Received 27 July 2018; | Revised 29 October 2018; | Accepted: 00 Month 0000

DOI: xxx/xxxx

ARTICLE TYPE

Studying the relationship between markers of glycemic control (based on)/ (using) flexible copula regression models // Assesing the relationship between markers of glycemic control through flexible copula regresion models

J. Espasandín-Domínguez*[1] | C. Cadarso-Suárez[1] | T. Kneib[2] | G. Marra[3] | N. Klein[4] | R. Radice[5] | O. Lado-Baleato[1] | A. González Quintela[6] | F. Gude[7]

[1]Unit of Biostatistics, Department of Statistics, Mathematical Analysis, and Optimization, Universidade de Santiago de Compostela, Spain

[2]Chair of Statistics, Georg-August-Universität Göttingen, Germany

[3]Department of Statistical Science, University College London, United Kingdom

[4]Humboldt-Universitat zu Berlin, Unter den Linden 6, Berlin, Germany

[5]Cass Business School, City, University of London, 106 Bunhill Row, EC1Y 8TZ London, United Kingdom

[6]Department of Internal Medicine, Complejo Hospitalario Universitario de Santiago de Compostela, Spain

[7]Clinical Epidemiology Unit, Complejo Hospitalario Universitario de Santiago de Compostela, Spain

Correspondence

*Corresponding author: Jenifer Espasandín-Domínguez, Email: jenifer.espasandin@usc.es

Present Address

Unit of Biostatistics, Department of Statistics, Mathematical Analysis, and Optimization, University of Santiago de Compostela, School of Medicine, C/San Francisco s/n, 15782-Santiago de Compostela, Spain

Summary

Glycated haemoglobin (HbA1c) is a sensitive marker of blood glucose in patients with diabetes. However, levels can vary considerably, even among individuals with similar mean blood glucose concentrations. Other glycated proteins, such as fructosamine, can also act as blood sugar markers, but estimating HbA1c and fructosamine via independent models may lead to errors of interpretation regarding disease severity. From a clinical standpoint, it would be of great interest to know the factors that affect the mean concentration of both HbA1c and fructosamine, that influence the variability in the concentrations of these glycated markers, and that cause HbA1c/fructosamine discordance. Flexible models are required that illustrate the behaviour of these variables as well as the association between them. The present work reviews existing models that might serve in this regard. Flexible copula regression models using P-splines, were used to provide a better understanding of the behaviour of both glycated proteins, and the relationship between them under the possible influence of different covariates. This work shows the usefulness of this type of models in practice, and provides a basis for its clinical interpretation by means of an understandable case study. Ultimately, to better understand the effects of each continuous covariate, they were represented at the true scale of the response variables.

KEYWORDS:

Bayesian inference, frequentist inference, penalised spline, bivariate copula, diabetes.

1 | INTRODUCTION

Diabetes is one of the most common human disorders. Early diagnosis and strict glucose control are crucial if serious complications are to be prevented or delayed. Prognoses for diabetes are based largely on determining the plasma glucose and glycated haemoglobin (HbA1c) concentrations. These tests are used to detect individuals with pre-diabetes, as well as to screen for and diagnose the disease. The results, however, are not foolproof, and the clinical usefulness of these tests is affected by a number of biological and analytical factors. In clinical practice, the introduction of other measures of glucose homeostasis, such as plasma fructosamine and glycated albumin, is attractive, especially when dealing with patients in whom the measurement of HbA1c may be biased (e.g., patients with kidney disease, anaemia, or disorders involving abnormal haemoglobin metabolism).[1]

Unfortunately, discordances are often seen among the results for HbA1c and other glycated proteins, and clinicians need to be aware of the conditions that might explain them.[2] Several authors have proposed metrics for quantifying the discrepancies between HbA1c and blood glucose in the form of glycation "gaps" or "indices", i.e the difference between the measured HbA1c and that which would be predicted from another measure of glycaemic control using a linear regression model.[3] However, both the glycation gap and index values correlate strongly with the concentration of HbA1c, and require that the distribution of this concentration be assumed Gaussian.

The biomedical aims of the present case study are: 1) to identify variables that might affect the mean concentrations of HbA1c and fructosamine, and which influence the variation in their concentrations, and 2) to identify the factors that may cause discordance between results for the concentrations of these glycated proteins. It is hoped that this will help improve the diagnosis and treatment of diabetes. Such aims require the use of statistical methods able to flexibly and simultaneously examine the mean concentrations of the above proteins, their variability, and the relationship between them. This paper thus reviews the available flexible regression models that meet these requirements. More specifically, in this manuscript, bivariate copula generalized additive models for location, scale and shape (CGAMLSS) were considered, based on either frequentist[4] or Bayesian[5] inference principles. These types of model extend univariate generalized additive models for location, scale and shape (GAMLSS),[6] as well as univariate distributional regression,[7] to the field of multivariate responses. More specifically, CGAMLSS estimates the joint multivariate distribution of a response vector where each parameter characterising the joint distribution is modelled simultaneously and is conditioned by covariates. The multivariate distribution is constructed from different copulas that allow for different dependence structures.[8] CGAMLSS enables the modelling of all distributional parameters using additive predictors that encompass several types of covariate effect, such as the non-linear effects of continuous covariates, random effects, and interactions.

Dependence modelling using copula functions is a useful multivariate modelling tool in situations in which multivariate dependence is of interest and multivariate normality is questionable.[9] In most published regression studies on multivariate responses, a specific distribution is assumed for the response variable. This is done mostly for the sake of convenience rather than any strong theoretical reason. In addition, most of the existing multivariate distributions are simple extensions of corresponding univariate distributions, and often suffer the restrictive property of all the marginal distributions being of the same type (by construction, all the marginal distributions of a multivariate normal are again normal). A major advantage of the copula approach is that the marginal distributions may belong to different, non-standard families.[4] Moreover, copulas can address non-symmetrical structures of dependencies rather than just those that are elliptical.

Other conditional copula regression techniques using copula functions have been proposed as alternatives to CGAMLSS, but they only provide some of the latter's flexibility - either because they only allow for the consideration of normal marginals,[10] or because they fail to consider additive predictors[11]. In the frequentist setting, attention must also be drawn to vector generalized additive models (VGAM) as an alternative to CGAMLSS.[12] The estimation of VGAM models is carried out by fitting a vector additive model at each iteraction of the IRLS algorithm (see for example[13] for more details). VGAMs permits each parameter of a bivariate non-standard response to be estimated in a flexible manner using an additive predictor. Thomas Yee[13] proposes the use of copulas as a special class of bivariate distributions. However, smoothing parameter selection is more difficult and thus leading to a potentially greater bias than in CGAMLSS estimations in terms of the accuracy of the estimated copula parameter.[5] In addition to VGAM, Vatter and Chavez proposed a two-stage approach in which the parameters of the marginal distributions and the copula are determined separately.[14] In contrast, in CGAMLSS regression models, all parameters are estimated simultaneously. Via simulation studies, Marra and Radice showed that CGAMLSS is slightly more efficient than a two step estimator.[4] A major advantage of the CGAMLSS algorithms proposed by Marra and Radice,[4] and Klein and Kneib,[5] is that they were created in a modular fashion, and therefore new parametric continuous marginal distributions and copula functions can be easily

included. Another is that copula regression parameters are integrated into the estimation of the model coefficients to allow for more flexible dependence modelling. Further, the smoothing parameters are selected automatically.

Despite the flexibility of the CGAMLSS framework, the lack of interpretability of the distributional parameters of the response variables reduces the use of this kind of modelling in the clinical setting. Indeed, clinicians prefer to use a generalized additive model requiring the assumption of a Gaussian response because the results are easier to understand. Further, with most additive models, the effects of continuous covariates on the results are centred.[15,16] As mentioned above, several papers has been published describing the statistical methodology of CGAMLSS in detail, including simulation studies and providing some example analysis (as per[4,5,17]), however there are very few manuscripts on real biomedical data. For this reason, in this case study we will give guidance for clinical researchers on how to apply these type of regression models. The present work also highlights a way to visualize and interpret the results obtained with the novel regression models used in practice.

The rest of the manuscript is organized as follows: In Section 2 we present the description of the database used in the study. Section 3 provides an overview of CGAMLSS regression models, including those involving Bayesian and frequentist inference. In Section 4, the model-building process is presented along with an analysis of the usefulness of the examined models in the clinical setting. Finally, Section 5 provides a discussion with some comments on directions of future research.

2 | GLYCATION DATA

The data used in this work comes from the The A-Estrada Glycation and Inflammation Study. This Section introduces this study and the dataset.

2.1 | The A-Estrada Glycation and Inflammation Study (AEGIS)

AEGIS is a cross-sectional, population-based study being performed in the municipality of A Estrada (NW, Spain); the data collection and recruiting phase were completed in March 2015. Its aim is to investigate the association between glycation, inflammation, lifestyles and their association with common diseases, and to study discordances between markers for glycaemia.[18] An outline of the AEGIS study is available at www.clinicaltrials.gov, code NCT01796184. The study was reviewed and approved by the Clinical Research Ethics Committee from Galicia, Spain (CEIC2012-025). Written informed consent was obtained from each participant in the study, which conformed to the current Helsinki Declaration.

A total of 1516 subjects agreed to participate in the study; their mean age was 52 years (range 18 to 91), 55% were females, and 187 (12%) had been previously diagnosed with diabetes. Among those with diabetes, 66.8% took oral anti-diabetics, 3.7% took insulin alone, and 13.3% took insulin plus oral drugs. The remaining 16.2% took none of these medications. Participants with elevated HbA1c and fructosamine levels were more likely to be older, to have fewer years of education, and were more likely to be current smokers than to have formerly used or never used tobacco. They were also less likely to undertake health enhancing physical activity and to be alcohol drinkers. HbA1c and fructosamine concentrations were highly correlated (Pearson correlation coefficient, $r = 0.72$), and the concentrations of both proteins correlated with fasting plasma glucose levels ($r = 0.72$ and $r = 0.56$ respectively).

In determining the factors that influence the HbA1c and fructosamine concentrations, their variability, and the relationship between them, the following variables were deemed covariates: fasting plasma glucose (glucose in formulae, mg/dL), age (in *years*), gender, body mass index (BMI, Kg/m^2), plasma albumin (Alb in formulae, f/L), and the mean corpuscular (red blood cell) volume (MCV, fL). The HbA1c and fructosamine concentrations were considered to be the bivariate response. All laboratory analyses were performed on the day of sample collection in the Clinical Biochemistry Laboratory of the Hospital Clínico Universitario de Santiago de Compostela, Spain. A more detailed description of participant's clinical characteristics and laboratory measurements is given in Table 1 .

