
This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: http://openaccess.city.ac.uk/id/eprint/20922/

Link to published version: http://dx.doi.org/10.1016/j.csda.2018.05.001

Copyright and reuse: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.
Copula based generalized additive models for location, scale and shape with non-random sample selection

Małgorzata Wojtyś\textsuperscript{a}, Giampiero Marra\textsuperscript{b}, Rosalba Radice\textsuperscript{c}

\textsuperscript{a}Centre for Mathematical Sciences, University of Plymouth
\textsuperscript{b}Department of Statistical Science, University College London
\textsuperscript{c}Department of Economics, Mathematics and Statistics, Birkbeck

Abstract

Non-random sample selection is a commonplace amongst many empirical studies and it appears when an output variable of interest is available only for a restricted non-random sub-sample of data. An extension of the generalized additive models for location, scale and shape which accounts for non-random sample selection by introducing a selection equation is discussed. The proposed approach allows for potentially any parametric distribution for the outcome variable, any parametric link function for the selection equation, several dependence structures between the (outcome and selection) equations through the use of copulae, and various types of covariate effects. Using a special case of the proposed model, it is shown how the score equations are corrected for the bias deriving from non-random sample selection. Parameter estimation is carried out within a penalized likelihood based framework. The empirical effectiveness of the approach is demonstrated through a simulation study and a case study. The models can be easily employed via the \texttt{gjrm()} function in the \texttt{R} package \texttt{GJRM}.

\textit{Keywords:} additive predictor, copula, marginal distribution, non-random sample selection, penalized regression spline, simultaneous equation estimation.

1. Introduction

Non-random sample selection arises when an output variable of interest is available only for a restricted non-random sub-sample of the data. This often occurs in sociological, medical and economic studies where individuals systematically select themselves into (or out of) the sample (e.g., Lennox et al., 2012; Vella, 1998; Collier and Mahoney, 1996, and references therein). If the aim is to model an outcome of interest in the entire population and the link between its availability in the sub-sample and its observed values is through factors which can not be accounted for then any analysis based on the available sub-sample will most likely lead to biased conclusions. Sample selection models allow one to use the entire sample whether or not observations on the output variable were generated. In its classical form, it consists of two equations which model the probability of inclusion in the sample and the outcome variable through a set of available covariates, and of a joint bivariate distribution linking the two equations.

\hspace{1cm}Email addresses: małgorzata.wojtys@plymouth.ac.uk (Małgorzata Wojtyś), giampiero.marra@ucl.ac.uk (Giampiero Marra), r.radice@bbk.ac.uk (Rosalba Radice)
The sample selection model was first discussed by Gronau (1974) and Lewis (1974). Heckman (1976) formulated a unified approach to estimating this model using a simultaneous equation system. In the classical version, the error terms of the two equations are assumed to follow a bivariate normal distribution where non-zero correlation indicates the presence of non-random sample selection. Heckman (1979) then translated the issue of sample selection into an omitted variable problem and proposed a simple and easy to implement estimation method known as two-step procedure.

Various modifications and generalizations of the classical sample selection model have been proposed in the literature and here we mention some of them. A non-parametric two-stage approach, which lifts the normality assumption, can be found in Das et al. (2003). Non-parametric methods are also considered in Lee (2008) and Chen and Zhou (2010). Semi-parametric approaches can instead be found in Gallant and Nychka (1987), Lee (1994), Powell (1994) and Newey (2009). In the Bayesian framework, Chib et al. (2009) dealt with non-linear covariate effects using Markov chain Monte Carlo estimation techniques and simultaneous equation systems. Wiesenfarth and Kneib (2010) further extended this approach by introducing a Bayesian algorithm based on low rank penalized B-splines for non-linear and varying-coefficient effects and Markov random-field priors for spatial effects. Frequentist counterparts of these Bayesian methods are given in Marra and Radice (2013b) in the context of binary responses and Marra and Radice (2013a) for continuous Gaussian outcomes. Zhelonkin et al. (2016) introduced a procedure for robustifying the Heckman’s two stage estimator by using M-estimators of Mallows’ type for both stages. Marchenko and Genton (2012) and Ding (2014) considered a bivariate Student-t distribution for the model’s errors as a way of tackling heavy-tailed data. Several authors proposed using copulae to model the joint distribution of the selection and outcome equations; see, e.g., Prieger (2002) who employed a Farlie-Gumbel-Morgenstern (FGM) bivariate copula. A more general copula approach, with a focus on Archimedean copulae, can be found in Smith (2003). As emphasized for instance by Genius and Strazzera (2008), copulae allow for the use of non-Gaussian distributions and have the additional benefit of making it possible to specify the marginal distributions independently of the dependence structure linking them. Importantly, while the copula approach is fully parametric, it is typically computationally more feasible than non/semi-parametric approaches and it still allows one to assess the sensitivity of results to different modeling assumptions. The aim of this work is to continue this stream of research.

In this paper, we introduce a generalized additive model for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005) which accounts for non-random sample selection. First, the clas-
sical GAMLSS is extended by introducing an extra equation which models the selection process. Specifically, the selection and outcome equations are linked by a joint probability distribution which is expressed in terms of a copula. Moreover, using the developments available in the spline literature (e.g., Ruppert et al., 2003; Wood, 2017), we model the relationship between covariates and responses by using penalized regression splines, thus capturing possibly complex relationships. Second, we show how the score equations are corrected for the bias deriving from non-random sample selection. To the best of our knowledge, this aspect has never been elucidated in the literature and provides an interesting insight into the correction mechanism underlying the selection approach. Third, we make the new developments available via the \texttt{gjrm()} function from the \texttt{R} package \texttt{GJRM} (Marra and Radice, 2018).

Note that the approach to estimating sample selection models using copulae and penalized regression splines has recently been adopted by Wojtyś et al. (2016), Marra and Wyszynski (2016), Wyszynski and Marra (2017) and Marra et al. (2017). The former only considers a Gaussian outcome, whereas the latter works deal with the cases of binary and discrete outcome distributions. This paper is concerned with providing a general modeling framework where any parametric link function and continuous distribution for the outcome can be utilized.

The remainder of the paper is organized as follows. Section 2 discusses the proposed sample selection GAMLSS as well as a special case which elucidates the nature of the non-random sample selection correction. Section 3 provides some estimation and inferential details. The finite-sample performance of the approach is investigated in Section 4, whereas a case study is presented in Section 5.

2. Sample selection GAMLSS

The proposed generalized additive sample selection model for location, scale and shape is structured as follows. We first assume that the outcome variable of interest can be described by a GAMLSS (Rigby and Stasinopoulos, 2005). Then, in order to take the selection process into account, we extend the model by adding the so-called selection equation, which is specified in terms of a binary regression that makes use of an arbitrary parametric link function. The two equations are linked by using a bivariate copula. Finally, all parameters of the marginal distributions as well as copula are specified as flexible functions of covariates.

2.1. Model definition

Let $Y_2^*$ denote the random variable of primary interest whose values are observed only for a subset of the individuals from a random sample. Moreover, let $Y_1^*$ be the latent random variable
that governs the selection process. The observed variables are

\[ Y_1 = (Y_1^* > 0), \]
\[ Y_2 = Y_2^* Y_1, \]

where symbol \((\cdot)\) denotes throughout an indicator function. Random variable \(Y_1\) indicates whether the value of \(Y_2^*\) is observed. Variable \(Y_2\) holds the observed value of \(Y_2^*\) and equals 0 if the observed value is missing.

Let \(F_1(y_1^*|\theta_1)\) and \(F_2(y_2^*|\theta_2)\) denote the cumulative distribution functions of the latent selection variable \(Y_1^*\) and of the output variable of interest \(Y_2^*\), which depend on vectors of parameters \(\theta_1 \in \mathbb{R}^{p_1}\) and \(\theta_2 \in \mathbb{R}^{p_2}\), respectively, where \(p_1, p_2 \in \mathbb{N}\). Analogically, \(f_1(y_1^*|\theta_1)\) and \(f_2(y_2^*|\theta_2)\) denote the probability (density) functions of \(Y_1^*\) and \(Y_2^*\). We specify the dependence structure between the two variables by taking advantage of Sklar’s theorem (Sklar, 1959). It states that for any two random variables there exists a two-place function, called copula, which represents the joint cumulative distribution function of the pair in a manner which makes a clear distinction between the marginal distributions and the form of dependence between them. An exhaustive introduction to copula theory can be found in Nelsen (2006) and Schweizer (1991). We use the symbol \(C_{\theta_3}(\cdot, \cdot)\) throughout to denote a copula parametrized with \(\theta_3 \in \mathbb{R}^{p_3}\), where \(p_3 \in \mathbb{N}\).

