\Gamma(-\frac{1}{\sigma}\Gamma(\frac{1}{\sigma}+\nu))}{\Gamma(\nu)}\}^2)$ if $\sigma > 2$.

Sigh-Maddala: defined by:

$$f_Y(y|\mu,\sigma,\nu) = \frac{\sigma\nu y^{\sigma-1}}{\mu^{\sigma}\{1+(\frac{y}{\mu})^2\}^{\nu+1}} \tag{A.11}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$, the mean and variance of this distribution is given by $E(Y) =$ and $Var(Y) = \mu^2\{\Gamma(1+\frac{2}{\sigma})\Gamma(\nu)\Gamma(-\frac{2}{\sigma}+\nu) - \Gamma(1+\frac{1}{\sigma})^2\Gamma(-\frac{1}{\sigma}+\nu)^2\}$ if $\sigma\nu > 2$.

FISK: defined by:

$$f_Y(y|\mu,\sigma,\nu) = \frac{\sigma y^{\sigma-1}}{\mu^{\sigma}\{1+(\frac{y}{\mu})^{\sigma}\}^2} \tag{A.12}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$, the mean and variance of this distribution is given by $E(Y) =$ and $Var(Y) = \mu^2\{\frac{2\pi/\sigma}{sin(2\pi/\sigma)} - \frac{(\pi/\sigma)^2}{sin(\pi/\sigma)^2}\}$ if $\sigma > 2$.

Bibliography

[1] Ali M. M., Mikhail N. N., and Haq M. S. (1978) A class of bivariate distributions including the bivariate logistic. Journal of Multivariate Analysis 8, 405-412.

[2] American Diabetes Association (2014) Standards of medical care in diabetes. Diabetes Care 37, S14-S80.

[3] Balakrishnan N. and Lai C. D. (2009) Continuous bivariate distributions. New York: Springer Science and Business Media.

[4] Brechmann E. C. and Schepsmeier U. (2013) Modeling dependence with C-and D-vine copulas: The R-package CDVine. Journal of Statistical Software 52, 1-27.

[5] Clayton D. G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65, 141-151.

[6] Conn A. R., Gould N. I., and Toint P. L. (2000) Trust region methods. Philadelphia: Society for Industrial and Applied Mathematics.

[7] Craig C.L., Marshall A.L., and Sjostrom M. (2003) International physical activity questionnaire: 12-country reliability and validity. Medical, Science Sports Exercise 35, 1381-1395.

[8] Danese E., Montagnana M., Nouvenne A., and Lippi G. (2015) Advantages and pitfalls of fructosamine and glycated albumin in the diagnosis and treatment of diabetes. Journal of Diabetes Science and Technology 9, 169-176.

[9] De Rekeneire N., Peila R., Ding J., Colbert L. H., Visser M., Shorr R. I., and Vellas B. (2006) Diabetes, hyperglycemia, and inflammation in older individuals. Diabetes Care 29, 1902-1908.

[10] Devroye L. (1986). Sample-based non-uniform random variate generation. In Proceedings of the 18th conference on Winter simulation, 206-265.

[11] Duarte E., de Sousa B., Cadarso-Surez C., Espasandn-Domnguez J., Lado-Baleato O., Marra G., Radice R., and Rodrigues V. (2017) Applying Spatial Copula Additive Regression to Breast Cancer Screening Data. In International Conference on Computational Science and Its Applications. Springer Cham, 586-599.

[12] Durante F. and Sempi C. (2015) Principles of copula theory. New York: Chapman and Hall/CRC.

[13] Diazyme Laboratories Protocol, http://www.diazyme.com/glycated-serum-protein-glycated-albumin, consulted August 2017.

[14] Dziedzic M., Petkowicz, D., Michalak M., and Solski J. (2012) Level of glycation gap in a healthy subject. Annals of Agricultural and Environmental Medicine 19, 842-845.

[15] Fahrmeir L., Kneib T., Lang S., and Marx B. (2013) Regression: models, methods and applications. Heidelberg, Berlin: Springer Science and Business Media.

[16] Enroth S., Johansson A., Enroth S. B., and Gyllensten U. (2014) Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. Nature communications, 5. Available from http://dx.doi.org/10.1038/ncomms5684.

[17] Filippou P., Marra G.,and Radice R. (2017) Penalized Likelihood Estimation of a Trivariate Additive Probit Model. Biostatistics 18, 569-568.