The main purpose of this biomedical study is to simultaneously study the glycation of HbA1c and fructosamine and the factors that could influence such glycation. Tackling this problem of multivariate response modelling thus requires new multivariate regression techniques. As has been justified in the introduction, we will consider the copula distributional regression models, introduced by Klein and Kneib[5] in Bayesian, and Marra and Radice[4] in the frequestist framework. These techniques will be compared for the first time via a real biomedical study in this work. In the following section, we present these type of regression models. More details and references can be found in[4,5,17].

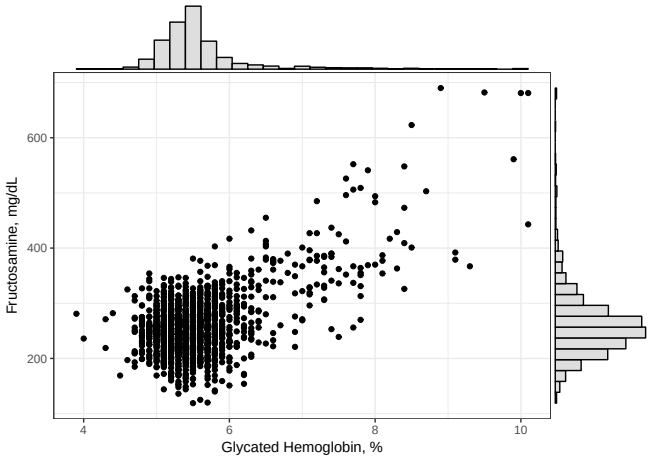

FIGURE 1 Scatterplot of fructosamine and HbA1c.

3 | MODERN BIVARIATE COPULA REGRESSION MODELS

CGAMLSS[4,5] model the joint distribution of a pair of response variables (y_1, y_2) given covariates based on a copula specification for the dependence structure between the two responses. Given the nature of the problem presented in the introduction, this case study focuses on the use of CGAMLSS with the pair of continuous random variables, $y_1 = HbA1c$ and $y_2 = Fructosamine$ as the response variables.

In the CGAMLSS approach, the joint cumulative distribution function (cdf) of y_1 and y_2, given the covariate information - colleted in $\boldsymbol{\nu}$-, is expressed in terms of the marginal cdfs and a copula function C that binds them together. More specifically, Sklar's theorem guarantees that we can write

$$F(y_1, y_2 \mid \boldsymbol{\nu}) = C(F_1(y_1 \mid \boldsymbol{\nu}), F_2(y_2 \mid \boldsymbol{\nu})), \tag{1}$$

where $F_1(y_1 \mid \boldsymbol{\nu})$ and $F_2(y_2 \mid \boldsymbol{\nu})$ are the marginal cdfs of $y_1 \mid \boldsymbol{\nu}$ and $y_2 \mid \boldsymbol{\nu}$ which take values of $(0, 1)$, $C(\cdot, \cdot \mid \boldsymbol{\nu})$ is a uniquely defined two-place copula function that does contain information about the association between the two outcomes solely.

The different parametric copula functions proposed in the literature allow different types of dependence structure between the response variables (see Figure 2 for a graphical illustration). For example, the Clayton copula allows one to consider asymmetric structures of dependence when two random variables show a stronger positive association at smaller values than at larger values. The Joe or Gumbel copula, in contrast, addresses the opposite situation, in which two random variables with positive dependence show a stronger association at higher values. Rotated versions of the Clayton, Gumbel and Joe copulas also exist for modelling negative structures of dependence.

3.1 | Model Formulation

CGAMLSS regression models combine flexibility in the specification of the marginal distributions of a bivariate response vector with additional flexibility in the dependence structure induced by a copula. Further, by modelling each parameter of the response at the same time - and not just the marginal means - they allow for the quantification of regression effects on basically all aspects of the bivariate response distribution, including the location, scale, or the shape parameters, among others, of the marginal distributions, as well as on the copula parameter. In this biomedical study, we will investigate the effects of covariates on the location and the scale parameters of the bivariate response's distribution (HbA1c and Fructosamine) and the association between them. For this reason, to simplify the notation, let us suppose that each marginal distribution has two parameters (corresponding to the mean and the scale parameter) and one copula parameter, i.e., $(\mu_{i1}, \mu_{i2}, \sigma_{i1}, \sigma_{i2}, \rho_i)$, where μ_{i1} and μ_{i2} are the location parameters of HbA1c and fructosamine, respectively. σ_{i1} and σ_{i2} are scale parameters of these margins, and ρ_i denotes the association parameter between HbA1c and fructosamine.

TABLE 1 Participant's clinical characteristics according to three different glycemic status: Normo-glycaemic (FPG < 100 mg/dL or HbA1c < 5.7%); Prediabetes (100 mg/dL ≤ FPG ≤ 125 mg/dL or 5.7% ≤ HbA1c < 6.5%); Diabetes (HbA1c ≥ 6.5% or FPG > 125 mg/dL).

Variable	Normo-glycaemic (n=1134)	Prediabetes (n=267)	Diabetes (n=115)	Overall Sample (n=1516)
Age, years	49 ± 17	63 ± 13	64 ± 13	52 ± 17
Gender				
Female	658 (58%)	131 (49%)	49(43%)	838(55%)
Male	476 (42%)	136 (51%)	66 (57%)	678 (45%)
BMI, kg/m^2				
Normal weight	386 (34%)	26 (10%)	26 (10%)	423(28%)
Overweight	447 (39%)	90 (34%)	37 (32%)	574(38%)
Obese	301 (27%)	151 (56%)	67 (58%)	519 (34%)
Physical activity				
Inactive	414 (37%)	125 (4 7%)	57 (50%)	596 (39%)
Minimally active	431 (38%)	81 (30%)	40 (35%)	552 (36%)
"HEPA active"	289 (25%)	61 (23%)	18 (15%)	368 (24%)
Alcohol consumption				
Abstainers	426 (38%)	80 (30%)	40 (35%)	546 (36%)
Light drinkers	479 (42%)	86 (32%)	33 (29%)	598 (39%)
Moderate drinkers	149 (13%)	67 (25%)	25 (22%)	241 (16%)
Heavy drinkers	80 (7%)	34 (13%)	17 (14%)	131 (9%)
Smoking				
Non-smokers	598 (62%)	166 (53%)	166 (53%)	825 (54%)
Ex-smokers	276 (30%)	79 (35%)	79 (35%)	395 (26%)
Smokers	260 (8%)	61 (23%)	22 (12%)	296 (20%)
FPG, mg/dL	85 ± 8	108 ± 7	157 ± 31	94 ± 23
HbA1c, %	5.4 ± 0.4	5.9 ± 0.6	7.3 ± 1.0	5.6 ± 0.7
Fructosamine, $\mu mol/L$	248 ± 44	272 ± 64	375 ± 95	262 ± 63
MCV, f/L	89.6 ± 4.7	90.4 ± 4.8	90.2 ± 5.5	89.9 ± 4.8
Albumin, f/L	4.4 ± 0.2	4.4 ± 0.2	4.4 ± 0.2	4.4 ± 0.2

In this table, continuous variables are summarize in terms of means ± standard deviation. Categorical variables are presented as absolute frequency (%). Here, FPG denotes Fasting Plasma Glucose, MCV (Mean Corpuscular Volume) and BMI (Body Mass Index). Physical activity was evaluated using *The International Physical Activity Questionnaire*[19]. The questionnaire records the time spent on different type of activities weighted according to some resting metabolic rates. Subjects were classified into three levels: inactive, minimal active and "HEPA active" (health-enhancing physical activity, the highest active category). Overweight ranging from a BMI of 25 kg/m^2 to 30 kg/m^2, Obese: BMI > 30 kg/m^2; Normal weight: BMI < 25 kg/m^2. Alcohol consumption was measured using the standard drinking unit system (see Gual et al.[20]). Individuals were classified into four categories according to their alcohol consumption: *abstainers* (individuals with a regular alcohol consumption of 0 g per week); *light drinkers* (alcohol consumption between 1 g to 139 g per week); *moderate drinkers* (alcohol consumption between 140 g to 279 g per week) and *heavy drinkers* (alcohol consumption ≥ 280 g per week). Tobacco consumption was assessed trough the number of cigarettes usually consumed per day, patients who smoke at least one cigarette by day or quit smoking during the previous year has been considered *smokers*.

Let us assume that observations on the response vector $\{\boldsymbol{y}_i = (y_{i1}, y_{i2}), i \in \{1, \dots, n\}\}$, where y_{i1}, y_{i2} are the marginal response variables corresponding to HbA1c and Fructosamine, respectively, and the generic covariate vector $\{\boldsymbol{v}_i, i = 1, \dots, n\}$ are available for n observational units. Let be p_1 and p_2 the marginal densities of y_1 and y_2, respectively,

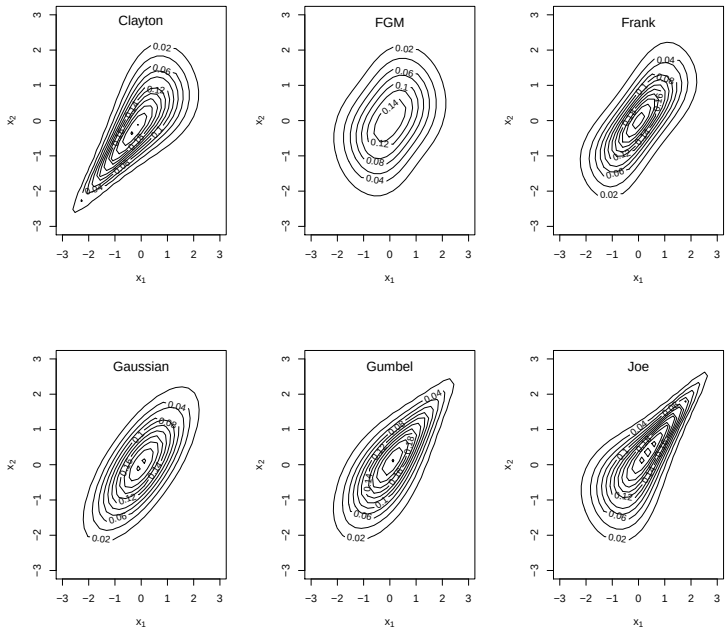

FIGURE 2 Contour plots of some of the classical copula functions with standard normal margins for data simulated using a Kendall's τ coefficient of 0.5.