Let \(C_{\theta_3}(\cdot, \cdot)\) be the copula such that the joint cdf of \((Y_1^*, Y_2^*)\) equals

\[ F(y_1^*, y_2^* | \theta_1, \theta_2, \theta_3) = C_{\theta_3} (F_1(y_1^*|\theta_1), F_2(y_2^*|\theta_2)) . \]  

(2.1)

Function \(C_{\theta_3}\) always exists and is unique for every \((y_1^*, y_2^*)\) in the support of the joint distribution \(F\). It is assumed that the unknown parameter vectors \(\theta_1 \in \mathbb{R}^{p_1}\), \(\theta_2 \in \mathbb{R}^{p_2}\) and \(\theta_3 \in \mathbb{R}^{p_3}\) can be linked to predictors (containing regression coefficients and covariates) via known monotonic link functions \(g_{1,j}(\cdot), g_{2,j}(\cdot)\) and \(g_{3,j}(\cdot)\) such that for \(k = 1, 2, 3\) it holds that

\[ g_{k,j}(\theta_{k,j}) = \eta^{(k,j)} \quad \text{for } j = 1, \ldots, p_k, \]

where the \(\theta_{k,j}\) are the components of the vectors \(\theta_k\), i.e. \(\theta_k = (\theta_{k,j})_{j=1,\ldots,p_k}\). The predictors \(\eta^{(k,j)}\) are assumed to depend on sets of covariates \(x^{(k,j)}\), so that \(\eta^{(k,j)} = \eta^{(k,j)}(x^{(k,j)})\), where \(x^{(k,j)} = (x_{1}^{(k,j)}, \ldots, x_{p_{k,j}}^{(k,j)})\). More details are given in Section 2.3.

The usual characterization of the GAMLSS model is achieved, for instance, by setting \(\theta_1 = (\mu_1, \sigma_1, \nu_1)\) and \(\theta_2 = (\mu_2, \sigma_2, \nu_2)\), where \(\mu_k, \sigma_k\) and \(\nu_k\), for \(k = 1, 2\), represent the parameters of the two marginal distributions. However, in general the marginals may depend on any number.
Table 1: Definition of copulae implemented in GJRM, with corresponding parameter range of association parameter $\theta_3$, link function of $\theta_3$, and relation between Kendall’s $\tau$ and $\theta_3$. $F_2(\cdot;\theta_3)$ denotes the cumulative distribution function (cdf) of a standard bivariate normal distribution with correlation coefficient $\theta_3$, and $\Phi(\cdot)$ the cdf of a univariate standard normal distribution. $t_{2,\zeta}(\cdot;\zeta,\theta_3)$ indicates the cdf of a standard bivariate Student-t distribution with correlation $\theta_3$ and fixed $\zeta \in (2, \infty)$ degrees of freedom, and $t_{\zeta}(\cdot)$ denotes the cdf of a univariate Student-t distribution with $\zeta$ degrees of freedom. $D_1(\theta_3) = \frac{1}{\sqrt{\pi}} \int_0^{\theta_3} e^{\frac{t^2}{2}} dt$ is the Debye function and $D_2(\theta_3) = \int_0^{\theta_3} t \log(t)(1-t)^{-3/2} dt$. Quantities $Q$ and $R$ are given by $1 + (\theta_3 - 1)(u + v)$ and $Q^2 - 4D_2(\theta_3 - 1)(u + v)$, respectively. The Kendall’s $\tau$ for “PL” is computed numerically as no analytical expression is available. Argument **null** of gjirm() in GJRM allows the user to employ the desired copula function and can be set to any of the values within brackets next to the copula names in the first column; for example, **null** = “J0”. For Clayton, Gumbel and Joe, the number after the capital letter indicates the degree of rotation required: the possible values are 0, 90, 180 and 270. 

$p_1$ and $p_2$ of population parameters. The distributions for $Y_2^*$ implemented for this work in the R package GJRM are the normal (“N”), log-normal (“LN”), Gumbel (“GU”), reverse Gumbel (“rGU”), logistic (“LO”), Weibull (“WEI”), inverse Gaussian (“iG”), gamma (“GA”), Dagum (“DAGUM”), Singh-Maddala (“SN”), beta (“BE”) and Fisk (“FISK”); their definitions can be found in Stasinopoulos et al. (2017b). For the binary selection variable $Y_1$, probit, logit and cloglog models can be employed. The choice of the link function for modelling $Y_1$ determines the type of the distribution assumed for the latent selection variable $Y_1^*$. For example, if $Y_1^*$ follows a normal distribution with mean $\theta_1 = \theta_{1,1} = \eta_{1,1}$ and variance equal to 1 then a probit regression model arises. In this case, $p_1 = 1$.

Argument **null** of gjirm() in GJRM allows the user to employ the desired link function and outcome distribution and can be set to any of the values indicated above within brackets. For example, **null** = c(“cloglog”, “GU”). The list of possible copulae, which are implemented in GJRM, is given in Table 1.

2.2. A special case: one-parameter exponential families

In this section, we assume that $Y_2^*$ has a density that belongs to the one-parameter exponential families, which is useful since it allows us to provide an interesting insight into the correction mechanism underlying the selection approach. In particular, $Y_2^*$ is assumed to have a density of

