Received: 22 June 2020 | Revised: 7 July 2021 | Accepted: 29 July 2021

DOI: 10.1002/sim.9163

RESEARCH ARTICLE

Statistics in Medicine WILEY

Modeling conditional reference regions: Application to glycemic markers

Óscar Lado-Baleato[1] | **Javier Roca-Pardiñas**[2] | **Carmen Cadarso-Suárez**[1] | **Francisco Gude**[3]

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

[2] Statistical Inference, Decision and Operations Research, Universidade de Vigo, Galicia, Spain

[3] Clinical Epidemiology Unit, Complexo Hospitalario de Santiago de Compostela, Galicia, Spain

Correspondence

Óscar Lado-Baleato, Department of Statistics, Mathematical Analysis, and Optimization, Universidade de Santiago de Compostela, Galicia, Spain.
Email: oscarlado.baleato@usc.es

Funding information

Consellería de Cultura, Educación e Ordenación Universitaria, Xunta de Galicia, Grant/Award Numbers: ED341D-R2016/032, ED431C2016-025, IN607B 2018-1; Instituto de Salud Carlos III, Grant/Award Numbers: PI16/01404, RD16/0017/0018; Ministerio de Economía y Competitividad, Grant/Award Number: MTM2017-83513-R

Many clinical decisions are taken based on the results of continuous diagnostic tests. Usually, only the results of one single test is taken into consideration, the interpretation of which requires a reference range for the healthy population. However, the use of two different tests, can be necessary in the diagnosis of certain diseases. This obliges a bivariate reference region be available for their interpretation. It should also be remembered that reference regions may depend on patient variables (eg, age and sex) independent of the suspected disease. However, few proposals have been made regarding the statistical modeling of such reference regions, and those put forward have always assumed a Gaussian distribution, which can be rather restrictive. The present work describes a new statistical method that allows such reference regions to be estimated with no insistence on the results being normally distributed. The proposed method is based on a bivariate location-scale model that provides probabilistic regions covering a specific percentage of the bivariate data, dependent on certain covariates. The reference region is estimated nonparametrically and the nonlinear effects of continuous covariates via polynomial kernel smoothers in additive models. The bivariate model is estimated using a backfitting algorithm, and the optimal smoothing parameters of the kernel smoothers selected by cross-validation. The model performed satisfactorily in simulation studies under the assumption of non-Gaussian conditions. Finally, the proposed methodology was found to be useful in estimating a reference region for two continuous diagnostic tests for diabetes (fasting plasma glucose and glycated hemoglobin), taking into account the age of the patient.

KEYWORDS

conditional reference regions, diabetes, flexible additive predictors, kernel smoothing, regression

1 | INTRODUCTION

The diagnosis and treatment of disease commonly rests on the results of measureable biomarker-based clinical laboratory tests. Indeed, some 70% of the decisions made in clinical practice are taken based on such results.[1] For every test result, the clinical laboratory provides comparator values to help the clinician understand in context the information provided.

These comparator values are often referred to as the reference interval,[2] they usually reflect the range of values within which 95% of the results of the normal healthy population falls.

When a single biomarker is examined, the reference interval is classically obtained using quantile estimation techniques,[3] or by conditional quantile regression if any variable modifies the distribution of the response variable (ie, the reference curve).[4,5] However, there is often more than one test for diagnosing a disease. For instance, the diagnosis of diabetes may be based on plasma glucose criteria, such as fasting plasma glucose (FPG) or the 2-hour plasma glucose value obtained during a 75-g oral glucose tolerance test, or the glycated hemoglobin (HbA1c) test.[6] The same tests may be used to screen for, diagnose and monitor the effects of treatment for diabetes. However, measuring glycemic control is not foolproof; its clinical usefulness is affected by a number of biological and analytical factors. Disagreement between glycemic control measurements are common, and clinicians need to know what might explain them.[7]

Following the guidelines of the American Diabetes Association, the diagnosis of diabetes is defined as FPG levels of ≥126 mg/dL, and of HbA1c ≥ 6.5%. The same cut-offs are used for children, adolescents, and adults. FPG and HbA1c levels are reported to be strongly correlated,[8] both in members of the general population and in patients with diabetes.[9,10] If two tests are indifferently used for the diagnosis of a disease, the correlation between them should be strong. Therefore, when diagnosing a disease using two markers, it might be reasonable to estimate their combined multivariate reference region instead of the reference interval for each test. The idea of combining reference regions for two or more laboratory tests has been discussed in the biomedical,[11-13] and statistical literature.[14] However, the proposals made have required responses that follow a multivariate Gaussian distribution, condition that are not fulfilled by many markers used in the clinical setting. For instance, FPG and HbA1c concentrations both show a skewed distribution. Hence, a more general method for estimating multivariate reference regions is needed. Moreover, since HbA1c increases with age even after adjusting for glucose levels,[15-17] this variable should be taken into account when establishing cut-offs or reference values.

The literature reports but a few attempts to define reference regions for non-Gaussian multivariate responses. Nonparametric reference regions for multivariate responses can be estimated using multivariate quantiles. However, there is no single definition of what a multivariate quantile is.[18] Most current definitions are based on a center-outward ordering of the data points,[19,20] defining a convex hull in which a proportion of the more central data points falls. Halfspace depth bivariate quantiles based on directional projections have recently been extended to the regression setting for estimating bivariate contours conditioned by covariates.[21] These conditional bivariate quantile models have been used in the study of anthropometric characteristics affected by age.[22] However, the clinical interpretation of the results is not clear. Easily interpretable conditional bivariate quantile contours can be estimated using the Wei model,[23] which is based on the estimation of i) the marginal stratified conditional quantile regression for each response, and ii) bivariate quantiles using simulated data points from the above marginal stratified conditional quantile regression. The main drawback of this proposal is the influence of the univariate quantile regression outcome on the performance of the final bivariate quantile contour. Another nonparametric conditional method for detecting extreme combinations of two variables has also been proposed.[24] This does not define a reference region, but four bivariate reference curves for detecting four possible atypical combinations of two variables. However, this alternative returns a higher false discovery rate for the reference region.

Non-Gaussian bivariate regions for detecting joint outliers may be estimated using conditional copula regression models.[25] This procedure was recently applied to identify children with abnormal vision.[26] Copula regression models allow a bivariate distribution to be constructed from a modular perspective, combining two univariate parametric distributions and the parametric copula joining them.[27] Finally, the response parameters are made dependent on the covariates using flexible additive models.[28] Given copula regression model structure, and the effect of the estimated covariate on the means, variances and correlation of the responses, a conditional bivariate region for exploring the atypical combinations of two measurements can be estimated. The parametric representation of these models means several choices have to be made in the model building process, which complicates the use of this methodology when dealing with real data problems.

The present work proposes a new regression model for estimating conditional bivariate reference regions. The reference region is defined as the density function contour level which contains the bivariate data points with a given probability depending on the covariates. Unlike existing methods, cited above, our statistical model places no parametric restrictions on the response, and the nonlinear effects of continuous covariates may be estimated using local polynomial regression smoothers. Our proposal is an extension to bivariate data of a previous work,[29] where the authors used a location-scale model to estimate univariate percentile curves. The final performance of our conditional reference regions depends heavily on a bivariate kernel density estimator. It is well known, that nonparametric density estimators are largely influenced by kernel bandwidth matrix choice. Therefore, in this work, we also propose a new kernel bandwidth estimation method in order to obtain a reference region with a coverage of the bivariate data points close to the nominal level.

The proposed statistical methodology was tested with simulated data, and applied to estimate an age dependent reference region for two glycemic markers (FPG and HbA1c), in a cohort of nominally normoglycemic persons.

The remainder of this article is organized as follows. Section 2 introduces the model structure and its estimation details. Section 3 evaluates the conditional reference region estimation, contemplating scenarios in which the model could be partially miss-specified. Section 4 makes use of the model for estimating an age-specific bivariate reference region for the diabetes markers FPG and HbA1c. Finally, Section 5 discusses the limitations and applicability of the proposed model, highlighting possible extensions.

2 | CONDITIONAL BIVARIATE REFERENCE REGION MODEL

This section presents a regression model for estimating a bivariate reference region conditioned by a set of covariates. The proposed model has no parametric restriction, and is based on the estimate of a location-scale regression, taking into account the effect of the covariate on the correlation between the response variables.

2.1 | Model formulation

Let $\mathbf{X} = (X_1, \dots, X_p)$ be a vector of p covariates, and let $\boldsymbol{Y} = (Y_1, Y_2)$ be a continuous bivariate response of interest. In this context, the aim is to obtain a bivariate region τ of $\mathbf{Y}$ conditioned by the covariates $\mathbf{X}$ denoted as $\mathbf{R}_\tau(\mathbf{X})$, and containing the $100\tau\%$ of the bivariate data points. The following structure is assumed:

$$\boldsymbol{Y} = \begin{pmatrix} \mu_1(\mathbf{X}) \\ \mu_2(\mathbf{X}) \end{pmatrix} + \boldsymbol{\Sigma}^{1/2}(\mathbf{X}) \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}, \tag{1}$$

where μ_1 and μ_2 represent the response means, and the variance-covariance matrix is given by:

$$\boldsymbol{\Sigma}(\mathbf{X}) = \begin{pmatrix} \sigma_1^2(\mathbf{X}) & \sigma_{12}(\mathbf{X}) \\ \sigma_{12}(\mathbf{X}) & \sigma_2^2(\mathbf{X}) \end{pmatrix}$$

The bivariate residuals $(\varepsilon_1, \varepsilon_2)$ are assumed to be independent of the covariates and to have a mean of zero, zero unit variance, zero correlation, and an unknown density function $f(\varepsilon_1, \varepsilon_2)$. Note that, $\Sigma^{1/2}(\boldsymbol{X})$ represents the Cholesky decomposition of the variance-covariance matrix $\Sigma(\boldsymbol{X})$ so that, $\Sigma^{1/2}(\boldsymbol{X})(\Sigma^{1/2}(\boldsymbol{X}))^T = \Sigma(\boldsymbol{X})$. Thus, for any given $\boldsymbol{X}$ the bivariate region for (Y_1, Y_2) is given by:

$$R_\tau(\mathbf{X}) = \begin{pmatrix} \mu_1(\mathbf{X}) \\ \mu_2(\mathbf{X}) \end{pmatrix} + \boldsymbol{\Sigma}^{1/2}(X) R_\tau \quad \text{for } \tau \in [0,1], \tag{2}$$

where the R_τ is the unconditional bivariate region containing the $100\tau\%$ of the model residuals $(\varepsilon_1, \varepsilon_2)$ defined as:

$$R_\tau = \{(u, v) \in \mathbb{R}^2 \lfloor f(u, v) \leq k\} \quad \text{for } \tau \in [0,1],$$

where k value is chosen so that $P((\varepsilon_1, \varepsilon_2) \in R_\tau) = \tau$ for $\tau \in [0,1]$.