TABLE 2 Some classic copulae functions, with corresponding parameter range of association parameter ρ and link function of ρ. $\Phi_2(\cdot,\cdot;\rho)$ denotes the cdf of a standard bivariate normal distribution with correlation coefficient ρ, and $\Phi(\cdot)$ the cdf of a univariate standard normal distribution. Finally, ϵ is set to 10^{-7} and is used to ensure that the restrictions on the space of ρ are maintained.

Copula	$C(u,v;\rho)$	Range of ρ	Link Function
Clayton	$(u^{-\rho}+v^{-\rho}-1)^{-1/\rho}$	$\rho\in(0,\infty)$	$\log(\rho-\epsilon)$
FGM	$uv\{1+\rho(1-u)(1-v)\}$	$\rho\in[-1,1]$	$\tanh^{-1}(\rho)$
Frank	$-\rho^{-1}\log\{1+(e^{-\rho u}-1)\ (e^{-\rho v}-1)/(e^{-\rho}-1)\}$	$\rho\in R\backslash\{0\}$	–
Gaussian	$\Phi_2\left(\Phi^{-1}(u),\Phi^{-1}(v);\rho\right)$	$\rho\in[-1,1]$	$\tanh^{-1}(\rho)$
Gumbel	$\exp\left[-\{(-\log u)^\rho+(-\log v)^\rho\}^{1/\rho}\right]$	$\rho\in(1,\infty)$	$\log(\rho-1)$
Joe	$1-\{(1-u)^\rho+(1-v)^\rho-(1-u)^\rho(1-v)^\rho\}^{1/\rho}$	$\rho\in(1,\infty)$	$\log(\rho-1-\epsilon)$

$$p_{1,i}\equiv p_1\left(Y_{1i}\mid\left(\mu_{1i},\sigma_{1i}\right)\right),$$
$$p_{2,i}\equiv p_2\left(Y_{2i}\mid\left(\mu_{2i},\sigma_{2i}\right)\right).$$

Note that $p_{1,i}$ and $p_{2,i}$ depend on 2 parameters each one. But in other studies involving other response variables it could be possible to contemplate more complex distributions.

In the CGAMLSS approach, all the parameters can be related to an additive predictor, η_i. As in a classical generalized linear regression model a suitable bijective link function, g, can be considered which ensures that the restrictions on the parameter spaces are maintained, as follows:

$$\eta_i^{\mu_1} = g_{\mu_1}\left(\mu_{1i}\right); \eta_i^{\sigma_1} = g_{\sigma_1}\left(\sigma_{1i}\right),$$
$$\eta_i^{\mu_2} = g_{\mu_2}\left(\mu_{2i}\right); \eta_2^{\sigma_2} = g_{\sigma_2}\left(\sigma_{2i}\right).$$

The choice of the link function is determined by the restrictions that apply to the parameter space of the corresponding parameter. For example, in this case, to model the standard deviation as a function of the covariates and regression coefficients, the link function $g_\sigma(\cdot)$ is equal to $\log(\cdot)$ to ensure positive values (see Marra and Radice,[4] or Klein et al.[7] for details on the link functions); as usually a identity link has been assumed to model the mean as a function of the covariate effects.

To simplify the notation, in this biomedical study, only copulas with one parameter are considered (see Table 2). Moreover, this copula parameter, say ρ_i, is also related to an additive predictor, η_i^ρ by assuming $\eta_i^\rho = g_\rho(\rho_i)$. The choice of the copula's link, $g_\rho(\cdot)$, depends on the copulae type. See Table 2 .

In a nutshell, the total number of response parameters that can be modelled in the CGAMLSS approach is the sum of the number of parameters of each marginal and the number of parameters of the copula function (i.e. in this study we model 5 parameters). Let us assumme $\boldsymbol{\vartheta}$ the 5-dimensional vector formed by all these parameters, $\boldsymbol{\vartheta} = \left(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\right)$. All these parameters comprised in $\boldsymbol{\vartheta}$, are also assumed to be related to regression coefficients and covariates (e.g., in this biomedical study we have considered binary coavariates - such as the gender of the participants - and continuous regressors - such as glucose or age -) collected in $\boldsymbol{v}_i$ via an additive predictor defined as:

$$\begin{aligned}
\eta_i^{\mu_k} &= \beta_0^{\mu_k} + \sum_{j=1}^{J_1} f_j^{\mu_k}(\boldsymbol{v}_i),\ k \in \{1,2\} \\
\eta_i^{\sigma_k} &= \beta_0^{\sigma_k} + \sum_{j=1}^{J_2} f_j^{\sigma_k}(\boldsymbol{v}_i),\ k \in \{1,2\} \\
\eta_i^{\rho} &= \beta_0^{\rho} + \sum_{j=1}^{J_3} f_j^{\rho}(\boldsymbol{v}_i),
\end{aligned} \tag{2}$$

where $\beta_0^{(\cdot)}$ is a general intercept of each predictor, and the functions $f_j^{(\cdot)}$ represent the different covariate effects.

Note that each distribution parameter of marginal distributions and copula function may depend on different covariates and a different number of effects (say J_1, J_2 or J_3). By dropping the parameter-dependence for notational simplicity, the following generic representation can be used to refer to equation (2):

$$\eta_i = \beta_0 + \sum_{j=1}^{J} f_j(\boldsymbol{v}_i). \tag{3}$$

As in generalized additive regression models each function f_j of equation (3) can be modelled by a linear combination of D_j appropriate basis functions:

$$f_j(\boldsymbol{v}_i) = \sum_{d_j=1}^{D_j} \beta_{j,d_j} \boldsymbol{B}_{j,d_j}(\boldsymbol{v}_i). \tag{4}$$

Equation (4) implies that the vector of evaluations $\left(f_j(\boldsymbol{v}_1), \ldots, f_j(\boldsymbol{v}_n)\right)^{\mathsf{T}}$ can be written as $\mathbf{Z}_j\boldsymbol{\beta}_j$ with $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,D_j})^{\mathsf{T}}$, where $\boldsymbol{\beta}_j$ consists of all the basis coefficients, and the entries $Z_j[i, d_j] = \boldsymbol{B}_{j,d_j}(\boldsymbol{v}_i)$ of the design matrix $\boldsymbol{Z}$ are the basis functions evaluated at the observed covariate values. Choices of the basis functions depend on the different effect types and we give specific examples at Section 4.1.2. Finally, Equation (3) can be written as:

$$\boldsymbol{\eta} = \beta_0\mathbf{1}_n + \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_J\boldsymbol{\beta}_J, \tag{5}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ represents the predictor vector for all observations, and $\mathbf{1}_n$ is an n-dimensional vector of one.

To regularize the estimation of the potentially high-dimensional vectors of basis coefficients, each vector $\boldsymbol{\beta}_j$ is associated with a quadratic penalty (in the penalized likelihood framework) or a multivariate Gaussian prior (in the Bayesian framework). More precisely, the quadratic penalties take the form $\lambda\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta}$ (dropping the parameter index and the function index for simplicity) where the positive semidefinite penalty matrix $\mathbf{K}$ is chosen to enforce the desirable properties of the corresponding functional

effect (e.g., smoothness or shrinkage). The smoothing parameter $\lambda \in [0, \infty)$ controls the trade-off between fit and smoothness, and plays a crucial role in the estimation of the shape of the corresponding effect.

In the Bayesian inference model, the penalty term is replaced by a partially improper Gaussian prior:

$$p(\boldsymbol{\beta}|\tau^2) \propto \exp(-\frac{1}{2\tau^2}\boldsymbol{\beta}^T \boldsymbol{K}\boldsymbol{\beta}), \tag{6}$$

where the matrix $\boldsymbol{K}$ is now the prior precision matrix, and τ^2 replaces the smoothing parameter from the penalized likelihood framework. The Bayesian posterior mode and penalized likelihood estimates correspond to each other via $\lambda = \frac{1}{2\tau^2}$.

Many types of effect can be modelled making different assumptions regarding the basis functions and the penalty/prior precision matrix.[16,21] Section 4.1.2. discusses only those effects contemplated in the present case study.

3.2 | Inference

The following lines discuss both frequentist and Bayesian inferences for the special case of the continuous-continuous copula regression models presented in Section 3.1. Thus, the log-likelihood of a CGAMLSS regression model with continuous margins can be written as:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} log\left\{c(F_{1i}(y_{1i}|\mu_{1i}, \sigma_{1i}), F_2(y_{2i}|\mu_{2i}, \sigma_{2i}); \rho_i)\right\} + \sum_{i=1}^{n}\sum_{d=1}^{2} log\left\{p_d(y_{di} \mid \mu_{di}, \sigma_{di}))\right\}, \tag{7}$$

for $d = 1, 2$, where $c(\cdot, \cdot, \rho)$ is the density function of the copula function, and $p_d(y_d|\mu_d, \sigma_d)$ the density of the d^{th} marginal. Parameter vector $\boldsymbol{\theta}$ is defined as $(\boldsymbol{\beta}_{\mu_1}^T, \boldsymbol{\beta}_{\mu_2}^T, \boldsymbol{\beta}_{\sigma_1}^T, \boldsymbol{\beta}_{\sigma_2}^T, \boldsymbol{\beta}_{\rho}^T)^T$ which refers to the coefficient vectors associated with $\eta_i^{\mu_1}$, $\eta_i^{\mu_2}$, $\eta_i^{\sigma_1}$, $\eta_i^{\sigma_2}$ and η_i^{ρ} respectively.

3.2.1 | Bayesian Inference

In the CGAMLSS framework, Bayesian inference is carried out using a generic algorithm based on Markov chain Monte Carlo simulations, via the iterative updating of all model parameters of the joint posterior. According to Klein and Kneib, the log-posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ is given by[5]:

$$\log\left((p(\boldsymbol{\theta}|\boldsymbol{y})\right) \propto \ell(\boldsymbol{\theta}) + \sum_{k=1}^{K}\sum_{j=1}^{J_k} \log(p(\boldsymbol{\beta}_{j,k}|\tau_{j,k}^2)p(\tau_{j,k}^2)),$$

where $\boldsymbol{\theta}$ is the complete parameter vector and $\boldsymbol{y}$ denotes the response matrix. Klein and Kneib resorted to a Metropolis-Hastings-algorithm.[5] For the variance parameters $\tau_{j,k}^2$, inverse gamma hyperpriors are assumed ($\tau_j^2 \sim IG(a_j, b_j)$, with $a_j = b_j = 0.001$) in order to obtain data-driven smoothness.

