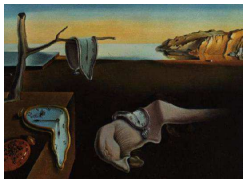


Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Copula Regression Spline Sample Selection Models: The R Package SemiParSampleSel

Małgorzata Wojtyś
University of Plymouth
Warsaw University of Technology

Giampiero Marra
University College London

Rosalba Radice
Birkbeck University of London

Abstract

Sample selection models deal with the situation in which an outcome of interest is observed for a restricted non-randomly selected sample of the population. The estimation of these models is based on a binary equation, which describes the selection process, and an outcome equation, which is used to examine the substantive question of interest. Classic sample selection models assume a priori that continuous covariates have a linear or pre-specified non-linear relationship to the outcome, and that the distribution linking the two equations is bivariate normal.

We introduce the R package **SemiParSampleSel** which implements copula regression spline sample selection models. The proposed implementation can deal with non-random sample selection, non-linear covariate-response relationships, and non-normal bivariate distributions between the model equations. We provide details of the model and algorithm and describe the implementation in **SemiParSampleSel**. The package is illustrated using simulated and real data examples.

Keywords: copula, non-random sample selection, penalized regression spline, selection bias, R.

1. Introduction

The sample selection model was introduced by Gronau (1974), Lewis (1974) and Heckman (1976) to deal with the situation in which the observations available for statistical analysis are not from a random sample of the population; the model was discussed by Heckman (1990) among others. This issue occurs when individuals have selected themselves into (or out of) the sample based on a combination of observed and unobserved characteristics. Estimates based on models that ignore such a non-random selection may be biased and inconsistent.

To fix ideas, let us consider the RAND Health Insurance Experiment (RHIE), a study con-

ducted in the United States between 1974 and 1982 (Newhouse 1999) which will also be analyzed in Section 5. The aim was to quantify the relationship between several socio-economic characteristics and annual health expenditures. Non-random selection arises if the sample consisting of individuals who used health care services differ in important characteristics from the sample of individuals who did not use them. If the link between the decision to use the services and health expenditure is through observables, then selection bias can be avoided by accounting for these variables. However, if the link is through unobservables as well then inconsistent parameter estimates are obtained when using a classic univariate equation method. There are two more aspects that may complicate modeling the relationship between covariates and annual health expenditure. Variables such as age and education are likely to have a non-linear relationship to both decision to use health services and amount to spend on them; this is because they embody productivity and life-cycle effects that are likely to have non-linear effects. Imposing a priori a linear relationship (or non-linear by simply using quadratic polynomials, for example) could mean failing to capture the true more complex relationships. Finally, the (often criticized) assumption of bivariate normality (employed in many sample selection models) between decision to use health services and expenditure may be too restrictive for applied work and is typically made for mathematical convenience.

The literature on sample selection models is vast and many variants of such models have been proposed. Chib, Greenberg, and Jeliazkov (2009) and Wiesenfarth and Kneib (2010) introduced two estimation methods to deal with non-linear covariate effects. Specifically, the approach of the former authors is based on Markov chain Monte Carlo simulation techniques and uses a simultaneous equation system that incorporates Bayesian versions of penalized smoothing splines. The latter 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. Recently, Marra and Radice (2013) proposed a frequentist counterpart which has the advantage of being computationally fast and can especially appeal to practitioners already familiar with traditional frequentist techniques.

Under the assumption of bivariate normality Heckman (1979) proposed a two-step estimator. However because the estimator is inconsistent under distributional misspecification various methods that relax the assumption of normality have been proposed over the years; these include semiparametric (e.g., Gallant and Nychka 1987; Powell, Stock, and Stoker 1989; Ahn and Powell 1993; Lee 1994a,b; Powell 1994; Andrews and Schafgans 1998; Newey 2009) and nonparametric methods (e.g., Das, Newey, and Vella 2003; Lee 2008; Chen and Zhou 2010). Another way to relax the normality assumption is to use non-normal parametric distributions. Recently, Marchenko and Genton (2012) and Ding (2014) extended the sample selection model to deal with heavy tailedness by using the bivariate Student-t distribution. Another parametric method, which includes as a subcase the above mentioned Student-t approach, is copula modeling. This allows for a great deal of flexibility in specifying the joint distribution of the selection and outcome equations (e.g., Smith 2003; Prieger 2002; Hasebe and Vijverberg 2012; Schwiebert 2013).

In summary, the numerous estimation approaches that deal with the assumption of normality in the sample selection model can be divided into two large groups: semi/non-parametric and flexible parametric estimators. The first relaxes the assumption of bivariate normality by using a general bivariate density function, whereas the second offers the possibility of replacing bivariate normality with an alternative parametric stochastic structure. There are advantages and disadvantages to both approaches (semi/non-parametric and flexible parametric). The

strongest point of the semi/non-parametric approach is the property of maintaining consistency of such estimators even disposing, in part or altogether, of distributional assumptions. In some cases, simplified versions of these methods are easy to implement (e.g., Das *et al.* 2003). However, these estimators do have shortcomings. Specifically, semi/non-parametric methods are usually restricted when it comes to including a large set of covariates in the model and the resulting estimates are inefficient relatively to fully parametrized models (e.g., Bhat and Eluru 2009). To date, packages implementing semi/non-parametric procedures are CPU-intensive and the set of options provided is often quite limited. In addition, convergence problems are likely to occur when using models which include, for instance, many discrete variables and interactions. As for the parametric approach, many scholars agree upon its greater computational feasibility as compared to semi/non-parametric approaches, which allows for the use of familiar tools such as maximum likelihood without requiring simulation methods or numerical integration. As pointed out by Smith (2003), maximum likelihood techniques allow for the simultaneous estimation of all model parameters, and such methods, if the usual regularity conditions hold and the model is correctly specified, ensure consistent, efficient and asymptotically normal estimators. In addition, when using copulas the practitioner has the possibility of a piece-wise model specification. This is because marginal distributions are not constrained to belong to the same family of the chosen bivariate copula distribution. Moreover, Genius and Strazzera (2008) argue that copula modeling allows for direct estimation of the dependence structure in the sample selection model while non-parametric methods do not. However, a crucial point stands on the correct specification of these models; maximum likelihood estimators are not consistent when the distributional assumption is not correct. Also, testing the distributional assumption is not straightforward. In the context of Heckman's two-step estimator, Lee (1982, 1984) presented misspecification tests based on bivariate Edgeworth expansions. Recently, Montes-Rojas (2011) proposed a similar methodology for testing normality in sample selection models. Specifically, he proposed Lagrange multiplier and Neyman's $C(\alpha)$ tests for the marginal normality and linearity of the conditional expectation of the error terms for the two-step estimator. Although these tests provided encouraging results, more research is necessary to construct likelihood ratio and Wald tests. As for the maximum likelihood approach, to date, all that can be done is a posteriori model selection using, for instance, traditional information criteria.

Some of the methods described above are implemented in popular software packages like SAS (SAS Institute Inc. 2011), Stata (StataCorp 2011) and R (R Development Core Team 2013). For example, the conventional Heckman sample selection model can be fitted in SAS using the **proc qlim** and in Stata using **heckman**. The non-parametric method by Lee (2008) can be employed using the Stata package **leebounds** and the bivariate Student-t distribution Heckman model using **heckt**. In R the sample selection packages are **sampleSelection** (Toomet and Henningsen 2008), **bayesSampleSelection** (Wiesenfarth and Kneib 2010), available from the first author's webpage, **ssmrob** (Zhelonkin, Genton, and Ronchetti 2013) and **SemiParBIVProbit** (Marra and Radice 2014). **sampleSelection** and **bayesSampleSelection** make the assumption of bivariate normality between the model equations. **sampleSelection** and **ssmrob** assume a priori that continuous regressors have linear or pre-specified non-linear relationships to the responses, whereas **ssmrob** relaxes the assumption of bivariate normality by providing a robust two-stage estimator of Heckman's approach. **sampleSelection** and **SemiParBIVProbit** support binary responses for the outcome equation, with the latter assuming bivariate normality but allowing for non-linear covariate effects. It is worth mentioning the

packages **censReg** (Henningsen 2012) which deals with censored dependent variables, and **intReg** (Toomet 2012) which implements interval regression models.

We introduce the R package **SemiParSampleSel** (Marra, Radice, and Wojtyś 2014) to deal simultaneously with non-random sample selection, non-linear covariate effects and non-normal bivariate distribution between the model equations. The problem of non-random sample selection is addressed using the conventional system of two equations: a binary selection equation determining whether a particular statistical unit will be available in the outcome equation. Covariate-response relationships are flexibly modeled using a spline approach whereas non-normal distributions are dealt with by using copula functions. The core algorithm is based on the penalized maximum likelihood framework proposed by Marra and Radice (2013) for the bivariate normal case. We further extend this by allowing for non-normal bivariate distributions using copulas. Note that if a normal copula is chosen and linear or pre-speficified covariate effects are assumed then, similarly to **sampleSelection**, **SemiParSampleSel** fits the classical Heckman sample selection model using the two-step and maximum likelihood approaches. We believe that when a practitioner faces a non-normality problem in the sample selection model, the option offered by the copula approach is worth pursuing whenever the accuracy of structural parameter estimates is the priority. Well motivated conjectures on the stochastic structure of the phenomenon may lead to specifications better fitting the data than the traditional sample selection model. Moreover, using different assumptions on the bivariate distribution, as it happens with copulas, allows the specification of the conditional mean to remain intact. This is crucial to the interpretability of the model parameters.

The paper is organized as follows. In the next section, we present the model, describe the algorithm used to estimate the model parameters and discuss inferential and numerical issues. Section 3 provides details on the implementation of the model in **SemiParSampleSel**. In Section 4, we illustrate the usage of the package on various simulated data sets, whereas Section 5 is devoted to an illustrative real data example.

2. Methodological and algorithmic details

2.1. Model definition

In the sample selection problem, our aim is to fit a regression model when some observations of the outcome variable are missing not at random. Thus assuming that y_{2i}^*, for $i = 1, \ldots, n$, is a random variable of our primary interest, we can represent the random sample using a pair of variables (y_{1i}, y_{2i}), such that $y_{i1} \in \{0, 1\}$ and $y_{2i} = y_{2i}^* y_{1i}$. The variable y_{1i} governs whether or not an observation on the variable of primary interest is generated and the unobserved values of the variable of interest are coded as 0. In the model statement, a latent continuous variable y_{1i}^* such that $y_{1i} = \mathbf{1}(y_{1i}^* > 0)$ is used, where $\mathbf{1}$ is the indicator function. Let F_i denote the joint cumulative distribution function (cdf) of (y_{1i}^*, y_{2i}^*) and let F_{1i} and F_{2i} be the marginal cdf's pertaining to y_{1i}^* and y_{2i}^*, respectively. We assume normality of the marginal distributions whilst the relationship between them is modeled using a copula approach. That is, $y_{1i}^* \sim \mathcal{N}(\mu_{1i}, 1)$ (which yields a probit model for y_{1i}) and $y_{2i}^* \sim \mathcal{N}(\mu_{2i}, \sigma)$, where $\mu_{1i}, \mu_{2i} \in \mathbb{R}$ are linear predictors defined in the next section and $\sigma > 0$, the standard deviation, is unknown. F_{1i} relates to the selection equation and F_{2i} to the outcome equation. The model is then

defined by using the copula representation

$$F_i(y_1^*, y_2^*) = C(F_{1i}(y_1^*), F_{2i}(y_2^*); \theta), \tag{1}$$

for some two-place function C which is unique, where θ is an association parameter measuring the dependence between the two marginal cdf's. In the package, the families implemented are normal, Clayton, Joe, Frank, Gumbel, Farlie-Gumbel-Morgenstern (FGM), and Ali-Mikhail-Haq (AMH); these are listed in Table 1. Rotations by 90, 180 and 270 degrees for Clayton, Joe and Gumbel can be obtained using the results reported in Brechmann and Schepsmeier (2013); these will be available in future releases. As it can be seen from Table 1, θ may be difficult to interpret in some cases. To this end, we can use the Kendall's τ coefficient which is a measure of association that lies in the customary range $[-1, 1]$. This is generally defined as $\tau = P\left((y_{11}^* - y_{12}^*)(y_{21}^* - y_{22}^*) > 0\right) - P\left((y_{11}^* - y_{12}^*)(y_{21}^* - y_{22}^*) < 0\right)$ for independent pairs (y_{1j}^*, y_{2j}^*), $j = 1, 2$, that are copies of (y_1^*, y_2^*). For each copula there exists a relation between θ and τ, as shown in Table 1. Testing the null hypothesis of absence of selection bias is an important issue as if the null hypothesis cannot be rejected then joint estimation of the two model equations can be avoided and consistent estimates for the parameters of the equation of interest can be obtained using a univariate equation model. In the context of the copula regression spline sample selection model, the absence of sample selection bias is equivalent to $\theta = 0$ which in turn is equivalent to the condition that the Kendall's τ coefficient equals 0. Thus the null hypothesis can, for instance, be tested by checking whether the confidence interval for the Kendall's τ includes 0. The problem of testing for sample selection bias is further addressed in Section 4.3. For a comprehensive introduction to the theory of copulas and their properties see the monographs of Nelsen (2006) and Joe (1997).

Copula likelihood

The log-likelihood function for the sample selection model can be expressed as a sum over two disjoint subsets of the sample: one for the observations with a missing value of the response of interest and the other for the remaining observations. In the first case, the likelihood for the i^{th} observation takes the simple form of $\mathbb{P}(y_{1i} = 0)$, which is equivalent to $F_{1i}(0)$. In the second case, the joint likelihood can be expressed, using a multiplication rule, as $\mathbb{P}(y_{1i}^* > 0) f_{2|1,i}(y_{2i}|y_{1i}^* > 0)$, where $f_{2|1,i}$ denotes the probability density function of y_{2i}^* given $y_{1i}^* > 0$. After substituting the conditional density $f_{2|1,i}(y_{2i}|y_{1i}^* > 0)$ by $\frac{1}{P(y_{1i}^* > 0)} \frac{\partial}{\partial y_2} \left(F_{2i}(y_2) - F_i(0, y_2)\right)\big|_{y_2 \to y_{2i}}$, we obtain the log-likelihood

$$\ell = \sum_{i=1}^{n} \left\{ (1 - y_{1i}) \log F_{1i}(0) + y_{1i} \log \left(f_{2i}(y_{2i}) - \frac{\partial}{\partial y_2} F_i(0, y_2)\big|_{y_2 \to y_{2i}} \right) \right\}.$$

Using (1), we then have

$$\ell = \sum_{i=1}^{n} \left\{ (1 - y_{1i}) \log F_{1i}(0) + y_{1i} \log \left(f_{2i}(y_{2i}) \left(1 - z_i\right) \right) \right\}, \tag{2}$$

where $z_i = \frac{\partial}{\partial v} C(F_{1i}(0), v; \theta)\big|_{v \to F_{2i}(y_{2i})}$. The normality of margins implies that $F_{1i}(0) = \Phi(-\mu_{1i})$ and $f_{2i}(y_{2i}) = \sigma^{-1} \phi\left((y_{2i} - \mu_{2i})\sigma^{-1}\right)$, where Φ and ϕ are used throughout to denote the standard normal distribution and density functions, respectively.

Copula	$C(u,v;\theta)$	Parameter space	Kendall's τ
Normal	$\Phi_2\left(\Phi^{-1}(u),\Phi^{-1}(v);\theta\right)$	$\theta\in[-1,1]$	$\frac{2}{\pi}\arcsin(\theta)$
Clayton	$\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta}$	$\theta\in(0,\infty)$	$\frac{\theta}{\theta+2}$
Joe	$1-\left[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}\right]^{1/\theta}$	$\theta\in(1,\infty)$	$1+\frac{4}{\theta^2}D_2(\theta)$
Frank	$-\theta^{-1}\log\left[1+(e^{-\theta u}-1)(e^{-\theta v}-1)/(e^{-\theta}-1)\right]$	$\theta\in\mathbb{R}\setminus\{0\}$	$1-\frac{4}{\theta}\left[1-D_1(\theta)\right]$
Gumbel	$\exp\left\{-\left[(-\log u)^{\theta}+(-\log v)^{\theta}\right]^{1/\theta}\right\}$	$\theta\in[1,\infty)$	$1-\frac{1}{\theta}$
FGM	$uv\left[1+\theta(1-u)(1-v)\right]$	$\theta\in[-1,1]$	$\frac{2}{9}\theta$
AMH	$uv/\left[1-\theta(1-u)(1-v)\right]$	$\theta\in[-1,1]$	$1-\frac{2}{3\theta^2}\left(\theta+(1-\theta)^2\ln(1-\theta)\right)$

Table 1: Families of copulas implemented in **SemiParSampleSel**, with corresponding parameter range of the association parameter θ and relation between Kendall's τ and θ. $\Phi_2(\cdot,\cdot;\theta)$ denotes the cumulative distribution function of a standard bivariate normal distribution with correlation coefficient θ. $D_1(\theta)=\frac{1}{\theta}\int_0^{\theta}\frac{t}{\exp(t)-1}dt$ is the Debye function and $D_2(\theta)=\int_0^1 t\log(t)(1-t)^{\frac{2(1-\theta)}{\theta}}dt$.