<table>
<thead>
<tr>
<th>Copula</th>
<th>$C(u, v; \theta_3)$</th>
<th>$\theta_3$</th>
<th>Link</th>
<th>Kendall’s $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMH (<em>AMH</em>)</td>
<td>$\frac{uv^{\theta_3}}{1-\theta_3(1-u)(1-v)}$</td>
<td>$\theta_3 \in [-1, 1]$</td>
<td>$\text{tanh}^{-1}(\theta_3)$</td>
<td>$-\frac{\theta_3}{\theta_3 + (1 - \theta_3)^2}$</td>
</tr>
<tr>
<td>Clayton (<em>CO</em>)</td>
<td>$\left(\frac{u^{\theta_3} + v^{\theta_3} - 1}{1} \right)^{-1/\theta_3}$</td>
<td>$\theta_3 \in (0, \infty)$</td>
<td>$\log(\theta_3)$</td>
<td>$\frac{\theta_3}{\theta_3 + (1 - \theta_3)^2}$</td>
</tr>
<tr>
<td>FGM (<em>FGM</em>)</td>
<td>$uv(1 + \theta_3(1-u)(1-v))$</td>
<td>$\theta_3 \in [-1, 1]$</td>
<td>$\text{tanh}^{-1}(\theta_3)$</td>
<td>$-\frac{\theta_3}{\theta_3 + (1 - \theta_3)^2}$</td>
</tr>
<tr>
<td>Frank (<em>FE</em>)</td>
<td>$-\frac{\theta_3^{-1}}{3} \log \left{ 1 + \left( \exp \left{ \theta_3 u \right} - 1 \right) \right}$</td>
<td>$\theta_3 \in \mathbb{R} \setminus {0}$</td>
<td>-</td>
<td>$1 - \frac{4}{3 \theta_3} [1 - D_1(\theta_3)]$</td>
</tr>
<tr>
<td>Hougaard (<em>HO</em>)</td>
<td>$\exp \left{ \frac{1}{3} \left[ \left( -\log u \right)^{\frac{1}{3}} + \left( -\log v \right)^{\frac{1}{3}} \right] \right}$</td>
<td>$\theta_3 \in (0, 1)$</td>
<td>$\log \left( \frac{\theta_3}{1 - \theta_3^3} \right)$</td>
<td>$1 - \theta_3$</td>
</tr>
<tr>
<td>Gaussian (<em>N</em>)</td>
<td>$\Phi_2 \left( \Phi^{-1}(u), \Phi^{-1}(v); \theta_3 \right)$</td>
<td>$\theta_3 \in [-1, 1]$</td>
<td>$\text{tanh}^{-1}(\theta_3)$</td>
<td>$\frac{2}{\pi} \arcsin(\theta_3)$</td>
</tr>
<tr>
<td>Gumbel (<em>GU</em>)</td>
<td>$\exp \left{ \frac{1}{3} \left[ \left( -\log u \right)^{\frac{1}{3}} + \left( -\log v \right)^{\frac{1}{3}} \right] \right}$</td>
<td>$\theta_3 \in [1, \infty)$</td>
<td>$\log(\theta_3 - 1)$</td>
<td>$1 - \frac{1}{\theta_3}$</td>
</tr>
<tr>
<td>Joe (<em>J0</em>)</td>
<td>$\frac{1 - \left{ (1 - u)^{\theta_3} + (1 - v)^{\theta_3} \right}}{2(\theta_3 - 1)}$</td>
<td>$\theta_3 \in (1, \infty)$</td>
<td>$\log(\theta_3 - 1)$</td>
<td>$1 + \frac{4}{\theta_3} D_2(\theta_3)$</td>
</tr>
<tr>
<td>Plackett (<em>PL</em>)</td>
<td>$\left( \sqrt{Q - R} \right) / {2(\theta_3 - 1)}$</td>
<td>$\theta_3 \in (0, \infty)$</td>
<td>$\log(\theta_3)$</td>
<td>-</td>
</tr>
<tr>
<td>Student-t (<em>T</em>)</td>
<td>$t_{2,\zeta} \left( t_{\zeta}^{-1}(u), t_{\zeta}^{-1}(v); \zeta, \theta_3 \right)$</td>
<td>$\theta_3 \in [-1, 1]$</td>
<td>$\text{tanh}^{-1}(\theta_3)$</td>
<td>$\frac{2}{\pi} \arcsin(\theta_3)$</td>
</tr>
</tbody>
</table>
the form
\[ f_2(y_2^*|\eta_2) = \exp \{y_2^*\eta_2 - b_2(\eta_2) + c_2(y_2^*)\} \]  
for some specific functions \( b_2(\cdot) \) and \( c_2(\cdot) \), where \( \eta_2 \) is the natural parameter. Here, \( \theta_2 = \eta_2 \) and \( p_2 = 1 \). It holds that \( \mathbb{E}(Y_2^*) = b_2'(\eta_2) \) and \( \text{Var}(Y_2^*) = b_2''(\eta_2) \), where \( b_2'(\cdot) \) and \( b_2''(\cdot) \) are the first and second derivatives of function \( b_2(\cdot) \), respectively (van der Vaart, 2000, p. 38).

Assume now that \((Y_1^*, Y_2^*)\) is an absolutely continuous random vector. Then the joint density of \((Y_1^*, Y_2^*)\) is
\[
f(y_1^*, y_2^*) = \frac{\partial^2}{\partial u \partial v} C_{\theta_3}(u, v) \bigg|_{u=F_2(y_1^*)} f_1(y_1^*) f_2(y_2^*) .
\]
The log-likelihood function for such defined sample selection model can be obtained by conditioning with respect to the value of the selection variable \( Y_1 \) (cf. Smith (2003), p. 108) and equals
\[
\ell = (1 - Y_1) \log F_1(0) + Y_1 \log \left( f_2(Y_2) - \frac{\partial}{\partial y_2^*} F(0, y_2^*) \bigg|_{y_2^*=y_2} \right) . \tag{2.3}
\]
Using (2.1), we obtain
\[
\frac{\partial}{\partial y_2^*} F(0, y_2^*) \bigg|_{y_2^*=y_2} = \frac{\partial}{\partial y_2^*} C_{\theta_3}(F_1(0), F_2(y_2^*)) \bigg|_{y_2^*=y_2} = \frac{\partial}{\partial v} C_{\theta_3}(F_1(0), v) \bigg|_{v=F_2(y_2)} f_2(Y_2) .
\]
Thus,
\[
\ell = (1 - Y_1) \log F_1(0) + Y_1 \log \left( f_2(Y_2) z(Y_2, \eta_1, \eta_2) \right),
\]
where \( z(y_2, \eta_1, \eta_2) = 1 - \frac{\partial}{\partial v} C_{\theta_3}(F_1(0), v) \bigg|_{v=F_2(y_2)} \). At the same time, \( f_2(y_2) z(y_2, \eta_1, \eta_2) = P(Y_1^* > 0, Y_2^* = y_2) \), which is implied by the very definition of likelihood. Hence
\[
z(y_2, \eta_1, \eta_2) = P(Y_1^* > 0 | Y_2^* = y_2),
\]
which has an intuitive interpretation: the probability of the output being observed given that its latent value is \( y_2 \).

Using (2.2), the log-likelihood can be written as
\[
\ell = (1 - Y_1) \log F_1(0) + Y_1 (\eta_2 Y_2 - b_2(\eta_2) + c_2(Y_2) + \log (z(Y_2, \eta_1, \eta_2))) . \tag{2.4}
\]
The fact that \( \mathbb{E}(Y_2) = b_2'(\eta_2) \) implies
\[
\frac{\partial}{\partial \eta_2} \ell = Y_1 (Y_2 - \mu_2) + Y_1 \frac{\partial}{\partial \eta_2} \log (z(Y_2, \eta_1, \eta_2)) ,
\]
where $\mu_2 = \mathbb{E}(Y_2)$. Note that the first term in the expression above is equal to the score for the standard model, when the sample selection does not appear and hence $Y_1$ always equals 1 and $z(Y_2, \eta_1, \eta_2) \equiv 1$. The second term corrects the score for sample selection bias. Using the fact that the expected value of the score is equal to 0 when evaluated at the true parameters $\eta_1, \eta_2, \theta_3$ or, more generally, at their values that minimize the Kullback-Leibler loss, we obtain

$$\mathbb{E} \left( Y_1 \frac{\partial}{\partial \eta_2} \log (z(Y_2, \eta_1, \eta_2)) \right) = -\text{Cov}(Y_1, Y_2).$$

Thus, the stronger the correlation between outcomes and selection mechanism, the further away from 0 the second term of the score is expected to be, hence implying greater influence on the estimates of $\eta_2$. The Fisher information $I(\eta_2) = -\mathbb{E} \left( \frac{\partial^2}{\partial \eta_2^2} \ell \right)$ for $\eta_2$ is

$$I(\eta_2) = -\mathbb{E} \left( -Y_1 b_2''(\eta_2) + \frac{\partial^2}{\partial \eta_2^2} \log (z(Y_2, \eta_1, \eta_2)) \right) = \text{Var}(Y_2^*) P(Y_1 = 1) - \mathbb{E} \left( \frac{\partial^2}{\partial \eta_2^2} \log (z(Y_2, \eta_1, \eta_2)) \right),$$

whereas the Fisher information for the model without sample selection is $I(\eta_2) = \text{Var}(Y_2^*)$.