In Equation (1), the response parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12})$ are related to the covariates vector $\boldsymbol{X}$ via additive predictors and known link functions G, which ensure that the restrictions on the parameter spaces are maintained. The following additive predictors are considered:

$$\mu_r(\boldsymbol{X}) = \alpha_r + \sum_{j=1}^{p} f_{jr}(X_j) \quad \text{and} \quad \sigma_r^2(\boldsymbol{X}) = G_\sigma\left(\beta_r + \sum_{j=1}^{p} g_{jr}(X_j)\right) \quad \text{for } r = 1, 2, \tag{3}$$

where f_{jr} and g_{jr} are smooth unknown functions, α_r and β_r are intercept coefficients and the link function $G_\sigma = \exp(\cdot)$ to ensure that $\sigma_r^2(\boldsymbol{X}) \geq 0$. Finally, the following additive structure is assumed for the responses' association with one another, expressed as a linear correlation coefficient (ρ):

$$\rho(\boldsymbol{X}) = G_\rho\left(\gamma + \sum_{j=1}^{p} m_j(X_j)\right), \tag{4}$$

where m_j are smooth unknown functions, γ is an intercept coefficient, and in this case the link function $G_\rho = \tanh(\cdot)$ to ensure that $\hat{\rho}(\boldsymbol{X}) \in [-1, 1]$. For the sake of mathematical notational simplicity, only a nonlinear effect of the continuous covariates is contemplated in Equations (3) and (4), but they could easily be adapted to incorporate factor effects. For instance, if the first p_1 covariates define categories, the expression of $\mu_r(\boldsymbol{X})$ in (3) can be replaced by the following semi-parametric structure $\mu_r(\boldsymbol{X}) = \alpha_r + \sum_{j=1}^{p_1}\alpha_{jr}X_j + \sum_{j=p_1+1}^{p} f_{jr}(X_j)$. The variances and correlation structures can be similarly treated.

2.2 | Estimation algorithm

This section discusses the procedure for estimating the conditional bivariate region presented in Equation (2). The methodology is based on the estimate of the covariate effects on the means of the responses via the use of a flexible additive predictor, and then on the variance-covariance matrix using the squared residuals of the conditional means estimates. Finally, the bivariate region τ is obtained using a bivariate kernel estimate of the standardized bivariate residuals density. Specifically, given a sample of $\{(Y_{i1}, Y_{i2}), \boldsymbol{X}_i\}_{i=1}^{n}$, where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})$, the proposed estimation algorithm is as follows:

Step 1: For $r = 1, 2$ additive predictor is fitted to the original sample $\{Y_{ir}, \boldsymbol{X}_i\}_{i=1}^{n}$ to obtain the estimates:

$$\hat{\mu}_r(\boldsymbol{X}_i) = \hat{\alpha}_r + \sum_{j=1}^{p} \hat{f}_{jr}(X_{ij}) \quad \text{for } i = 1, \dots, n. \tag{5}$$

Step 2: For $r = 1, 2$ the squared residuals of the previous models $\hat{\mu}_r(\boldsymbol{X}_i)$ are obtained and additive predictor fitted to the sample $\{(Y_{ir} - \hat{\mu}_r(\boldsymbol{X}_i))^2, \boldsymbol{X}_i\}_{i=1}^{n}$ and obtain the estimates:

$$\hat{\sigma}_r^2(\boldsymbol{X}_i) = G_\sigma\left(\hat{\beta}_r + \sum_{j=1}^{p} \hat{g}_{jr}\left(X_{ij}\right)\right) \quad \text{for } i = 1, \dots, n. \tag{6}$$

Step 3: For $i = 1, \dots, n$ the standardized residuals are computed:

$$\hat{r}_i = \frac{(Y_{i1} - \hat{\mu}_1(\boldsymbol{X}_i))\,(Y_{i2} - \hat{\mu}_2(\boldsymbol{X}_i))}{\hat{\sigma}_1(\boldsymbol{X}_i)\hat{\sigma}_2(\boldsymbol{X}_i)}$$

and the correlation model $\rho(\boldsymbol{X})$ using the sample $\{\hat{r}_i, \boldsymbol{X}_i\}_{i=1}^{n}$ is fitted as:

$$\hat{\rho}(\boldsymbol{X}_i) = G_\rho\left(\hat{\gamma} + \sum_{j=1}^{p} \hat{m}_j(X_{ij})\right). \tag{7}$$

Step 4: For $i = 1, \dots, n$ the estimated standardized bivariate residuals are then obtained:

$$\begin{pmatrix} \hat{\varepsilon}_{i1} \\ \hat{\varepsilon}_{i2} \end{pmatrix} = \hat{\Sigma}^{-1/2}(\boldsymbol{X}) \begin{pmatrix} Y_{i1} - \hat{\mu}_1(\mathbf{X}_i) \\ Y_{i2} - \hat{\mu}_2(\mathbf{X}_i) \end{pmatrix},$$

where $\hat{\Sigma}^{-1/2}(\boldsymbol{X})$ is the inverse of the Cholesky decomposition of $\Sigma(\boldsymbol{X})$. Using the sample $\{(\hat{\varepsilon}_{i1}, \hat{\varepsilon}_{i2})\}_{i=1}^{n}$ the kernel estimation of the bivariate density $\hat{f}(\varepsilon_1, \varepsilon_2)$ is obtained. Then, the bivariate region on the residual scale (R_τ) can be obtained as

$$\hat{R}_\tau = \{(u,v) \in \mathbb{R}^2 | \hat{f}(u,v) \leq \hat{k}\}, \tag{8}$$

where $\hat{k}$ is the quantile τ of $\hat{f}(\hat{\varepsilon}_1, \hat{\varepsilon}_2)$.

Step 5: Finally, the conditional bivariate region for each $\mathbf{X}_i$ value is given by:

$$\hat{R}_\tau(\mathbf{X}_i) = \begin{pmatrix} \hat{\mu}_1(\mathbf{X}_i) \\ \hat{\mu}_2(\mathbf{X}_i) \end{pmatrix} + \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X}_i)\hat{R}_\tau.$$

Full details about nonlinear functions ($\hat{f}_{jr}$, $\hat{g}_{jr}$, $\hat{m}_j$) estimation can be found in Appendixes A and B. Moreover, in Appendix D, we explain how to obtain confidence intervals for the additive predictors covariates effects through bootstrap.

In addition to the $R_\tau(\boldsymbol{X})$ estimate, the proposed algorithm can be used to obtain the univariate conditional reference curves for each response variable, applying the following expression:

$$\hat{Q}_{\tau r}(\boldsymbol{X}) = \hat{\mu}_r(\boldsymbol{X}) + \hat{\sigma}_r(\boldsymbol{X})\hat{\varepsilon}_{\tau r} \quad \text{for } r = 1, 2, \tag{9}$$

where $\hat{\varepsilon}_{\tau r}$ is the empirical τ-quantile of the univariate errors $\varepsilon_{1r}, \dots, \varepsilon_{nr}$.

2.3 | Bivariate kernel estimation: Bandwidth selection problem

The covariate dependent bivariate reference region $R_\tau(\boldsymbol{X}_i)$ requires the estimation of a region containing the standardized bivariate residuals with a given probability τ derived from the bivariate residuals' density (see Equations (11) and (12)). Given the sample $\{(\hat{\varepsilon}_{i1}, \hat{\varepsilon}_{i2})\}_{i=1}^n$, the f density estimator at a given point (u, v) is given by:

$$\hat{f}\,((u,v), \mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}} \begin{pmatrix} u - \hat{\varepsilon}_{i1} \\ v - \hat{\varepsilon}_{i2} \end{pmatrix},$$

where $K(\cdot)$ represents the kernel (a bivariate symmetric probability density function, usually the standard bivariate Gaussian distribution), and $\mathbf{H}$ a diagonal matrix defining the kernel bandwidth, the selection of which is crucial for obtaining good estimate of R_τ and hence the final region $R_\tau(X_i)$.

For the selection of the bandwidth $\mathbf{H}$, a plug-in[30] or cross-validation[31] estimator can be used, as in any density estimation problem. However, the optimal bandwidth for density estimation might not be optimal for the coverage properties of the conditional bivariate region $\hat{R}_\tau(\mathbf{X}_i)$ (see Figure 1 and Appendix C). Thus, a bandwidth needs to be chosen such that it minimizes the difference between the estimated and nominal coverage of $\hat{R}_\tau(\mathbf{X}_i)$. This is achieved here by basing the selection on the expression $\hat{H} = \hat{h}\lambda$, where $\hat{h}$ is the former bandwidth estimate obtained using the plug-in estimator, and λ is a parameter modulating the final shape of the estimated region. This parameter is estimated as

$$\hat{\lambda} = \arg\min_{\lambda} \left| \left(n^{-1} \sum_{i=1}^{n} I\{(Y_{i1}, Y_{i2}) \in R^{(-i)}(\boldsymbol{X}_i)\} \right) - \tau \right|, \tag{10}$$

where τ is the desired coverage and $\hat{R}_\tau^{(-i)}(X_i)$ is the estimated bivariate region without the ith observation. Given the high-computational cost of (10), a k-fold cross-validation scheme could be used instead.

3 | SIMULATION STUDY

In this section, the estimated conditional bivariate regions $\hat{R}_\tau(\boldsymbol{X})$ are evaluated in terms of the data point coverage and its proximity to the theoretical $R_\tau(\boldsymbol{X})$. This coverage is estimated using an out-sample design, using one dataset in the estimation and the other in the evaluation. The comparison between the estimated and theoretical bivariate regions was performed using the root mean square error (RMSE) distance. Taking into account that both the theoretical and the estimated bivariate reference regions are defined by a set of bivariate points, $R_\tau(\boldsymbol{X}_i) = \{(u_k^i, v_k^i)\}_{k=1}^B$, and $\hat{R}_\tau(\boldsymbol{X}_i) = \{(\hat{u}_j^i, \hat{v}_j^i)\}_{j=1}^b$,

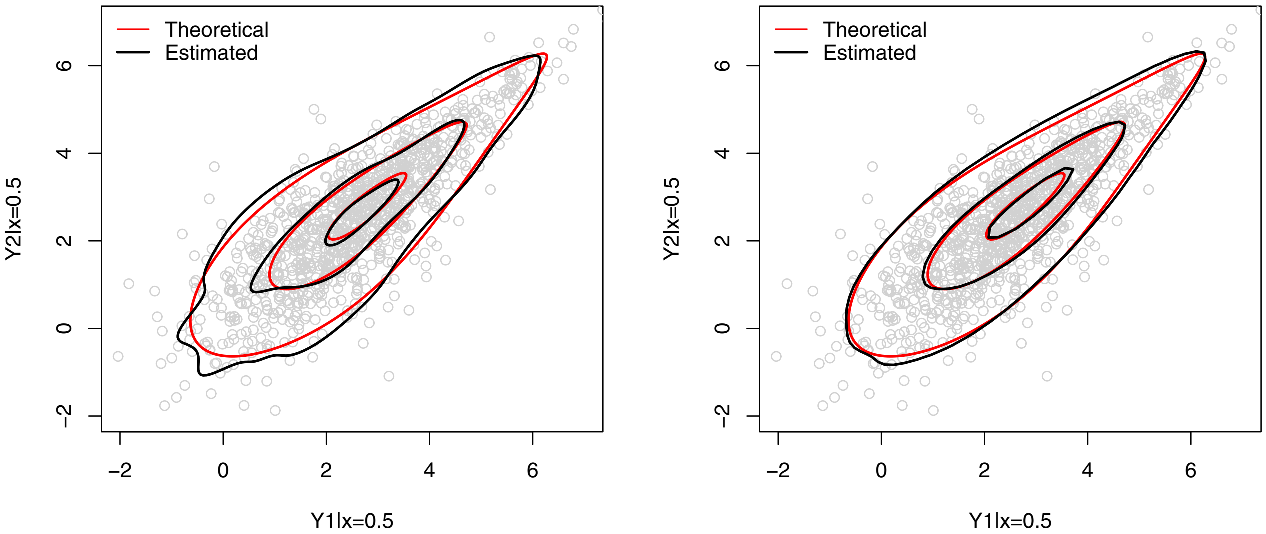

FIGURE 1 Example estimating the $R_\tau(\boldsymbol{X}_i)$ for $\tau = 0.10, 0.50$ and 0.90, using the plug-in bandwidth estimator (left) and following the method proposed in Equation (10) (right). The theoretical bivariate regions were obtained using the parametric density function of the response variable, and 100 000 simulated data points [Colour figure can be viewed at wileyonlinelibrary.com]

the RMSE distance was defined as follows:

$$\mathrm{RMSE}\left(\hat{R}_\tau(\boldsymbol{X}_i), R_\tau(\boldsymbol{X}_i)\right) = b^{-1}\sum_{j=1}^{b}\min_{1\le k\le B}\sqrt{(\hat{u}_j^i - u_k^i)^2 + (\hat{v}_j^i - v_k^i)^2}. \tag{11}$$

In practical applications, the bivariate kernel contour lines are used to estimate $(\hat{u}_j^i, \hat{v}_j^i)$.