Linear predictor specification

We assume that the expected values μ_{1i} and μ_{2i} of variables y_{1i}^* and y_{2i}^*, respectively, are linked with the predictors, i.e., $\mu_{1i}=\eta_{1i}$ and $\mu_{2i}=\eta_{2i}$, where the linear predictor of the selection equation can be written as

$$\eta_{1i}=\mathbf{u}_{1i}^{\mathsf{T}}\boldsymbol{\alpha}_1+\sum_{k_1=1}^{K_1}s_{1k_1}(z_{1k_1i}),\quad i=1,\ldots n, \tag{3}$$

and that of the outcome equation as

$$\eta_{2i}=\mathbf{u}_{2i}^{\mathsf{T}}\boldsymbol{\alpha}_2+\sum_{k_2=1}^{K_2}s_{2k_2}(z_{2k_2i}),\quad i\in\{j:\,y_{1j}=1\}, \tag{4}$$

where vector $\mathbf{u}_{1i}^{\mathsf{T}}=(1,u_{12i},\ldots,u_{1P_1i})$ is the i^{th} row of $\mathbf{U}_1=(\mathbf{u}_{11},\ldots,\mathbf{u}_{1n})^{\mathsf{T}}$, the $n\times P_1$ model matrix containing P_1 parametric model components (e.g., intercept, dummy and categorical variables), $\boldsymbol{\alpha}_1$ is a parameter vector, and the s_{1k_1} are unknown smooth functions of the K_1 continuous covariates z_{1k_1i}. Our implementation supports varying coefficients' models, obtained by multiplying one or more smooth terms by some predictor(s) (Hastie and Tibshirani 1993), and smooth functions of two or more (e.g., spatial) covariates as described in Wood (2006). Similarly, $\mathbf{u}_{2i}^{\mathsf{T}}=(1,u_{22i},\ldots,u_{2P_2i})$ is the i^{th} row vector of the $n_s\times P_2$ model matrix $\mathbf{U}_2=(\mathbf{u}_{21},\ldots,\mathbf{u}_{2n_s})^{\mathsf{T}}$, where n_s is the size of the selected sample, $\boldsymbol{\alpha}_2$ is a parameter vector, and the s_{2k_2} are unknown smooth terms of the K_2 continuous regressors z_{2k_2i}. The smooth functions are subject to the centering (identifiability) constraint $\sum_i s_{vk_v}(z_{vk_vi})=0$ for $v=1,2$, $k_v=1,\ldots,K_v$ (Wood 2006).

The smooth functions are represented using regression splines, where, in the one-dimensional case, a generic $s_k(z_{ki})$ is approximated by a linear combination of known spline basis functions, $b_{kj}(z_{ki})$, and regression parameters, β_{kj}, i.e., $s_k(z_{ki})=\sum_{j=1}^{J_k}\beta_{kj}b_{kj}(z_{ki})=\boldsymbol{\beta}_k^{\mathsf{T}}\mathbf{B}_k(z_{ki})$, where J_k is the number of spline bases used to represent s_k, $\mathbf{B}_k(z_{ki})$ is the i^{th} vector of dimension J_k containing the basis functions evaluated at the observation z_{ki}, i.e., $\mathbf{B}_k(z_{ki})=$

$\{b_{k1}(z_{ki}), b_{k2}(z_{ki}), \ldots, b_{kJ_k}(z_{ki})\}^\mathsf{T}$, and $\boldsymbol{\beta}_k$ is the corresponding parameter vector. The subscript indicating which equation each smooth component belongs to has been suppressed for simplicity. Calculating $\mathbf{B}_k(z_{ki})$ for each i yields J_k curves (encompassing different degrees of complexity) which multiplied by some real valued parameter vector $\boldsymbol{\beta}_k$ and then summed will give a (linear or non-linear) estimate for $s_k(z_k)$ (see, for instance, Marra and Radice (2010) for a more detailed overview). Basis functions should be chosen to have convenient mathematical and numerical properties. B-splines, cubic regression and low rank thin plate regression splines are supported in our implementation (see Wood (2006) for full details on these spline bases). The cases of smooths of more than one variable and of varying-coefficient smooth functions follow a similar construction. For instance, in the case of a smooth of two variables z_{1i} and z_{2i} we would have $s_{12}(z_{1i}, z_{2i}) = \sum_{j=1}^{J_{12}} \beta_{12j} b_{12j}(z_{1i}, z_{2i})$, where the specification of the basis functions depends again on the kind of spline chosen (Wood 2006). Linear predictors (3) and (4) can, therefore, be written as $\eta_{vi} = \mathbf{u}_{vi}^\mathsf{T}\boldsymbol{\alpha}_v + \mathbf{B}_{vi}^\mathsf{T}\boldsymbol{\beta}_v$, where $\mathbf{B}_{vi}^\mathsf{T} = \left\{\mathbf{B}_{v1}(z_{v1i})^\mathsf{T}, \ldots, \mathbf{B}_{vK_v}(z_{vK_vi})^\mathsf{T}\right\}$ and $\boldsymbol{\beta}_v^\mathsf{T} = (\boldsymbol{\beta}_{v1}^\mathsf{T}, \ldots, \boldsymbol{\beta}_{vK_v}^\mathsf{T})$, for $v = 1, 2$. In principle, the parameters of the sample selection model are identified even if the same regressors appear in both linear predictors (e.g., Wiesenfarth and Kneib 2010). However, better estimation results are generally obtained when the set of regressors in the selection equation contains at least one or more regressors (usually known as exclusion restrictions) that are not included in the outcome equation (e.g., Marra and Radice 2013).

2.2. Estimation approach

Denote the log-likelihood function as $\ell(\boldsymbol{\delta})$, where $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \sigma, \theta)$ and $\boldsymbol{\delta}_v^\mathsf{T} = (\boldsymbol{\alpha}_v^\mathsf{T}, \boldsymbol{\beta}_v^\mathsf{T})$, for $v = 1, 2$. Given the flexible structure of the linear predictors considered here, unpenalized estimation can result in smooth term estimates that are too rough to produce practically useful results. This issue is dealt with by using the penalty term $\sum_{v=1}^{2}\sum_{k_v=1}^{K_v} \lambda_{vk_v} \int s''_{vk_v}(z_{vk_v})^2 dz_{vk_v}$ which measures the (typically, second-order) roughness of the smooth terms in the model. For a smooth of two variables generically written as $s_{12}(z_1, z_2)$ and represented using thin plate regression splines the integral would look like $\int\int \left(\frac{\partial^2 s_{12}}{\partial z_1^2}\right)^2 + 2\left(\frac{\partial^2 s_{12}}{\partial z_1 \partial z_2}\right)^2 + \left(\frac{\partial^2 z_{12}}{\partial z_2^2}\right)^2 dz_1 dz_2$, where the subscripts have been dropped to avoid clutter. The λ_{vk_v} are smoothing parameters controlling the trade-off between fit and smoothness. Since regression splines are linear in their model parameters, the overall penalty can be written as $\boldsymbol{\beta}^\mathsf{T}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$ where $\boldsymbol{\beta}^\mathsf{T} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T})$, $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{v=1}^{2}\sum_{k_v=1}^{K_v} \lambda_{vk_v}\mathbf{S}_{vk_v}$ and the $\mathbf{S}_{vk_v}$ are positive semi-definite known square matrices expanded with zeros everywhere except for the elements which correspond to the coefficients of the vk_v^{th} smooth term. Because of the restrictions on the values that θ can take, we use a proper transformation of it, θ^*, in order to avoid the use of a constraint when estimating this parameter (see Table 2 for the list of transformations used). Similarly, since σ can only take positive real values, we use $\sigma^* = \log(\sigma)$. So, in optimization, we use $\boldsymbol{\delta}_*^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \sigma^*, \theta^*) \in \mathbb{R}^p$, where p is the total number of parameters. Therefore, the function to maximize is

$$\ell_p(\boldsymbol{\delta}_*) = \ell(\boldsymbol{\delta}_*) - \frac{1}{2}\boldsymbol{\beta}^\mathsf{T}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}. \tag{5}$$

Given a parameter vector value for $\hat{\boldsymbol{\lambda}}^\mathsf{T} = (\hat{\lambda}_{1k_1}, \ldots, \hat{\lambda}_{1K_1}, \hat{\lambda}_{2k_2}, \ldots, \hat{\lambda}_{2K_2})$, we seek to maximize (5). The issues with this maximization problem are that $\ell_p(\boldsymbol{\delta}_*)$ is not globally concave and the penalized Hessian may be non-positive definite on some occasions (Toomet and Henningsen 2008; Marra and Radice 2013). To this end, we use a trust region approach which is typically believed to be more stable than its line-search counterparts, particularly for functions that are,

for example, non-concave and/or exhibit regions that are close to flat (Nocedal and Wright 2006, Chapter 4). Let a be an iteration index. Intuitively speaking, line search methods choose a direction to move from, say, m_a to m_{a+1} and find the distance along that direction which gives the best improvement in the objective function. If the function is, for instance, non-convex or has long plateaus, the optimizer may search far away from m_a but choose an m_{a+1} that is close to m_a and that offers marginal improvement in the objective function. In some cases, the function will be evaluated so far away from m_a that it will not be finite and the algorithm will fail. Trust region methods choose a maximum distance for the move from m_a to m_{a+1}, defining a "trust region" around m_a that has a radius of that maximum distance, and then let a candidate for m_{t+1} be the minimum of a quadratic approximation of the objective function. Since points outside of the trust region are not considered, the algorithm never runs too far and/or too fast from the current iteration. The trust region is shrunken if the proposed point in the region is worse/not better than the current point. The new problem with smaller region is then solved. If a point close to the boundary of the trust region is accepted and it gives a large enough improvement in the function then the region for the next iteration is expanded. If a point along a search path causes the objective function to be undefined or indeterminate, most implementations of line search methods will fail and user intervention is required. In a trust region approach, the search for m_{t+1} is always a solution to the trust region problem; if the function at the proposed m_{t+1} is not finite or not better than the value at m_t, then the proposal is rejected and the trust region shrunken. Finally, a line search approach requires repeated estimation of the objective function, while trust region methods evaluate the objective function only after solving the trust region problem. Hence, trust region methods can be considerably faster when the objective function is expensive to compute. Full details can be found in (Nocedal and Wright 2006, Chapter 4).

In practice, we adopt a trust region Newton method (Nocedal and Wright 2006, Chapter 4) which, in our case, solves the problem

$$\min_{\mathbf{p}} \ \breve{\ell}_p(\boldsymbol{\delta}_*^{[a]}) \stackrel{\text{def}}{=} -\left\{ \ell_p(\boldsymbol{\delta}_*{}^{[a]}) + \mathbf{p}^\mathsf{T}(\mathbf{g}^{[a]} - \mathbf{S}^*_{\hat{\boldsymbol{\lambda}}}\hat{\boldsymbol{\delta}}^{[a]}) + \frac{1}{2}\mathbf{p}^\mathsf{T}(\boldsymbol{\mathcal{H}}^{[a]} - \mathbf{S}^*_{\hat{\boldsymbol{\lambda}}})\mathbf{p} \right\} \ \text{ so that } \ \|\mathbf{p}\| \leq r^{[a]},$$
$$\boldsymbol{\delta}_*{}^{[a+1]} = \arg\min_{\mathbf{p}} \ \breve{\ell}_p(\boldsymbol{\delta}_*^{[a]}) + \boldsymbol{\delta}_*^{[a]},$$

where $\|\cdot\|$ denotes the Euclidean norm and $r^{[a]}$ represents the radius of the trust region. $\mathbf{S}^*_{\hat{\boldsymbol{\lambda}}}$ is the overall block-diagonal penalty matrix which is made up of $\hat{\lambda}_{vk_v}\mathbf{S}_{vk_v}$ and $\mathbf{0}$ components. After dropping the iteration index, the score vector $\mathbf{g}$ is defined by two subvectors $\mathbf{g}_1 = \partial\ell(\boldsymbol{\delta}_*)/\partial\boldsymbol{\delta}_1$ and $\mathbf{g}_2 = \partial\ell(\boldsymbol{\delta}_*)/\partial\boldsymbol{\delta}_2$ and two scalars $g_3 = \partial\ell(\boldsymbol{\delta}_*)/\partial\sigma^*$ and $g_4 = \partial\ell(\boldsymbol{\delta}_*)/\partial\theta^*$, while the Hessian matrix has a 4×4 matrix block structure with $(r,h)^{th}$ element $\boldsymbol{\mathcal{H}}_{r,h} = \partial^2\ell(\boldsymbol{\delta}_*)/\partial\boldsymbol{\delta}_r\partial\boldsymbol{\delta}_h^\mathsf{T}$, $r,h = 1,\ldots,4$, where $\boldsymbol{\delta}_3 = \sigma^*$ and $\boldsymbol{\delta}_4 = \theta^*$. The expressions of $\mathbf{g}$ and $\boldsymbol{\mathcal{H}}$ for all copulas are given in Appendix A; these have been derived analytically and verified using numerical derivatives.

At each iteration of the algorithm, $\breve{\ell}_p(\boldsymbol{\delta}_*^{[a]})$ is minimized subject to the constraint that the solution falls within a trust region with radius $r^{[a]}$. The proposed solution is then accepted or rejected and the trust region expanded or shrunken based on the ratio between the improvement in the objective function when going from $\boldsymbol{\delta}_*{}^{[a]}$ to $\boldsymbol{\delta}_*{}^{[a+1]}$ and that predicted by the quadratic approximation. Note that, near the solution, the trust region Newton algorithm typically behaves as a Newton algorithm.

Copula	θ^*
Normal	$\tanh^{-1}(\theta)$
Clayton	$\log(\theta - \epsilon)$
Frank	$\theta - \epsilon$
Joe	$\log(\theta - 1 - \epsilon)$
Gumbel	$\log(\theta - 1)$
FGM	$\tanh^{-1}(\theta)$
AMH	$\tanh^{-1}(\theta)$

Table 2: Transformations, θ^*, of the dependence parameter, θ, used in optimization. Quantity ϵ is set to the machine smallest positive floating-point number multiplied by 10^6 and is used to ensure that the dependence parameters lie in the ranges reported in Table 1.