[18] Forbes J. M. and Cooper M. E. (2013) Mechanisms of diabetic complications. Physiological Reviews 93, 137-188.

[19] Gumbel E. J. (1960) Bivariate exponential distributions. Journal of the American Statistical Association 55, 698-707.

[20] Giugliano D., Ceriello A., and Esposito K. (2008) Glucose metabolism and hyperglycemia. The American Journal of Clinical Nutrition 87, 217S-222S.

[21] Hastie, T. and Robert T. (1990) Generalized additive models. London: John Wiley and Sons.

[22] Hofert M., Kojadinovic I., Maechler M., and Yan J. (2017) Copula: Multivariate Dependence with Copulas. R package version 0.999-16 URL https://CRAN.R-project.org/package=copula.

[23] Hong J. W., Ku C. R., Noh J. H., Ko K. S., Rhee B. D., and Kim D. J. (2015) Association between Self-Reported Smoking and Hemoglobin A1c in a Korean population without Diabetes: the 2011-2012 Korean National Health and Nutrition Examination Survey. PLoS One 10, e0126746.

[24] Huh J. H., Kim K. J., Lee B. W., Kim D. W., Kang E. S., Cha B. S., and Lee H. C. (2014) The Relationship between BMI and Glycated Albumin to Glycated Hemoglobin (GA/A1c) Ratio According to Glucose Tolerance Status. PloS One 9, e89478.

[25] Inaba M., Okuno S., Kumeda Y., Yamada S., Imanishi Y., Tabata T., and Nishizawa, Y. (2007) Glycated albumin is a better glycemic indicator than glycated hemoglobin values in hemodialysis patients with diabetes: effect of anemia and erythropoietin injection. Journal of the American Society of Nephrology 18, 896-903.

[26] IPAQ-Group International Physical Activity Questionnaire. Available at: https://sites.google.com/site/theipaq/home. Accessed June 2017.

[27] Joe H., and Hu T. (1996). Multivariate distributions from mixtures of max-infinitely divisible distributions. Journal of multivariate analysis 57, 240-265.

[28] Kilpatrick E.S., Dominiczak M.H. and, Small M. (1996) The effects of ageing on glycation and the interpretation of glycaemic control in type 2 diabetes. QJM: An International Journal of Medicine 89, 307-312.

[29] Kleiber C. and Kotz S. (2003) Statistical size distributions in economics and actuarial sciences. New Jersey: Wiley Series in Probability and Statistics.

[30] Klein N., Kneib T., Klasen S., and Lang S. (2015a) Bayesian structured additive distributional regression for multivariate responses. Journal of the Royal Statistical Society: Series C (Applied Statistics) 64, 569-591.

[31] Klein N., Kneib T., and Lang S. (2015b) Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. Journal of the American Statistical Association 110, 405-419.

[32] Klein N., Kneib T., Lang S., and Sohn A. (2015c) Bayesian structured additive distributional regression with an application to regional income inequality in Germany. The Annals of Applied Statistics 9, 1024-1052.

[33] Klein N., and Kneib T. (2016) Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. Statistics and Computing 26, 841-860.

[34] Kobayashi H., Abe M., Yoshida Y., Suzuki H., Maruyama N., and Okada K. (2016) Glycated albumin versus glycated hemoglobin as a glycemic indicator in diabetic patients on peritoneal dialysis. International Journal of Molecular Sciences 17, 619.

[35] Koga M., Saito H., Mukai M., Otsuki M., and Kasayama S. (2009) Serum glycated albumin levels are influenced by smoking status, independent of plasma glucose levels. Acta diabetologica 46, 141-144.

[36] Kolev N., and Paiva D. (2009) Copula-based regression models: A survey. Journal of Statistical Planning and Inference 139, 3847-3856.

[37] Laakso M. (2010) Cardiovascular disease in type 2 diabetes from population to man to mechanisms. Diabetes Care 33, 442-449.

[38] Lee J. E. (2015) Alternative biomarkers for assessing glycemic control in diabetes: fructosamine, glycated albumin, and 1, 5-anhydroglucitol. Annals of Pediatric Endocrinology and Metabolism 20, 74-78.

[39] Liamis G., Liberopoulos E., Barkas F., and Elisaf M. (2014) Diabetes mellitus and electrolyte disorders. World Journal of Clinical Cases: WJCC 2, 488-490.