### 2.3. Additive predictors and penalized regression spline representation

In line with the latest developments in the spline literature (e.g., Wood, 2017), we assume the additive form for the model’s predictors. That is,

$$\eta^{(k,j)}(x^{(k,j)}) = \eta_1^{(k,j)}(x_1^{(k,j)}) + \eta_2^{(k,j)}(x_2^{(k,j)}) + \ldots + \eta_{D_{k,j}}^{(k,j)}(x_{D_{k,j}}^{(k,j)}). \quad (2.5)$$

To flexibly represent the components in (2.5), we employ the penalized regression spline approach (Eilers and Marx, 1996). Specifically, for each $v = 1, \ldots, D_{k,j}$ we approximate $\eta_v^{(k,j)}(x)$ by a linear combination of basis functions $B_{v,j}^{(k,j)}(x)$ and coefficients,

$$\sum_{j=1}^{K_v^{(k,j)}} \beta_{v,j}^{(k,j)} B_j(x). \quad (2.6)$$

In the following equations, we drop the superscript of $K_v^{(k,j)}$ to avoid an over-complicated display. However, we have to bear in mind that $K_v = K_v^{(k,j)}$ still depends on $k$ and $j$. We define vectors $\beta_v^{(k,j)} \in \mathbb{R}^{K_v}$ as

$$\beta_v^{(k,j)} = (\beta_{v,1}^{(k,j)}, \ldots, \beta_{v,K_v}^{(k,j)})^T \quad \text{for } v = 1, \ldots, D_{k,j},$$

d and $\beta = (\beta_v^{(k,j)})_{v=1,\ldots,D_{k,j}; j=1,\ldots,p_k; k=1,2,3} \in \mathbb{R}^q$ where $q = \sum_{k=1}^3 \sum_{j=1}^{p_k} \sum_{v=1}^{D_{k,j}} K_v$. Thus, equa-
tion (2.6) implies that the vector of evaluations \( \left( \eta_{v}^{(k,j)}(x_{v}^{(k,j)}) \right)_{v=1,\ldots,D_{k,j}; j=1,\ldots,p_{k}; k=1,2,3} \) can be written as \( x \beta \) where the row vector \( x \) holds the values of \( B_{j}(x_{v}^{(k,j)}) \) for \( v = 1, \ldots, D_{k,j}, \ j = 1, \ldots, p_{k}, \) and \( k = 1,2,3 \). The row vectors \( x \) evaluated for each one of \( n \) observations in the random sample will form the design matrix \( X \).

Each \( \beta_{v}^{(k,j)} \) has an associated quadratic penalty \( \lambda_{v}^{(k,j)} \left( \beta_{v}^{(k,j)} \right)^{T} D_{v}^{(k,j)} \beta_{v}^{(k,j)} \) whose role is to enforce, during fitting, specific properties of the function \( \eta_{v}^{(k,j)}(x) \), such as smoothness or shrinkage. Matrix \( D_{v}^{(k,j)} \) depends on the choices made to implement equation (2.6), such as \( K_{v}^{(k,j)} \) and the definition adopted for \( B_{j}(x) \). The smoothing parameter \( \lambda_{v}^{(k,j)} \geq 0 \) controls the trade-off between fit and smoothness, and plays a crucial role in determining the shape of the estimates of smooth functions \( \eta_{v}^{(k,j)}(x) \). The overall penalty can be defined as \( \beta^{T} Q(\lambda) \beta \), where \( Q(\lambda) = \text{diag}(\lambda_{v}^{(k,j)} D_{v}^{(k,j)}), \ v = 1, \ldots, D_{k,j}; j = 1, \ldots, p_{k}, k = 1,2,3 \).

The set up described above can allow one to account for several types of covariate effects such as linear, non-linear, spatial, random and functional effects. We refer the reader to Wood (2017) for the exact definitions of the spline bases and penalties for the above mentioned cases.

3. Some estimation and inferential details

For a given \( n \in \mathbb{N} \), assume that \( (Y_{1i}, Y_{2i})_{i=1}^{n} \) are independent random variables related to covariate values \( x_{i}^{(k,j)} \) for \( i = 1, \ldots, n \) such that \( Y_{1i} = (Y_{1i}^{*} > 0) \) and \( Y_{2i} = Y_{2i}^{*} Y_{1i} \), where \( Y_{1i}^{*} \) and \( Y_{2i}^{*} \) are distributed according to (2.1). Let \( F_{1i} \) and \( F_{2i} \) denote the distribution functions of \( Y_{1i}^{*} \) and \( Y_{2i}^{*} \), and let \( F_{1}(\cdot, \cdot) \) be the joint cdf of the pair \( (Y_{1i}^{*}, Y_{2i}^{*}) \).

In order to estimate the overall vector of parameters \( \beta \), we employ a penalized likelihood approach to avoid overfitting. The log-likelihood given the observed random sample \( (y_{1i}, y_{2i})_{i=1}^{n} \) is given by

\[
\ell(\beta) = \sum_{i=1}^{n} (1 - y_{1i}) \log F_{1i}(0) + \sum_{i=1}^{n} y_{1i} \log \{P(Y_{1i} = 1, Y_{2i} = y_{2i})\}
\]

if \( Y_{2i} \) is and

\[
\ell(\beta) = \sum_{i=1}^{n} (1 - y_{1i}) \log F_{1i}(0) + \sum_{i=1}^{n} y_{1i} \log \left( f_{2i}(y_{2i}) - \frac{\partial}{\partial y_{2}^{*}} F_{i}(0, y_{2}^{*}) \bigg|_{y_{2}^{*} = y_{2i}} \right)
\]

if the outcome is continuous, based on (2.3). The penalized log-likelihood is given by

\[
\ell_{p}(\beta) = \ell(\beta) - \frac{1}{2} \beta^{T} Q(\lambda) \beta.
\]
Estimation of $\beta$ and $\lambda$ is achieved by adapting to this context the stable and efficient trust region algorithm with integrated automatic multiple smoothing parameter selection by Marra et al. (2017). This required working with first and second order analytical derivatives which have been tediously derived and verified using numerical derivatives. All relevant quantities have been implemented in a modular way. This means that no substantial programming work will be required to incorporate copulae and marginal distributions not considered in this article, as long as their cumulative and probability density functions are known and their derivatives with respect to their parameters exist.

At convergence, reliable point-wise confidence intervals for linear and non-linear functions of the model coefficients can be obtained using the Bayesian large sample approximation $\beta \sim \mathcal{N}({\hat{\beta}}, -H_p(\hat{\beta})^{-1})$, where $H_p$ is the penalized model’s Hessian (Marra et al., 2017). Intervals derived using this result have good frequentist properties since they account for both sampling variability and smoothing bias. Furthermore, intervals for any non-linear function of the model’s coefficients can be conveniently obtained by simulation from the posterior distribution of $\beta$.

The theoretical properties of the proposed estimator could be studied by considering a fixed number of knots for the basis functions, in which case $n^{1/2}$ asymptotic results can be straightforwardly obtained.

3.1. Software

The models can be employed via the gjrm() function in the R package GJRM (Marra and Radice, 2018). An example of call is

```r
fl <- list(y1 ~ x1 + s(x2) + s(x3),
           y2 ~ x1 + s(x2),
           ~ x1 + s(x3),
           ~ x1 + s(x2))
md <- gjrm(fl, margins = c("logit", "WEI"), BivD = "PL", Model = "BSS")
```

where `fl` is a list containing four equations (the first for the selection equation, the second and third for the two parameters of the response distribution, and the fourth for the copula dependence parameter), `margins` specifies the marginal distributions and `BivD` the copula. Argument `Model = "BSS"` means that a bivariate model with sample selection will be employed.

4. Simulation study

The aim of this section is to assess the empirical properties of the proposed modelling approach. To this end, we consider three main scenarios. In scenario I, we use logistic and gamma
margins which are linked via the Clayton copula with parameter $\theta_3 = 3$ (equivalently, Kendall’s $\tau = 0.6$). Here, only the means of the marginal distributions are specified as functions of additive predictors. Specifically,

$$
\mu_1 = \alpha_0 + \alpha_1 x_1 + s_1(x_2), \\
\log\mu_2 = \beta_0 + s_2(x_2) + \beta_1 x_3,
$$

where $\alpha_0 = -0.8$, $\alpha_1 = -1.3$, $\beta_0 = 0.1$, $\beta_1 = -0.9$, $s_1(x) = x + \exp(-30(x - 0.5)^2)$ and $s_2(x) = \sin(2\pi x)$.

Scenario II is essentially the same as scenario I but both margins are Gaussian, with identity link functions, and the copula employed is the Gumbel with $\theta_3 = 2.5$ (Kendall’s $\tau = 0.6$).