Smoothing parameter estimation

Multiple smoothing parameter estimation by direct grid search optimization of, for instance, a prediction error criterion can be computationally expensive, especially if the model has more than one smooth term per equation. This section briefly describes the automatic approach employed by Marra and Radice (2013) to estimate $\boldsymbol{\lambda}$. Note that joint estimation of $\boldsymbol{\delta}_*$ and $\boldsymbol{\lambda}$ via maximization of (5) would clearly lead to overfitting since the highest value for $\ell_p(\boldsymbol{\delta}_*)$ would be obtained when $\boldsymbol{\lambda} = \mathbf{0}$. Parameter vector $\hat{\boldsymbol{\lambda}}$ is the solution to the problem

$$\text{minimize} \quad \frac{1}{n_*}\|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\delta}_*)\|^2 - 1 + \frac{2}{n_*}\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}}) \quad \text{w.r.t.} \quad \boldsymbol{\lambda}, \tag{6}$$

where $\sqrt{\mathbf{W}}$ is a weight non-diagonal matrix square root, $\mathbf{z}_i$ is the 4-dimensional vector $\mathbf{z}_i = \mathbf{X}_i\boldsymbol{\delta}_*^{[a]} + \mathbf{W}_i^{-1}\mathbf{d}_i$, $\mathbf{d}_i = \{\partial\ell(\boldsymbol{\delta}_*)_i/\partial\eta_{1i}, \partial\ell(\boldsymbol{\delta}_*)_i/\partial\eta_{2i}, \partial\ell(\boldsymbol{\delta}_*)_i/\partial\eta_{3i}, \partial\ell(\boldsymbol{\delta}_*)_i/\partial\eta_{4i}\}^\mathsf{T}$, $\eta_{3i} = \sigma^*$, $\eta_{4i} = \theta^*$, $\mathbf{W}_i$ is a 4×4 matrix with $(r,h)^{th}$ element $(\mathbf{W}_i)_{rh} = -\partial^2\ell(\boldsymbol{\delta}_*)_i/\partial\eta_{ri}\partial\eta_{hi}$, $r, h = 1, \ldots, 4$, $\mathbf{X}_i = \text{diag}\left\{\left(\mathbf{u}_{1i}^\mathsf{T}, \mathbf{B}_{1i}^\mathsf{T}\right), \left(\mathbf{u}_{2i}^\mathsf{T}, \mathbf{B}_{2i}^\mathsf{T}\right), 1, 1\right\}$, $n_* = 4n$, $\mathbf{A}_{\boldsymbol{\lambda}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}}^*)^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}$ is the hat matrix, and $\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}})$ the estimated degrees of freedom (*edf*) of the penalized model. The iteration index has been dropped to avoid clutter. Note that the working linear model quantities are constructed for a given estimate of $\boldsymbol{\delta}_*$. Iteration (6) will produce an updated estimate for $\boldsymbol{\lambda}$ which will then be used to obtain a new parameter vector estimate for $\boldsymbol{\delta}_*$. The two steps, one for $\boldsymbol{\delta}_*$ and the other for $\boldsymbol{\lambda}$, are iterated until convergence.

To speed up the algorithm the sparse structure of $\mathbf{W}$ is exploited. This allows us to set up the working linear model quantities $\mathbf{W}^{-1}\mathbf{d}$, $\sqrt{\mathbf{W}}\mathbf{z}$ and $\sqrt{\mathbf{W}}\mathbf{X}$ in $O(n_*(m+2))$ rather than $O(n_*^2(m+2))$ operations, where m is the number of columns of $\mathbf{X}$. Furthermore, since σ^* and θ^* are not penalized, the working linear model is constructed for fixed values of these two parameters. In this way, the computational load and storage demand of the algorithm is reduced considerably since in this case $n_* = 2$. The *leapfrog* algorithm described in Appendix C of Marra and Radice (2013) is employed to achieved this.

2.3. Confidence intervals, variable selection and model selection

Inferential theory for penalized estimators is complicated by the presence of smoothing penalties which undermines the usefulness of classic frequentist results for practical modeling.

As shown in Marra and Radice (2013), reliable pointwise confidence intervals for the terms

of a regression spline sample selection model can be constructed using

$$\boldsymbol{\delta}_*|\mathbf{y} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\delta}}_*, \mathbf{V}_{\boldsymbol{\delta}_*}), \tag{7}$$

where $\mathbf{y}$ refers to the response vectors, $\hat{\boldsymbol{\delta}}_*$ is an estimate of $\boldsymbol{\delta}_*$ and $\mathbf{V}_{\boldsymbol{\delta}_*} = (-\boldsymbol{\mathcal{H}} + \mathbf{S}^*_{\hat{\boldsymbol{\lambda}}})^{-1}$. The structure of $\mathbf{V}_{\boldsymbol{\delta}_*}$ is such that it includes both a bias and variance component in a frequentist sense, which is why such intervals exhibit close to nominal coverage probabilities (Marra and Wood 2012). Given (7), confidence intervals for linear and non-linear functions of the model parameters can be easily obtained. For instance, for a generic $\hat{s}_k(z_{ki})$ these can be obtained using

$$\hat{s}_k(z_{ki}) \dot{\sim} \mathcal{N}(s_k(z_{ki}), \mathbf{B}_k(z_{ki})^\mathsf{T} \mathbf{V}_{\boldsymbol{\delta}_{*k}} \mathbf{B}_k(z_{ki})), \tag{8}$$

where $\mathbf{V}_{\boldsymbol{\delta}_{*k}}$ is the submatrix of $\mathbf{V}_{\boldsymbol{\delta}_*}$ corresponding to the regression spline parameters associated with k^{th} function. Intervals for non-linear functions of the estimated model coefficients (i.e., σ, θ and Kendall's τ) can be conveniently obtained by simulation from the posterior distribution of $\boldsymbol{\delta}_*$. As for the parametric model components, using (7) is equivalent to using classic likelihood results because such terms are not penalized.

Result (8) can be used to find intervals for $s_k(z_{ki})$ for each k and i but cannot be used to test whether smooth terms are equal to zero (e.g., Ruppert, Wand, and Carroll 2003, Chapter 6). For this purpose, p-values or shrinkage methods may be employed. To test smooth components for equality to zero we use the results by Wood (2013). Define $\hat{\mathbf{s}}_k = \mathbf{B}_k(\mathbf{z}_k)\hat{\boldsymbol{\beta}}_k$, where $\mathbf{B}_k(\mathbf{z}_k)$ denotes a full column rank matrix and $\mathbf{z}_k = (z_{k1}, z_{k2}, \ldots, z_{kn})^\mathsf{T}$, and $\mathbf{V}_{\mathbf{s}_k} = \mathbf{B}_k(\mathbf{z}_k)\mathbf{V}_{\boldsymbol{\delta}_{*k}}\mathbf{B}_k(\mathbf{z}_k)^\mathsf{T}$. It is then possible to obtain approximate p-values for testing smooth components for equality to zero based on

$$T_{r_k} = \hat{\mathbf{s}}_k^\mathsf{T} \mathbf{V}_{\mathbf{s}_k}^{r_k -} \hat{\mathbf{s}}_k \dot{\sim} \chi^2_{r_k},$$

where $\mathbf{V}_{\mathbf{s}_k}^{r_k -}$ is the rank r_k Moore-Penrose pseudoinverse of $\mathbf{V}_{\mathbf{s}_k}$. Parameter r_k is selected using the established notion of *edf* used in (6). Because *edf* is not an integer, it can be rounded as follows (Wood 2013)

$$r_k = \begin{cases} \text{floor}(edf_k) & \text{if } \ edf_k < \text{floor}(edf_k) + 0.05 \\ \text{floor}(edf_k) + 1 & \text{otherwise} \end{cases},$$

which proved effective in semiparametric bivariate probit models (Marra 2013).

As an alternative, the shrinkage single penalty approach presented in Marra and Wood (2011) can be adopted. Specifically, the generic second-order smoothing penalty matrix $\mathbf{S}_k$ can be decomposed as $\mathbf{U}_k \Lambda_k \mathbf{U}_k^\mathsf{T}$, where $\mathbf{U}_k$ is an eigenvector matrix associated with the k^{th} smooth function, and Λ_K the corresponding diagonal eigenvalue matrix. Because a part of the spline basis deals with the penalty null space, Λ_k contains zero eigenvalues. So even if λ_k goes to infinity the smooth term of a nuisance variable may still be estimated as non-zero, because the function component in the null space (i.e., the linear term) is unpenalized. This can be fixed by replacing Λ_k with $\tilde{\Lambda}_k$, where the latter is the same as the former except that the zero eigenvalues are set to a small proportion, typically 0.1, of the smallest strictly positive eigenvalue of $\mathbf{S}_k$. This forces the eigenvalues of the new penalty matrix, $\tilde{\mathbf{S}}_k$, associated with the penalty null space to be different from zero. Hence a smooth component can in principle be removed from the model altogether.

Copula models with a single dependence parameter can be thought of as non-nested models. As suggested by Zimmer and Trivedi (2006) among others, one approach for choosing between copula models is to use either the Akaike or (Schwarz) Bayesian information criterion (AIC and BIC, respectively). In our case, $AIC = -2\ell(\hat{\boldsymbol{\delta}}_*) + 2edf$ and $BIC = -2\ell(\hat{\boldsymbol{\delta}}_*) + \log(n)edf$, where the log-likelihood is evaluated at the penalized parameter estimates and $edf = \text{tr}(\hat{\mathbf{A}}_{\hat{\boldsymbol{\lambda}}})$.

2.4. Numerical considerations

As explained in Section 2.2, a trust region Newton algorithm is a more reliable choice to estimate the model parameters. As for the initial values, they are provided by using an extension of the Heckman (1979) procedure detailed in Appendix B of Marra and Radice (2013). The adopted approach proved to be fast and reliable in most cases, with occasional convergence failure for small values of n and n_s.

As the analytical expressions for $\mathbf{g}$ and $\boldsymbol{\mathcal{H}}$ of the copula log-likelihood functions are very complicated, numerical issues may be encountered in some cases when certain quantities take values which lie nearby their boundaries. Firstly, this may occur when the dependence between the margins is very strong or very weak, i.e., when θ takes extreme values (for example, association tending to 1 implies $\theta \to \infty$ for a number of copulas). This leads to expressions which are equal to `Inf` during the numerical evaluations, especially the Frank copula where the exponential transformation of θ appears in the expressions for the gradient and Hessian. Secondly, data points which lie in the tails of F_{1i} and F_{2i} will lead to their values equal to 0 or 1. Also, the value of z_i appearing in log-likelihood (2) may be approximately equal to 1, hence producing `-Inf`. These numerical problems are dealt with by truncating the values of F_{1i}, F_{2i}, f_{2i} and z_i to the interval $(\varepsilon, 1-\varepsilon)$ with $\varepsilon = 10^{-10}$. Moreover, the ratio $\phi(x)/\Phi(x)$ appearing in the expressions for $\mathbf{g}$ and $\boldsymbol{\mathcal{H}}$ is defined using the approximation $\phi(x)/\Phi(x) \sim -x$ for $x < -35$ in order to avoid `NaN`.

If a given model cannot be fitted due to numerical issues then the user receives the message `Ill-conditioned task`. It is worth noting that numerical problems that arise when fitting a model may be also a hint that the chosen model is not appropriate to fit the data at hand.

3. Overview of the package

The **SemiParSampleSel** package is available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/web/packages/SemiParSampleSel/index.html. The package depends on **magic** (Hankin 2005), **mgcv** (Wood 2006), **trust** (Geyer 2013), **mvtnorm** (Genz and Bretz 2009), **Matrix** (Bates and Maechler 2014) and **copula** (Yan 2007). The main function in **SemiParSampleSel** is `SemiParSampleSel()`, which fits copula regression spline sample selection models as described in the previous section. The function can be called using the following syntax:

```
SemiParSampleSel(formula.eq1, formula.eq2, data = list(), BivD = "N",
                 margins = c("N", "N"), gamma = 1, ...)
```

The two required arguments are `formula.eq1` and `formula.eq2` which are the formulas for the selection and outcome equations, respectively. These are `glm` like formulas except that smooth terms can be included in the equations as for `gam` in **mgcv**. For instance, the selection equation may look like:

```
y.sel ~ as.factor(x1) + s(x2, bs = "cr", k = 10, m = 2) + s(x3, x4) + ...,
```

where `y.sel` represents the binary selection variable, `x1` is a categorical predictor, and the `s` terms are used to specify smooth functions of the continuous predictors `x2`, `x3` and `x4`. Argument `bs` specifies the spline basis; possible choices include `cr` (cubic regression spline), `cs` (shrinkage version of `cr`), `tp` (thin plate regression spline) and `ts` (shrinkage version of `tp`). Bivariate smoothing, e.g., `s(x3, x4)`, is achieved using `bs = "tp"`. `k` is the basis dimension (default is 10) and `m` the order of the penalty (default is 2). More details and options on smooth term specification can be found in the documentation of **mgcv**. **SemiParSampleSel** does not currently support the use of tensor product smooths.

Optional arguments of the function `SemiParSampleSel` include `data` which is a data frame, list or environment containing the variables in the model, and `gamma` which is an inflation factor for the model degrees of freedom used in the smoothing step. Smoother models can be obtained setting this parameter to a value greater than `1`. In our experience, `gamma = 1.4` typically achieves this; this was also found by Kim and Gu (2004) in a different context. The type of bivariate copula linking the two model equations can be specified through `BivD`. Possible choices are `"N"`, `"C"`, `"J"`, `"FGM"`, `"F"`, `"AMH"` and `"G"` which stand for bivariate normal, Clayton, Joe, Farlie-Gumbel-Morgenstern, Frank, Ali-Mikhail-Haq and Gumbel. The argument `margins` specifies the marginal distributions of the selection and outcome equations, given in the form of a two-dimensional vector which is equal to `c("N","N")` for normal margins. Details on all the other arguments, including starting value and control options, and the fitted-object list that the function returns can be found in (Marra *et al.* 2014).

Other available functions are:

`plot(x, eq, pages = 0, scale = -1, shade = FALSE, seWithMean = FALSE, ...)`. This function takes a fitted object `x` as produced by `SemiParSampleSel()` and plots the component smooth functions that make it up on the scale of the linear predictor. `eq` denotes the equation from which smooth terms should be considered for printing, `pages` is the number of pages over which to produce the plots (e.g., if `pages = 1` then all terms will be plotted on one page), and `scale` is the y-axis scale to use for each plot (`scale = 0` gives a different axis for each plot). If `shade` is set to `TRUE` then shaded regions as confidence bands for smooth terms are produced. Of interest is the argument `seWithMean` which indicates whether the component smooth should be shown with confidence intervals that include the uncertainty about the overall mean. Marra and Wood (2012) showed that `seWithMean = TRUE` results in intervals with better nominal frequentist coverage probabilities. This function is based on `plot.gam()` in **mgcv** to which the reader is referred for full details.

`predict(object, eq, ...)`. This function takes a fitted `SemiParSampleSel` object and produces predictions for a new set of values of the model covariates or the original values used for the model fit. Standard errors of predictions can be produced. These are based on the posterior distribution of the model coefficients. This function is based on `predict.gam()` in **mgcv**.

`summary(object, n.sim = 1000, s.meth = "svd", prob.lev = 0.05, ...)`. This function produces some summaries of a fitted `SemiParSampleSel` object. `n.sim` indicates the number of simulated coefficient vectors from the posterior distribution of the

estimated model parameters, which are used to calculate 'confidence' intervals for σ, θ and Kendall's τ, for instance. `s.meth` is the matrix decomposition used to determine the matrix root of the covariance matrix (see the documentation of **mvtnorm** for further details). `prob.lev` is the probability of the left and right tails of the posterior distribution used for interval calculations. The object list returned includes, for instance, summary tables for the selection and outcome equations for the parametric and nonparametric components, and the estimated standard deviation and association coefficient.

`ss.checks(x)` which produces some diagnostic information about the fitting procedure for a `SemiParSampleSel` object.

These functions will be illustrated in Section 5.

4. Simulations

In this section, we conduct a Monte Carlo simulation study to evaluate the empirical effectiveness of the copula regression spline sample selection models implemented in the package. For convenience, all the tables and figures of results are given in Appendix B.