3.1 | Scenario 1: Response with Gaussian error

In the first simulation scenario, the datasets were generated according to Equation (1). The means, the variances, and correlation structures of the responses were made dependent on a continuous and a binary regressor. Given a continuous $X_1 \in U[0,1]$, and a binary covariate X_2, the following predictor structure was considered:

$$\begin{cases} \mu_1(\boldsymbol{X}) = 1 + X_2 + f_1(X_1) \\ \sigma_1^2(\boldsymbol{X}) = 1 + X_2 \\ \mu_2(\boldsymbol{X}) = 1 + 0.5X_2 + f_2(X_1) \\ \sigma_2^2(\boldsymbol{X}) = 1 + 0.5X_2 \\ \rho(\boldsymbol{X}) = 0.3 + 0.2X_1 + 0.3X_2 \end{cases}, \tag{12}$$

where $f_1(X_1) = X_1\sin(3X_1)$ and $f_2(X_1) = \sin(2\pi X_1)$ represent nonlinear effects. Finally the error term $(\varepsilon_1, \varepsilon_2)$ was simulated from a standard bivariate Gaussian distribution. The sample size were set to $n = 500, 1000, 2000$, and the evaluation was done using 1000 replicates.

In Figure 2, the $R_{0.95}$ bivariate region estimates are depicted along with the theoretical estimates for $X_1 = 0.5$ and $X_2 = 0$. The estimated bivariate region shows a similar shape to the theoretical situation, becoming even closer to it and showing smaller variability as the sample size increases. In Figure 3, the RMSE for $\hat{R}_{0.95}$ is presented for a sequence of continuous predictor values with the binary covariate fixed at zero, for varying sample sizes. As expected, the median RMSE value and variability decreased with increasing sample size. Moreover, the RMSE was higher for the X_1 values located at the covariate range boundaries (0 and 1); this is justified by the frontier effect associated with nonlinear regression models.

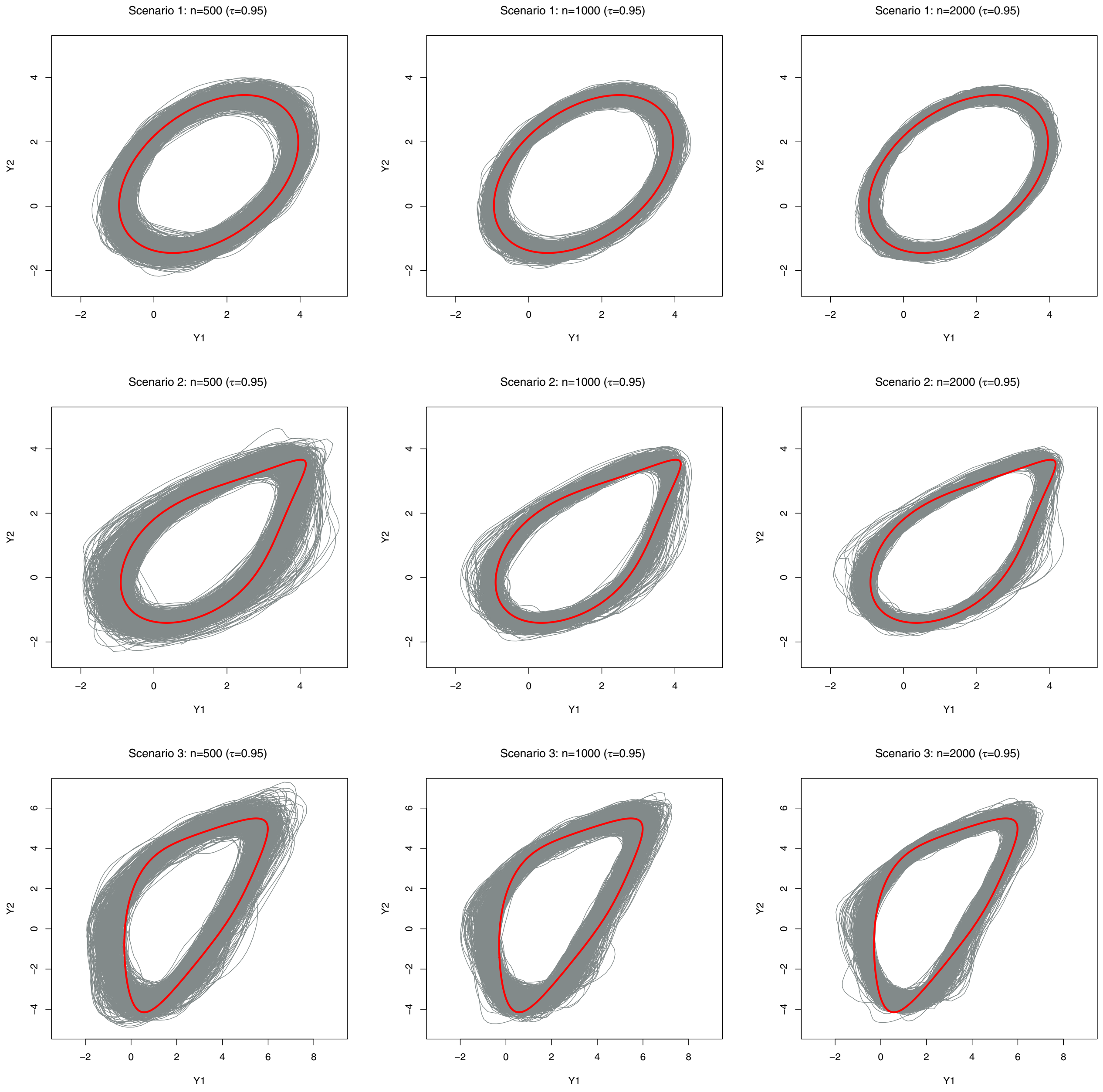

FIGURE 2 Estimated bivariate regions for $\tau = 0.95$, $X_1 = 0.5$ and $X_2 = 0$ along with the theoretical ones (red), for every sample size considered (500, 1000, and 2000) and three simulation scenarios [Colour figure can be viewed at wileyonlinelibrary.com]

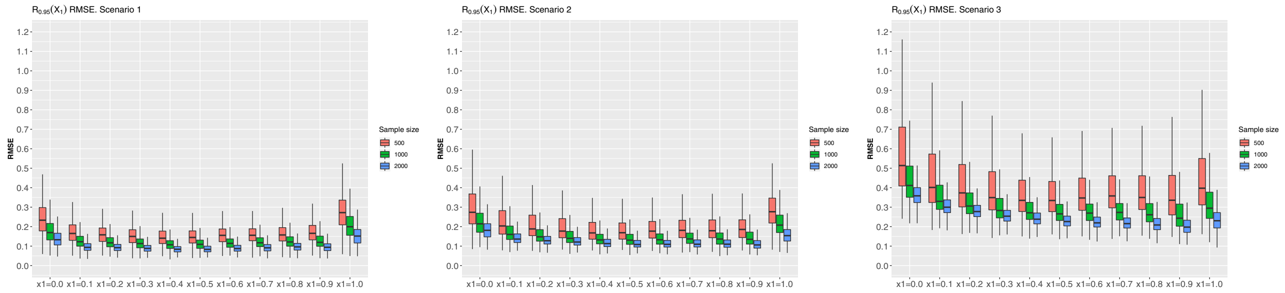

FIGURE 3 Root mean square error of the estimated bivariate regions for $\tau = 0.95$, for different sample sizes (500, 1000, and 2000), and with $X_2 = 0$ and $X_1 = (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)$ for the three simulation scenarios [Colour figure can be viewed at wileyonlinelibrary.com]

Table 1 shows the percentage of bivariate data points contained within the estimated bivariate region to be similar to that seen for the theoretical situation, whether for different nominal levels ($\tau = 0.05, 0.50, 0.90, 0.95$), continuous predictor variable values, or sample sizes. As for the RMSE, the coverage probability approaches the nominal level as the sample size increases, and becomes worse for values of X_1 located at the covariate range boundaries (0 and 1).

3.2 | Scenarios 2 and 3: Non-Gaussian dependencies

In the previous simulation, the proposed model obtained good estimates of the conditional bivariate region. However, one possible problem lies in that, besides the lack of parametric assumptions, the correlation estimate is based on the linear correlation coefficient which is not optimal under some circumstances (non-Gaussian margins and/or nonelliptical structures of dependence). The model was therefore used in two more scenarios contemplating nonstandard bivariate distributions generated from a parametric copula representation. This representation allows complex bivariate distributions to be simulated since the different parametric copula functions proposed in the literature allow for different types of dependence structure between the responses, and each variable may follow any parametric distribution. Given two continuous variables (Y_1, Y_2), their joint distribution function $F_{1,2}$ conditioned on $\boldsymbol{X}$ may be represented in terms of their univariate distributions functions (F_1, F_2) and a copula C joining both, as

$$F_{1,2}\left(Y_1, Y_2|\boldsymbol{X}\right) = C\left(F_1\left(Y_2|\mu_1(\boldsymbol{X}), \sigma_2(\boldsymbol{X})\right), F_2\left(Y_2|\mu_1(\boldsymbol{X}), \sigma_2(\boldsymbol{X})\right); \theta(\boldsymbol{X})\right), \tag{13}$$

where F_1, F_2 are two univariate parametric marginal distributions, defined by a location (μ) and the scale (σ) parameter and C represents a parametric copula function defined by an association parameter θ measuring the correlation between both responses. In the present case, the following bivariate response structures were contemplated:

- Scenario 2: F_1 and F_2 were considered to be two Gaussian distributions and C a Gumbel copula, defining a nonelliptical structure of dependence (upper-tail dependence).
- Scenario 3: F_1 was simulated from reverse Gumbel distribution and F_2 from a logistic one.[32] As in the former scenario, both are joined by a Gumbel copula.

The predictor structures for the bivariate distribution parameters were the same as presented in Equation (12), but the association was expressed in terms of Kendall's correlation coefficient. The datasets of both scenarios were generated using the gamlss[33] and copula[34] packages in R.