In scenario III, data are generated using Gaussian and Gumbel margins, and the Joe copula. In this case, all the parameters of the bivariate distribution depend on additive predictors. That is,

$$
\mu_1 = \alpha_0 + \alpha_1 x_1 + s_1(x_2) + \alpha_2 x_3, \\
\mu_2 = \beta_0 + s_2(x_2) + \beta_1 x_3, \\
\log(\sigma^2) = \beta_{0\sigma}^\sigma + \beta_{1\sigma}^\sigma x_3, \\
\log(\theta_3) = \beta_{0\theta}^\theta + \beta_{1\theta}^\theta x_1 + s_3(x_2),
$$

where $\alpha_0 = -0.8$, $\alpha_1 = -1.3$, $\alpha_2 = 1$, $\beta_0 = 0.1$, $\beta_1 = -0.9$, $\beta_{0\sigma}^\sigma = 0.5$, $\beta_{1\sigma}^\sigma = 1$, $\beta_{0\theta}^\theta = 1.1$, $\beta_{1\theta}^\theta = -1.4$, $s_1(x) = x + \exp(-30(x - 0.5)^2)$, $s_2(x) = \sin(2\pi x)$ and $s_3(x) = 0.6(e^x + \sin(2.9x))$.

Note that the marginal distributions are parametrised according to Stasinopoulos et al. (2017b).

The simulated data-sets consist of two continuous outcomes, one binary covariate and two continuous regressors. The first continuous response is dichotomised since it refers to the selection equation. Sample sizes are set to 1000 and 5000, the number of replicates to 1000, and the models fitted using gjrm() in GJRM. Each smooth function is represented using a penalized low rank thin plate spline with second order penalty and 10 basis functions. For each replicate, smooth function estimates are constructed using 200 equally spaced fixed values in the $(0,1)$ range (e.g., Radice et al., 2016). Exact details on the generation of the simulated datasets are given in Appendix A.

4.1. Results

In this section we focus on the results obtained for the outcome equation, which is the one of interest, as well as for the Kendall’s $\tau$. Figure 1 displays the findings for case of data generated according to scenario I. In this case, estimates are shown for the models based on:

- logit and gamma margins with a Clayton copula (the correct model);
logit and inverse Gaussian margins with a Clayton copula (the outcome distribution is misspecified);

logit and gamma margins with the classic Gaussian copula (the dependence structure is misspecified).

We chose the inverse Gaussian since it has the same mean as that of the gamma (Stasinopoulos et al., 2017b), hence facilitating the comparison of estimates. When the model is correctly specified, all mean estimates are very close to the true values and, as expected, their variability decreases as the sample size increases. Misspecifying the marginal outcome distribution has a substantial detrimental impact on all the parameter estimates, hence stressing the importance of choosing a suitable outcome distribution in practical situations. Using the incorrect dependence also affects the estimates (although in a less pronounced manner), hence emphasizing the potential benefits of allowing for non-Gaussian structures. We also fitted models based on other copulae (such as Frank, FGM, AMH and Joe available in GJRM) and the findings were similar. Moreover, the correct model was always selected by criteria such as AIC and BIC. Misspecifying the link function (using probit and cloglog links) for the selection equation did not significantly affect the results. Perhaps this is not surprising given that all links produced very similar predicted probabilities for the selection response variable. Nevertheless, the availability of different link functions allowed us to assess the impact of this misspecification on the parameters of interest. Using a 2.20-GHz Intel(R) Core(TM) computer running Windows 7, model fitting took on average 2 seconds for \( n = 1000 \), and 7 seconds for \( n = 5000 \). Increasing the number of basis functions to 20 did not have a noticeable impact on the results but increased computing time by about 20% on average. Moreover, using other spline definitions (such as penalized cubic regression splines and P-splines) virtually led to identical results. These findings were somewhat expected and have also been documented in similar contexts by (Wood, 2017).

The results for scenario II are given in Figure 2. Estimates are shown for the models based on:

probit and Gaussian margins with a Gumbel copula (the correct model);

probit and Gaussian margins with the classic Gaussian copula (the dependence structure is misspecified).

The conclusions are similar to those obtained for scenario I. Specifically, for the correctly specified model the mean estimates are close to the true values and the variability of the estimates decreases as the sample size grows large, whereas using the incorrect dependence structure affects negatively...
all the parameter estimates. Using various copulae the correct model was always picked by AIC and BIC, link function misspecification did not really alter the estimates, computing times were similar to those found for scenario I, and increasing the number of basis functions and using different spline’s definitions did not have a tangible impact on the results. We have not reported the results obtained when misspecifying the marginal outcome distribution as these were nearly identical to those obtained for scenario I.

The results for scenario III are given in Figure 3 and are based on probit and Gumbel margins with a Joe copula (the correct model). This scenario is more complex than the previous ones in that all distributional parameters are specified as functions of covariates. The findings show that the approach can estimate all the model components fairly well, and that the estimates improve as the sample size increases. The components in the additive predictor of the dependence parameter are estimated less precisely than those of the others. This indicates that the effects of covariates on the association between the selection and outcome equations may be more difficult to estimate. This is reasonable given that the likelihood contributions for the association parameter come from the selected sample of observations only. Average computing times were about 16 seconds for \( n = 1000 \), and 42 seconds for \( n = 5000 \). We also tested the models under misspecification of the dependence structure and marginal outcome distribution. In the former case, the findings were similar to those for scenarios I and II; using the incorrect copula affects adversely the parameter estimates in terms of bias and efficiency. In the latter case, the models failed to converge in many of the iterations (55% for \( n = 1000 \) and 43% for \( n = 5000 \)) and for the converged models computing times were between 20 and 30 times those reported above. This highlighted the importance of choosing an appropriate distribution for the outcome variable, especially when the model specification is complex.
(a) Estimates for $\beta_1 = -0.9$.  

(b) Estimates for Kendall’s $\tau = 0.6$. 

(c) Smooth curve estimates for $s_2(\cdot)$. 

Figure 1: Scenario I. In Figures (a) and (b), black circles and vertical bars refer to the results obtained under the correct model, dark grey circles and bars to those obtained when misspecifying the outcome distribution, and light grey circles and bars to those obtained when the dependence structure is misspecified. Circles indicate mean estimates while bars represent the estimates’ ranges resulting from 5% and 95% quantiles. In Figure (c), mean estimates are represented by dashed lines and point-wise ranges resulting from 5% and 95% quantiles by shaded areas. The top plots refer to $n = 1000$ and the bottom ones to $n = 5000$. In all figures, true values are given by the black solid lines.
(a) Estimates for $\beta_1 = -0.9$.

(b) Estimates for Kendall’s $\tau = 0.6$.

(c) Smooth curve estimates for $s_2(\cdot)$.

Figure 2: Scenario II. In Figures (a) and (b), black circles and vertical bars refer to the results obtained under the correct model, and grey circles and bars to those obtained when the dependence structure is misspecified. Circles indicate mean estimates while bars represent the estimates’ ranges resulting from 5% and 95% quantiles. In Figure (c), mean estimates are represented by dashed lines and point-wise ranges resulting from 5% and 95% quantiles by shaded areas. The top plots are obtained from the correct model and bottom ones from the misspecified model. In all figures, true values are given by the black solid lines.
(a) Estimates for the parametric components in the model.

(b) Smooth curve estimates for \( s_1(\cdot) \), \( s_2(\cdot) \) and \( s_3(\cdot) \).

Figure 3: Scenario III. In Figure (a), black circles and vertical bars refer to the results obtained when \( n = 1000 \), and grey circles and bars to those obtained when \( n = 5000 \). Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. In Figure (b), mean estimates are represented by dashed lines and point-wise ranges resulting from 5% and 95% quantiles by shaded areas. The top plots are obtained when \( n = 1000 \) and bottom ones when \( n = 5000 \). In all figures, true values are given by the black solid lines.
5. Empirical application

As a real world application, we consider the study of the effects of insurance status and managed care on hospitalization spells previously analysed by Prieger (2002). The data set is based on a nationally representative survey of US medical care (Medical Expenditure Panel Survey) and it contains information about the length of individuals’ hospital stays in 1996 along with factors such as membership in health maintenance organization, type of insurance, health status, demographic variables, sex, race, marriage, employment status and quantitative variables including age, years of education, number of self-reported medical conditions and number of conditions on the priority list. A detailed description of the variables can be found in Table 3 given in Appendix B. The sample used in the analysis consists of 14,946 observations. The response variable for the selection equation is whether an individual had a hospital stay. If the link between hospital admittance and the spell of hospital stay is not through observables alone then sample selection bias arises and using a univariate regression approach is not adequate.