As in Marra and Radice (2013), the sampling experiments were based on the equations

$$\begin{aligned} \eta_{1i} &= \alpha_{11} + \alpha_{12} u_i + s_{11}(z_{1i}) + s_{12}(z_{2i}) \\ \eta_{2i} &= \alpha_{21} + \alpha_{22} u_i + s_{21}(z_{1i}) \end{aligned}, \tag{9}$$

where y_{1i} and y_{2i} were determined as described in Section 2.1. The test functions are displayed in Figure 1 and are defined as $s_{11}(z_{1i}) = -0.7\left\{4z_{1i} + 2.5z_{1i}^2 + 0.7\sin(5z_{1i}) + \cos(7.5z_{1i})\right\}$, $s_{12}(z_{2i}) = -0.4\left\{-0.3 - 1.6z_{2i} + \sin(5z_{2i})\right\}$, and $s_{21}(z_{1i}) = 0.6\left\{\exp(z_{1i}) + \sin(2.9z_{1i})\right\}$. Parameter vector $(\alpha_{12}, \alpha_{21}, \alpha_{22})$ and σ were set to $(2.5, -0.68, -1.5)$ and 1. Binary values for y_{1i} were generated so that approximately 50% of the total number of observations were selected to fit the outcome equation; this was achieved by setting α_{11} to 0.58. Regressors u_i, z_{1i} and z_{2i} were generated as three uniform covariates on $(0, 1)$ with correlation approximately equal to 0.5. This was achieved using `rmvnorm()` in **mvtnorm**, generating standardized multivariate random draws with correlation 0.5 and then applying `pnorm()` (e.g., Marra and Radice 2013). Regressor u_i was eventually dichotomized using `round()`. As joint distribution of $(y_{1i}^*, y_{2i})_{i=1}^n$ the following copulas were considered: normal, Clayton, Joe, FGM, AMH, Frank and Gumbel, each with normal margins. The sample size n was set to 1000. For each copula, different values of the association parameter were considered:

normal copula: $\theta = 0.16$ ($\tau = 0.1$), $\theta = 0.71$ ($\tau = 0.5$), $\theta = 0.89$ ($\tau = 0.7$),

Clayton copula: $\theta = 0.22$ ($\tau = 0.1$), $\theta = 2$ ($\tau = 0.5$), $\theta = 57$ ($\tau = 0.7$),

Joe copula: $\theta = 1.31$ ($\tau = 0.15$), $\theta = 2.86$ ($\tau = 0.5$), $\theta = 6.78$ ($\tau = 0.75$),

FGM copula: $\theta = -0.9$ ($\tau = -0.2$), $\theta = 0.68$ ($\tau = 0.15$),

AMH copula: $\theta = -0.62$ ($\tau = -0.12$), $\theta = 0.4$ ($\tau = 0.1$), $\theta = 0.9$ ($\tau = 0.28$),

Frank copula: $\theta = 1.86$ ($\tau = 0.2$), $\theta = 5.74$ ($\tau = 0.5$), $\theta = 11.41$ ($\tau = 0.7$),

Gumbel copula: $\theta = 1.25$ ($\tau = 0.2$), $\theta = 2$ ($\tau = 0.5$), $\theta = 5$ ($\tau = 0.8$).

In Tables 4 - 10 the association parameter used to generate the data is expressed in terms of Kendall's τ coefficient. For each combination of parameter settings, the number of simulated datasets was set to 250. We also explored the performance of the models in the absence of an exclusion restriction as detailed in Section 4.2.

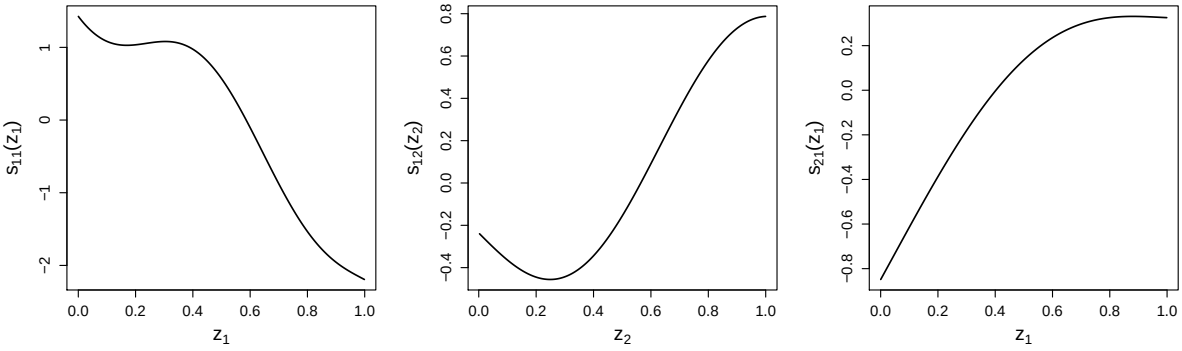


Figure 1: The test functions used in the simulation studies.

4.1. Main results

Since the selection equation is not in principle affected by non-random sample selection bias, we focus on the estimation results for the outcome equation only. Tables 4 - 10 report the percentage relative bias and root mean squared error (RMSE) calculated for the estimators of α_{21}, α_{22}, σ, τ, and the RMSE for that of $s_{21}(z_1)$, calculated as $\sqrt{\frac{1}{200}\sum_{b=1}^{200}\{\hat{s}(z_{1b}) - s(z_{1b})\}^2}$, based on the estimates for 200 fixed covariate values. The tables also report the percentage frequency at which each copula model was selected by AIC and BIC.

The results presented in the tables show overall that the model employing the true copula achieves the lowest bias and/or RMSE of the estimators of all considered parameters in most cases. We can particularly observe this for data generated using the Clayton copula (see Table 5), where the estimators of α_{21}, α_{22}, σ, τ and s_{22} obtained from the Clayton model outperform in terms of bias and RMSE those yielded by the other copula models. Using the right model is particularly important for estimating τ when its true value falls outside the dependence range covered by a given copula, as some of them allow only for a restricted interval of dependence (here, this is the case for AMH and FGM). The results also show that, for data generated using the Frank or normal copulas, both models yield comparably good results, hence reflecting the similarity between these two copulas (see Tables 4 and 9 for $\tau = 0.7$). We observe a similar effect for data generated using the Joe and Gumbel copulas. The findings also suggest that in some cases for small values of τ the choice of the correct copula model does not seem to play an important role in estimation (see Table 4 for $\tau = 0.1$, Table 7 and Table 8), and often the Clayton and Gumbel models yield estimators with a relatively low bias and RMSE for such data regardless of the true copula.

As for copula model selection, the two criteria work overall well. The case of very weak dependence is the most difficult one as the underlying distribution converges to the normal product distribution when $\tau \to 0$. Thus in this situation all copulas entail very similar distributions.

As an example, Figure 2 presents contour plots of FGM, Clayton and Joe copulas with normal margins for small values of the dependence parameter. For those distributions the choice of the correct copula based on an empirical sample is extremely difficult and the selection criteria appear to select an arbitrary model as can be seen in Table 7. At the same time, the finite sample performance of the estimators is unaffected by the wrong choice of a copula in those border cases as, again, all copulas tend to the same (normal product) distribution here. Also, even in this difficult situation, AIC seems to be successful for some copulas (see Tables 5, 6 and 9). For medium and large values of τ, the true copula model is the most frequent choice with all model selection criteria, with AIC performing much better than BIC and achieving a hit rate of more than 90% in some cases (see Table 5). It is also worth noting that in general, the accuracy of the choice of the copula improves with the sample size as can be seen in Tables 1 and 2 of supplementary materials where the experiment was repeated for samples of size $n = 3000$ and $n = 5000$ pertaining to bivariate normal distribution. There we can also observe consistency of the estimators when the right copula is chosen. In the case of a wrong copula the estimators are inconsistent.

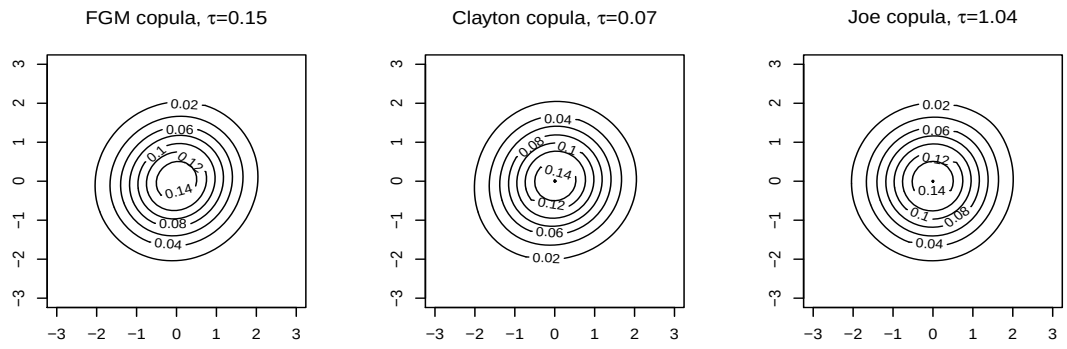


Figure 2: Contour plots of FGM, Clayton and Joe copulas with normal margins for small values of the dependence parameter.

4.2. Absence of an exclusion restriction

Sometimes the same regressors have to be used in both selection and outcome equations as an exclusion restriction is not available. To investigate the performance of the copula sample selection models in this situation, the simulation study described above was repeated for the case in which system (9) did not include $s_{12}(z_{2i})$. The sampling experiments were based on

$$\begin{aligned} \eta_{1i} &= \alpha_{11} + \alpha_{12} u_i + s_{11}(z_{1i}) \\ \eta_{2i} &= \alpha_{21} + \alpha_{22} u_i + s_{21}(z_{1i}) \end{aligned}, \tag{10}$$

where functions s_{11} and s_{21} and parameters α_{11}, α_{12}, α_{21}, α_{22} were the same as in Section 4 and the predictors u_i and z_{1i} were generated in the same way.

Following a reviewer's suggestion we also considered the harder scenario in which the same functional form of the effect of variable z_1 was present in both model equations. Thus the

simulated data were based on the equations

$$\begin{aligned} \eta_{1i} &= \alpha_{11} + \alpha_{12} u_i + s_{21}(z_{1i}) \\ \eta_{2i} &= \alpha_{21} + \alpha_{22} u_i + s_{21}(z_{1i}) \end{aligned}. \tag{11}$$

Figures 5 and 6 demonstrate the influence of the lack of exclusion restriction on the estimators of the model parameters in terms of their mean squared error and bias, for a choice of copulas: normal, Clayton, Joe and FGM. The solid lines correspond to root mean squared errors of the estimators $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and the smooth function $\hat{s}_{21}$ (upper panels) and absolute values of percentage bias of estimators $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$ and $\hat{\tau}$ (lower panels) for model (10) without the exclusion restriction. The corresponding lines for model (9) in which the selection equation contains an additional term $s_{12}(z_2)$ are added for comparison as dotted lines. Analogically, Figures 7 and 8 demonstrate the influence of the lack of exclusion restriction on the estimators for model (11) in comparison to model (9).

We observe that the quality of the estimator $\hat{\sigma}$ is practically unaffected by the lack of exclusion restriction for both scenarios considered in terms of root mean squared error and bias. For the remaining parameters, we observe that removing $s_{12}(z_2)$ from the selection equation increases the bias and RMSE of the estimators in most of the cases considered. We also observe a larger variance and more cases of lack of convergence when the exclusion restriction is not available. The lack of exclusion restriction leads to particularly unstable estimators of the Kendall's τ in terms of the relative bias in cases where this parameter is close to zero, as can be seen in figures 5 a) and 6 b) where the relative percentage bias exceeds 160% and 110%, respectively, while the RMSEs of $\hat{\tau}$ in those cases do not indicate any particularly bad performance. The above values of relative bias imply however that the average estimated values of τ equal approximately -0.06 and -0.018, respectively, which in turn has a major importance while testing the absence of sample selection bias as it affects the size and power of the test. This issue is further discussed in section 4.3. However, for scenario (10) the influence of the lack of exclusion restriction is usually much less significant than for the more difficult scenario (11). Moreover, in some cases the differences between the RMSEs of the model parameters for the cases with and without the exclusion restriction are rather negligible (see Figure 5 b) and Figure 6 a)).

4.3. Testing the absence of sample selection bias

The key issue while fitting a sample selection model is testing the null hypothesis of absence of selection bias. If the variables y_{1i}^* and y_{2i}^* are correlated then the sample selection bias occurs and it is necessary to consider both outcome equation and selection equation together with the dependence structure between the two of them while estimating the model. Otherwise, the model can be much simplified by dropping the selection equation (and consequently the copula function) from the analysis.

In general, the approach to testing for sample selection bias relies heavily on the specific sample selection model assumed. In the Heckman's two step procedure (Heckman (1979)) sample selection bias is tested using the t-test related with the significance of the omitted variable. Dubin and Rivers (1989) considered likelihood ratio, Wald and Lagrange multiplier tests in the context of a censored probit model. Moreover, Vella (1992) proposed a conditional moment test.

In the context of copula regression spline sample selection models, testing for sample selection

bias can be based on the dependence parameter θ as absence of sample selection bias is equivalent to the condition $\theta = 0$. However, because of the restrictions on the values of the copula association parameters, the use of classic testing approaches may yield unreliable results in some copula cases. As a practical alternative, the Kendall's τ coefficient can be employed. Hence the null hypothesis can be tested by checking whether the confidence interval for τ includes 0. In this section, results of a Monte-Carlo study of the finite-sample performance of such approach are presented.

For data sets generated using the equations (9), the empirical size of the test for absence of selectivity bias has been calculated based on 99%, 95% and 90% confidence intervals for the parameter τ. As before, for every data set different copulas were considered while fitting the spline sample selection models (normal, FGM, AMH, Frank and Gumbel). The Clayton and Joe copulas are not considered in the study as they allow only strictly positive dependence implying that the size of the test always equals 1 in this case. The results of the Monte Carlo simulations based on 250 repetitions are presented in table 11. The smallest values of the empirical size α can be observed while using copulas FGM and AMH, which can be intuitively explained by the fact that both of those copulas allow very restricted range of the parameter τ ($\tau \in [-2/9, 2/9]$ for FGM and $\tau \in [-0.1817, 1/3]$ for AMH) which implies that models based on FGM and AMH copulas are well suited for data with very weak dependence. Among the three remaining copulas, which allow the full range of dependence, i.e. $\tau \in (-1, 1)$, Frank copula performs the best achieving the empirical size of the test very close to the theoretical value for $n = 5000$. The poorest performance occurs in the case of Gumbel copula for which the empirical size converges to the theoretical value at a much slower rate, as the sample size increases.

Tables 12 and 13 present the empirical size for the sample selection bias test for data without exclusion restriction, generated using the equations (10) and (11), respectively. In both cases, a negative influence of lack of an the exclusion restriction can be observed as the values of empirical sizes are larger than those presented in table 11 where the exclusion restriction is used. Moreover, the effect of lack of exclusion restriction is more severe for data generated using equations (11) where the variable z_1 enters both, the selection and the outcome equations, in the same functional form. However, in Tables 12 and 13 the same tendency regarding to the comparison between different copulas can be observed with FGM and AMH achieving the smallest empirical sizes and Gumbel copula displaying the worst performance.

Moreover, a study of empirical power of the test for sample selection bias has been conducted, with results reported in Tables 14-15. As expected, the power of all tests drops when the dependence parameter is close to 0, i.e., when the selection and outcome equations are close to independence. In this difficult scenario, using Gumbel copula leads to the most powerful tests. A particularly poor performance can be observed when fitting the FGM and AMH copulas. Those are the copulas allowing very limited scope of parameter τ which makes them perform poorly not only in the proximity of the null hypothesis but also when a strong dependence holds. For the three remaining copulas, a very good performance can be observed when the coefficient τ equals at least 0.5 as the power usually exceeds the level of 97% in those cases. Overall, the largest power is usually achieved for Gumbel and normal copulas. It should be noted that developing alternative methods of testing sample selection bias in the context of copula regression spline sample selection models is going to be the subject of future research.

Tables 16-19 present powers of the test for sample selection bias in the absence of an exclusion restriction. In most of the cases considered, the powers of the test are smaller than when the exclusion restriction is present. However, in some cases larger powers also can be observed for data generated using FGM and AMH copulas. Thus the absence of an exclusion restriction appears to affect the testing process more severely when the correlation between the two model equations is moderate or strong.

5. Real data example

The copula regression spline sample selection models presented in this paper are illustrated using data from the RAND Health Insurance Experiment (RHIE) which was a comprehensive study of health care cost, utilization and outcome conducted in the United States between 1974 and 1982 (Newhouse 1999). As explained in the introductory section, the aim was to quantify the relationship between various covariates and annual health expenditures in the population as a whole.

Variable	Definition
`lnmeddol`	log of the medical expenses of the individual (*outcome variable*)
`binexp`	binary variable indicating whether the medical expenses are positive (*selection variable*)
`logc`	log of the coinsurance rate (`coins`) plus 1
`idp`	binary variable for individual deductible plans
`pi`	participation incentive payment
`fmde`	is 0 if `idp`=1, and $\log[\max\{1, \text{maximum expenditure offer}/(0.01 * \texttt{coins})\}]$ otherwise
`physlm`	physical limitations
`disea`	number of chronic diseases
`hlthg`	binary variable for good self-rated health (the baseline is excellent self-rated health)
`hlthf`	binary variable for fair self-rated health
`hlthp`	binary variable for poor self-rated health
`inc`	family income
`fam`	family size
`educdec`	education of household head in years
`xage`	age of the individual in years
`female`	binary variable for female individuals
`child`	binary variable for individuals younger than 18 years
`fchild`	binary variable for female individuals younger than 18 years
`black`	binary variable for black household heads

Table 3: Description of the outcome and selection variables, and of the regressors.