Figure 3 shows the RMSE of the estimated bivariate reference region for both scenarios to decrease with increasing sample size. However, the estimate error is clearly higher than in the first scenario. Nonetheless, the proposed model provides a good approximation to the shape of the theoretical region (see Figure 2), and the percentage of bivariate data points contained within the estimated region is close to the nominal level in both scenarios, becoming better as the sample size increases (see Table 1). These results suggest that the proposed model is quite robust in the face of possible miss-specifications.

4 | AGE-SPECIFIC BIVARIATE REFERENCE REGION FOR THE GLYCEMIC TESTS

4.1 | Motivating database

The A-Estrada Glycation and Inflammation Study (AEGIS) is a cross-sectional, population-based study that was performed in the municipality of A Estrada (Galicia, NW Spain). The study objective was to investigate the association between glycation, inflammation status, lifestyle, and common diseases, and to investigate any discordance between glycemic marker results.[35] An age-stratified random sample of the population aged ≥18 years was drawn from Spain's National Health System Registry. From November 2012 until March 2015, all subjects were successively convened for one day at the A Estrada Primary Care Centre for an evaluation which comprised fasting venous blood sampling, questionnaire interviews, and the description of subjects' lifestyles. FPG was determined in subjects' plasma samples using

TABLE 1 Percentage coverage of the estimated conditional bivariate region for different sample sizes (500, 1000, and 2000), and nominal levels (5%, 50%, 90% and, 95%) in the three simulation scenarios

				X_1										
	Sample size	**Nominal**	**Global**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Scenario 1	n = 500	5	4.6	4.4	4.8	4.7	4.7	4.8	4.7	4.6	4.6	4.7	4.7	4.3
		50	47.9	46.7	48.5	48.3	48.3	48.6	48.6	48.4	48.4	48.2	48.0	45.3
		90	88.3	87.3	88.8	88.6	88.7	88.8	88.8	88.7	88.7	88.4	88.2	86.0
		95	93.6	92.9	93.9	93.9	94.0	93.9	94.0	93.9	93.9	93.7	93.5	91.9
	n = 1000	5	4.7	4.7	4.8	4.8	4.7	4.7	4.6	4.7	4.6	4.6	4.7	4.3
		50	48.8	49.1	49.3	49.4	49.1	49.0	49.0	48.8	48.7	48.5	48.6	46.9
		90	88.9	89.2	89.2	89.2	89.1	89.1	89.2	88.9	89.0	88.8	88.6	87.6
		95	94.1	94.4	94.3	94.4	94.3	94.3	94.3	94.1	94.1	94.0	93.9	93.2
	n = 2000	5	4.7	4.7	4.8	4.8	4.7	4.8	4.7	4.7	4.7	4.7	4.8	4.6
		50	49.2	48.9	49.7	49.4	49.3	49.4	49.5	49.3	49.3	49.1	49.1	48.0
		90	89.2	89.1	89.3	89.5	89.3	89.4	89.5	89.3	89.2	89.4	89.1	88.4
		95	94.3	94.3	94.3	94.6	94.5	94.4	94.5	94.4	94.4	94.4	94.2	93.7
Scenario 2	n = 500	5	4.6	4.3	4.6	4.6	4.7	4.7	4.8	4.7	4.6	4.7	4.6	4.0
		50	47.6	45.4	47.9	48.3	48.6	48.8	48.7	48.4	48.1	48.0	47.9	43.7
		90	88.7	87.0	88.8	89.1	89.6	89.6	89.7	89.5	89.3	89.0	88.7	85.7
		95	94.1	92.9	94.2	94.4	94.7	94.7	94.8	94.7	94.6	94.4	94.0	92.0
	n = 1000	5	4.5	4.4	4.6	4.5	4.6	4.6	4.5	4.6	4.5	4.5	4.4	4.3
		50	47.0	46.1	46.9	47.1	47.3	47.7	47.4	47.3	47.5	47.3	47.0	45.9
		90	88.5	87.3	88.1	88.4	88.7	88.9	88.8	89.0	88.9	88.8	88.8	88.3
		95	94.0	93.0	93.6	93.8	94.1	94.2	94.2	94.4	94.3	94.3	94.2	93.9
	n = 2000	5	4.6	4.4	4.7	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.7	4.2
		50	48.3	47.0	48.2	48.6	48.8	49.0	49.0	48.6	48.7	48.7	48.8	46.0
		90	89.3	88.2	89.2	89.6	89.6	89.8	89.8	89.7	89.6	89.5	89.3	87.7
		95	94.5	93.7	94.4	94.7	94.7	94.9	94.9	94.9	94.7	94.6	94.6	93.5
Scenario 3	n = 500	5	4.6	4.4	4.7	4.7	4.8	4.7	4.7	4.7	4.6	4.6	4.5	3.9
		50	47.0	44.9	47.2	47.7	48.2	48.4	48.2	47.9	47.7	47.5	46.9	42.3
		90	88.3	86.1	87.9	88.7	89.2	89.4	89.4	89.3	89.1	88.7	88.3	85.6
		95	93.8	92.1	93.4	94.0	94.4	94.5	94.5	94.5	94.4	94.1	93.8	92.1
	n = 1000	5	4.5	4.4	4.6	4.5	4.6	4.6	4.5	4.6	4.5	4.5	4.4	4.3
		50	47.0	46.1	46.9	47.1	47.3	47.7	47.4	47.3	47.5	47.3	47.0	45.9
		90	88.5	87.3	88.1	88.4	88.7	88.9	88.8	89.0	88.9	88.8	88.8	88.3
		95	94.0	93.0	93.6	93.8	94.1	94.2	94.2	94.4	94.3	94.3	94.2	93.9
	n = 2000	5	4.9	4.4	4.6	4.5	4.6	4.6	4.5	4.6	4.5	4.5	4.4	4.3
		50	49.5	46.1	46.9	47.1	47.3	47.7	47.4	47.3	47.5	47.3	47.0	45.9
		90	89.6	87.3	88.1	88.4	88.6	88.9	88.8	89.0	88.9	88.8	88.8	88.3
		95	94.6	93.0	93.6	93.8	94.1	94.2	94.2	94.2	94.3	94.3	94.2	93.9

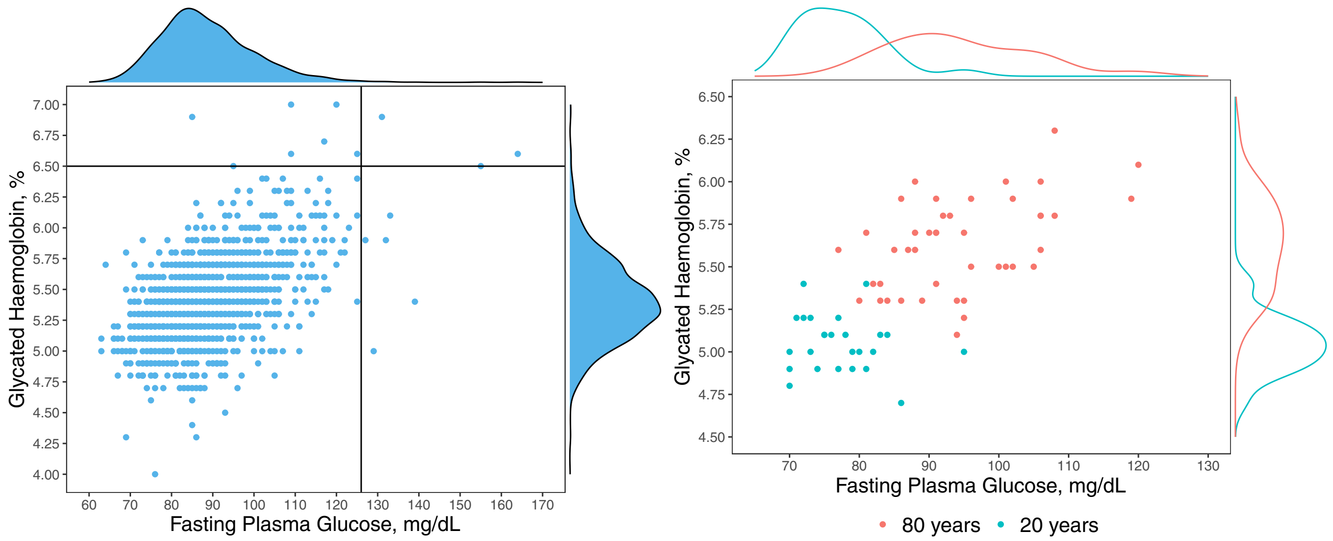

FIGURE 4 In the left plot an scatter plot and univariate density estimations of the glycemic test results for subjects not previously diagnosed with diabetes is given, the black lines represent the current diagnostic criteria. In the right plot a comparison of the test results between the younger (20 years) and older patients (80 years) is depicted [Colour figure can be viewed at wileyonlinelibrary.com]

the glucose oxidase-peroxidase method. HbA1c was determined by high performance liquid chromatography using a Menarini Diagnosticcs HA-8160 analyzer; all HbA1c values were converted to DCCT-aligned values.[36]

A total of 1516 subjects (55% female) agreed to participate in the study; their mean age was 52 years, (range 18 to 91). Among them, 187 (12%) had been diagnosed with diabetes, and among these 66.8% took oral antidiabetics, 3.7% took insulin alone, and 13.3% took insulin and oral drugs. The remaining 16.2% took none of these medications.

Figure 4 shows the glycemic marker concentrations for the subjects along with the current diagnosis cut-off points. As can be seen, according to the criteria in current use, a physician may encounter discordant FPG and HbA1c results, that is, i) an FPG value outside its reference interval but the HbA1c concentration within the normal range, ii) an HbA1c value outside its reference interval but the FPG level inside the normal range, or iii) both values inside their reference intervals but representing an unlikely combination. To the best of our knowledge, there is no criterion for interpreting such results. Moreover, the use of the same diagnosis cut-off points for the younger and older patients seem unreasonable since the mean values, the variability in the results, and the correlation between the glycemic markers, increased in older, diabetes-free subjects.

4.2 | Conditional bivariate reference region estimation

This section examines the 1329 AEGIS subjects with no diagnosis of diabetes. The age-specific bivariate region containing 95% of these subjects was estimated using the formula below:

$$\begin{pmatrix} FPG \\ HbA1c \end{pmatrix} = \begin{pmatrix} \mu_1(age) \\ \mu_2(age) \end{pmatrix} + \begin{pmatrix} \sigma_1^2(age) & \sigma_{12}(age) \\ \sigma_{12}(age) & \sigma_2^2(age) \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}, \tag{14}$$

where the effect of age on expectations, the variance, and the correlation between the markers was estimated using polynomial kernel smoothers to account for possible nonlinear trends. As shown in Figure 5, the mean HbA1c and FPG levels, and their variability, increase with age, while the strength of their correlation appears not to change.