[40] Marra G., and Radice R. (2017a) Bivariate Copula Additive Models for Location, Scale and Shape. Computational Statistics and Data Analysis 112, 99-113.

[41] Marra G., and Radice R. (2017b) GJRM: generalized joint regression modelling. Available from: https://cran.r-project.org/web/packages/GJRM/GJRM.pdf.

[42] McCullagh, J. and Nelder A. (1989) Generalized Linear Models. New York: Chapman and Hall/CRC.

[43] Morais M. P. P., Marshall D., Flower S. E., Caunt C. J., James T. D., Williams R. J., and Van Den Elsen J. M. (2013) Analysis of protein glycation using fluorescent phenylboronate gel electrophoresis. Scientific Reports, 3. Available from: http://dx.doi.org/10.1038/srep01437.

[44] Nakashima K., Nishizaki O., and Andoh Y. (1993) Acceleration of hemoglobin glycation with aging. Clinical Chemical Acta 215, 111-118.

[45] Nelsen R. B. (2007). An introduction to copulas. New York: Springer Science and Business Media.

[46] Nocedal J., and Wright S. (2006). Numerical optimization. New York: Springer-Verlag.

[47] Plackett R. L. (1965) A class of bivariate distributions. Journal of the American Statistical Association 60, 516-522.

[48] Radice R., Marra G. and Wojtyś, M. (2016) Copula regression spline models for binary outcomes. Statistics and Computing 26, 981-995.

[49] Radin M. S. (2014) Pitfalls in hemoglobin A1c measurement: when results may be misleading. Journal of General Internal Medicine 29, 388-394.

[50] Rahbar S. (1968) An abnormal hemoglobin in red cells of diabetics. Clinica Chimica Acta 22, 296-298.

[51] Rahbar S. (2005) The discovery of glycated hemoglobin: a major event in the study of nonenzymatic chemistry in biological systems. Annals of the New York Academy of Sciences 1043, 9-19.

[52] 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, 507-554.

[53] Roder P. V., Wu B., Liu Y., and Han W. (2016) Pancreatic regulation of glucose homeostasis. Experimental and molecular medicine 48, 219-222.

[54] Serrano-Rios M. and Gutierrez-Fuentes J. (2009) Type 2 Diabetes Mellitus. Madrid: Elsevier España.

[55] Shrayyef, M. Z. and Gerich, J. E. (2010) Normal glucose homeostasis. In Principles of diabetes mellitus. New York: Springer.

[56] Sklar, M. (1959) Fonctions de répartition a n dimensions et leurs marges. Université Paris 8, 229-231.

[57] Stasinopoulos M. D., Rigby R. A., Heller G. Z., Voudouris V., and De Bastiani F. (2017) Flexible Regression and Smoothing: Using GAMLSS in R. New York: Chapman and Hall/CRC.

[58] Szablewski L. (2011) Glucose homeostasis-mechanism and defects. INTECH Open Access Publisher.

[59] Trivedi P. K. and Zimmer D. M. (2007) Copula modeling: an introduction for practitioners. Foundations and Trends in Econometrics 1, 1-111.

[60] Ulrich P. and Cerami, A. (2001) Protein glycation, diabetes, and aging. Recent progress in hormone research 56, 1-22.

[61] Vatter T. and Chavez-Demoulin V. (2015) Generalized additive models for conditional dependence structures. Journal of Multivariate Analysis 141, 147-167

[62] Vatter T. and Nagler T. (2016) Generalized additive models for pair-copula constructions, ArXiv e-prints. URL arXiv:1608.01593.

[63] World Health Organization. (2016) Global report on diabetes. Available from: http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257eng.pdf

[64] Wood, S. N. (2003) Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65, 95-114.

[65] Wood S. N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673-686.

[66] Wood S.N. (2006) Generalized additive models: an introduction with R. London: Chapman and Hall.

[67] Wood S. N., Pya N., and Safken B. (2016) Smoothing parameter and model selection for general smooth models. Journal of the American Statistical Association 111, 1548-1563.

[68] Wu W. C., Ma W. Y., Wei J. N., Yu T. Y., Lin M. S., Shih S. R., and Li H. Y. (2016) Serum glycated albumin to guide the diagnosis of diabetes mellitus. PloS One 11, e0146780.

[69] Yee T. W. (2015) Vector generalized linear and additive models, with an implementation in R. New York: Springer.