In this context, non-random sample selection arises because the sample consisting of individuals who used health care services differ in important characteristics from the sample of individuals who did not use them. Because some characteristics cannot be observed, traditional regression modeling is likely to deliver inconsistent estimates, hence the need to correct parameter estimates for non-random sample selection. We use the same subsample as in Cameron and Trivedi (2005, p. 553), and model annual health expenditures. The sample size and number of selected observations are 5574 and 4281. The variables are defined in Table 3. Additional information can be found in Cameron and Trivedi (2005, Table 20.4) and Newhouse (1999).

Following Cameron and Trivedi (2005) the outcome and the selection equations include the same set of regressors. As in Marra and Radice (2013) the two equations include `logc`, `idp`,

fmde, physlm, disea, hlthg, hlthf, hlthp, female, child, fchild and black as parametric components, and smooth functions of pi, inc, fam, educdec and xage, represented using thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives (which are the default options in the package). Specifically, after reading the dataset, called ND, we load the package and specify the selection and outcome equations.

```
R> library(SemiParSampleSel)
R> SE <- binexp ~ logc + idp + fmde + physlm + disea + hlthg + hlthf +
                  hlthp + female + child + fchild + black + s(pi) + s(inc) +
                  s(fam) + s(educdec) + s(xage)
R> OE <- lnmeddol ~ logc + idp + fmde + physlm + disea + hlthg + hlthf +
                    hlthp + female + child + fchild + black + s(pi) + s(inc) +
                    s(fam) + s(educdec) + s(xage)
```

We then estimate the copula regression spline sample selection models by penalized likelihood, as described in Section 2.2, setting gamma = 1.4 to obtain smoother models.

```
R> out_N   <- SemiParSampleSel(SE, OE, data = ND, gamma = 1.4)
R> out_C   <- SemiParSampleSel(SE, OE, data = ND, BivD = "C", gamma = 1.4)
R> out_J   <- SemiParSampleSel(SE, OE, data = ND, BivD = "J", gamma = 1.4)
R> out_FGM <- SemiParSampleSel(SE, OE, data = ND, BivD = "FGM", gamma = 1.4)
R> out_F   <- SemiParSampleSel(SE, OE, data = ND, BivD = "F", gamma = 1.4)
R> out_AMH <- SemiParSampleSel(SE, OE, data = ND, BivD = "AMH", gamma = 1.4)
R> out_G   <- SemiParSampleSel(SE, OE, data = ND, BivD = "G", gamma = 1.4)
```

Given the superior performance of AIC on BIC shown in the simulation study, we use the AIC to select a model.

```
R> AIC_N   <- AIC(out_N)
R> AIC_C   <- AIC(out_C)
R> AIC_J   <- AIC(out_J)
R> AIC_FGM <- AIC(out_FGM)
R> AIC_F   <- AIC(out_F)
R> AIC_AMH <- AIC(out_AMH)
R> AIC_G   <- AIC(out_G)

R> AIC_N

[1] 20294.88

R> AIC_C

[1] 20298.31

R> AIC_J

[1] 20339.53
```

```
R> AIC_FGM

[1] 20287.24

R> AIC_F

[1] 20281.23

R> AIC_AMH

[1] 20282.8

R> AIC_G

[1] 20293.89
```

The Frank copula model is chosen. Before looking at the results, we check that the algorithm has found a solution.

```
R> ss.checks(out_F)

Largest absolute gradient value: 2.726033e-07
Information matrix is positive definite

Trust region Newton iterations before smoothing parameter estimation: 6
Smoothing parameter/leapfrog loops: 4
Trust region Newton iterations after smoothing parameter/leapfrog step: 2
```

We can now look at the results.

```
R> set.seed(1)
R> summary(out_F)

Family: SAMPLE SELECTION Frank Copula with normal margins

SELECTION EQ.: binexp ~ logc + idp + fmde + physlm + disea + hlthg + hlthf +
    hlthp + female + child + fchild + black + s(pi) + s(inc) +
    s(fam) + s(educdec) + s(xage)

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.652639   0.082962   7.867 3.64e-15 ***
logc        -0.090549   0.028381  -3.191 0.001420 **
idp         -0.175910   0.055540  -3.167 0.001539 **
fmde        -0.006780   0.017808  -0.381 0.703393
physlm       0.253415   0.074348   3.408 0.000653 ***
disea        0.021558   0.003718   5.798 6.70e-09 ***
```

```
hlthg        0.083350   0.044482   1.874 0.060956 .
hlthf        0.190734   0.083414   2.287 0.022220 *
hlthp        0.590048   0.208658   2.828 0.004687 **
female       0.463019   0.054950   8.426  < 2e-16 ***
child        0.252364   0.146261   1.725 0.084449 .
fchild      -0.457734   0.080554  -5.682 1.33e-08 ***
black       -0.591727   0.054067 -10.944  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

Smooth components  approximate significance:
            edf Est.rank Chi.sq  p-value
s(pi)     8.035        8 34.022 4.03e-05 ***
s(inc)    2.485        3 30.990 8.54e-07 ***
s(fam)    1.805        2  2.071 0.355006
s(educdec) 1.623       2 14.962 0.000564 ***
s(xage)   6.921        7 51.414 7.62e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1    1


OUTCOME EQ.: lnmeddol ~ logc + idp + fmde + physlm + disea + hlthg + hlthf +
    hlthp + female + child + fchild + black + s(pi) + s(inc) +
    s(fam) + s(educdec) + s(xage)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.007309   0.092939  43.118  < 2e-16 ***
logc         0.008707   0.033452   0.260 0.794634
idp         -0.068785   0.064412  -1.068 0.285574
fmde        -0.028829   0.020067  -1.437 0.150831
physlm       0.198391   0.070378   2.819 0.004818 **
disea        0.015770   0.003622   4.354 1.34e-05 ***
hlthg        0.142250   0.049331   2.884 0.003932 **
hlthf        0.327870   0.089692   3.656 0.000257 ***
hlthp        0.631742   0.174589   3.618 0.000296 ***
female       0.214176   0.059725   3.586 0.000336 ***
child        0.017721   0.170062   0.104 0.917008
fchild      -0.229596   0.091147  -2.519 0.011771 *
black       -0.024775   0.070988  -0.349 0.727087
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

Smooth components  approximate significance:
            edf Est.rank Chi.sq  p-value
s(pi)     4.077        5  5.254 0.385675
s(inc)    3.651        4  8.784 0.066723 .
s(fam)    1.000        1  8.128 0.004358 **
```

```
s(educdec) 1.966        2  2.436 0.295892
s(xage)    7.084        8 31.027 0.000139 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1

n = 5574  n.sel = 4281  sigma = 1.422(1.386,1.46)
theta = -3.043(-3.942,-2.268)  Kendall s Tau = -0.311(-0.384,-0.24)
total edf = 66.648
```

Notice that we set a seed before `summary()`. This allows us to recover the same results for the confidence intervals of the quantities reported at the bottom of the summary output; recall that intervals for such components are calculated using Bayesian posterior simulation as mentioned in Section 2.3.

As for the selection equation, the results show that all variables, which enter the model parametrically, are statistically significant at the 10% level, except for `fmde`. The p-values for the smooth terms, calculated as discussed in Section 2.3, indicate that `fam` does not have an impact on the response. Regarding the outcome equation, health status variables (such as `physlm` and `disea`) have an effect on annual health expenses, whereas health insurance variables (`logc` and `idp`) seem not to determine the medical expenses. The p-values for the estimated smooths indicate that `inc`, `fam` and `xage` are significantly different from zero. The estimate for σ is 1.42 and is significantly different from zero. The estimate for θ is negative and statistically different from zero. `Kendall's Tau` is also negative and significantly different from zero. This indicates that the unobserved factors which affect the use of health services also affect medical expenses. The estimated degrees of freedom (`total edf`) of the penalized model, calculated as described in Section 2.2, is 66.648.

Using `plot()`, we produce the smooth function estimates for the outcome equation obtained from the Frank copula model; these are displayed in Figure 3.

```
R> plot(out_F, eq = 2, pages = 1, scale = 0, shade = TRUE, seWithMean = TRUE,
        cex.axis = 1.6, cex.lab = 1.6)
```

The shaded regions represent 95% confidence bands calculated from the posterior distribution, as described in Section 2.3. The 'rug plot', at the bottom of each graph, shows the covariate values. The numbers shown on the y-axis in each plot indicate the estimated degrees of freedom (`edf`). Due to the identifiability constraints, the estimated curves are centered around zero. The results for `xage` and `fam` are consistent with the interpretation that health expenditure increases non-linearly as people become older, and that individual health expenditure decreases as family size increases.

We re-fit the Frank copula regression model by using the shrinkage option `bs = "ts"` in `s()`.

```
R> SE_s <- binexp ~ logc + idp + fmde + physlm + disea + hlthg + hlthf +
                    hlthp + female + child + fchild + black +
                    s(pi, bs = "ts") + s(inc, bs = "ts") + s(fam, bs = "ts") +
                    s(educdec, bs = "ts") + s(xage, bs = "ts")
R> OE_s <- lnmeddol~ logc + idp + fmde + physlm + disea + hlthg + hlthf +
                     hlthp + female + child + fchild + black +
```

```
                    s(pi, bs = "ts") + s(inc, bs = "ts") + s(fam, bs = "ts") +
                    s(educdec, bs = "ts") + s(xage, bs = "ts")
R> out_F_s <- SemiParSampleSel(SE_s, OE_s, data = ND, BivD = "F", gamma = 1.4)
R> plot(out_F_s, eq = 2, pages = 1, scale = 0, shade = TRUE, seWithMean = TRUE,
        cex.axis = 1.6, cex.lab = 1.6)
```

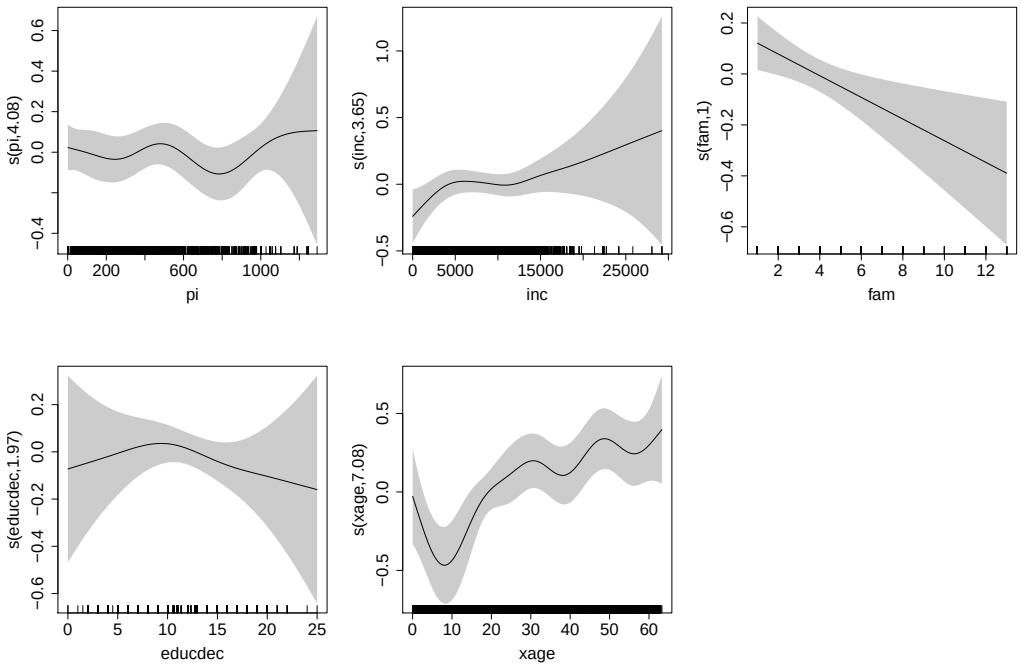


Figure 3: Smooth function estimates and 95% confidence bands obtained applying the Frank copula regression spline sample selection model on the RAND RHIE dataset described in Section 4.

We obtain the fitted smooth functions depicted in Figure 4; regressor `pi` has been suppressed, whereas the other covariate effects exhibit patterns similar to those reported in Figure 3.

Finally, we use `predict()` to produce a prediction for new values of the model covariates. Specifically, we predict the log of medical expenditure for a typical individual.

```
R> new_data <- data.frame(logc = median(ND$logc), idp = median(ND$idp),
                          fmde = median(ND$fmde), physlm = median(ND$physlm),
                          disea = median(ND$disea), hlthg = median(ND$hlthg),
                          hlthf = median(ND$hlthf), hlthp = median(ND$hlthp),
                          female = median(ND$female), child = median(ND$child),
                          fchild = median(ND$fchild), black = median(ND$black),
                          pi = mean(ND$pi), inc = mean(ND$inc), fam = mean(ND$fam),
                          educdec = mean(ND$educdec), xage = mean(ND$xage))
R> pred <- predict(out_F, eq = 2, new_data)
R> pred
```

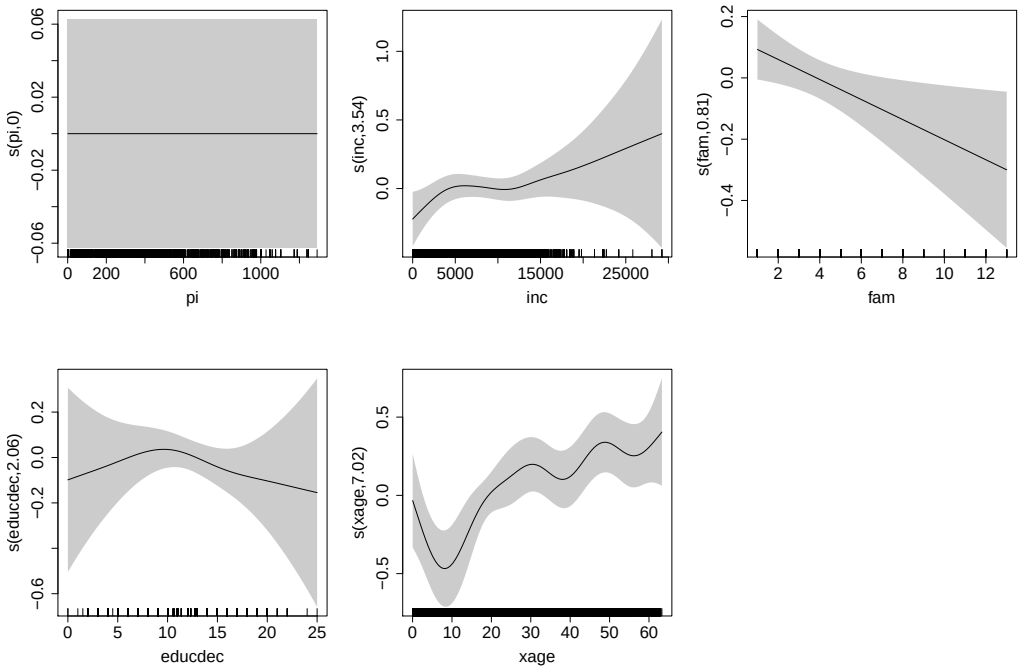


Figure 4: Smooth function estimates and 95% confidence bands obtained applying the Frank copula model with shrinkage option on the RAND RHIE dataset.

```
       1
4.430282
```

6. Discussion

We introduced flexible continuous response sample selection models and discussed the R package **SemiParSampleSel** which implements them. The package can be used to fit models where the linear predictors are flexibly specified using parametric and non-parametric components, and the dependence between the selection and outcome equations is modeled through the use of copulas. The developments and implementation proposed here extend and complement previous R implementations of sample selection models. Allowing for non-normal bivariate distributions between the model equations is important since the assumption of bivariate normality is often criticized.

A large number of copulas have been proposed in the literature and our selection aims to reflect the most commonly used bivariate copulas in empirical applications as well as different types of dependence in the data. Copulas such as normal and Frank allow for equal degrees of positive and negative dependence and are comprehensive (they cover the whole range of Kendall's τ). On the other hand, copulas such as Clayton, Joe and Gumbel only account for positive dependence but capture a type of structure which is not reflected by Frank or normal. Specifically, the Clayton copula exhibits a strong left tail dependence and a relatively weak right tail dependence, and vice versa for Gumbel and Joe.