3.2.2 | Penalized Maximum Likelihood Inference

In the frequentist setting, regression parameter estimations are based on direct optimization of the penalized likelihood with automatic selection of the smoothing parameter. In this type of model, the use of an unpenalized optimization algorithm could produce unduly wiggly estimates. Marra and Radice proposed maximizing the expression[4]:

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{S}\boldsymbol{\theta}, \tag{8}$$

where $\ell(\boldsymbol{\theta})$ is the log-likelihood of a CGAMLSS regression model with continuous margins (defined in equation 7); ℓ_p denotes the penalized log-likelihood of the model, and $\mathbf{S} = \text{diag}(\boldsymbol{K}_{\mu_1}, \boldsymbol{K}_{\mu_2}, \boldsymbol{K}_{\sigma_1}, \boldsymbol{K}_{\sigma_2}, \boldsymbol{K}_{\rho})$. The smoothing parameters contained in the $\boldsymbol{K}$ components make up the overall vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mu_1}^T, \boldsymbol{\lambda}_{\mu_2}^T, \boldsymbol{\lambda}_{\sigma_1}^T, \boldsymbol{\lambda}_{\sigma_2}^T, \boldsymbol{\lambda}_{\rho}^T)^T$.[4]

Marra and Radice proposed estimating $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ using a stable and efficient trust region algorithm with integrated automatic multiple smoothing parameter selection.[4]

3.3 | Software

The reviewed methods can be undertaken using free software. The frequentist approach with CGAMLSS can be performed in R using the `GJRM` package.[4] CGAMLSS from a Bayesian perspective can be undertaken using BayesX open-source software.[22]

The two R-packages BayesX[23] and R2BayesX[24,22] can be used to provide graphic interfaces in the Bayesian setting. See the appendix for more details.

4 | JOINT MODELLING OF GLYCATION DATA

This section describes the construction of a bivariate model for studying the relationship between HbA1c (Y_1) and fructosamine (Y_2). This involved making the choice of appropriate marginal distributions, additive predictors and a suitable copula function.

4.1 | Model Building

The Akaike and Bayesian Information Criteria (AIC/BIC) can be used to deal with model selection in the frequentist setting,[4] while the Deviance Information Criterion (DIC)[25] and the Widely Applicable Information Criterion (WAIC)[26] can be used to choose the best response distributions and a suitable copula function in the Bayesian framework. The DIC is a commonly used criterion for model choice in Bayesian regression models. It became popular partly because of its easy implementation from the Markov chain Monte Carlo (MMMC) output. The performance of the DIC was evaluated positively by Klein and Kneib, who compared several misspecified models with the true model using this criterion.[5] In addition, Marra and Radice[4] showed via simulation studies that in the frequestist framework AIC and BIC are able to identify the correct copula function and hence the correct type of dependence structure.

4.1.1 | Marginal distributions

If the user chooses the wrong marginal distribution this can also affect the choice of the copula but this will depend on the severity of marginal misspecification. To avoid this situation, we propose to start with the selection of two adequate marginal distributions that fit each response satisfactorily as two different and independent GAMLSS regression models. In this univariate framework, model selection was extendedly studied in the statistical literature. See for example Klein et. al.[7] for a detailed guide for dealing with model choice.

In this case study, AIC/BIC (in the frequentist approach) and WAIC/DIC (in the Bayesian approach) showed log-normal distributions to provide the best fit for both margin distributions within a set of candidate distributions (data not shown). The log-normal distribution is characterized by a location parameter μ plus a scale parameter σ. For a log-normally distributed random variable y with a density function $f(y|\mu,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left[-\frac{\{\log(y)-\mu\}^2}{2\sigma^2}\right]$, the expectation and the variance can be expressed as $E(y) = e^{\frac{\sigma^2}{2}} e^{\mu}$ and $V(y) = e^{\sigma^2}\left(e^{\sigma^2}-1\right)e^{2\mu}$.

4.1.2 | Additive Predictors

In CGAMLSS, the analyst has to define regression predictors for each parameter of the response distribution. Good knowledge of the biological process, plus the information criterion, can be used to guide the selection of the covariates for each predictor component. For example, in this study MCV was selected as a covariate in the first marginal (HbA1c), and albumin in the second (fructosamine). The MCV is the mean volume of red blood cells, and is useful in classifying the type of anaemia based on red cell morphology. A higher than normal MCV indicates that the red blood cells are too big, and could reflect folic acid or vitamine B12 deficiencies, conditions frequently associated with alcohol consumption. Conversely, a low MCV reveals the volume of red blood cells to be below normal - a very common condition that usually reflects iron deficiency, especially in women with heavy menstrual bleeding. The reasons for such influence of MCV on the concentration of HbA1c are largely unknown. In both of these anaemia conditions, the blood circulation time of the red cells is reduced, and since haemoglobin is found inside red cells, the lifespan of the red cells could be reduced too. Serum albumin is the main substrate to which glucose binds to, forming serum fructosamine.

Age and BMI were included as covariates since they can be expected to modify the glycation processes.[27,28] For the scale parameter, age and glucose were considered since they are involved in the variability of the glycation rate (See Figure 1). Glucose, age and MCV were contemplated as covariates for the same reasons. In addition, values for AIC/BIC and DIC/WAIC

were taken into account when selecting the final model. Finally, the additive predictors for the parameters of the joint distribution were specified as

$$\begin{cases} \eta_i^{\mu_1} = \beta_{0i}^{\mu_1} + f_i^{\mu_1}(Glucose) + f_i^{\mu_1}(Age) + Gender_i\beta_{1i}^{\mu_1} + f_i^{\mu_1}(BMI)_i + f_i^{\mu_1}(MCV)_i \\ \eta_i^{\sigma_1} = \beta_{0i}^{\sigma_1} + f_i^{\sigma_1}(Glucose) + f_i^{\sigma_1}(Age) + Gender_i\beta_{1i}^{\sigma_1} \\ \eta_i^{\mu_2} = \beta_{0i}^{\mu_2} + f_i^{\mu_2}(Glucose) + f_i^{\mu_2}(Age) + Gender_i\beta_{1i}^{\mu_2} + f_i^{\mu_2}(BMI) + f_i^{\mu_2}(Alb), \\ \eta_i^{\sigma^2} = \beta_{0i}^{\sigma_2} + f_i^{\sigma_2}(Glucose) + f_i^{\sigma_2}(Age) + Gender_i\beta_{1i}^{\sigma_2} \\ \eta_i^{\rho} = \beta_{0i}^{\rho} + f_i^{\rho}(Glucose) + f_i^{\rho}(Age) + Gender_i\beta_{1i}^{\rho} + f_i^{\rho}(MCV), \end{cases} \tag{9}$$

Hereafter the parameter index is eliminated for the sake of simplicity. The predictors (η_i) are formed through the additive composition of an intercept β_0 representing the overall level of the predictor, linear effects for $gender$, and functions f reflecting the non-linear effects of the continuous covariates (glucose, age, BMI, albumin and MCV). The first and third equations of Formula 8 refer to the location parameters μ_1 and μ_2 of HbA1c and fructosamine respectively, while the second and fourth equations refer to the scale parameters σ_1 and σ_2. The eighth equation refers to the association parameter of the copula ρ.

For gender, no penalty is (as usually) assigned (i.e, $\mathbf{K}_j = \mathbf{0}$) which corresponds to flat priors from a Bayesian point of view. To render the Bayesian and frequentist modelling comparable in our empirical analysis, the smooth functions, f, of the continuous covariates (glucose, age, BMI, MCV and Albumin) were estimated using penalized splines. For Bayesian inference, cubic B-splines with a 10 equidistant inner knot grid were used such that dim($\boldsymbol{\beta}$)=12. The prior for $\boldsymbol{\beta}$ is based on a second or random walk prior with inverse gamma hyperpriors for τ^2.[29] For frequentist CGAMLSS, penalized splines with second order penalties were also considered, and the number of basis functions fixed to 10 (as in the Bayesian approach). There are several reasons for the the penalized spline specifications, i) For practical application, the use of penalizations relaxed the election of the number of knots[30]; ii) The number and placement of the knots has only a very minor impact on the fit if the number of knots chosen is not too small,[31,29,32] iii) All continuous covariates included in the model have a lot of different values so it makes sense to have a few knots; iv) We have made several tries with different number of knots and 10 equidistant knots yield sufficient flexibility for all the covariates included in the model (data not shown); v) Finally, in Bayesian inference, second order random walk priors leave a linear effect unpenalized which is in analogy to the common penalty for smoothing splines. In an analogous framework, second order penalizations are considered in frequentist inference. On the other hand, first order differences often yield more wiggly estimates than second order differences.

4.1.3 | Selection of Copula Function

As for the choice of the copula, different copula functions available in both the frequentist and the Bayesian model formulations were contemplated. In this work, we focus on the Gaussian, Gumbel and Clayton copulas to make our analysis concise and justified. These represent the classes of copulas with no (Gaussian), upper (Gumbel) and lower (Clayton) tail dependence as yielded the best results in terms of our model selection criteria and with respect to numerical stability and convergence. The best fit was provided by the Gumbel copula in both frameworks (frequentist and Bayesian). See Table 3 .

TABLE 3 Comparison of model choice criteria under different copula assumptions.

	GJRM		**BayesX**	
Copula	AIC	BIC	DIC	WAIC
Gumbel	16008.93	16328.37	5905.2	5960.84
Gaussian	16073.56	16486.07	5939.74	5993.23
Clayton	17466.43	18185.62	6047.67	6130.85

Note in that respect that while it is often possible to identify whether the conditional responses between these three types of dependence structures (no, upper, and lower tail dependence) it is sometimes hard to identify the best copula within one of these classes. We should also note that, the software BayesX does not allow for estimation of all the copulas that are currently available in GJRM. To a more detailed study of the dependence structure, we estimated the models also with further copulas implemented in the GJRM package and report the results in Table 4 . We have observed numerical issues with the Joe copula.

But as can be seen in this table, Frank and FGM copula yield higher AIC/BIC values than the Gumbel copula such that the copula choice does not change.

Residuals can be used to check the performance of the selected model. If the estimated model is close to the true model, the normalized quantile residuals approximately follow a Gaussian distribution. Note that the residual checks are only for the margins and not the whole model. Figure 3 shows quantile-quantile plots for the margin models for HbA1c (top) and fructosamine (bottom). Reference bands are included for judging the departures of the quantile-quantile plots from the ideal (red line). The log-normal distribution appears appropriate for residuals in the range -2.5 to 2.4 for both margin models, but deviates from the diagonal for extreme values. Even so, these extreme values are a 2.15% of the total ($n = 1516$) for first marginal (1.4% of the database to the left and 0.75% to the right) and a 3.02% of the total data for first marginal (1.68% of the database to the left and 1.34% to the right) for the second marginal. Consequently, log-normal distribution margins explains the vast majority of observations properly. This was the best-fitting distribution.