These data were studied by Prieger (2002) who motivates the use of the gamma distribution to model the length of hospital stay, uses a probit selection equation, and fits three models based on the assumption of independence, and on the Gaussian and FGM copulae. All the covariates entered the selection and outcome equations parametrically. Prieger found, for instance, that non-random sample selection was present and, based on various criteria, chose the FGM copula which produced a negative and significant estimated dependence between the two equations.

We re-analyse these data by considering a wider set of marginal outcome distributions, link functions and copulae. We also employ smooth functions of age and years of education (using the same set up described in the simulation study), and specify all parameters of the marginal distributions as functions of additive predictors.

Regarding the marginals, we chose the probit link and found that the inverse Gaussian instead of the gamma distribution provides the best fit as judged by the plots of normalised quantile residuals (Stasinopoulos et al., 2017b) and information criteria (see Figure 4). Using the logit and cloglog links for the selection equation did not affect the results. As for the choice of copula, we started off with the Gaussian, Frank, FGM, AMH, Student-t and Plackett (since they allow for both positive and negative dependence) and then employed all of the remaining copulae that were consistent with the sign of dependence found. For this empirical application, we tried all copulae available as there was not a clear indication of positive or negative dependence. In all cases, the values for $\tau$ were very close to zero as well as non significantly different from zero for those copulae admitting both positive and negative association. The AIC and BIC values across
Figure 4: Histograms and Q-Q plots of normalized quantile residuals for length of hospital stay produced after fitting copula models based on the inverse Gaussian (top) and gamma (bottom) distributions. The Q-Q-plots also exhibit reference bands for judging the relevance of departures from the red reference lines. AIC and BIC values are 15413.15 and 16101.04 for the model with inverse Gaussian outcome distribution and 15616.74 and 16351.86 for the model with gamma distribution.
Table 2: Comparison of AIC and BIC values under different copula assumptions, and probit and inverse Gaussian margins.

<table>
<thead>
<tr>
<th>Copula</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>15413.15</td>
<td>16101.04</td>
</tr>
<tr>
<td>F</td>
<td>15413.02</td>
<td>16101.95</td>
</tr>
<tr>
<td>FGM</td>
<td>15412.96</td>
<td>16101.96</td>
</tr>
<tr>
<td>AMH</td>
<td>15412.94</td>
<td>16101.96</td>
</tr>
<tr>
<td>T</td>
<td>15469.05</td>
<td>16222.56</td>
</tr>
<tr>
<td>Pl</td>
<td>15413.04</td>
<td>16101.94</td>
</tr>
<tr>
<td>Ho</td>
<td>15414.38</td>
<td>16091.43</td>
</tr>
<tr>
<td>C0</td>
<td>15412.53</td>
<td>16100.63</td>
</tr>
<tr>
<td>J0</td>
<td>15414.40</td>
<td>16090.91</td>
</tr>
<tr>
<td>G0</td>
<td>15414.51</td>
<td>16100.78</td>
</tr>
<tr>
<td>C90</td>
<td>15413.36</td>
<td>16101.37</td>
</tr>
<tr>
<td>J90</td>
<td>15411.32</td>
<td>16092.09</td>
</tr>
<tr>
<td>G90</td>
<td>15411.32</td>
<td>16092.09</td>
</tr>
<tr>
<td>C180</td>
<td>15418.19</td>
<td>16093.86</td>
</tr>
<tr>
<td>J180</td>
<td>15413.31</td>
<td>16101.55</td>
</tr>
<tr>
<td>G180</td>
<td>15411.32</td>
<td>16092.09</td>
</tr>
<tr>
<td>C270</td>
<td>15411.32</td>
<td>16092.09</td>
</tr>
<tr>
<td>J270</td>
<td>15412.86</td>
<td>16102.81</td>
</tr>
<tr>
<td>G270</td>
<td>15414.87</td>
<td>16103.51</td>
</tr>
</tbody>
</table>

Copula assumptions were fairly close in most cases (see Table 2). This was somewhat expected given that no significant association between the equations was detected with all copulae.

Appendix B shows the summary output obtained from the final model which is based on the $270^\circ$ Clayton copula and probit and inverse Gaussian margins. Employing other copulae (for instance, G180, G90, J90) produced nearly identical results. The main findings can be summarised as follows:

- As argued by Prieger (2002), the association (positive or negative) between admittance and length of stay may suggest the presence of specific selection mechanisms. He also states that there is no a priori expectation on the sign of the dependence. As opposed to Prieger’s finding of a negative association between the selection and outcome equations, we found that non-random sample selection is not present when using the inverse Gaussian (the distribution supported by the data). However, when using the gamma as outcome distribution and the FGM copula (as well as other copulae such as Gaussian, Frank, Student-t and Plackett) we found that the association parameter is negative and significant (e.g., $\hat{\tau} = -0.514$ with $(-0.589, -0.428)$ as 95% confidence interval for $\tau$), which is line with Prieger’s result. Our simulations show that misspecifying the outcome distribution can have a severe detrimental impact on the parameter estimates including the Kendall’s $\tau$. This all suggests that Prieger’s finding is biased by the choice of gamma distribution for the outcome equation.

- Selection equation: from the summary output for equation 1, reported in Appendix B, we
observe for instance that the insurance variables privins, medicare and medicaid increase the probability of hospital admittance, that such effects are either reinforced or tempered by privmcare, privmcaid and mcaremcaid, and that covariates hmopriv, hmomcare, hmomcaid have no significant effect on hospital admittance. Moreover, variables condn, priolist, adlhelp and poorhlth increase the probability of hospital admittance. These findings are consistent with those of Prieger (2002) to which the reader is referred to for a more thorough discussion.

The estimated smooth functions for age and years of education are displayed in Figure 6, Appendix B. The effect of education is not significant and nearly linear (see also respective p-value reported in the summary output). On the other hand, the effect of age is non-linear and significant; its shape suggests that age decreases the probability of hospital admittance up to about 45 years and then increases such probability afterward. This may be due to the fact that age embodies productivity and life-cycle effects that are likely to affect the responses considered in this study non-linearly.

- Outcome equation: from the summary output for equation 2 we observe for example that poorhlth and adlhelp significantly lengthen the stay in hospital, hmopriv decreases the stay, that privins does not influence the outcome, and that medicare decreases the duration of stay. Our findings are in agreement with those by Prieger (2002). Note, however, that different distributions and parametrizations are employed in two analyses, hence an exact comparison is not possible.

The estimated smooth function for years of education (not shown here) is linear and non-significant (as also supported by the respective p-value in the summary output). Figure 5 shows the effect of age on the average hospital stay duration. It suggests that as age increases the average length of hospital stay increases up to 35, decreases and then increases again after 45. It may be argued that, given the width of the confidence intervals, a straight line relationship is also suitable here. In the absence of a formal test of linearity of a smooth function for the current modelling framework, an informal indication of whether a simpler model would be appropriate can be obtained using information criteria. Specifically, AIC and BIC values are 15411.32 and 16092.09 for the model with non-linear effect for age, and 15420.02 and 16064.05 for the model with linear effect; the conclusions reached by the two criteria are discordant and a definitive answer cannot be provided in this case.

A univariate analysis using the selected sample of observations only led to the same results.
as those for the copula model’s outcome equation. This is not surprising given that, as discussed previously, no significant association between the selection and outcome equations was found.

- Copula models where dependence parameter $\theta_3$ was specified as a function of various combinations of covariates were also fitted. This allowed us to capture potential heterogeneity in the selection process, hence possibly justifying the overall non-significance of the dependence parameter potentially due to compensating effects. However, the results consistently pointed to the lack of significant association between the selection and outcome equations.

The analysis presented in this section has extended Prieger’s one by considering a wider set of marginal distributions and copulae as well as non-linear covariate effects. Using the proposed modeling framework we found evidence of non-linearity for some covariate effects and that non-random sample selection does not seem to be present when employing the outcome distribution that is most supported by the data (inverse Gaussian in this case). Although the absence of selec-
tion bias may be regarded as a ‘non-finding’ at first, we argue that our result still has important implications for the study of selection bias since using a more restrictive set of modelling choices may lead to unfounded speculations on the presence of certain selection mechanisms.