The reader is cautioned that the class of models presented here is not intended to be exhaustive; as the majority of the methods, under model misspecification the proposed approach does not provide consistent estimates. For example, if the marginals are non-normal (e.g., they exhibit a heavy-tailed behavior or can be modeled using skewed, contaminated and mixture distributions), biased estimates should be expected. The extent of the bias cannot be predicted a priori and it depends on the application at hand. In light of this, possible generalizations of the methods implemented in **SemiParSampleSel** are to extend the scope of the marginal distribution for the outcome equation, using for instance the gamma, Poisson and Student-t distributions, and that of the available copulas in the package, using for example the Plackett and rotated copulas. Future research will also concern the development of model checking tools.

Acknowledgments

The first two authors were supported by the Engineering and Physical Sciences Research Council, UK (Grant EP/J006742/1). We are indebted to two anonymous reviewers for their constructive criticism which helped to improve considerably the presentation of the article.

References

Ahn H, Powell JL (1993). "Semiparametric Estimation of Censored Selection Models With a Nonparametric Selection Mechanism." *Econometrics*, **58**, 3–29.

Andrews DWK, Schafgans MMA (1998). "Semiparametric Estimation of the Intercept of a Sample Selection Model." *Review of Economic Studies*, **65**, 497–517.

Bates D, Maechler M (2014). ***Matrix:*** *Sparse and Dense Matrix Classes and Methods*. R package version 1.1-2, URL http://cran.r-project.org/web/packages/Matrix/index.html.

Bhat CR, Eluru N (2009). "A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel behavior modeling." *Transportation Research Part B: Methodological*, **43**, 749–765.

Brechmann EC, Schepsmeier U (2013). "Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine." *Journal of Statistical Software*, **52**(3), 1–27. ISSN 1548-7660. URL http://www.jstatsoft.org/v52/i03.

Cameron A, Trivedi P (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Chen S, Zhou Y (2010). "Semiparametric and Nonparametric Estimation of Sample Selection Models under Symmetry." *Journal of Econometrics*, **157**, 143–150.

Chib S, Greenberg E, Jeliazkov I (2009). "Estimation of semiparametric models in the presence of endogeneity and sample selection." *Journal of Computational and Graphical Statistic*, **18**, 321–348.

Das M, Newey WK, Vella F (2003). "Nonparametric Estimation of Sample Selection Models." *Review of Economic Studies*, **70**, 33–58.

Ding P (2014). "Bayesian Robust Inference of Sample Selection using Selection-Models." *Journal of Multivariate Analysis*, **124**, 451–464.

Dubin JA, Rivers D (1989). "Selection Bias in Linear Regression, Logit and Probit Models." *Sociological Methods and Research*, **18**, 360–390.

Gallant RA, Nychka DW (1987). "Semi-Nonparametric Maximum Likelihood Estimation." *Econometrica*, **55**, 363–390.

Genius M, Strazzera E (2008). "Applying the copula approach to sample selection modelling." *Applied Economics*, **40**, 1443–1455.

Genz A, Bretz F (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg. ISBN 978-3-642-01688-2.

Geyer CJ (2013). ***trust**: Trust Region Optimization*. R package version 0.1-4, URL http://cran.r-project.org/web/packages/trust/index.html.

Gronau R (1974). "Wage Comparisons: A Selectivity Bias." *Journal of Political Economy*, **82**, 1119–1143.

Hankin RKS (2005). "Recreational Mathematics with R: Introducing the 'magic' Package." *R News*, **5**.

Hasebe T, Vijverberg WP (2012). "A Flexible Sample Selection Model: A GTL-Copula Approach." *IZA Discussion Papers 7003*, Institute for the Study of Labor (IZA). URL http://ideas.repec.org/p/iza/izadps/dp7003.html.

Hastie T, Tibshirani R (1993). "Varying-Coefficient Models." *Journal of the Royal Statistical Society B*, **55**, 757–796.

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.

Heckman J (1979). "Sample Selection Bias as a Specification Error." *Econometrica*, **47**, 153–162.

Heckman J (1990). "Varieties of selection bias." *The American Economic Review*, **80**, 313–318.

Henningsen A (2012). ***censReg**: Censored Regression (Tobit) Models*. R package version 0.5-16, URL http://cran.r-project.org/package=censReg.

Joe H (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall Ltd., London.

Kim YJ, Gu C (2004). "Smoothing spline Gaussian regression: More scalable computation via efficient approximation." *Journal of the Royal Statistical Society Series B*, **66**, 337–356.

Lee DS (2008). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies*, **76**(11721), 1071–1102.

Lee LF (1982). "Some Approaches to the Correction of Selectivity Bias." *Review of Economic Studies*, **49**, 355–72.

Lee LF (1984). "Tests for the Bivariate Normal Distribution in Econometric Models with Selectivity." *Econometrica*, **52**, 843–63.

Lee LF (1994a). "Semiparametric instrumental variable estimation of simultaneous equation sample selection models." *Journal of Econometrics*, **63**, 341–388.

Lee LF (1994b). ""Semiparametric two-stage estimation of sample selection models subject to Tobit-type selection rules." *Journal of Econometrics*, **61**, 305–344.

Lewis HG (1974). "Comments on Selectivity Biases in Wage Comparisons." *Journal of Political Economy*, **82**, 1145–1155.

Marchenko YV, Genton MG (2012). "A Heckman Selection-t Model." *Journal of the American Statistical Association*, **107**, 304–317.

Marra G (2013). "On P-values for Semiparametric Bivariate Probit Models." *Statistical Methodology*, **10**, 23–28.

Marra G, Radice R (2010). "Penalised Regression Splines: Theory and Application to Medical Research." *Statistical Methods in Medical Research*, **19**, 107–125.

Marra G, Radice R (2013). "Estimation of a Regression Spline Sample Selection Model." *Computational Statistics and Data Analysis*, **61**, 158–173.

Marra G, Radice R (2014). ***SemiParBIVProbit**: Semiparametric Bivariate Probit Modelling*. R package version 3.2-10, URL http://CRAN.R-project.org/package=SemiParBIVProbit.

Marra G, Radice R, Wojtyś M (2014). *Semiparametric Sample Selection Modelling with Continuous Response*. R package version 1.1, URL http://cran.r-project.org/web/packages/SemiParSampleSel/index.html.

Marra G, Wood S (2011). "Practical Variable Selection for Generalized Additive Models." *Computational Statistics and Data Analysis*, **55**, 2372–2387.

Marra G, Wood S (2012). "Coverage Properties of Confidence Intervals for Generalized Additive Model Components." *Scandinavian Journal of Statistics*, **39**, 53–74.

Montes-Rojas G (2011). "Robust Misspecification Tests for the Heckman's Two-Step Estimator." *Econometric Reviews*, **30**, 154–172.

Nelsen R (2006). *An Introduction to Copulas*. second edition. Springer-Verlag, New York.

Newey W (2009). "Two-step Series Estimation of Sample Selection Models." *Econometrics Journal*, **12**, S217–S229.

Newhouse JP (1999). *RAND Health Insurance Experiment [in Metropolitan and Non-Metropolitan Areas of the United States], 1974–1982*. Inter-university Consortium for Political and Social Research.

Nocedal J, Wright S (2006). *Numerical Optimization*. Springer-Verlag, New York.

Powell JL (1994). *Handbook of Econometrics*, chapter Estimation of Semiparametric Models, pp. 5307–5368. Elsevier, Amsterdam.

Powell JL, Stock JH, Stoker TM (1989). "Semiparametric Estimation of Index Coefficients." *Econometrica*, **57**, 1403–30.

Prieger JE (2002). "A Flexible Parametric Selection Model for Non-Normal Data with Application to Health Care Usage." *Journal of Applied Econometrics*, **17**, 367–392.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ruppert D, Wand M, Carroll R (2003). *Semiparametric Regression*. Cambridge University Press, New York.

SAS Institute Inc (2011). *SAS/STAT Software, Version 9.3*. URL http://www.sas.com/.

Schwiebert J (2013). "Sieve Maximum Likelihood Estimation of a Copula-Based Sample Selection Model." *Iza discussion papers*, Institute for the Study of Labor (IZA). URL http://www.iza.org/conference_files/SUMS_2013/schwiebert_j8731.pdf.

Smith MD (2003). "Modelling sample selection using Archimedean copulas." *Econometrics Journal*, **6**, 99–123.

StataCorp (2011). *Stata Statistical Software: Release 12*.

Toomet O (2012). ***intReg**: Interval Regression*. R package version 0.1-2, URL http://cran.r-project.org/package=intReg.

Toomet O, Henningsen A (2008). "Sample Selection Models in R: Package **sampleSelection**." URL http://www.jstatsoft.org/v27/i07/.

Vella F (1992). "Simple tests for sample selection bias in censored and discrete choice models." *Journal of Applied Econometrics*, **7**, 413–421.

Wiesenfarth M, Kneib T (2010). "Bayesian Geoadditive Sample Selection Models." *Journal of the Royal Statistical Society C*, **59**, 381–404.

Wood S (2013). "On P-values for Smooth Components of an Extended Generalized Additive Model." *Biometrika*, **100**, 221–228.

Wood SN (2006). *Generalized Additive Models: An Introduction With R*. Chapman & Hall/CRC, London.

Yan J (2007). "Enjoy the Joy of Copulas: With a Package **copula**." *Journal of Statistical Software*, **21**(4), 1–21. URL http://www.jstatsoft.org/v21/i04/.

Zhelonkin M, Genton MG, Ronchetti E (2013). *Robust Estimation and Inference in Sample Selection Models*. R package version 0.3, URL http://cran.r-project.org/web/packages/ssmrob/index.html.

Zimmer DM, Trivedi PK (2006). "Using trivariate copulas to model sample selection and treatment effects: application to family health care demand." *Journal of Business & Economic Statistics*, **24**, 63–76.

Appendix A - Analytical expressions for g and $\mathcal{H}$

In this section, we present expressions for the gradient vector and Hessian matrix of sample selection log-likelihood function (2) for the Clayton, Joe, FGM, AMH, Frank and Gumbel copulas, with normal margins. The expressions for the normal case can be found in Marra and Radice (2013). We use the notation $F_1 = \Phi(-\eta_{1i})$, $F_2 = \Phi(\tilde{e}_{2i})$ and $f_2 = \sigma^{-1}\phi(\tilde{e}_{2i})$, where $\eta_{vi} = \mathbf{X}_{vi}\boldsymbol{\delta}_v$, $\mathbf{X}_{vi} = \left(\mathbf{u}_{vi}^{\mathsf{T}}, \mathbf{B}_{vi}^{\mathsf{T}}\right)$, for $v = 1, 2$, $\tilde{e}_{2i} = \sigma^{-1}(y_{2i} - \eta_{2i})$, and Φ and ϕ are the standard normal distribution and density functions, respectively.

The elements of the gradient can be expressed as

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1} &= \sum_{i=1}^{n} \left\{(y_{1i}-1)F_1^{-1} + y_{1i}p_i\right\}\phi(-\eta_{1i})\mathbf{X}_{1i},\\
\frac{\partial \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2} &= \sum_{i=1}^{n} y_{1i}\left(h_i + \sigma^{-1}\tilde{e}_{2i}\right)\mathbf{X}_{2i},\\
\frac{\partial \ell(\boldsymbol{\delta}_*)}{\partial \sigma^*} &= \sum_{i=1}^{n} y_{1i}\left(h_i\sigma\tilde{e}_{2i} + \tilde{e}_{2i}^2 - 1\right),\\
\frac{\partial \ell(\boldsymbol{\delta}_*)}{\partial \theta^*} &= \sum_{i=1}^{n} y_{1i}b_i,
\end{aligned}
$$

whereas those of the Hessian as

$$
\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^T} &= \sum_{i=1}^{n} \left\{(y_{1i}-1)F_1^{-1}(F_1^{-1}\phi(-\eta_{1i}) - \eta_{1i}) + y_{1i}P_i\right\}\phi(-\eta_{1i})\mathbf{X}_{1i}^T\mathbf{X}_{1i},\\
\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^T} &= \sum_{i=1}^{n} y_{i1}A_i\phi(-\eta_{1i})\mathbf{X}_{1i}^T\mathbf{X}_{2i},\\
\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \sigma^*} &= \sum_{i=1}^{n} y_{1i}\sigma A_i\phi(-\eta_{1i})\tilde{e}_{2i}\mathbf{X}_{1i},\\
\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \theta^*} &= \sum_{i=1}^{n} y_{1i}h_{14}\phi(-\eta_{1i})\mathbf{X}_{1i},\\
\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^T} &= \sum_{i=1}^{n} y_{1i}(h_iE_i - \sigma^{-2})\mathbf{X}_{2i}^T\mathbf{X}_{2i},\\
\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2 \partial \sigma^*} &= \sum_{i=1}^{n} y_{1i}\sigma\left[h_i(E_i\tilde{e}_{2i} - \sigma^{-1}) - 2\sigma^{-2}\tilde{e}_{2i}\right]\mathbf{X}_{2i},\\
\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2 \partial \theta^*} &= \sum_{i=1}^{n} y_{1i}B_i\mathbf{X}_{2i},\\
\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \sigma^{*2}} &= \sum_{i=1}^{n} y_{1i}\sigma^2\tilde{e}_{2i}\left[h_i\left(E_i\tilde{e}_{2i} - \sigma^{-1}\right) - 2\sigma^{-2}\tilde{e}_{2i}\right],\\
\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \sigma^* \partial \theta^*} &= \sum_{i=1}^{n} y_{1i}\sigma B_i\tilde{e}_{2i},
\end{aligned}
$$

$$\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial {\theta^*}^2} = \sum_{i=1}^{n} y_{1i} h_{44},$$

where p_i, h_i, b_i, A_i, B_i, E_i, h_{14} and h_{44} are defined below for each copula.
For Clayton,

$$\begin{aligned}
\sigma &= \exp(\sigma^*),\\
\theta &= \exp(\theta^*) + \epsilon,\\
u_i &= F_1^{-\theta} + F_2^{-\theta} - 1,\\
z_i &= F_2^{-\theta-1} u^{-\frac{1+\theta}{\theta}},\\
p_i &= (\theta+1)\frac{z}{1-z} F_1^{-\theta-1} u^{-1},\\
h_i &= (\theta+1)\frac{z}{1-z} f_2 F_2^{-\theta-1}(u^{-1} - F_2^{\theta}),\\
C_i &= F_1^{-\theta}\log F_1 + F_2^{-\theta}\log F_2,\\
\tilde{C}_i &= F_1^{-\theta}(\log F_1)^2 + F_2^{-\theta}(\log F_2)^2,\\
b_i &= \frac{z}{1-z}\left(\theta \log F_2 - \frac{\log u_i}{\theta} - (1+\theta)\frac{C_i}{u_i}\right),\\
A_i &= -p_i\left(\frac{h_i}{z_i} + \frac{\theta}{u} f_2 F_2^{-\theta-1}\right),\\
B_i &= h_i\left[\frac{\theta}{\theta+1} - \frac{b_i}{z_i} + \theta\left(F_2^{\theta}u_i - 1\right)^{-1}\left(\log F_2 - \frac{C_i}{u_i}\right)\right],\\
E_i &= f_2 F_2^{-\theta-1}\left(F_2^{\theta} - \frac{\theta}{u_i}\right) - \frac{h_i}{z_i} + \sigma^{-1}\tilde{e}_{2i},\\
P_i &= p_i\left[F_1^{-1}\phi(-\eta_{1i})(\theta + 1 - F_1^{-\theta}u^{-1}(\theta + (1+\theta)(1-z)^{-1})) - \eta_{1i}\right],\\
h_{14} &= \theta p_i\left(\frac{1}{\theta+1} + \frac{C_i}{u} - \frac{b_i}{z_i\theta} - \log F_1\right),\\
h_{44} &= \frac{z}{1-z}\left[\theta\log F_2 + \frac{\log u_i}{\theta} + (1-\theta)\frac{C_i}{u_i} + \theta(\theta+1)\left(\frac{\tilde{C}_i}{u_i} - \frac{C_i^2}{u_i^2}\right)\right] - \frac{b_i^2}{z_i}.
\end{aligned}$$