Figure 6 shows the bivariate reference region displayed in the standardized residuals scale, after adjusting for age, including approximately 95% of the disease-free subjects. From a clinical point of view, these subjects would have “normal” values for both glycemic tests taking into account their age. The other 5% of the participants might be classified in four different groups: (I) first quadrant, individuals with high values for both tests; (II) second quadrant, discordant individuals with high HbA1c concentrations and low/medium FPG; (III) third quadrant, individuals with low values for

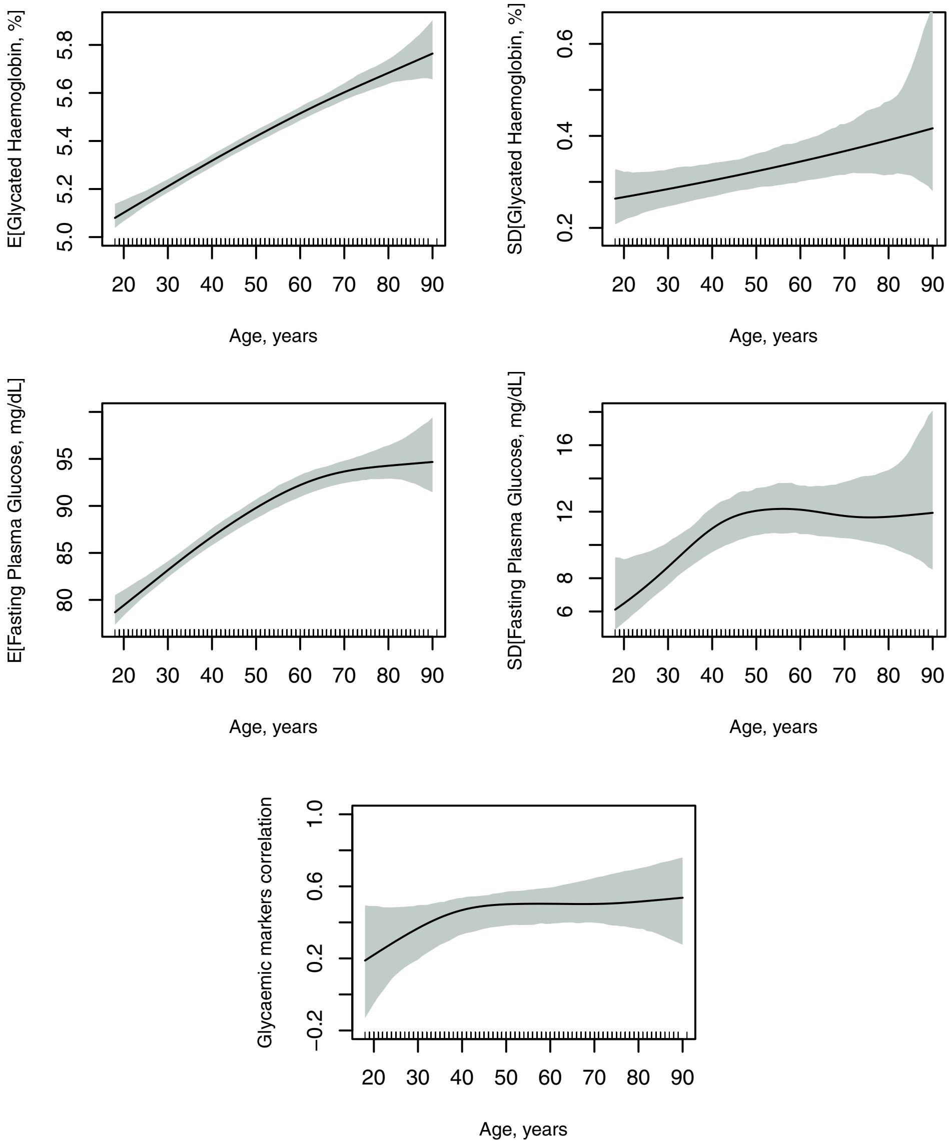

FIGURE 5 Estimated effect of age (black) and the corresponding 95% confidence interval (gray) on the mean (E) and standard deviation (SD) of both glycemic markers and the correlation between them [Colour figure can be viewed at wileyonlinelibrary.com]

both tests; and (IV) fourth quadrant, individuals with low/medium HbA1c concentrations and high FPG values. These results may have clinical implications, especially for subjects with both markers showing high values, but also for those with discordant results. Subjects returning high values for both (first quadrant) very likely have undiagnosed diabetes. Individuals who fall outside of the reference region in the second quadrant could be labeled as high glycators, that is, people with normal glucose values but who are at higher risk of cardiovascular disease[37,38] because of their glycation rate. In contrast, individuals in the fourth quadrant, who show high glucose levels but normal glycated hemoglobin levels, could be labeled as low glycators.

In Figure 7 the bivariate reference region is depicted for several ages. These regions shift toward the upper right corner and expand as age increase. This agrees with the nonlinear effects of age on the expected means and variability of both markers (Figure 5). This may also have clinical implications. For instance, a subject older than 40 years of age with FPG = 100 mg/dL and HbA1c = 6.0% should be considered diabetes-free, while a younger subject with the same levels of both markers should be considered to have glycemic dysregulation.

The performance of the proposed reference region was evaluated in terms of coverage for different τ using a leave-one-out cross-validation scheme and three age groups. Table 2 shows the estimated region to have a coverage close to the nominal level for every τ and age group considered.

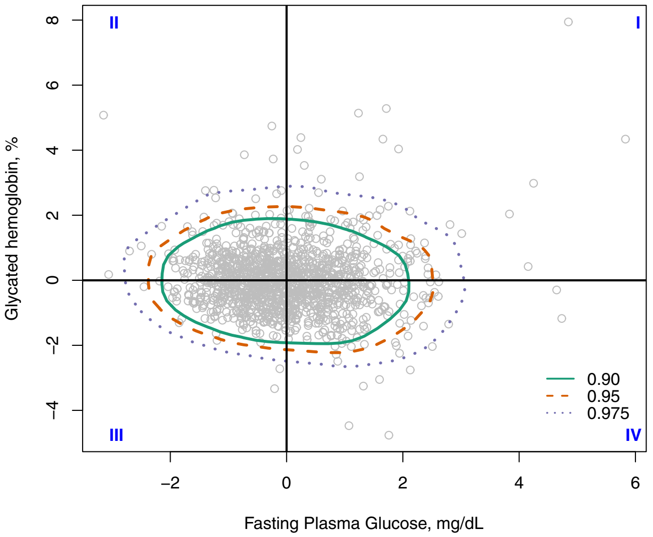

FIGURE 6 Bivariate reference region for the glycemic markers (FPG, HbA1c) using standarized bivariate residuals for $\tau = 0.90, 0.95, 0.975$ [Colour figure can be viewed at wileyonlinelibrary.com]

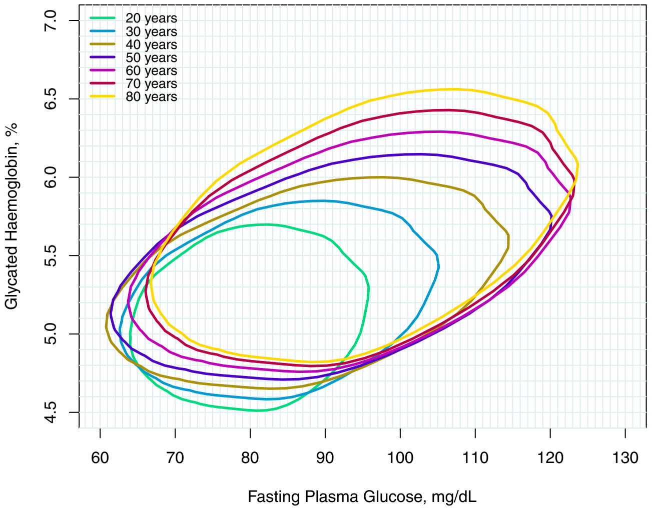

FIGURE 7 Reference region ($\tau = 0.95$) for age decades [Colour figure can be viewed at wileyonlinelibrary.com]

5 | DISCUSSION

This work proposes a new means of estimating reference regions under the assumption of nonparametric conditions, that can be used to help diagnose and treat patients with diseases for which the results of two different tests are considered. The proposed method overcomes the problem of the Gaussian distribution restriction of previously introduced multivariate reference regions. Moreover, it can estimate nonlinear effects of continuous covariates using polynomial kernel smoothers. In simulation studies, it was shown that the procedure for estimating the conditional bivariate reference region was efficient, even for datasets with complex bivariate response distributions.

The use of the proposed model revealed that the two biomarkers routinely used in diabetes screening and control (FPG and HbA1c) are better interpreted jointly. Patient age should be taken into account in all interpretations. Disagreements between the measured concentrations of different biomarkers can hinder decision-making when they are so-examined, but they may also provide insight into the future progress of the disease.[39-42]

Multiple test diagnosis accuracy might be investigated under ROC curves perspective, pondering both their joint specificity and sensitivity.[43] However, in our diabetes research application, reference region diagnosis sensitivity was not

TABLE 2 Percentage of test results contained within the estimated bivariate reference regions for different nominal levels

		Cross-validation evaluation			
Nominal	Apparent	Global	(18, 40]	(40, 60]	(60, 91]
5	4.9	4.9	5.7	5.1	4.0
50	50.0	49.7	50.9	49.6	48.5
90	90.0	89.7	91.6	87.5	90.2
95	95.0	94.1	95.0	93.1	94.3
97	97.1	96.7	97.1	96.2	96.9
98	97.8	97.6	97.6	97.7	97.6
99	99.0	98.2	98.1	98.1	98.5

Note: Cross-validation evaluation refers to a leave-one-out cross validation evaluation. Coverage probability is presented for the entire dataset (Global) and for three age groups.

formally evaluated. Preliminary results, with AEGIS's patients previously diagnosed with diabetes, indicates a 92% diagnosis accuracy, with an estimated sensitivity equal to 72% (data not shown). However, discussing these results are beyond the scope of this article. In addition, we cannot provide a clear interpretation of the estimated sensitivity, because most patients suffering from diabetes receive antidiabetic treatments. Indeed, future research from a purely clinical point of view must be conducted in order to answer questions arising from our novelty interpretation of the (FPG, HbA1c) values. Some of these open questions are i) how many patients will be classified as healthy, based on the FPG and HbA1c independent interpretation, but diseased based on their bivariate distribution? ii) which diabetes complications are more likely among patients with FPG and HbA1c disagreements? iii) are the terms "low glycator" and "high glycator" applicable after our proposal? or iv) which exogenous glycemic control measures maintain FPG and HbA1c in balance, and at the same time close to healthy patients' results?

Screening for, and the control of, diabetes mellitus is based largely on the results for two biomarkers, but in other diseases three or more biomarkers may be taken into consideration. For example, thyroid dysfunction is assessed by measuring blood concentrations of thyroid stimulating hormone, tri-iodothyronine (T3) and tetraiodothyronine (T4). Currently, the results for each test are compared with their respective univariate reference intervals. Bivariate and trivariate reference regions for a thyroid-healthy control group and for a sample of patients was previously reported,[44] comparing their diagnostic efficiency with the standard assessment method. However, these authors applied the current definition of reference regions that required multivariate Gaussianity, and the variation of thyroid hormone concentrations caused by a number of biological factors[45] was not taken into account. Other clinical studies have estimated trivariate reference regions based on the Gaussian distribution for cancer[46] and cardiovascular disease.[47] Given the need for reference regions for more than two tests, it would be of interest to extend the proposed model beyond the bivariate case. The lack of parametric restrictions in the proposed model, and the possibility of estimating the nonlinear effects of the predictor variables would be advantageous in the clinical setting.

The proposed model suffers the limitation that the correlation between the response variables is measured using a linear correlation coefficient. This is not optimal for non-Gaussian margins and nonelliptical dependence structures. Although in the simulation studies the proposed model was shown to be quite robust for this type of misspecification, a slightly increased error in estimates might be expected. More general correlation measures, such as those derived from copula functions,[48,49] might help overcome this problem.