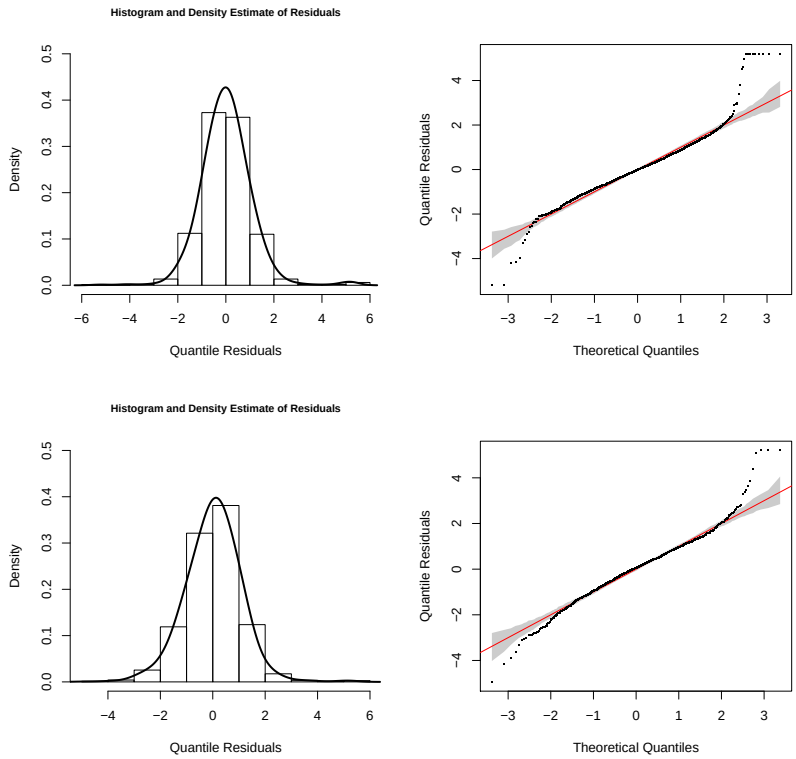

FIGURE 3 Histograms and Quantile-quantile plots of normalized quantile residuals for glycated haemoglobin (top) and fructosamine (bottom) for the selected model. The closer the residuals to the bisecting line, the better the fit to the data. Note that residuals are only indicating the goodness of fit in the marginals.

The non-linear effects of continuous variables are typically represented as centred in additive regression models. It should be noted that, in this study, new functions were developed to plot these effects on the real scale for each parameter of the bivariate responses. Figures 4 , 5 and 6 show the results obtained. Variability is expressed as the standard deviation, and the association parameter with Kendall's τ. Given that the magnitudes of the copulas' dependence parameters (ρ) are not comparable between copulas, it is normal to use association measurements such as Kendall's τ to facilitate interpretation. Kendall's τ is a well known statistical coefficient that allows one to study the strength of the relationship between two variables.[33,34] Further, for each copula function a relation exists between ρ and Kendall's $\tau \in [-1, 1]$. More specifically, for the Gumbel copula, it can be shown that: $\tau = 1 - \frac{1}{\rho}$, and the copula parameter needs to be greater than 1 such that $\log(\rho - 1)$ is an appropriate link function. Before describing the results obtained, it should be understood that the Bayesian and frequentist approaches returned very similar results. Thus, a single set of results are presented.

TABLE 4 Comparison of model choice criteria under different copula assumptions using GJRM.

	Copula model selection		
Copula	AIC	BIC	Convergence Warnings
Joe	113456.20	113663.10	Warnings obtained
Frank	16181.07	16530.73	No Warnings obtained
FGM	16194.27	16588.31	No Warnings obtained

TABLE 5 Summary of estimated linear effects for model (9) obtained from BayesX software. The results were analogous in the frequentist framework (data not shown).

Parameter	**mean**	**2.5% quantile**	**median**	**97.5% quantile**
$\beta_0^{\mu_1}$ (intercept)	1.01	0.91	1.01	1.11
$\beta_0^{\sigma_1^2}$ (intercept)	-0.71	-0.89	-0.71	-0.52
$\beta_1^{\mu_1}$ (gender)	-0.04	-0.09	-0.04	0.02
$\beta_1^{\sigma_1^2}$ (gender)	0.32	0.17	0.33	0.47
$\beta_0^{\mu_2}$ (intercept)	0.84	0.71	0.84	1.00
$\beta_0^{\sigma_2^2}$ (intercept)	-0.27	-0.45	-0.28	-0.11
$\beta_1^{\mu_2}$ (gender)	-0.26	-0.34	-0.26	-0.18
$\beta_1^{\sigma_2^2}$ (gender)	0.04	-0.11	0.04	0.19
β_0^{ρ} (intercept)	-0.18	-2.37	-1.82	-1.33
β_1^{ρ} (gender)	0.14	-0.41	0.14	0.67

4.2 | Results

4.2.1 | Marginal Expectations

Gender had no influence on the mean concentration of HbA1c. However, it had some influence on the fructosamine concentration, although this would seem to be clinically irrelevant -the differences between men and women were minimal (smaller than 0.1 mg/dL). See, Table 5 .

Fasting plasma glucose was the main covariate influencing the concentrations of both HbA1c and fructosamine. The functional form of the effect of glucose levels on these proteins was similar (Figures 4). The relationship between HbA1c and glycaemia has been extensively explored in studies by other authors, the results of which support an association between HbA1c and the glucose concentration during the preceding 5-12 weeks.[35,36]

The mean HbA1c concentration increased almost linearly with age, but the fructosamine concentration only did so in elderly people (> 50 years). These findings are consistent with the view that glycation is accelerated by ageing.[37] The age-related increase in HbA1c is similar in magnitude to that reported in the Framingham Offspring Study (FOS), which examined data from 2473 non-diabetic participants, as well as that reported in the National Health and Nutrition Examination Survey NHANES, 2001-2004 which involved 3270 non-diabetic participants. Since the HbA1c concentration increased with age after adjusting for glucose levels, factors unrelated to glucose metabolism must be involved. One such factor may be the ageing-related change in the rate of glycation. Pani et al. also reported clear differences between HbA1c and fructosamine concentrations in subjects of different ages.[27]

In individuals suffering from overweight or obesity (BMI > 25 Kg/m^2 and BMI > 30 Kg/m^2 respectively), higher BMI values were associated with a higher mean concentration of HbA1c and a lower mean concentration of fructosamine. Several studies have suggested a negative correlation between BMI and serum glycated proteins in people with and without diabetes.[38,39,28] Some authors suggest that the inverse association between extra-intravascular glycated proteins and BMI is the result of different mechanisms coming into play depending on the glucose tolerance status. In people without diabetes it would appear to be due to a direct association between BMI and glycated proteins, while in people with diabetes, glycated proteins are influenced by plasma glucose values as well.[28]

Both higher and lower MCVs appear associated with lower levels of glycated haemoglobin. It is well known that the formation of HbA1c increases in erythrocytes over their lifetime: younger cells contain smaller amounts and older cells larger amounts. Since the circulation time of the red cells is reduced under conditions of anaemia, and given that haemoglobin is found inside red cells, it might be expected that the concentration of HbA1c should fall in people suffering from anaemia. The results also show that the higher the concentration of albumin, the higher that of fructosamine. Serum albumin is the main substrate to which glucose binds, forming serum fructosamine.

4.2.2 | Marginal variances

Gender had no influence on the variability of fructosamine. HbA1c variability, however, was significantly greater in men. See Table 5 .

The variability plots suggest that variations in HbA1c and fructosamine are greater at higher glucose concentrations, identifying people with diabetes and prediabetes. The wide variability in the glycated proteins at lower glucose concentrations might also identify people with diabetes who are being treated with anti-diabetic drugs, and who have low fasting glucose concentrations (Figure 5). The variability of both HbA1c and fructosamine increased with age.

4.2.3 | Dependence

Gender had no influence on the association between HbA1c and fructosamine (see Table 5).

As expected, the association between HbA1c and fructosamine is strengthened with increasing fasting plasma glucose, and with age. It is worth noting that Kendall's τ is high for high fasting plasma glucose. In normoglycaemic subjects (plasma glucose < 100 mg/dL), no association was seen between the response variables. See Figure 6 .

The strength of the association between the glycated proteins was variable, reflecting how the lifespan of the red blood cells may be shortened in people with anaemia.

To sum up, glucose and age are identified to be possible factors that cause discordance between HbA1c and fructosamine. More specifically, Figure 6 suggests that the association (in terms of τ) is smaller and close to zero for small values of glucose and age. As a whole, the above findings could have important clinical implications when diagnosing prediabetes (a condition characterized by slightly elevated blood glucose concentrations, and indicative of risk of progression to diabetes), and it should be taken into account that discordances in these markers may be common, especially in young people or in those with prediabetes or early stage diabetes.

One of the advantages of copula regression models is the possibility of deriving further interpretable results from the fitted models. For example, Figure 7 shows the contours of the fitted bivariate distribution for different plasma glucose conditions -normo-glycaemic, prediabetes, and diabetes; criteria from American Diabetes Association.[1] It can be seen that the relationship between markers of glycemic control is varying according the levels of plasma glucose values. Correlation is high in patients with diabetes and there is no correlation in normo-glycemics.

Figure 8 shows the joint probability of exceeding certain thresholds for some covariates at different ages. It can be displayed that the probability of finding both markers below the diagnostic threshold decreases with age.

Note that, in Figure 7 the effect of gender was set to women while all continuous covariates except for glucose were fixed at their mean values for the entire data set. In Figure 8 , the effect of gender was also set to women while all continuous covariates except for glucose and age were fixed at their mean values for the entire data set.