For Joe,

$$\begin{aligned}
\sigma &= \exp(\sigma^*),\\
\theta &= 1 + \exp(\theta^*) + \epsilon,\\
u_i &= \bar{F}_1^{\theta} + \bar{F}_2^{\theta} - (\bar{F}_1\bar{F}_2)^{\theta},\\
z_i &= (1-\bar{F}_1^{\theta})\bar{F}_2^{\theta-1} u^{\frac{1-\theta}{\theta}},\\
p_i &= \frac{1}{1-z}(\bar{F}_1\bar{F}_2)^{\theta-1} u^{\frac{1-2\theta}{\theta}}(u+\theta-1),\\
b_i &= (\theta-1)\frac{z}{1-z}\left(\frac{\log u_i}{\theta^2} - \log\bar{F}_2 - \frac{1-\theta}{\theta}\frac{C_i}{u_i} + \frac{\bar{F}_1^{\theta}}{1-\bar{F}_1^{\theta}}\log\bar{F}_1\right),\\
h_i &= (\theta-1)\frac{z}{z-1} f_2 \bar{F}_2^{-1}\bar{F}_1^{\theta} u^{-1},\\
C_i &= \bar{F}_1^{\theta}\log\bar{F}_1 + \bar{F}_2^{\theta}\log\bar{F}_2 - (\bar{F}_1\bar{F}_2)^{\theta}\log(\bar{F}_1\bar{F}_2),\\
\tilde{C}_i &= \bar{F}_1^{\theta}(\log\bar{F}_1)^2 + \bar{F}_2^{\theta}(\log\bar{F}_2)^2 - (\bar{F}_1\bar{F}_2)^{\theta}(\log(\bar{F}_1\bar{F}_2))^2,
\end{aligned}$$

$$
\begin{aligned}
B_i &= h_i\left[1-\frac{b_i}{z_i}+(\theta-1)\left(\log\bar{F}_1-\frac{C_i}{u_i}\right)\right],\\
E_i &= f_2\bar{F}_2^{-1}\left(\theta\bar{F}_1^{\theta}u^{-1}-\theta-1\right)-\frac{h_i}{z_i}+\sigma^{-1}\tilde{e}_{2i},\\
A_i &= (\theta-1)p_if_2\bar{F}_2^{-1}\left(1+\frac{z}{1-z}\bar{F}_1^{\theta}u^{-1}\right)+\bar{F}_2^{\theta}\frac{2\theta+u-1}{u}h_i\bar{F}_1^{-1},\\
P_i &= \phi(-\eta_{1i})\bar{F}_1^{-1}\left[p_i(\theta-1-\bar{F}_1p_i)+(1-\theta)\left(1-\frac{\bar{F}_1^{\theta}}{u}\right)\left(2p_i+\frac{z}{1-z}\bar{F}_1^{-1}\left(1-\frac{\bar{F}_2^{\theta}}{u}\right)\right)\right]\\
&\quad -p_i\eta_{1i},\\
h_{14} &= p_i\left[(\theta-1)\left(\frac{1-(\theta-1)C_iu^{-1}}{u_i+\theta-1}+\frac{\log\bar{F}_1}{1-\bar{F}_1^{\theta}}\right)-\frac{b_i}{z_i}\right],\\
h_{44} &= b_i(1-b_i)+(\theta-1)\left[\left(\log\bar{F}_2+\frac{1-\theta}{\theta}\frac{C_i}{u_i}-\frac{\log u_i}{\theta^2}\right)\left(b_i-(\theta-1)\frac{z_i}{z_i-1}\frac{\log\bar{F}_1}{1-\bar{F}_\theta}\right)\right.\\
&\quad + \left. b_i\log\bar{F}_1+(\theta-1)\frac{z_i}{z_i-1}\left(\frac{1-\theta}{\theta}\left(\frac{\tilde{C}_i}{u}-\frac{C_i^2}{u^2}\right)-\frac{2}{\theta^2}\left(\frac{C_i}{u}-\frac{\log u}{\theta}\right)\right)\right].
\end{aligned}
$$

For FGM,

$$
\begin{aligned}
\sigma &= \exp(\sigma^*),\\
\theta &= \tanh(\theta^*),\\
u_i &= 1-\theta F_1(1-2F_2),\\
z_i &= 1-u_i(1-F_1),\\
p_i &= (1-F_1)^{-1}+\theta(1-2F_2)u^{-1},\\
b_i &= (\theta^2-1)F_1(1-2F_2)u^{-1},\\
h_i &= -2\theta F_1f_2u^{-1},\\
B_i &= (\theta^2-1)2F_1f_2u^{-2},\\
E_i &= -h_i+\tilde{e}_{2i}\sigma^{-1},\\
A_i &= 2\theta f_2u^{-2},\\
P_i &= \phi(-\eta_{1i})\left(p_i^2-2\theta u^{-1}(1-2F_2)(1-F_1)^{-1}\right)+\eta_{1i}p_i,\\
h_{14} &= (1-\theta^2)(1-2F_2)u^{-2},\\
h_{44} &= -b_i(2\theta+b_i).
\end{aligned}
$$

For Frank,

$$
\begin{aligned}
\sigma &= \exp(\sigma^*),\\
\theta &= \theta^*+\text{sign}(\theta^*)\epsilon,\\
u_i &= e^{\theta(F_1+F_2)}-e^{\theta(1+F_2)},\\
z_i &= 1-u\left(u-e^{\theta(1+F_1)}+e^{\theta}\right)^{-1},\\
p_i &= \theta(e^{\theta}-1)\frac{1-z}{u^2}e^{\theta(1+F_1+F_2)},\\
b_i &= \frac{1-z}{u}\left[\frac{z}{1-z}\left(u(F_1+F_2-1)+(F_1-1)e^{\theta(1+F_2)}\right)+F_1e^{\theta(1+F_1)}\right],
\end{aligned}
$$

$$
\begin{aligned}
h_i &= \theta e^{\theta}(e^{\theta F_1}-1)\frac{1-z}{u}f_2,\\
E_i &= \theta(1-z)f_2+\tilde{e}_{2i}\sigma^{-1},\\
A_i &= (1-e^{\theta})^{-1}p_i^2 f_2 u_i(e^{-\theta F_1}-e^{-\theta}),\\
B_i &= f_2\frac{1-z}{u}e^{\theta}\left\{(1+\theta+\theta F_1)e^{\theta F_1}-\theta-1\right.\\
&- \left.\theta(e^{\theta F_1}-1)\frac{1-z}{u}\left[u(F_1+F_2)+(F_1-1)e^{\theta(1+F_2)}-(1+F_1)e^{\theta(1+F_1)}+e^{\theta}\right]\right\},\\
P_i &= p_i\left[\phi(-\eta_{1i})(e^{\theta}-1)^{-1}p_i\left(e^{\theta F_1}\left(e^{\theta(F_2-1)}-1\right)-e^{\theta(1-F_1)}\left(e^{\theta F_2}-1\right)\right)-\eta_{1i}\right],\\
h_{14} &= \theta^{-1}(e^{\theta}-1)^{-1}p_i\left\{(1+\theta)e^{\theta}-1-p_i\left[(F_1+F_2-1)e^{\theta(F_1+F_2-1)}-2F_2e^{\theta F_2}-F_1e^{\theta F_1}\right.\right.\\
&+ \left.\left.(1-F_1+F_2)e^{\theta(1-F_1+F_2)}-(1-F_1)e^{\theta(1-F_1)}+e^{\theta}\right]\right\},\\
h_{44} &= b_i\left[b_i-u^{-1}\left(u(F_1+F_2)+e^{\theta(1+F_2)}(F_1-1)\right)\right]+\frac{1-z}{u}F_1(1+F_1)e^{\theta(1+F_1)}\\
&+ (F_1+F_2-1)\left(z(F_1+F_2)-b_i\right)+u^{-1}e^{\theta(1+F_2)}(F_1-1)\left(z(F_1+2F_2)-b_i\right).
\end{aligned}
$$

For AMH,

$$
\begin{aligned}
\sigma &= \exp(\sigma^*),\\
\theta &= \tanh(\theta^*),\\
u_i &= 1-\theta(1-F_1)(1-F_2),\\
z_i &= F_1(1-\theta+\theta F_1)u^{-2},\\
p_i &= \frac{1}{z-1}u^{-2}\left[2\theta(zu(1-F_2)-F_1)+\theta-1\right],\\
b_i &= \frac{1-\theta^2}{z-1}(1-F_1)\left(2zu^{-1}(1-F_2)-F_1u^{-2}\right),\\
h_i &= 2\theta\frac{z}{z-1}u^{-1}(1-F_1)f_2,\\
E_i &= \theta u^{-1}(1-F_1)f_2(z-3)(z-1)^{-1}+\sigma^{-1}\tilde{e}_{2i},\\
A_i &= 2\theta(z-1)^{-1}u^{-1}f_2\left[zu^{-1}-p_i(1-F_1)\right],\\
B_i &= 2(1-F_1)f_2u^{-1}(z-1)^{-1}\left((1-\theta^2)zu^{-1}-b_i\theta\right),\\
P_i &= \phi(-\eta_{1i})\left[\frac{2\theta}{u^2}\left((1-F_2)\left(2p_iu_i-\frac{z}{z-1}\theta(1-F_2)\right)+\frac{1}{z-1}\right)-p_i^2\right]-p_i\eta_{1i},\\
h_{14} &= \frac{1-\theta^2}{u^2}\left[2p_iu_i(1-F_1)(1-F_2)+\frac{z}{z-1}2(2u_i-1)(1-F_2)-\frac{2F_1-1}{z-1}\right]+2\theta(1-F_2)\frac{b_i}{u_i}-b_ip_i,\\
h_{44} &= -b_i(b_i+2\theta)+2\frac{1-\theta^2}{u}(1-F_1)(1-F_2)\left(2b_i-\frac{1-\theta^2}{u}(1-F_1)(1-F_2)\frac{z}{z-1}\right).
\end{aligned}
$$

For Gumbel,

$$
\begin{aligned}
\sigma &= \exp(\sigma^*),\\
\theta &= 1+\exp(\theta^*),\\
u_i &= (-\log F_1)^{\theta}+(-\log F_2)^{\theta},\\
z_i &= \exp\{-u^{1/\theta}\}F_2^{-1}u^{\frac{1-\theta}{\theta}}(-\log F_2)^{\theta-1},
\end{aligned}
$$

$$
\begin{aligned}
p_i &= \frac{z}{z-1}(-\log F_1)^{\theta-1}F_1^{-1}u^{-1}(1-\theta-u^{1/\theta}),\\
C_i &= (-\log F_1)^{\theta}\log(-\log F_1)+(-\log F_2)^{\theta}\log(-\log F_2),\\
\tilde{C}_i &= (-\log F_1)^{\theta}\left(\log(-\log F_1)\right)^2+(-\log F_2)^{\theta}\left(\log(-\log F_2)\right)^2,\\
b_i &= \frac{z}{z-1}(\theta-1)\left[\frac{1}{\theta}\frac{C_i}{u}\left(1-\theta-u^{1/\theta}\right)+\log(-\log F_2)-\frac{C_i}{u}\right],\\
h_i &= \frac{z}{z-1}f_2F_2^{-1}\left[1+(\theta-1)(-\log F_2)^{-1}+u^{-1}(-\log F_2)^{\theta-1}(1-\theta-u^{1/\theta})\right],\\
B_i &= (\theta-1)\frac{z}{z-1}f_2F_2^{-1}(-\log F_2)^{\theta-1}u^{-1}\left[(\theta+u^{1/\theta})\left(\frac{C_i}{u_i}-\log(-\log F_2)\right)-\frac{C_i}{u_i}\theta^{-1}-1\right.\\
&+ \left.\frac{\log u}{\theta^2}+\frac{z}{z-1}\frac{b_i}{\theta-1}\right]-\frac{h_ib_i}{z_i}-(\theta-1)\frac{z}{z-1}f_2F_2^{-1}(\log F_2)^{-1},\\
E_i &= \frac{\tilde{e}_{2i}}{\sigma}-\frac{h_i}{z_i}+f_2F_2^{-1}\left\{1+(\theta-1)(\log F_2)^{-2}\frac{z}{z-1}f_2F_2^{-1}h_i^{-1}\cdot\right.\\
&\left.\cdot\left[(-\log F_2)^{\theta}\left(\frac{1-\theta-u^{1/\theta}}{u}+\frac{\theta+u^{1/\theta}}{u^2}(-\log F_2)^{\theta}\right)-1\right]\right\},\\
A_i &= (\theta-1)f_2F_2^{-1}p_i(\theta+u^{1/\theta})u^{-1}(-\log F_2)^{\theta-1}(1-\theta-u^{1/\theta})^{-1}-\frac{h_ip_i}{z_i},\\
P_i &= p_i\phi(-\eta_{1i})\left[F_1^{-1}-F_1^{-1}(\log F_1)^{-1}\left(\theta-1-\frac{(-\log F_1)^{\theta}}{u}\left(\theta+\frac{u^{1/\theta}}{1-\theta-u^{1/\theta}}\right)\right)-\frac{p_i}{z_i}\right]\\
&-p_i\eta_{1i},\\
h_{14} &= p_i\left\{\frac{\theta-1}{1-\theta-u^{1/\theta}}\left[\frac{\theta-1}{\theta}\frac{C_i}{u_i}(\theta+u^{1/\theta})+u^{1/\theta}\frac{\log u}{\theta^2}-1\right]-\frac{b_i}{z_i}+(\theta-1)\log(-\log F_1)\right\},\\
h_{44} &= (\theta-1)^2\frac{z_i}{z_i-1}\left[\frac{1-u^{1/\theta}}{\theta}\left(\frac{\tilde{C}_i}{u}-\frac{C_i^2}{u^2}-\frac{2}{\theta}\left(\frac{C_i}{u}-\frac{\log u}{\theta}\right)\right)-u^{1/\theta}\frac{1}{\theta^2}\left(\frac{C_i}{u}-\frac{\log u}{\theta}\right)^2\right.\\
&- \left.\frac{\tilde{C}_i}{u}+\frac{C_i^2}{u^2}\right]+b_i-\frac{b_i^2}{z_i}.
\end{aligned}
$$

Appendix B - Tables and figures of simulation results

For convenience, in this section we report all the tables and figures of results which are commented in Sections 4.1 and 4.2.

		$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		
		Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	*AIC* (%)	*BIC* (%)
$\tau = 0.1$	Normal	10.2	0.336	5.7	0.292	0.1	0.034	-62.7	0.18	0.118	10	6.8
	Clayton	6.4	0.174	4.5	0.171	0.5	0.04	-40.5	0.087	0.093	33.6	38.8
	Joe	4.7	0.193	2.9	0.188	-2	0.038	-48.4	0.095	0.096	10.8	14
	FGM	14.2	0.278	7.3	0.248	-0.3	0.032	-78.2	0.156	0.109	12.8	12.8
	AMH	12.5	0.263	6.9	0.236	0.3	0.037	-66.1	0.147	0.108	8.4	4.8
	Frank	13.2	0.309	6.9	0.271	-0.1	0.033	-74.4	0.173	0.114	7.2	7.2
	Gumbel	0.1	0.22	1.2	0.203	-1.4	0.034	-25.9	0.108	0.098	17.2	15.6
$\tau = 0.5$	Normal	-4.5	0.169	-0.4	0.161	-0.5	0.046	1	0.078	0.086	56.9	22
	Clayton	33	0.308	17.5	0.313	0.9	0.049	-25.6	0.161	0.123	3.7	3.3
	Joe	4.3	0.329	1.2	0.29	-5.2	0.079	-14.5	0.189	0.114	1.6	2.8
	FGM	59.2	0.464	26	0.408	-8	0.085	-56.6	0.285	0.161	1.6	26.8
	AMH	45.9	0.368	22.3	0.356	-2.2	0.042	-38	0.196	0.138	10.2	28.5
	Frank	5.2	0.223	3.6	0.207	-1.3	0.052	-7.3	0.118	0.093	8.5	7.7
	Gumbel	-9.3	0.226	-3.5	0.204	-1.8	0.056	3.2	0.111	0.093	17.5	8.9
$\tau = 0.7$	Normal	-6	0.127	-1	0.13	-0.1	0.042	2.4	0.049	0.076	62.3	52.9
	Clayton	11.1	0.177	8.6	0.204	2.4	0.052	-4.2	0.084	0.086	9.4	12.1
	Joe	-12.2	0.159	-5.1	0.154	-1.5	0.049	0.7	0.059	0.081	0.9	1.3
	FGM	89.9	0.687	38.8	0.591	-14.1	0.143	-68.3	0.478	0.219	0	0
	AMH	78.3	0.599	36	0.549	-8.6	0.091	-52.7	0.369	0.198	0	4.5
	Frank	-2.9	0.126	0.7	0.132	-0.1	0.043	-0.2	0.052	0.075	6.7	10.3
	Gumbel	-10.6	0.144	-3.5	0.139	-0.2	0.043	5	0.058	0.078	20.6	18.8

Table 4: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}(z_1)$, and percentage frequency at which each copula model was selected by *AIC* and *BIC* for data simulated using a normal bivariate distribution, when employing the normal, Clayton, Joe, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Number of simulated datasets is equal to 250. τ denotes the association between the selection and outcome equations. See Section 4 for further details.