ACKNOWLEDGEMENTS

Óscar Lado-Baleato is funded by a predoctoral grant (ED481A-2018) from the Galician Government (Plan I2C)-Xunta de Galicia. This research was also supported by grants from the Carlos III Health Institute, Spain (PI16/01404 and RD16/0017/0018), and by the project 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). Javier Roca-Pardiñas acknowledges financial support by the Grant MTM2017-89422-P (MINECO/AEI/FEDER, UE).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

Our data are freely available online. In the refreg R package (https://CRAN.R-project.org/package=refreg) the reader can find a dataset called AEGIS which contains the glycemic markers data.

ORCID

Óscar Lado-Baleato https://orcid.org/0000-0001-9592-4879
Javier Roca-Pardiñas https://orcid.org/0000-0003-3107-4515

REFERENCES

1. Hallworth MJ. The '70% claim': what is the evidence base? *Ann Clin Biochem.* 2011;48(6):487-488.
2. Siest G, Henny J, Gräsbeck R, et al. The theory of reference values: an unfinished symphony. *Clin Chem Lab Med.* 2013;51(1):47-64.
3. Wright EM, Royston P. Calculating reference intervals for laboratory measurements. *Stat Method Med Res.* 1999;8(2):93-112.
4. Koenker R, Bassett G Jr. Regression quantiles. *Econometrica.* 1978;46(1):33-50.
5. Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalized likelihood. *Stat Med.* 1992;11(10):1305-1319.
6. American Diabetes Association. Classification and diagnosis of diabetes: standards of medical care in diabetes-2019. *Diabetes Care.* 2019;42(Suppl 1):13-27.
7. Sacks DB. A1C versus glucose testing: a comparison. *Diabetes Care.* 2011;34(2):518-523.
8. Aleyassine H, Gardiner R, Tonks D, Koch P. Glycosylated hemoglobin in diabetes mellitus: Correlations with fasting plasma glucose, serum lipids and glycosuria. *Diabetes Care.* 1980;3(4):508-514.
9. Van't Riet E, Alssema M, Rijkelijkhuizen JM, Kostense PJ, Nijpels G, Dekker JM. Relationship between A1C and glucose levels in the general Dutch population: the new Hoorn study. *Diabetes Care.* 2010;33(1):61-66.
10. Ramachandran A, Riddle MC, Kabali C, Gerstein HC, ORIGIN Investigators. Relationship between A1C and fasting plasma glucose in dysglycemia or type 2 diabetes: an analysis of baseline data from the ORIGIN trial. *Diabetes Care.* 2012;35(4):749-753.
11. Boyd JC, Lacher DA. The multivariate reference range: an alternative interpretation of multi-test profiles. *Clin Chem.* 1982;28(2):259-265.
12. Harris EK, Yasaka T, Horton MR, Shakarji G. Comparing multivariate and univariate subject-specific reference regions for blood constituents in healthy persons. *Clin Chem.* 1982;28(3):422-426.
13. Slotnick H, Etzell P. Multivariate interpretation of laboratory tests used in monitoring patients. *Clin Chem.* 1990;36(5):748-751.
14. Dong X, Mathew T. Central tolerance regions and reference regions for multivariate normal populations. *J Multivar Anal.* 2015;134:50-60.
15. 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. *Metab Clin Exp.* 1979;28(6):688-705.
16. 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.
17. Espasandín-Domínguez J, Cadarso-Suárez C, Kneib T, et al. Assessing the relationship between markers of glycemic control through flexible copula regression models. *Stat Med.* 2019;38(27):5161-5181.
18. Serfling R. Quantile functions for multivariate analysis: approaches and applications. *Stat Neerl.* 2002;56(2):214-232.
19. Chaudhuri P. On a geometric notion of quantiles for multivariate data. *J Am Stat Assoc.* 1996;91(434):862-872.
20. Chakraborty B. On affine equivariant multivariate quantiles. *Ann Inst Stat Math.* 2001;53(2):380-403.
21. Hallin M, Paindaveine D, Siman M. Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth [with Discussion and Rejoinder]. *Ann Stat.* 2010;38(2):635-703.
22. Geraci M, Boghossian NS, Farcomeni A, Horbar JD. Quantile contours and allometric modelling for risk classification of abnormal ratios with an application to asymmetric growth-restriction in preterm infants. *Stat Methods Med Res.* 2019;29(7):1769-1786.
23. Wei Y. An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *J Am Stat Assoc.* 2008;103(481):397-409.
24. Petersen JH. A non-parametric conditional bivariate reference region with an application to height/weight measurements on normal girls. *Biom J.* 2009;51(4):697-709.
25. Patton AJ. Modelling asymmetric exchange rate dependence. *Int Econ Rev.* 2006;47(2):527-556.
26. Stander J, Dalla Valle L, Taglioni C, Liseo B, Wade A, Cortina-Borja M. Analysis of pediatric visual acuity using Bayesian copula models with sinh-arcsinh marginal densities. *Stat Med.* 2019;38(18):3421-3443.
27. Sklar A. Random variables, joint distribution functions, and copulas. *Kybernetika.* 1973;9(6):449-460.
28. Klein N, Kneib T. Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Stat Comput.* 2016;26(4):841-860.

29. Martínez-Silva I, Roca-Pardiñas J, Ordóñez C. Forecasting SO_2 pollution incidents by means of quantile curves based on additive models. *Environmetrics.* 2016;27(3):147-157.
30. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc Series B Stat Methodol.* 1991;53(3):683-690.
31. Bowman AW. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika.* 1984;71(2):353-360.
32. Johnson NL, Kotz S, Balakrishnan N. *Continuous Univariate Distributions*. 2nd ed. New York, NY: Wiley; 1995.
33. Stasinopoulos DM, Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. *J Stat Softw.* 2007;23(7):1-46.
34. Kojadinovic I, Yan J. Modeling multivariate distributions with continuous margins using the copula R package. *J Stat Softw.* 2010;34(9):1-20.
35. 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.
36. Hoelzel W, Weykamp C, Jeppsson JO, et al. IFCC reference system for measurement of hemoglobin A1c in human blood and the national standardization schemes in the United States, Japan, and Sweden: a method-comparison study. *Clin Chem.* 2004;50(1):166-174.
37. McCarter RJ, Hempe JM, Gomez R, Chalew SA. Biological variation in HbA1c predicts risk of retinopathy and nephropathy in type 1 diabetes. *Diabetes Care.* 2004;27(6):1259-1264.
38. Cohen RM. A1C: does one size fit all? *Diabetes Care*. 2007;30(10):2756-2758.
39. Rodríguez-Segade S, Rodríguez J, Cabezas-Agricola JM, Casanueva FF, Camina F. Progression of nephropathy in type 2 diabetes: the glycation gap is a significant predictor after adjustment for glycohemoglobin (HbA1c). *Clin Chem.* 2011;57(2):264-271.
40. Nayak AU, Nevill AM, Bassett P, Singh BM. Association of glycation gap with mortality and vascular complications in diabetes. *Diabetes Care.* 2013;36(10):3247-3253.
41. Kim MK, Jeong JS, Yun JS, et al. Hemoglobin glycation index predicts cardiovascular disease in people with type 2 diabetes mellitus: a 10-year longitudinal cohort study. *J Diabetes Complicat.* 2018;32(10):906-910.
42. Nayak AU, Singh BM, Dunmore SJ. Potential clinical error arising from use of HbA1c in diabetes: effects of the Glycation Gap. *Endocr Rev.* 2019;40(4):988-999.
43. Tang LL, Zhou XH. A semiparametric separation curve approach for comparing correlated ROC data from multiple markers. *J. Comput. Graph. Stat.* 2012;21(3):662-676.
44. Hoermann R, Larisch R, Dietrich JW, Midgley JE. Derivation of a multivariate reference range for pituitary thyrotropin and thyroid hormones: diagnostic efficiency compared with conventional single-reference method. *Eur J Endocrinol.* 2016;174(6):735-743.
45. Jonklaas J, Razvi S. Reference intervals in the diagnosis of thyroid dysfunction: treating patients not numbers. *Lancet Diabetes Endocrinol.* 2019;7(6):473-483.
46. Mattsson A, Svensson D, Schuett B, Osterziel KJ, Ranke MB. Multidimensional reference regions for IGF-I, IGFBP-2 and IGFBP-3 concentrations in serum of healthy adults. *Growth Horm IGF Res.* 2008;18(6):506-516.
47. Selmeryd J, Henriksen E, Dalen H, Hedberg P. Derivation and evaluation of age-specific multivariate reference regions to aid in identification of abnormal filling patterns: the HUNT and VaMIS studies. *JACC Cardiovasc Imaging.* 2018;11(3):400-408.
48. Gijbels I, Veraberbeke N, Omelka M. Conditional copulas, association measures and their applications. *Comput Stat Data Anal.* 2011;55(5):1919-1932.
49. Veraverbeke N, Omelka M, Gijbels I. Estimation of a conditional copula and association measures. *Scand J Stat.* 2011;38(4):766-780.
50. Hastie T, Tibshirani R. *Generalized Additive Models*. London, UK: Chapman & Hall; 1990:175-186.
51. Wood S. *Generalized Additive Models: An Introduction with R*. 2nd ed. London, UK: Chapman & Hall/CRC; 2017.
52. Lado-Baleato O, Roca-Pardiñas J, Cadarso-Suárez C, Gude F. *refreg: Conditional Multivariate Reference Regions*. R package version 0.1-0. 2021.

How to cite this article: Lado-Baleato Ó, Roca-Pardiñas J, Cadarso-Suárez C, Gude F. Modeling conditional reference regions: Application to glycemic markers. *Statistics in Medicine*. 2021;40(26):5926-5946. https://doi.org/10.1002/sim.9163

APPENDIX A. FLEXIBLE ADDITIVE MODELS ESTIMATION

In order to obtain the estimated additive models in Equations (5), (6), and (7), we have used a backfitting algorithm based on local polynomial kernel smoothers. For mathematical notation simplicity, we denote in this section Y as our response variable, and $\boldsymbol{X} = (X_1, \dots, X_p)$ the p vector of covariates. In this regression framework, we consider the transformed additive model:

$$E[Y|\boldsymbol{X}] = G\left(\alpha + \sum_{j=1}^{p} f_j(X_j)\right),$$

where $G(\cdot)$ is a known link function, α is a constant, and f_j unknown functions representing the effects of continuous covariates. Given a sample $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^{n}$, this model can be estimated using the following iterative process based on a Newton Raphson procedure which extends the ACE (Alternating Conditional Expectation) algorithm.[50]

Initialize: compute the initial estimates, $\hat{\alpha} = G^{-1}(\overline{Y})$ with $\overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i, \hat{f}_1^0, \dots, \hat{f}_p^0 = 0$.

Step 1: for $i = 1, \dots, n$ construct the linearized response $\tilde{Y}$ and the weights W so that:

$$\tilde{Y}_i = \hat{\eta}_i^0 + \frac{Y_i - G(\hat{\eta}_i^0)}{G'(\hat{\eta}_i^0)} \quad \text{and} \quad W_i = \frac{G'(\hat{\eta}_i^0)^2}{\hat{\sigma}_i^2},$$

where $\hat{\eta}_i^0 = \hat{\alpha} + \sum_{j=1}^{p} \hat{f}_j^0(X_j)$, $G'(\eta) = \frac{\delta G}{\delta \eta}$, and $\hat{\sigma}_i^2$ is an estimation of the variance $\sigma^2(Y_i|G(\hat{\eta}_i^0))$. The estimated $\hat{\sigma}_i^2$ can be obtained fitting an additive model to $(Y_i - G(\hat{\eta}_i^0))^2$.