In this work, the Gumbel copula provides the best statistical fit (see Section 3.2.3 of the manuscript) and the best clinical explanation. Figure 1 shows the dispersion diagram for HbA1c and fructosamine; one can see how this copula captures the relationship between the variables better than either the Guassian or Clayton copulae (see also Figure 2). The biomedical literature reports that the strongest association between glycated proteins is seen when glucose values are higher.[40] Indeed, one of the aims of the paper was to demonstrate that the relationship between HbA1c and fructosamine changes with the blood glucose concentration (see Figures 7 and 8). At lower glucose values, no, or only a weak, relationship exists between these variables, while at higher values the relationship is strong. In different biomedical studies in which the correlation between glucose, HbA1c and fructosamine has been examined, a strong correlation has been reported. However, this relationship was usually studied in persons with diabetes, in whom both biomarkers are present at higher concentration. When it is studied in persons who are normo-glycaemic, in whom the concentrations of both biomarkers are lower, no relationship is seen. This type of dependence - weaker at lower biomarker concentrations - may explain the discrepancies encountered when these biomarkers

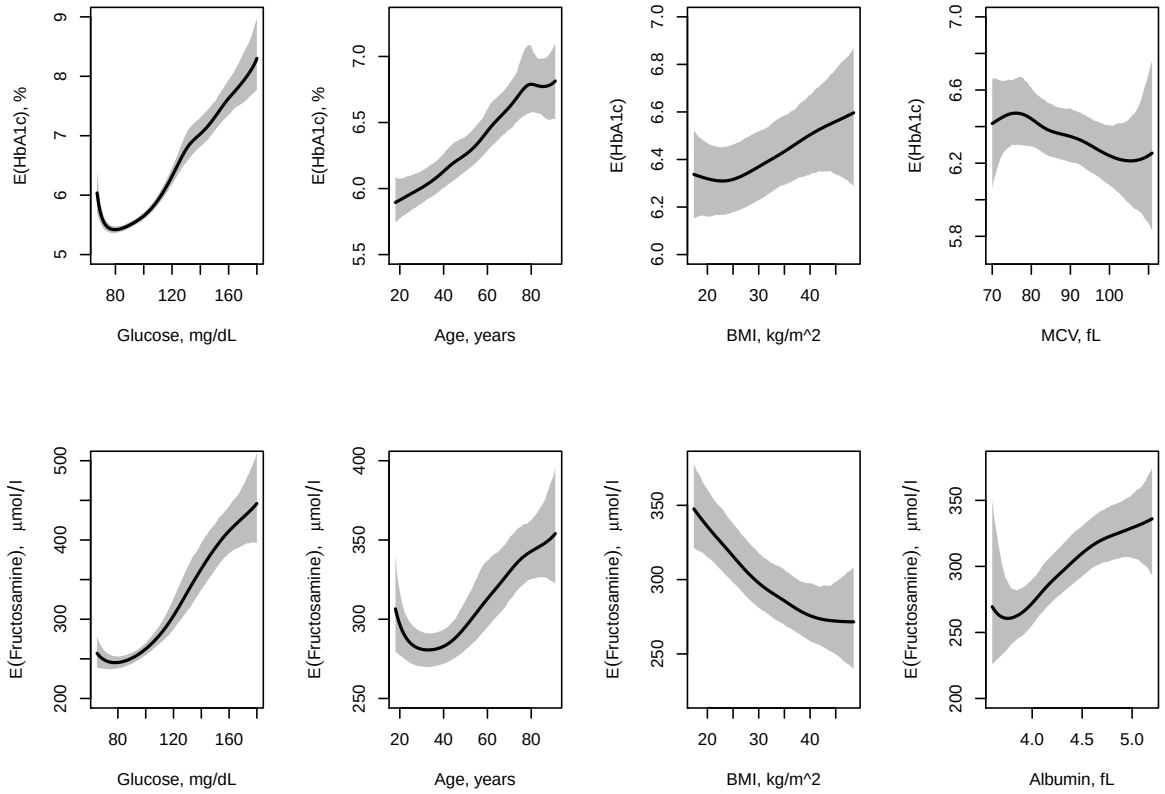

FIGURE 4 Smooth effect of Glucose, Age, BMI, albumin and MCV on the mean of HbA1c and fructosamine levels.

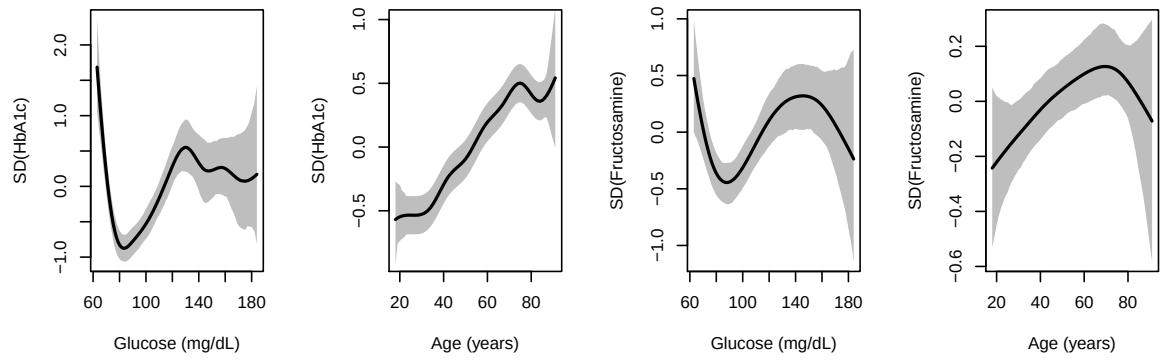

FIGURE 5 Smooth effect of Glucose and Age on the standard deviation of the HbA1c and fructosamine levels.

are used indiscriminately for the diagnosis and monitoring of a particular disease. Therefore, the use of a Gumbel-like copula becomes important.

In studies in which the correlation between the response variables is strong for lower values of a determined covariate, a Clayton-like copula may provide a better fit than a Gumbel copula. In contrast, in those in which there are no tail dependences, the Gaussian copula might be a better option for modelling the relationship between the response variables.

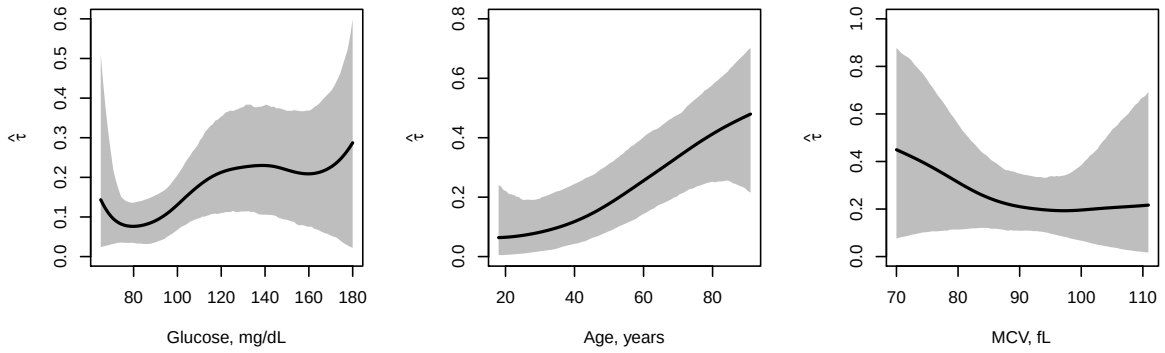

FIGURE 6 Estimates for τ from a Gumbel copula model with log-normal margins for both, HbA1c and fructosamine.

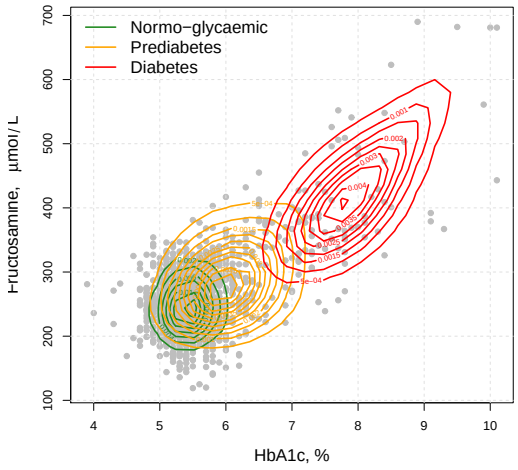

FIGURE 7 Contour lines of densities for three different glucose levels: Normo-glycaemic (FPG $<$ 100 mg/dL or HbA1c $<$ 5.7%); Prediabetes (100 mg/dL $\leq$ FPG $\leq$ 125mg/dL or 5.7% $\leq$ HbA1c $<$ 6.5%); Diabetes (HbA1c $\geq$ 6.5% or FPG $>$ 125 mg/dL).[1] In this Figure all remaining non-linear effects (except Glucose) are kept constant at $f(\overline{x})$ (estimated functions evaluated at mean covariate values). Gender has fixed to women.

5 | DISCUSSION

This study reviews flexible modern strategies for simultaneously investigating factors that influence the discordance between markers used to screen for, and diagnose, individuals with diabetes. The value of different bivariate copula generalized additive models for location, scale and shape, based on either frequentist or Bayesian inferential principles, is examined.

CGAMLSS is a new methodology that until now has been little employed in the biomedical setting. One of the reasons for this is that the results it provides are hard to interpret. The present case study, however, highlights the value of CGAMLSS to medical researchers when dealing with datasets in which multivariate dependence is of interest and marginal distributions may come from different non-standard families. This work also shows how to visualize the results of CGAMLSS at the real scale of the response variables.

From a statistical standpoint, the CGAMLSS regression models reviewed provide a generic framework for performing regression analyses in which the parameters of a potentially non-standard multivariate response distribution are related to flexible

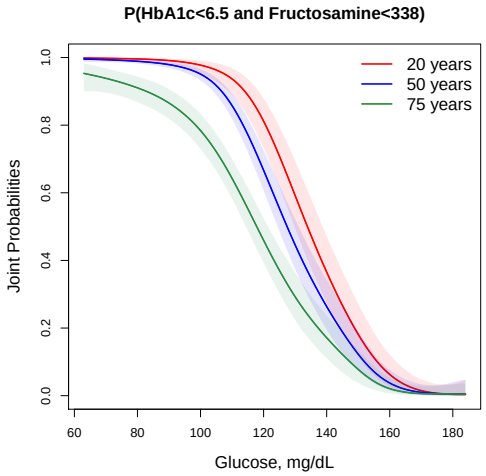

FIGURE 8 Joint Probabilities with confidence bands in terms of the glucose values for three different age levels.

regression predictors. An important feature of this type of model is the possibility of modelling, in a flexible and unified manner, different types of effect, such as spatio-temporal trends, interactions and random effects, as covariates. In addition, and as shown in the present case study, they allow the non-linear effects of continuous covariates to be investigated. In the present work, penalized splines were employed, using the same number of knots, to model continuous covariates in both the frequentist and Bayesian approach.[31] However, other penalized spline definitions could be employed in the frequentist method, such as penalized low rank thin plate splines[41] or cubic regression splines.[16] In the present case study, the use of additive instead of linear predictors was particularly useful in detecting the effect of age and glucose concentration on the variability of the HbA1c and fructosamine values.