6. Discussion

We have introduced an extension of GAMLSS which accounts for non-random sample selection. The proposed approach is flexible in that it allows for different parametric distributions of the selection and outcome variables, several types of dependence structures between the model’s equations, and for various types of covariate effects. Using the special case of one-parameter exponential families, we have elucidated the nature of the correction mechanism underlying the selection approach. Parameter estimation is carried out within a penalized likelihood framework based on a trust region algorithm with integrated smoothing parameter selection. The approach has been illustrated in simulation and through a case study. All new developments have been incorporated in the R package GJRM (Marra and Radice, 2018).

Many marginal distributions and copulae have been considered in this work and we plan on extending the set of choices available. Future research will look into generalising the proposed sample selection GAMLSS framework to empirical situations where rules of double selection exist (e.g., Smith, 2003; Zhang et al., 2015), exploiting for instance C- and D-Vine constructions.

Acknowledgment

We would like to thanks the two anonymous reviewers for their well thought out comments and suggestions which have helped to improve considerably the quality and massage of the article.

Appendix A: R codes to generate data for scenarios I, II and III

For the first two scenarios, data were generated using the following R code.

```r
library(copula); library(gamlss.dist)
library(GJRM)

cor.cov <- matrix(0.5, 3, 3); diag(cor.cov) <- 1

s1 <- function(x) x + exp(-30*(x - 0.5)^2)
s2 <- function(x) sin(2*pi*x)

datagen12 <- function(cor.cov, s1, s2, scen = 1)(
```
cov <- rMVN(1, rep(0,3), cor.cov)
cov <- pnorm(cov)

x1 <- cov[, 1]
x2 <- cov[, 2]
x3 <- round(cov[, 3])

eta_mu1 <- -0.8 - 1.3*x1 + s1(x2) + x3
eta_mu2 <- 0.1 + s2(x2) - 0.9*x3

speclist1 <- list( mu = eta_mu1, sigma = 1)

if(scen == 1){
  speclist2 <- list( mu = exp(eta_mu2), sigma = 3)
  spec <- mvdc(copula = Cop, c("LO", "GA"), list(speclist1, speclist2) )
  Cop <- archmCopula(family = "clayton", dim = 2, param = 3)
}

else{
  speclist2 <- list( mu = eta_mu2, sigma = 2)
  spec <- mvdc(copula = Cop, c("NO", "NO"), list(speclist1, speclist2) )
  Cop <- archmCopula(family = "gumbel", dim = 2, param = 2.5)
}

resp <- rMvdc(1, spec)
c(resp, x1, x2, x3)

Package copula (Yan, 2007) contains functions archmCopula(), mvdc() and rMvdc() which allow one to simulate from the desired copula. Package gamlss.dist (Stasinopoulos et al., 2017a) contains all the functions required to simulate the marginals adopted here, and rMVN() (from GJRM) allows one to simulate Gaussian correlated variables. The correlation matrix used to associate the three simulated Gaussian covariates is cor.cov, whereas cov <- pnorm(cov) allows one to obtain Uniform(0,1) correlated covariates (e.g., Gentle, 2003). A balanced binary regressor is created using round(cov[, 3]). Functions s1 and s2 produce curves with different degrees of complexity. The various eta refer to the model’s additive predictors. If necessary,
these are transformed in `speclist1` and `speclist2` (and also `archmCopula` for scenario III below) to ensure that the restrictions on the parameters’ spaces of the bivariate distributions are maintained. In the first two scenarios the copula dependence parameters are set to 3 and 2.5 which correspond to a Kendall’s $\tau$ of 0.6.

The code used to generate data for scenario III is given below.

```r
datagen3 <- function(cor.cov, s1, s2, s3)
{
  cov <- rMVN(1, rep(0,3), cor.cov)
  cov <- pnorm(cov)

  z1 <- cov[, 1]
  z2 <- cov[, 2]
  z3 <- round(cov[, 3])

  eta_mu1 <- -.8 - 1.3*z1 + s1(z2) + z3
  eta_mu2 <- 0.1 + s2(z2) - 0.9*z3
  eta_si2 <- 0.5 + z3
  eta_the <- 1.1 - 1.4*z1 + s3(z2)

  Cop <- archmCopula(family = "joe", dim = 2,
                     param = exp(eta_the) + 1 + 1e-07)

  speclist1 <- list( mu = eta_mu1, sigma = 1)
  speclist2 <- list( mu = eta_mu2, sigma = sqrt(exp(eta_si2)))

  spec <- mvdc(copula = Cop, c("NO", "GU"), list(speclist1, speclist2) )

  resp <- rMvdc(1, spec)

  c(resp, z1, z2, z3)
}
```
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>hospstay</td>
<td>Binary variable: 1 = individual had hospital stay</td>
<td>0.09</td>
<td>0.29</td>
</tr>
<tr>
<td>hospdur</td>
<td>Length of all hospitalizations, given HOSPSTAY = 1</td>
<td>7.43</td>
<td>11.36</td>
</tr>
<tr>
<td>hospnum</td>
<td>Number of hospital stays, given HOSPSTAY = 1</td>
<td>1.42</td>
<td>0.85</td>
</tr>
<tr>
<td>adlhelp</td>
<td>1 = requires assistance with daily living tasks</td>
<td>0.04</td>
<td>0.20</td>
</tr>
<tr>
<td>age</td>
<td>Age</td>
<td>44.40</td>
<td>17.31</td>
</tr>
<tr>
<td>Black</td>
<td>1 = black (not hispanic)</td>
<td>0.12</td>
<td>0.33</td>
</tr>
<tr>
<td>condn</td>
<td>Number of self-reported medical conditions</td>
<td>1.68</td>
<td>1.91</td>
</tr>
<tr>
<td>educ</td>
<td>Years of education</td>
<td>12.38</td>
<td>3.16</td>
</tr>
<tr>
<td>employed</td>
<td>Employment status: 1 = currently employed</td>
<td>0.65</td>
<td>0.48</td>
</tr>
<tr>
<td>exclhlth</td>
<td>1 = individual reports health to be ‘excellent’</td>
<td>0.29</td>
<td>0.45</td>
</tr>
<tr>
<td>female</td>
<td>1 = female</td>
<td>0.54</td>
<td>0.60</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1 = of hispanic ethnicity</td>
<td>0.18</td>
<td>0.38</td>
</tr>
<tr>
<td>hmomcaid</td>
<td>1 = enrolled in a HMO and covered by Medicaid</td>
<td>0.03</td>
<td>0.18</td>
</tr>
<tr>
<td>hmomcare</td>
<td>1 = enrolled in a HMO and covered by Medicare</td>
<td>0.04</td>
<td>0.19</td>
</tr>
<tr>
<td>hmopriv</td>
<td>1 = enrolled in a HMO and covered by private insurance</td>
<td>0.33</td>
<td>0.47</td>
</tr>
<tr>
<td>married</td>
<td>Marital status: 1 = currently married</td>
<td>0.57</td>
<td>0.49</td>
</tr>
<tr>
<td>mcaremcaid</td>
<td>1 = currently covered by Medicaid and Medicare</td>
<td>0.02</td>
<td>0.16</td>
</tr>
<tr>
<td>medicaid</td>
<td>1 = currently covered by Medicaid</td>
<td>0.09</td>
<td>0.28</td>
</tr>
<tr>
<td>medicare</td>
<td>1 = currently covered by Medicare</td>
<td>0.17</td>
<td>0.38</td>
</tr>
<tr>
<td>MidWest</td>
<td>Regional indicator (EAST is the excluded dummy)</td>
<td>0.22</td>
<td>0.42</td>
</tr>
<tr>
<td>poorhlth</td>
<td>1 = individual reports health to be ‘poor’</td>
<td>0.04</td>
<td>0.20</td>
</tr>
<tr>
<td>priolist</td>
<td>Number of conditions on the priority list</td>
<td>0.54</td>
<td>1.00</td>
</tr>
<tr>
<td>privins</td>
<td>1 = covered by private insurance of any type</td>
<td>0.66</td>
<td>0.47</td>
</tr>
<tr>
<td>privmcaid</td>
<td>1 = covered by private insurance and Medicaid</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>privmcare</td>
<td>1 = covered by private insurance and Medicare</td>
<td>0.10</td>
<td>0.29</td>
</tr>
<tr>
<td>South</td>
<td>Regional indicator (EAST is the excluded dummy)</td>
<td>0.35</td>
<td>0.48</td>
</tr>
<tr>
<td>West</td>
<td>Regional indicator (EAST is the excluded dummy)</td>
<td>0.23</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Table 3: MEPS data: variable definitions and summary statistics. All hospitalization variables are for 1996. This table is from Prieger (2002).