Step 2: fit an additive model to $\tilde{Y}$ weighted by W and compute the updates $\hat{\alpha}$ and $\hat{f}_j$ for $j = 1, \dots, p$. At this step we have used an inner backfitting algorithm based on a local polynomial kernel smoother:

Step 2.0: update the constant $\hat{\alpha} = \left(\sum_{i=1}^{n} W_i\right)^{-1} \sum_{i=1}^{n} W_i \tilde{Y}$

Step 2.1: for $j = 1, \dots, p$ calculate the partial residuals

$$R_i^j = \tilde{Y}_i - \hat{\alpha} - \sum_{k=1}^{j-1} \hat{f}_k\,(X_{ik}) - \sum_{k=j+1}^{p} \hat{f}_k^0(X_{ik})$$

and for $i = 1, \dots, n$, compute the polynomial kernel estimator updates:

$$\hat{f}_j(X_{ij}) = \hat{\psi}\left(X_{ij}, \left\{(X_{lj}, R_l^j, W_i)\right\}_{l=1}^{n}, h_j^f\right)$$

being h_j^f the smoothing bandwidth associated with the estimation of f_j.

Step 2.2: Repeat **Step 2.1** replacing $\hat{f}_j^0(X_{ij})$ by $\hat{f}_j(X_{ij})$ for $j = 1, \dots, p$ and $i = 1, \dots, n$, until the convergence criterion

$$\frac{\sum_{i=1}^{n}\left(\hat{f}_j(X_{ij}) - \hat{f}_j^0(X_{ij})\right)^2}{\sum_{i=1}^{n}\left(\hat{f}_j^0(X_{ij})\right)^2 + 0.001} \leq \varepsilon \quad \text{for all } j = 1, \dots, p$$

is reached.

Step 3: repeat the **Steps 1** and **2** with $\hat{\eta}_i^0$ being replaced by $\hat{\eta}_i = \hat{\alpha} + \sum_{j=1}^{p} \hat{f}_j\left(X_{ij}\right)$ for $i = 1, \dots, n$ until:

$$\frac{|MSE(\hat{\eta}, Y) - MSE(\hat{\eta}_0, Y)|}{MSE(\hat{\eta}_0, Y)} \leq \epsilon,$$

where ϵ is a small and the mean squared error $MSE(\hat{\eta}, Y)$ is defined as $MSE(\hat{\eta}, Y) = n^{-1}\sum_{i=1}^{n} W_i(Y_i - G(\hat{\eta}_i))^2$

The proposed algorithm use two loops: (i) an external loop, for adjusting the transformed response models (Steps 1 and 3), (ii) an internal loop which estimates the nonlinear effect of continuous covariates using a backfitting method (Step 2). This algorithm was used to estimate the additive models for means (Equation (5)) with identity link , the additive models for variances (Equation (6)) with exponential link and the correlation additive model (Equation (7)) with $\tanh(\cdot)$ link. Note that in the first case the algorithm is reduced to the internal loop. Finally, the polynomial kernel smoother used in Step 2 may be replaced by an alternative estimator (eg, penalized splines).[51]

APPENDIX B. LOCAL POLYNOMIAL KERNEL SMOOTHERS

Given a sample $\{(X_i, Y_i)\}_{i=1}^{n}$ with a vector of weights $\{W_i\}_{i=1}^{n}$ the local linear kernel smoother at a location x, $\hat{\psi}(x) = \hat{\psi}\left(x, \{(X_i, Y_i, W_i)\}_{l=1}^{n}, h\right)$ is defined as $\hat{\psi}(x) = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_q)$ is a vector which minimizes:

$$\sum_{i=1}^{n} W_i \left[Y_i - \sum_{j=0}^{q} \beta_j (X_i - x)^j \right]^2 K\left(\frac{X_i - x}{h} \right),$$

where $K(\cdot)$ denotes a kernel function (a symmetric density), q is the polynomial degree, and $h > 0$ is the smoothing parameter, chosen using:

$$CV = \frac{1}{n} \sum_{i=1}^{n} W_i \left(Y_i - \hat{\psi}^{(-i)}(X_i) \right)^2,$$

where $\hat{\psi}^{(-i)}(X_i)$ indicates the fit at X_i leaving out the ith data vector.

APPENDIX C. BIVARIATE KERNEL BANDWIDTH SELECTION

In this section, the bivariate kernel bandwidth estimator proposed in Equation (10) is evaluated and compared with the well-known plug-in, and least-square cross-validation estimators, and an arbitrary large bandwidth (5***H***). The influence of the kernel bandwidth on the shape of the final conditional reference region is determined by examining the RMSEs (see Equation (11)), and the region perimeter with respect to the theoretical regions. Moreover, the coverage properties of the estimated region is assessed using an out-sample design. This evaluation was performed under two scenarios:

- Scenario 1: involving a bivariate response simulated from the model structure (Equation (1)). The bivariate error $(\varepsilon_1, \varepsilon_2)$ was drawn from a bivariate standard Gaussian distribution.
- Scenario 2: involving a bivariate response simulated from a parametric copula representation (see Equation (13)), considering a reverse Gumbel and logistic distribution joined by a Gumbel copula.

In both scenarios the bivariate response parameters were made dependent on a continuous ($X_1 \in U[0, 1]$) and a binary regressor (X_2), specifically:

$$\begin{cases} \mu_1(\boldsymbol{X}) = 1 + X_2 + f_1(X_1) \\ \sigma_1^2(\boldsymbol{X}) = 1 + X_2 \\ \mu_2(\boldsymbol{X}) = 1 + 0.5X_2 + f_2(X_1) \\ \sigma_2^2(\boldsymbol{X}) = 1 + 0.5X_2 \\ \rho(\boldsymbol{X}) = 0.3 + 0.2X_1 + 0.3X_2 \end{cases},$$

where $f_1(X_1) = X_1 \sin(3X_1)$ and $f_2(X_1) = \sin(2\pi X_1)$ represent nonlinear effects. The sample size were set to $n = 500, 1000,$ and 2000 and the evaluation performed using 1000 replicates.

Figure C1 shows the main results for both simulation scenarios. With Gaussian data (scenario 1), the RMSE of the estimated bivariate reference region ($\tau = 0.95$) is higher for the plug-in, and cross-validation methods for every sample size, and continuous covariate value. Moreover, estimated regions' perimeters were higher than the real one for classical bandwidth estimators. Accordingly, a kernel bandwidth wider than plug-in would obtain a better estimate of the bivariate region. Given that the bivariate error is represented using a standard bivariate Gaussian distribution, which is also used as a kernel function in the density estimator, this result is not surprising. This also explains the good coverages obtained for a large kernel bandwidth (see Table C1). Hence, if the standardized bivariate residuals are Gaussian-distributed, a parametric expression would be better used in the estimation of R_τ. For non-Gaussian data (scenario 2), our proposed method obtains a smaller estimation error than that returned by the plug-in, and cross-validation estimators. Moreover, the use of a large bandwidth returns a worse estimation error. In addition, theoretical region perimeters are better approximated by our kernel bandwidth estimator. Figure C2 shows the estimated bivariate reference region for every kernel bandwidth. Plug-in, and cross-validation estimators returned greater variabilities in their estimates, while the large bandwidth ignored the true shape of the region. Moreover, the best coverages were obtained using the bandwidth selector described in Section 2.3 (see Table C1). Specifically, the plug-in, and cross-validation methods seems to overfit the training dataset, resulting in a coverage below 95% for the estimated reference region, while the large bandwidth (5H) offer a coverage of

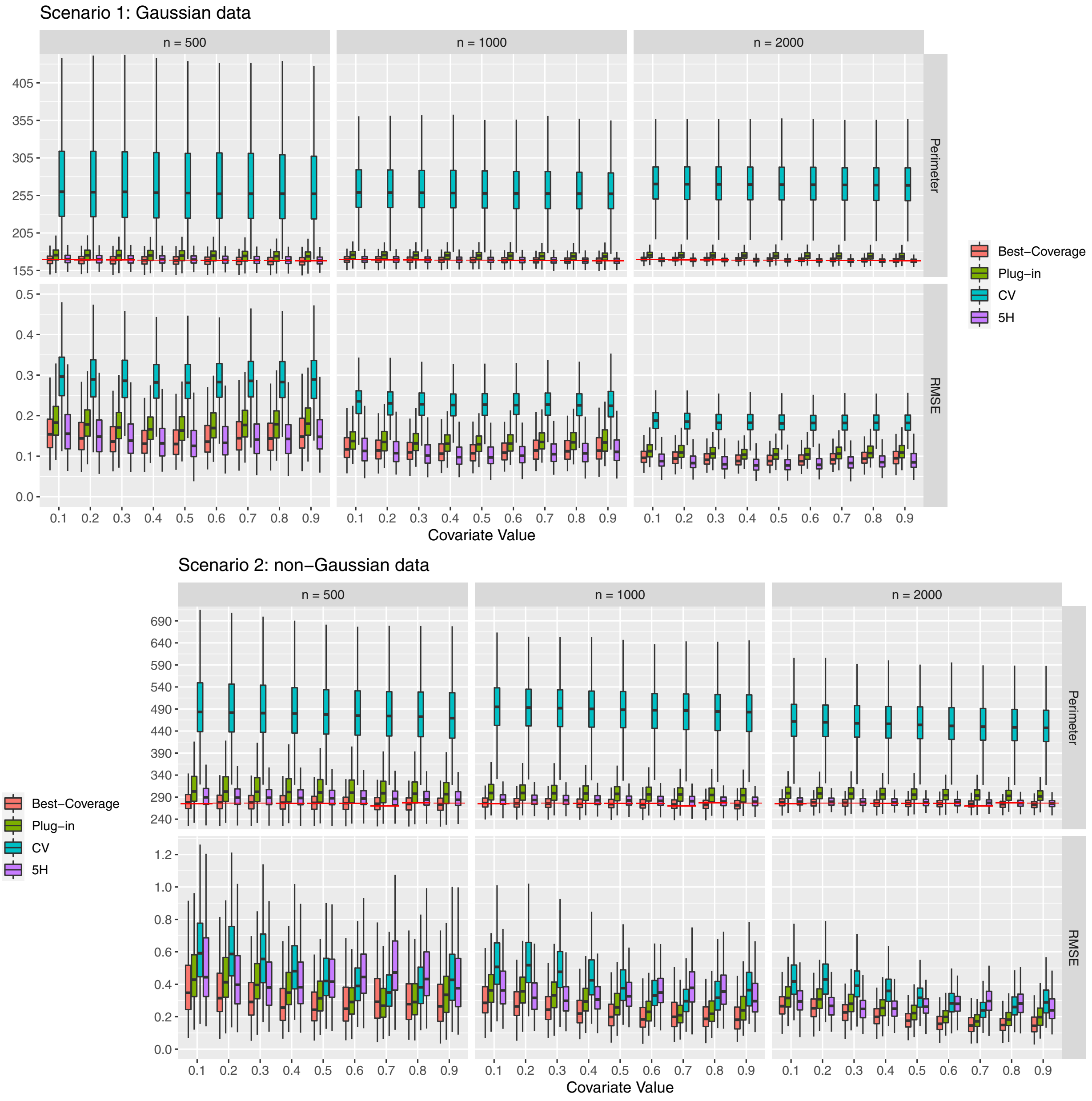

FIGURE C1 Bivariate reference region performance depending on kernel bandwidth estimator for different X_1 predictor variable values, with X_2 fixed at zero, sample sizes (500, 1000, and 2000), for Gaussian and non-Gaussian data. Red line represents the theoretical region's perimeter. Best-Coverage represents kernel bandwidth estimator proposed in Equation (10), CV is the least-square cross-validation method, and 5$\boldsymbol{H}$ an arbitrary large bandwidth [Colour figure can be viewed at wileyonlinelibrary.com]

over 95%. Hence, when dealing with nonstandard responses, the density estimator plays a key role in the performance of the bivariate reference region. The present method shows better performance than the plug-in, and cross-validation estimators, but the bandwidth selection problem requires further work.