Therefore, this methodology is potentially very useful in biomedical research. In the case examined here, we have found that two biomarkers that are indistinctly used in diabetes control (HbA1c and Fructosamine) can diverge in their results depending on the characteristics of the patients. The most important clinical contribution of these models arises when studying the association (dependency) of these two response variables in light of the covariates. For example, thanks to the models, our results demonstrate that these two diagnostic tests show bigger discrepancies between them when the patients are young and their glucose levels are normal (Figures 6 and 7). This means that the interpretation of the results should take into account the individual characteristics of the patient under examination. In other words, any of the two biomarkers could be indistinctly used provided that the association between them is high and that there are no discrepancies when taking the covariates into account.

Another benefit of CGAMLSS is that it can contemplate a broad family of non-standard response variables. Although this study focuses on two continuous response variables, this framework allows one to estimate bivariate regression models with binary responses (where link functions are not restricted to probit alone) or bivariate models with binary/discrete/continuous margins in the presence of associated responses/endogeneity. In fact, the authors are currently investigating the concordance of the different diagnostic criteria for diabetes. Two of these criteria, established by American Diabetes Association, are fasting plasma glucose levels ($\geqslant$ 126 mg/dL) and ($HbA1c \geqslant$ 6.5%).[1] We are working on the development of a bivariate binary model that allows one to investigate whether concordance exists between these two diagnostic criteria, and whether the threshold levels used are the most appropriate. Other types of diagnostic criteria have also been defined by the ADA such as the 2-hours plasma glucose value during a 75-g oral glucose tolerance test. We are also working on the development of regression models for trivariate copulas with a view to simultaneously studying the effectiveness of these three criteria, in the line with the recent proposal of Filippou et al.[42] Preliminary results suggest that these approximations may help improve the diagnosis of diabetes.

Finally, we would like to point that these modern regression techniques may be useful for clinicians since they allow for simultaneously explain which mechanisms are affecting on multivariate responses and therefore using these models could shed lights on certain important biological processes. In this case study, the usefulness of this methodology was tested in the setting of diabetes research, but it could be similarly used in studies on the markers of cancer, cardiovascular disease, and for instance in other studies of health related quality of life.[43] Further work is needed to bring this type of analysis into the clinic.

ACKNOWLEDGEMENTS

Jenifer Espasandín-Domínguez is funded by a pre-doctoral grant (ED481A-2015/113) from the Galician Government (Plan I2C)-Xunta de Galicia. This research was also supported by grants from the Carlos III Health Institute, Spain (PI16/01395; PI16/01404; RD16/0007/0006 and RD16/0017/0018), and by the projects MTM2014-52975-C2-1-R and MTM2017-83513-R cofinanced by the Ministry of Economy and Competitiveness (SPAIN) and the European Regional Development Fund (FEDER). This work was also supported by grants from the Galician Government: RED INBIOEST (ED341D-R2016/032), Grupo de Referencia Competitiva (ED431C 2016-025) and Grupo de Potencial Crecimiento (IN607B 2018-1). Nadja Klein gratefully acknowledges financial support by the Alexander von Humboldt foundation. Thomas Kneib acknowledges financial support by the German Science Foundation (DFG), grant KN 922/9-1.

Conflict of interest

The authors declare no potential conflict of interests.

APPENDIX

A - FREQUENTIST CGAMLSS (MARRA AND RADICE, 2017)

The CGAMLSS regression models introduced by Klein and Kneib (2016a) and Marra and Radice (2017) can be estimated using open-source software. The following shows the most important parts of the code employed in the present work.

To estimate the model, the R software GJRM package (Marra and Radice, 2017) was used.

```
> mu1 <-   HbA1c      ~ s(Glucose, bs = "ps", k=10) + s(Age, bs = "ps", k=10) + factor(Gender)
+ s(Bmi, bs = "ps", k=10) + s(Mcv, bs = "ps", k=10)
> mu2 <- Fructosamine ~ s(Glucose, bs = "ps", k=10) + s(Age, bs = "ps", k=10) + factor(Gender)
+ s(Bmi, bs = "ps", k=10) + s(Albumine, bs = "ps", k=10)
> sd1   <-            ~ s(Glucose, bs = "ps", k=10) + s(Age, bs = "ps", k=10) + factor(Gender)
> sd2   <-            ~ s(Glucose, bs = "ps", k=10) + s(Age, bs = "ps", k=10) + factor(Gender)
> theta <-            ~ s(Glucose, bs = "ps", k=10) + s(Age, bs = "ps", k=10) + factor(Gender)
+ s(Mcv, bs = "ps", k=10)
> f  <- list(mu1, mu2, sd1, sd2, theta)
> m1 <- gjrm(f, data = data, margins = c("LN", "LN"), Model = "B",  BivD="G0")
```

In the GJRM package, the `plot` function represents the centred effects of the response variables. To represent the effects of the continuous covariates at the true scale of the response variables, the function `pred.mvt()` can be used. This function takes into account the link functions contemplated, and allows for the fact that the mean and variability of a distribution may depend on the latter's parameters. (See Marginal Distribution, beginning of page 7). By way of example, the following shows how to represent the effect of glucose on the mean concentration of HbA1c. The effect of Gender was set to Women (zero), while all continuous covariates but the one being visualised (glucose) were fixed at the average values for the entire data set.

```
> glucoses <- seq(min(Glucose), max(Glucose), 1)
> nw       <- data.frame(Glucose = glucoses, Gender = 0, Age=mean(data$Age), Bmi=mean(data$Bmi),
Mcv=mean(data$Mcv, na.rm=T))
> res      <- pred.mvt(m1, eq = 1, fun = "mean", newdata = nw, n.sim = 10000, prob.lev = 0.05)
> minimum  <- min(as.numeric(res$CIpred))
> maximum  <- max(as.numeric(res$CIpred))

> plot(glucoses, res$pred, type = "l", ylab = "E(HbA1c)", xlab = "Glucose (mg/dl)",
ylim=c(minimum,maximum))
> polygon(c(glucoses,rev(glucoses)),c(res$CIpred[, 1],rev(res$CIpred[, 2])),col="gray80",border=NA)
> lines(glucoses, res$pred, type = "l")
```

In the function `pred.mvt`, "eq" can take values of 1 (referring to the first marginal) or 2 (second marginal). The user must also specify the effect to be visualized with the option "fun" which can take the value "mean", "variance" or "tau".

B - BAYESIAN CGAMLSS (KLEIN AND KNEIB, 2016A)

BayesX software[22] was used for this process:

```
% Dataset
dataset d
d.infile using /home/jdomin/SMMR/data.raw % The correct path must be written here
d.replace HbA1c        = log(HbA1c)
d.replace Fructosamine = log(Fructosamine)
mcmcreg yreg
yreg.outfile = /home/jdomin/SMMR/model

%Model Estimation
yreg.hregress HbA1c  = const+Gender+Glucose(pspline, nrknots=10, lambda=1000)
+ Age(pspline, nrknots=10, lambda=1000),  copula family=normal equationtype=sigma2
iterations=12000 step=10 burnin=2000 using d
yreg.hregress HbA1c  = const+Gender+Glucose(pspline, nrknots=10, lambda=1000)
+Age(pspline, nrknots=10, lambda=1000)+Bmi(pspline, nrknots=10, lambda=1000)
+Mcv(pspline,nrknots=10, lambda=1000), family=normal equationtype=mu using d


yreg.hregress Fructosamine  = const+Gender+Glucose(pspline, nrknots=10, lambda=1000)
+ Age(pspline, nrknots=10, lambda=1000),
family=normal equationtype=sigma2  using d
yreg.hregress Fructosamine  = const+Gender+Glucose(pspline, nrknots=10, lambda=1000)
+ Age(pspline, nrknots=10, lambda=1000) + Bmi(pspline, nrknots=10, lambda=1000)
+ Albumine(pspline,nrknots=10, lambda=1000),
family=normal equationtype=mu using d
yreg.hregress Fructosamine  = const+Gender+Glucose(pspline, nrknots=10, lambda=1000)
+ Age(pspline, nrknots=10, lambda=1000)
+ Mcv(pspline, nrknots=10, lambda=1000), predict=light family=clayton_copula
equationtype=rho setseed=123  using d
drop yreg
```