Appendix B: summary results from model selected in empirical application

COPULA: 270 Clayton
MARGIN 1: Bernoulli
MARGIN 2: inverse Gaussian

EQUATION 1
Link function for mu.1: probit

Formula: y1 ~ privins + medicare + medicaid + hmopriv + hmomcare + hmomcaid +
privmcaid + privmcare + mcaremcaid + condn + priolist + exclhlth +
poorhlth + adlhelp + MidWest + South + West + female + s(age) +
Black + Hispanic + s(educ) + married + employed

Parametric coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | -1.850351 | 0.069180 | -26.747 | < 2e-16 *** |
| privins  | 0.188455  | 0.054576 | 3.453   | 0.000554 *** |
| medicare | 0.301470  | 0.092851 | 3.247   | 0.001167 **  |
| medicaid | 0.470590  | 0.078475 | 5.997   | 2.0e-09 ***  |
hmopriv 0.030331 0.042296 0.717 0.473299
hmomcare 0.026591 0.076898 0.346 0.729490
hmomcaid -0.037339 0.092341 -0.404 0.685947
privmcare -0.208284 0.080720 -2.580 0.009870**
privmcaid 0.355882 0.157468 2.260 0.023820 *
mcaremcaid -0.493067 0.111832 -4.409 1.04e-05 ***
condn 0.082802 0.010335 8.012 1.13e-15 ***
priolist 0.065037 0.019043 3.415 0.000637 ***
exclhlth -0.149980 0.039539 -3.793 0.000149 ***
poorhlth 0.219394 0.064518 3.400 0.000673 ***
adlhelp 0.338518 0.064002 5.289 1.23e-07 ***
MidWest 0.020925 0.047101 0.444 0.656850
South 0.009535 0.043148 0.220 0.826183
West -0.085877 0.048660 -1.765 0.077589 .
female 0.149305 0.032633 4.575 4.76e-06***
Black -0.028681 0.049880 -0.575 0.565286
Hispanic 0.084497 0.046239 1.827 0.067643 .
marrried 0.105973 0.034991 3.029 0.002457**
employed -0.163872 0.040729 -4.023 5.74e-05 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
edf Ref.df Chi.sq  p-value
s(age) 6.008 7.176 37.165 5.33e-06 ***
s(educ) 1.785 2.236 1.369 0.496
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

EQUATION 2

Link function for mu.2: log

Formula: y2 ~ privins + medicare + medicaid + hmopriv + hmomcare + hmomcaid +
privmcare + privmcaid + mcaremcaid + condn + priolist + exclhlth +
poorhlth + adlhelp + MidWest + South + West + female + s(age) +
Black + Hispanic + s(educ) + married + employed

Parametric coefficients:

| Estimate  | Std. Error | z value | Pr(>|z|) |
|-----------|------------|---------|---------|
| (Intercept) | 2.244987   | 0.169074 | 13.278  | < 2e-16 *** |
| privins   | 0.167173    | 0.129127 | 1.295   | 0.195444   |
| medicare  | -0.331687   | 0.185638 | -1.787  | 0.073979   |
medicaid -0.015643 0.156081 -0.100 0.920165
hmopriv -0.443685 0.106618 -4.161 3.16e-05***
hmomcare 0.523395 0.216804 2.414 0.015772 *
hmomcaid -0.080578 0.186132 -0.433 0.665082
privmcare -0.096168 0.177819 -0.541 0.588631
privmcaid 0.509183 0.242352 2.101 0.035640 *
mcaremcaid 0.143036 0.220478 0.649 0.516496
condn 0.004007 0.022184 0.181 0.856672
priolist 0.082884 0.043742 1.895 0.058118 .
exclhlth -0.047858 0.089880 -0.532 0.594402
poorhlth 0.434519 0.143721 3.023 0.002500**
adhelp 0.355209 0.145970 2.433 0.014956 *
MidWest -0.188903 0.111958 -1.687 0.091552 .
South -0.096852 0.105099 -0.922 0.356777
West -0.383978 0.112817 -3.404 0.000665***
female -0.484550 0.092887 -5.217 1.82e-07 ***
Black 0.366785 0.114616 3.200 0.001374 **
Hispanic 0.067553 0.093491 0.723 0.469948
married -0.123221 0.078603 -1.568 0.116966
employed -0.113859 0.081580 -1.396 0.162810

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:

<table>
<thead>
<tr>
<th>edf</th>
<th>Ref.df</th>
<th>Chi.sq</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>6.791</td>
<td>7.908</td>
<td>52.609</td>
</tr>
<tr>
<td>educ</td>
<td>1.000</td>
<td>1.000</td>
<td>0.007</td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

EQUATION 3

Link function for sigma2: log

Formula: -privins + medicare + medicaid + hmopriv + hmomcare + hmomcaid + privmcare + privmcaid + mcaremcaid + condn + priolist + exclhlth + poorhlth + adhelp + MidWest + South + West + female + s(age) + Black + Hispanic + s(educ) + married + employed

Parametric coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|
| (Intercept) | -1.437645 | 0.198799 | -7.232  | 4.77e-13  *** |
| privins    | 0.412330  | 0.156008 | 2.643   | 0.00822   ** |
medicare  0.004227 0.207456  0.020 0.98375
medicaid  0.365449 0.207283  1.763 0.07789 .
hmopriv  -0.270515 0.118284 -2.287 0.02220 *
hmomcare  0.505104 0.190237  2.655 0.00793 **
hmomcaid  0.218626 0.215001  1.017 0.30922
privmcare  0.049511 0.198518  0.249 0.80305
privmcaid -0.845025 0.326819 -2.586 0.00972 **
mcaremcaid -0.460742 0.261357 -1.763 0.07792 .
condn     0.032792 0.022851  1.435 0.15128
priolist  -0.072283 0.039988 -1.808 0.07067 .
exclhlth  0.047090 0.116003  0.406 0.68479
poorhlth  -0.088651 0.143464 -0.618 0.53662
adlhelp   -0.223425 0.132612 -1.685 0.09203 .
MidWest   -0.159945 0.124918 -1.280 0.20040
South     -0.167665 0.114324 -1.467 0.14249
West      -0.039339 0.127541 -0.308 0.75774
female   -0.100680 0.092344 -1.090 0.27559
Black     -0.235727 0.134401 -1.754 0.07945 .
Hispanic  0.018131 0.124514  0.146 0.88423
married  -0.045687 0.092462 -0.494 0.62123
employed  0.136274 0.109383  1.246 0.21282

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:

<table>
<thead>
<tr>
<th>edf</th>
<th>Ref.df</th>
<th>Chi.sq</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>s(age)</td>
<td>2.255</td>
<td>2.868</td>
<td>4.803</td>
</tr>
<tr>
<td>s(educ)</td>
<td>1.000</td>
<td>1.000</td>
<td>0.322</td>
</tr>
</tbody>
</table>

EQUATION 4

Link function for theta: log(- .)
Formula: -1

Parametric coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | -20.68 | 348.64 | -0.059 | 0.953 |

n = 14946  n.sel = 1346
sigma^2 = 0.293(0.207,0.42)
theta = -4.14e-08(-100,-4.14e-08)  tau = -2.07e-08(-0.98,-2.07e-08)
total edf = 87.8

Figure 6: Selection equation: smooth effects for age and years of education and associated 95% point-wise intervals obtained from the final model which is based on the 270° Clayton copula and probit and inverse Gaussian margins. The rug plot, at the bottom of each graph, shows the covariate values. The number in brackets in the y-axis of each plot’s caption represents the effective degrees of freedom of the respective smooth curve.


Heckman, J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5, 475–492.