APPENDIX D. BOOTSTRAP INFERENCE

In this section we present a bootstrap procedure to obtain punctual confidence intervals, given a specific vector of covariates $\mathbf{X}_0$, for the components (mean, deviation and correlation components) of the model presented in Equation (1). The steps for construction of the bootstrap confidence intervals are:

Step 1. From the sample data $\{(Y_{i1}, Y_{i2}), \boldsymbol{X}_i\}_{i=1}^{n}$ obtain the estimates $\hat{\mu}_r(\boldsymbol{X}_0)$, $\hat{\sigma}_r(\boldsymbol{X}_0)$ $(r = 1, 2)$ and $\hat{\rho}(\boldsymbol{X}_0)$.

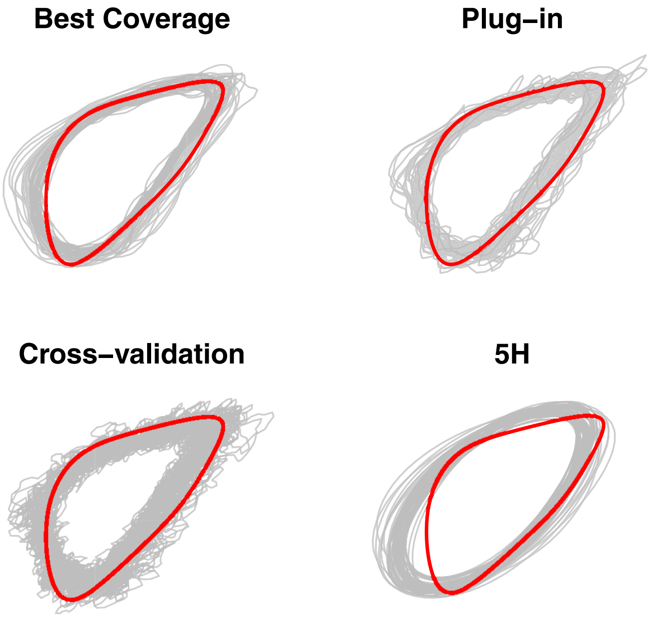

FIGURE C2 Estimation of the bivariate reference region for 50 replicates (gray) along with the theoretical region (red), for n = 1000, $X_1 = 0.5$ and $X_2 = 0$, and the bivariate kernel bandwidths used in the estimation of the model's residual density function. Best coverage represents our Equation (10) proposal, and 5$\boldsymbol{H}$ an arbitrary large bandwidth [Colour figure can be viewed at wileyonlinelibrary.com]

Step 2. For $b = 1, \dots, B$ generate bootstrap samples $\{(Y^{\bullet}_{i1}, Y^{\bullet}_{i2}), \boldsymbol{X}_i\}_{i=1}^{n}$ with

$$\begin{pmatrix} Y^{\bullet}_{i1} \\ Y^{\bullet}_{i2} \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1(\mathbf{X}_i) \\ \hat{\mu}_2(\mathbf{X}_i) \end{pmatrix} + \hat{\Sigma}^{1/2}(\mathbf{X}_i) \begin{pmatrix} \hat{\varepsilon}^{\bullet}_{i1} \\ \hat{\varepsilon}^{\bullet}_{i2} \end{pmatrix},$$

where $\left\{(\hat{\varepsilon}^{\bullet}_{i1}, \hat{\varepsilon}^{\bullet}_{i2})\right\}_{i=1}^{n}$ is a sample of size n from the residuals $\{(\hat{\varepsilon}_{i1}, \hat{\varepsilon}_{i2})\}_{i=1}^{n}$ with replacement, and compute $\hat{\mu}_r^{\bullet b}(\boldsymbol{X}_0)$, $\hat{\sigma}_r^{\bullet b}(\boldsymbol{X}_0)$ and $\hat{\rho}^{\bullet b}(\boldsymbol{X}_0)$ as in Step 1.

The limits for the $100(1-\alpha)\%$ confidence intervals of the true components $\mu_r(\boldsymbol{X}_0)$, $\sigma_r(\boldsymbol{X}_0)$, and $\rho(\boldsymbol{X}_0)$ are given respectively by $\left(\hat{\mu}_r^{\alpha/2}(\boldsymbol{X}_0), \hat{\mu}_r^{1-\alpha/2}(\boldsymbol{X}_0)\right)$, $\left(\hat{\sigma}_r^{\alpha/2}(\boldsymbol{X}_0), \hat{\sigma}_r^{1-\alpha/2}(\boldsymbol{X}_0)\right)$, and $\left(\hat{\rho}^{\alpha/2}(\boldsymbol{X}_0), \hat{\rho}^{1-\alpha/2}(\boldsymbol{X}_0)\right)$, where $\hat{\mu}_r^{p}(\boldsymbol{X}_0)$ represents the p-percentile of $\hat{\mu}_r^{\bullet 1}(\boldsymbol{X}_0), \dots, \hat{\mu}_r^{\bullet B}(\boldsymbol{X}_0)$, $\hat{\sigma}_r^{p}(\boldsymbol{X}_0)$ represents the p-percentile of $\hat{\sigma}_r^{\bullet 1}(\boldsymbol{X}_0), \dots, \hat{\sigma}_r^{\bullet B}(\boldsymbol{X}_0)$, and $\hat{\rho}^{p}(\boldsymbol{X}_0)$ is the p-percentile of $\hat{\rho}^{\bullet 1}(\boldsymbol{X}_0), \dots, \hat{\rho}^{\bullet B}(\boldsymbol{X}_0)$.

APPENDIX E. R CODE AND DATA ANALYSIS

The statistical methods developed in this article, and the AEGIS dataset, are available in the R package `refreg`.[52] In the following we present the R code necessary to replicate Section 4.2 results.

```
install.packages("refreg")
library(refreg)

#-- AEGIS dataset
?aegis
head(aegis)
summary(aegis)

dm_no = subset(aegis, aegis$dm=="no") # healthy patients sample

#-- Fitting bivariate location-scale model
mu1 = fpg ~ s(age)
```

TABLE C1 Coverage probability of bivariate data points for different sample sizes (500, 1000, and 2000), covariate X_1 values, with X_2 fixed at zero, and bivariate kernel bandwidths estimators

			X_1										
	Sample size	**Bandwidth**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Scenario 1	n = 500	Best coverage	93.0	93.9	93.9	94.0	94.0	94.0	93.8	93.7	93.6	93.4	91.7
		Plug-in	91.5	92.5	92.6	92.6	92.7	92.6	92.5	92.3	92.3	92.1	90.1
		CV	87.2	88.4	88.5	88.5	88.6	88.5	88.4	88.2	88.1	87.9	85.6
		5***H***	93.6	94.4	94.4	94.5	94.5	94.5	94.3	94.2	94.2	94.0	92.3
	n = 1000	Best coverage	93.8	94.3	94.3	94.3	94.3	94.3	94.2	94.1	94.1	94.0	93.2
		Plug-in	93.1	93.6	93.6	93.6	93.6	93.6	93.5	93.4	93.4	93.3	92.4
		CV	89.8	90.5	90.5	90.5	90.5	90.5	90.4	90.3	90.3	90.2	89.0
		5***H***	94.3	94.7	94.7	94.7	94.7	94.7	94.6	94.6	94.5	94.4	93.7
	n = 2000	Best coverage	94.2	94.5	94.5	94.5	94.5	94.4	94.4	94.4	94.3	94.3	93.8
		Plug-in	93.8	94.1	94.1	94.1	94.1	94.1	94.0	94.0	93.9	93.9	93.4
		CV	91.6	91.9	91.9	91.9	91.9	91.9	91.8	91.8	91.7	91.7	91.1
		5***H***	94.5	94.8	94.8	94.8	94.8	94.8	94.7	94.7	94.7	94.6	94.2
Scenario 2	n = 500	Best coverage	92.5	93.6	94.0	94.3	94.3	94.2	94.0	94.2	93.6	93.4	91.7
		Plug-in	89.9	91.1	91.8	92.0	92.2	92.2	92.0	92.1	91.5	91.1	88.4
		CV	86.1	87.5	88.3	88.6	88.8	89.0	88.7	88.7	88.2	87.6	84.2
		5***H***	94.0	94.8	95.2	95.4	95.3	95.2	94.9	95.1	94.5	94.4	93.5
	n = 1000	Best coverage	92.7	93.5	93.9	94.0	94.0	94.0	94.1	94.4	94.0	94.1	93.5
		Plug-in	91.2	92.1	92.5	92.7	92.9	92.9	93.0	93.4	92.9	92.9	91.6
		CV	87.4	88.5	89.1	89.3	89.5	89.6	89.7	89.9	89.6	89.5	87.6
		5***H***	94.2	95.0	95.2	95.2	95.1	95.0	95.0	95.4	95.0	95.1	95.3
	n = 2000	Best coverage	93.1	93.6	93.9	94.1	94.1	94.1	94.2	94.6	94.1	94.1	93.7
		Plug-in	92.2	92.7	93.0	93.3	93.3	93.5	93.5	93.9	93.4	93.4	92.8
		CV	90.0	90.5	91.0	91.2	91.3	91.5	91.5	91.9	91.4	91.4	90.6
		5***H***	94.4	94.9	95.1	95.2	95.1	95.1	95.0	95.4	94.9	95.0	95.1

Note: Best coverage represents our Equation (10) proposal, CV is the least-square cross-validation method, and 5***H*** an arbitrary large bandwidth.

```
mu2 = hba1c ~ s(age)
var1 = ~ s(age)
var2 = ~ s(age)
rho = ~ s(age)
formula = list(mu1,mu2,var1,var2,rho)

fit = bivRegr(formula,data=dm_no)
s_b = summary_boot(fit,B=250,parallel = T) # bootstrap 95% CI

plot(s_b,eq=1)
plot(s_b,eq=2)
plot(s_b,eq=3)
plot(s_b,eq=4)
plot(s_b,eq=5,select=1)+ggplot2::geom_hline(yintercept=0,color = "red")
```

```
#-- Bivariate reference region
region0 = bivRegion(fit,H_choice = "plug.in",tau=c(0.50,0.90,0.95))
plot(region0,tau=0.95,col="gray",reg.lwd=2) # overfitted region

set.seed(719)
region1 = bivRegion(fit,H_choice = "Hcov",tau=0.95,k=50) # best coverage region
plot(region1,tau=0.95,col="gray",reg.lwd=2.75, legend=T,xlab="Fasting Plasma
Glucose, mg/dL",
ylab="Glycated hemoglobin, %" # standarized region)
summary(region1) # identify outside patients

#-- Depict conditional region
plot(region1,cond=T,newdata=data.frame(age=c(20,30,40,50,60,70,80)),col=NA,
reg.lwd=2)
```