R software was used to represent the results obtained with BayesX. In R, the `plot2d` functions of the `R2BayesX` package, and the `plotnonp` function of the `BayesX` package, allow the effects of the centred continuous covariates to be represented. To do this at the true scale of the response variables, some code was created by the authors of this manuscript. By way of example, the following shows how to represent the effect of glucose on the mean concentration of HbA1c:

```
library("BayesX")
library("splines")

fixed_mu1_samples <- read.table("model_MAIN_mu_REGRESSION_HbA1c_LinearEffects_sample.raw",
header=TRUE)[,-1,drop=FALSE]
fixed_sigma1_samples <- read.table("model_MAIN_sigma2_REGRESSION_HbA1c_LinearEffects_sample.raw",
header=TRUE)[,-1,drop=FALSE]

f11 <- read.table("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Glucose_sample.raw",
header=TRUE)[,-1]
f12 <- read.table("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Age_sample.raw",
header=TRUE)[,-1]
f13 <- read.table("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Bmi_sample.raw",
header=TRUE)[,-1]
f14 <- read.table("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Mcv_sample.raw",
header=TRUE)[,-1]

v11 <- read.table("model_MAIN_sigma2_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Glucose_sample.raw",
header=TRUE)[,-1]
v12 <- read.table("model_MAIN_sigma2_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Age_sample.raw",
header=TRUE)[,-1]

fixed_B1 <- cbind(rep(1,100), rep(1,100))
fixed_B0 <- cbind(rep(1,100), rep(0,100))

Glucose_seq <- seq(min(data$Glucose), max(data$Glucose), length=100)
source("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Glucose_basisR.res")
Glucose_B1=BayesX.design.matrix(Glucose_seq)
dim(Glucose_B1)

Age_seq <- rep(mean(unique(data$Age)), 100)
```

```
source("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Age_basisR.res")
Age_B1=BayesX.design.matrix(Age_seq)

Bmi_seq <- rep(mean(unique(data$Bmi)), 100)
source("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Bmi_basisR.res")
Bmi_B1=BayesX.design.matrix(Bmi_seq)

Mcv_seq <- rep(mean(unique(data$Mcv)), 100)
source("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Mcv_basisR.res")
Mcv_B1=BayesX.design.matrix(Mcv_seq)

Glucose_seq <- seq(min(data$Glucose)+2, max(data$Glucose)-4, length=100)
source("model_MAIN_sigma2_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Glucose_basisR.res")
Glucose_B2=BayesX.design.matrix(Glucose_seq)

Age_seq <- rep(mean(unique(data$Age)), 100)
source("model_MAIN_sigma2_REGRESSION_HbA1c_nonlinear_pspline_effect_of_Age_basisR.res")
Age_B2=BayesX.design.matrix(Age_seq)


niter <- 1000
eta_mu_Glucose1 <- matrix(0, nrow=100, ncol=niter)
eta_sigma2_Glucose1 <- matrix(0, nrow=100, ncol=niter)
for(i in 1:niter)
{
eta_mu_Glucose1[,i] <- fixed_B0 %*% t(fixed_mu1_samples[i,]) + Glucose_B1 %*% t(f11[i,])
+  Age_B1 %*% t(f12[i,])+  Bmi_B1 %*% t(f13[i,])+  Mcv_B1 %*% t(f14[i,])
eta_sigma2_Glucose1[,i] <- fixed_B0 %*% t(fixed_sigma1_samples[i,]) + Glucose_B2 %*% t(v11[i,])
+ Age_B2 %*% t(v12[i,])
}

lnmean <- function(eta_mu, eta_sigma2){exp(eta_mu+0.5*exp(eta_sigma2))}
lnvar <- function(eta_mu, eta_sigma2){sqrt((exp(exp(eta_sigma2))-1)*(exp(2*eta_mu+exp(eta_sigma2))))}

mean_Glucose_HbA1c1 <- lnmean(eta_mu_Glucose1, eta_sigma2_Glucose1)
std_Glucose_HbA1c1 <- lnvar(eta_mu_Glucose1, eta_sigma2_Glucose1)

mean_Glucose1_mean <- apply(mean_Glucose_HbA1c1 , 1, mean)
mean_Glucose1_q2p5 <- apply(mean_Glucose_HbA1c1 , 1, quantile, prob=0.025)
mean_Glucose1_q97p5 <- apply(mean_Glucose_HbA1c1 , 1, quantile, prob=0.975)

std_Glucose1_mean <- apply(std_Glucose_HbA1c1 , 1, mean)
std_Glucose1_q2p5 <- apply(std_Glucose_HbA1c1 , 1, quantile, prob=0.025)
std_Glucose1_q97p5 <- apply(std_Glucose_HbA1c1 , 1, quantile, prob=0.975)


plot(Glucose_seq, mean_Glucose1_mean, type="l", xlab="Glucose", ylab="E(HbA1c)")
polygon(c(Glucose_seq,rev(Glucose_seq)),c(mean_Glucose1_q2p5,rev(mean_Glucose1_q97p5)),border=NA)
lines(Glucose_seq, mean_Glucose1_mean, type="l", xlab="Glucose", ylab="E(HbA1c)")

plot(Glucose_seq, std_Glucose1_mean, type="l", xlab="Glucose", ylab="SD(HbA1c)")
polygon(c(Glucose_seq,rev(Glucose_seq)),c(std_Glucose1_q2p5,rev(std_Glucose1_q97p5)),border=NA)
lines(Glucose_seq, std_Glucose1_mean*, type="l", xlab="Glucose", ylab="E(HbA1c)")
```

In the latter code, the effect of Gender was set to Women (zero), while all continuous covariates but the one being visualised were fixed at the average values for the entire data set.

References

1. American Diabetes Association. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes - 2018. *Diabetes Care.* 2018;41(Supplement 1):S13-S27.

2. Sacks DB. A1C versus glucose testing: a comparison. *Diabetes Care.* 2011;34(2):518-523.

3. Cohen RM, Holmes YR, Chenier TC, Joiner CH. Discordance between HbA1c and fructosamine: evidence for a glycosylation gap and its relation to diabetic nephropathy. *Diabetes Care.* 2003;26(1):163-167.

4. Marra G, Rosalba R. Bivariate copula additive models for location, scale and shape. *Comput Stat Data Anal.* 2017;112:99-113.
5. Klein N, Kneib T. Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Stat Comput.* 2016;26(4):841-860.
6. Rigby A, Stasinopoulos DM. Generalized additive models for location, scale and shape (with discussion). *J R Stat Soc Ser C Appl Stat.* 2005;54(3):507-554.
7. Klein N, Kneib T, Klasen S, Lang S. Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann Appl Stat.* 2015;9(2):1024-1052.
8. Sklar M. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris.* 1959;8:229-231.
9. Yan J. Enjoy the joy of copulas: with a package copula. *J Stat Softw.* 2007;21(4):1-21.
10. Sabeti A, Wei M, Craiu RV. Additive models for conditional copulas. *Stat.* 2014;3(1):300-312.
11. Acar EF, Craiu RV, Yao F. Statistical testing of covariate effects in conditional copula models. *Electron J Stat.* 2013;7(4):2822-2850.
12. Yee TW, Wild CJ. Vector generalized additive models. *58.* 1996;(3):481-493.
13. Yee TW. *Vector generalized linear and additive models: with an implementation in R*. New York: Springer; 2015.
14. Vatter T, Chavez-Demoulin V. Generalized additive models for conditional dependence structures. *J Multivar Anal.* 2015;141:147-167.
15. Hastie TJ, Tibshirani RJ. *Generalized Additive Models.* London: Chapman-Hall; 1990.
16. Wood SN. *Generalized additive models: an introduction with R*. London: Chapman and Hall; 2006.
17. Radice Rosalba, Marra Giampiero, Wojtyś Małgorzata. Copula regression spline models for binary outcomes. *Stat Comp.* 2016;26(5):981-995.
18. Gude F, Díaz-Vidal P, Rúa-Pérez C, et al. Glycemic Variability and Its Association With Demographics and Lifestyles in a General Adult Population. *J Diabetes Sci Technol.* 2017;11(4):780-790.
19. Craig Cora L, Marshall Alison L, Sjorstrom Michael, et al. International physical activity questionnaire: 12-country reliability and validity. *Medicine and science in sports and exercise.* 2003;35(8):1381-1395.
20. Gual A, Martos A Rodriguez, Lligoña A, Llopis JJ. Does the concept of a standard drink apply to viticultural societies?. *Alcohol and Alcoholism (Oxford, Oxfordshire).* 1999;34(2):153-160.
21. Fahrmeir L, Kneib T, Lang S, Marx B. *Regression. Models, Methods and Applications*. Heidelberg/Berlin: Springer; 2013.
22. Belitz C., Brezger A., Klein N., Kneib T., Lang S., Umlauf N.. *BayesX: Software for Bayesian Inference in Structured Additive Regression Models, Version 3.0.2*. Available online on http://www.BayesX.org/; 2015.
23. Kneib T, Heinzl F, Brezger A, Bové D Sabanés, Klein N. *BayesX: R utilities accompanying the software package BayesX*. R package version 0.2-9; 2014.
24. Umlauf N, Adler D, Kneib T, Lang S, Zeileis A. Structured Additive Regression Models: An R Interface to BayesX. *J Stat Softw.* 2015;63(21):1-46.
25. Spiegelhalter DJ, Best NJ, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Series B Stat Methodol.* 2002;64(4):583-639.
26. Watanabe S. A widely applicable Bayesian information criterion. *J Mach Learn Res.* 2013;14(4):867-897.

27. Pani LN, Korenda L, Meigs JB, et al. Effect of aging on A1C levels in individuals without diabetes: evidence from the Framingham Offspring Study and the National Health and Nutrition Examination Survey 2001-2004. *Diabetes Care.* 2008;31(10):1991-1996.

28. Huh JH, Kim KJ, Lee Byung-Wan, et al. The relationship between BMI and glycated albumin to glycated hemoglobin (GA/A1c) ratio according to glucose tolerance status. *PloS one.* 2014;9(2):e89478.

29. Lang S, Brezger A. Bayesian P-splines. Journal of Computational and Graphical Statistics. *J Comput Graph Stat.* 2004;13(1):183-212.

30. Rice JA, Wu CO. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics.* 2001;57(1):253-259.

31. Eilers PH, Marx BD. Flexible Smoothing with B-splines and Penalties. *Stat Sci.* 1996;11(2):89-121.

32. Brezger A, Lang S. Generalized structured additive regression based on Bayesian P-splines. *Comput Stat Data Anal.* 2006;50(4):967-991.

33. Joe H. *Multivariate models and multivariate dependence concepts*. London, New York: Chapman and Hall/CRC; 1997.

34. Nelsen RB. *An Introduction to Copulas*. Berlin, Heidelberg: Springer-Verlag; 2nd ed.; 2006.

35. Koenig RJ, Peterson CM, Jones RL, Saudek C, Lehrman M, Cerami A. Correlation of glucose regulation and hemoglobin AIc in diabetes mellitus. *N Engl J Med.* 1976;295(8):417-420.

36. Nathan DM, Turgeon H, Regan S. Relationship between glycated haemoglobin levels and mean glucose levels over time. *Diabetologia.* 2007;50(11):2239-2244.

37. Davidson MB. The effect of aging on carbohydrate metabolism: a review of the English literature and a practical approach to the diagnosis of diabetes mellitus in the elderly. *Metabolism.* 1979;28(6):688-705.

38. Miyashita Y, Nishimura R, Morimoto A, Matsudaira T, Sano H, Tajima N. Glycated albumin is low in obese, type 2 diabetic patients. *Diabetes Res Clin Pract.* 2007;78(1):51-55.

39. Koga M, Matsumoto S, Saito H, Kayasama S. Body mass index negatively influences glycated albumin, but not glycated hemoglobin, in diabetic patients. *Endocr J.* 2006;53(3):387-391.

40. Juraschek SP, Steffes MW, Selvin E. Associations of alternative markers of glycemia with hemoglobin A(1c) and fasting glucose. *Clin Chem.* 2012;58(12):1648-1655.

41. Wood SN. Thin Plate Regression Splines. *J R Stat Soc Series B Stat Methodol.* 2003;65(1):375-379.

42. Filippou P, Marra G, Radice R. Penalized Likelihood Estimation of a Trivariate Probit Model. *Biostatistics.* 2017;18(3):569-585.

43. Espasandín-Domínguez J, Carollo-Limeres C, Coladas-Uría L, Cadarso-Suárez C, Lado-Baleato O, Gude F. *Bivariate Copula Additive Models for Location, Scale and Shape with Applications in Biomedicine. In: Gil E, Gil E, Gil J, et al. (eds). The Mathematics of the Uncertain*. Verlag: Springer International Publishing; 2018.