		$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		
		Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	*AIC* (%)	*BIC* (%)
$\tau = 0.1$	Normal	23.2	0.431	10.8	0.373	-1.7	0.044	-125.2	0.242	0.139	13.6	5.6
	Clayton	-3.3	0.18	0.6	0.164	-0.8	0.048	-4	0.093	0.09	36.8	31.2
	Joe	6.7	0.204	3.6	0.192	-3.9	0.052	-61.8	0.108	0.1	12	23.2
	FGM	25.9	0.333	11.9	0.294	-2.5	0.043	-137.5	0.2	0.125	13.2	12.4
	AMH	22.5	0.31	10.8	0.277	-1.8	0.044	-117	0.185	0.121	9.2	7.6
	Frank	27.6	0.379	12.6	0.33	-2.2	0.043	-144.4	0.226	0.132	7.2	8
	Gumbel	3.3	0.225	2.3	0.205	-3.5	0.049	-45.6	0.12	0.102	8	12
$\tau = 0.5$	Normal	-14.7	0.35	-6.6	0.304	-4.1	0.069	2.3	0.192	0.099	4.4	3.6
	Clayton	-3.8	0.156	0.3	0.147	-0.8	0.047	0.9	0.078	0.077	81.2	33.2
	Joe	5.8	0.519	0.6	0.45	-7.6	0.108	-24.8	0.312	0.155	0	0.8
	FGM	63.7	0.564	26	0.466	-12.3	0.127	-72.9	0.4	0.174	0	2.4
	AMH	46.5	0.449	20.7	0.388	-7.3	0.083	-48.1	0.297	0.142	14	56
	Frank	6.1	0.517	2.2	0.433	-4.8	0.079	-19.5	0.32	0.13	0.4	2
	Gumbel	-15.1	0.416	-7.4	0.361	-4.8	0.084	0.3	0.233	0.119	0	2
$\tau = 0.7$	Normal	-18.6	0.184	-7.9	0.17	-2	0.05	4.9	0.067	0.076	2.8	8.1
	Clayton	-6.3	0.128	-0.9	0.126	-0.4	0.042	2.7	0.061	0.069	96.3	80.9
	Joe	-28.2	0.264	-13.3	0.249	-2.1	0.06	3.5	0.092	0.095	0	0
	FGM	82.4	0.642	34.3	0.534	-16.1	0.164	-70.3	0.498	0.209	0	0
	AMH	69.6	0.532	31	0.473	-11.3	0.117	-52.4	0.367	0.187	0	7.7
	Frank	-19.7	0.189	-8	0.171	-1.5	0.048	3.4	0.061	0.076	0.4	2.4
	Gumbel	-24.7	0.222	-10.9	0.204	-1.3	0.049	7.8	0.08	0.082	0.4	0.8

Table 5: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}(z_1)$, and percentage frequency at which each copula model was selected by *AIC* and *BIC* for data simulated using a bivariate Clayton copula with normal margins, when employing the normal, Clayton, Joe, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Number of simulated datasets is equal to 250. τ denotes the association between the selection and outcome equations. See Section 4 for further details.

		$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		
		Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	*AIC* (%)	*BIC* (%)
$\tau = 0.15$	Normal	6.5	0.231	5.2	0.218	3.3	0.049	-16.5	0.114	0.1	6.4	5.2
	Clayton	31.9	0.278	16	0.275	2.9	0.046	-81.6	0.133	0.115	18	51.2
	Joe	-2.1	0.189	0.5	0.185	-0.2	0.033	-5.8	0.083	0.095	57.2	28.8
	FGM	23.2	0.244	12.2	0.243	2.5	0.042	-57.8	0.116	0.106	2.4	4.8
	AMH	25.8	0.247	13.4	0.248	2.7	0.043	-64.1	0.117	0.108	0.8	0.4
	Frank	18.9	0.238	10.4	0.235	2.7	0.043	-45.8	0.117	0.104	5.2	2.4
	Gumbel	-5.3	0.196	-0.5	0.184	1.4	0.036	8.2	0.091	0.093	10	7.2
$\tau = 0.5$	Normal	14.4	0.182	9.8	0.202	2.1	0.049	-10.5	0.096	0.093	1.8	0.9
	Clayton	78.4	0.623	37.8	0.591	0	0.051	-62.3	0.331	0.195	0.5	1.4
	Joe	-3.6	0.119	-0.1	0.121	-0.8	0.043	1.4	0.061	0.082	82.4	66.2
	FGM	68.4	0.527	32.3	0.495	-4.1	0.053	-55.8	0.279	0.174	0.5	14.4
	AMH	70	0.54	33.6	0.516	-2	0.042	-53.7	0.271	0.176	0	0.5
	Frank	21.7	0.209	12.9	0.231	1	0.042	-11.2	0.09	0.096	3.6	4.5
	Gumbel	-2.8	0.117	1.4	0.121	1.6	0.045	3.6	0.063	0.082	11.3	12.2
$\tau = 0.75$	Normal	3.8	0.107	5.2	0.141	1.4	0.046	-1.7	0.054	0.081	0.6	2.8
	Clayton	27.8	0.303	17.9	0.336	3.9	0.068	-12.3	0.168	0.115	0	0.6
	Joe	-4.1	0.099	0	0.109	-0.2	0.042	2.8	0.043	0.074	82.8	65.6
	FGM	99.9	0.762	44.7	0.677	-13.3	0.138	-70.4	0.528	0.239	0	0
	AMH	92	0.703	43.1	0.653	-8.9	0.096	-58.2	0.437	0.224	0	0
	Frank	7.2	0.112	6.3	0.146	0.2	0.042	0.3	0.044	0.076	3.3	10
	Gumbel	-3	0.096	1.5	0.111	1.1	0.044	3.4	0.045	0.075	13.3	21.1

Table 6: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}(z_1)$, and percentage frequency at which each copula model was selected by *AIC* and *BIC* for data simulated using a bivariate Joe copula with normal margins, when employing the normal, Clayton, Joe, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Number of simulated datasets is equal to 250. τ denotes the association between the selection and outcome equations. See Section 4 for further details.

		$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		
		Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	*AIC* (%)	*BIC* (%)
$\tau = -0.2$	Normal	0.3	0.294	1.9	0.259	0.5	0.039	10.6	0.149	0.105	10.8	2.8
	Clayton	-64.1	0.511	-24.4	0.393	-1.1	0.038	-126.5	0.265	0.14	22.4	30.8
	Joe	-55.3	0.444	-21.3	0.352	-2.6	0.041	-104.2	0.215	0.123	4.8	7.6
	FGM	-11.6	0.2	-3	0.178	-0.7	0.035	-12.6	0.089	0.087	29.2	31.2
	AMH	-18.4	0.22	-5.5	0.184	-0.4	0.035	-30	0.097	0.086	13.6	12
	Frank	0.9	0.251	2.3	0.231	0.3	0.039	14.5	0.129	0.102	15.6	7.6
	Gumbel	-56	0.454	-21.6	0.358	-2.5	0.041	-106	0.222	0.124	3.6	8
$\tau = 0.15$	Normal	13.5	0.352	7.5	0.312	-0.3	0.035	-51	0.189	0.117	8.4	4
	Clayton	16.6	0.21	9.3	0.206	0.2	0.039	-54.3	0.117	0.106	28	34.8
	Joe	7.2	0.287	4.2	0.264	-2.3	0.041	-41.5	0.145	0.113	17.2	23.2
	FGM	12.3	0.267	6.9	0.245	-0.8	0.035	-45.7	0.144	0.103	16.4	14.8
	AMH	14.8	0.269	8.3	0.249	0	0.038	-48.5	0.145	0.105	8.8	6.4
	Frank	9.7	0.322	5.9	0.289	-0.4	0.036	-38.3	0.171	0.111	11.2	6.8
	Gumbel	1.9	0.294	2.2	0.266	-1.5	0.038	-23.2	0.148	0.112	10	10

Table 7: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}(z_1)$, and percentage frequency at which each copula model was selected by *AIC* and *BIC* for data simulated using a bivariate FGM copula with normal margins, when employing the normal, Clayton, Joe, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Number of simulated datasets is equal to 250. τ denotes the association between the selection and outcome equations. See Section 4 for further details.

		$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		
		Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	*AIC* (%)	*BIC* (%)
$\tau = -0.12$	Normal	6.1	0.344	3.7	0.295	0.6	0.042	35.6	0.176	0.112	11.2	4.8
	Clayton	-50.1	0.413	-19.3	0.324	-1	0.038	-133.6	0.202	0.12	26.8	33.6
	Joe	-42.3	0.349	-16.5	0.287	-2.3	0.039	-105.7	0.154	0.106	5.6	8.8
	FGM	-4.5	0.213	-0.6	0.193	-0.6	0.034	7.7	0.101	0.092	17.2	23.2
	AMH	-10.7	0.227	-2.9	0.197	-0.1	0.036	-15.6	0.105	0.089	16.8	13.2
	Frank	7.6	0.303	4.5	0.267	0.4	0.041	44.4	0.159	0.108	17.6	9.2
	Gumbel	-43.3	0.358	-16.9	0.292	-2.3	0.039	-108.8	0.161	0.107	4.8	7.2
$\tau = 0.1$	Normal	16.5	0.382	7.9	0.327	0	0.035	-93.2	0.204	0.122	9.2	7.6
	Clayton	6.9	0.162	4.3	0.157	0	0.036	-49.5	0.088	0.091	37.6	43.2
	Joe	2.7	0.23	1.8	0.218	-2	0.038	-46.2	0.108	0.098	12	14
	FGM	13.8	0.306	6.9	0.269	-0.4	0.033	-82.1	0.164	0.111	10.8	9.6
	AMH	13.1	0.293	6.8	0.259	0.3	0.037	-72.8	0.157	0.108	11.2	7.6
	Frank	11.8	0.339	6	0.294	-0.2	0.034	-73.9	0.183	0.114	10	7.6
	Gumbel	-0.9	0.245	0.4	0.22	-1.4	0.035	-27.7	0.119	0.099	9.2	10.4
$\tau = 0.28$	Normal	16.9	0.493	6.6	0.409	-4.1	0.06	-40.7	0.29	0.134	7.2	2.8
	Clayton	11.6	0.204	5.9	0.184	-1.7	0.049	-25	0.128	0.093	26.4	10
	Joe	25.8	0.439	9.6	0.375	-7	0.082	-59	0.265	0.147	2.8	14.8
	FGM	46.2	0.493	18.8	0.411	-6.3	0.071	-83.8	0.313	0.145	12.8	30
	AMH	32.7	0.434	14.1	0.366	-3.4	0.056	-57.6	0.268	0.128	35.2	22.8
	Frank	34.4	0.58	13.9	0.484	-4.3	0.062	-66.6	0.35	0.159	11.2	8
	Gumbel	11.2	0.414	3.7	0.354	-5.7	0.074	-35.2	0.243	0.134	4.4	11.6

Table 8: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}(z_1)$, and percentage frequency at which each copula model was selected by *AIC* and *BIC* for data simulated using a bivariate AMH copula with normal margins, when employing the normal, Clayton, Joe, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Number of simulated datasets is equal to 250. τ denotes the association between the selection and outcome equations. See Section 4 for further details.

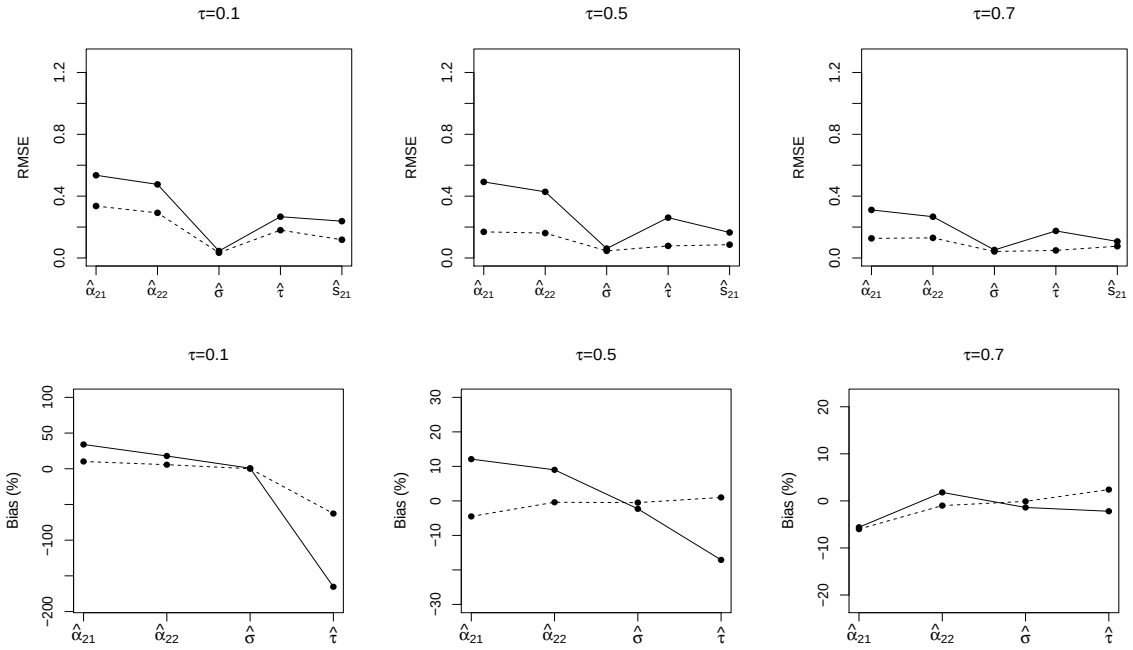
		$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		
		Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	*AIC* (%)	*BIC* (%)
$\tau = 0.2$	Normal	7.8	0.345	5	0.3	-0.3	0.039	-30.3	0.189	0.117	9.6	4.4
	Clayton	22.2	0.242	11.3	0.233	0.2	0.044	-55.1	0.145	0.112	19.2	24.8
	Joe	1.8	0.33	1.1	0.299	-2.8	0.046	-26.6	0.166	0.119	12	18.8
	FGM	11.3	0.229	6	0.211	-1.2	0.035	-35.6	0.132	0.1	19.2	18.4
	AMH	13.7	0.238	7.5	0.221	-0.1	0.041	-36.4	0.136	0.103	6.4	8
	Frank	3.2	0.282	2.6	0.246	-0.5	0.037	-18	0.157	0.106	23.6	12.8
	Gumbel	-3.7	0.324	-0.8	0.286	-1.5	0.042	-10.3	0.167	0.115	10	12.8
$\tau = 0.5$	Normal	-2.7	0.191	0.4	0.18	-0.7	0.044	-2.8	0.103	0.092	9.6	7.2
	Clayton	40.7	0.376	20.6	0.362	-0.1	0.053	-34.7	0.218	0.141	3.2	2.8
	Joe	-18.6	0.283	-8.6	0.259	-2.6	0.057	3.4	0.131	0.103	4	2.4
	FGM	56.5	0.441	24.9	0.39	-7.9	0.084	-56.3	0.284	0.158	4.4	33.2
	AMH	49.5	0.392	23.3	0.369	-3.8	0.051	-43.2	0.222	0.146	4.4	10.8
	Frank	-3.9	0.191	-0.3	0.174	-0.6	0.042	0.2	0.103	0.086	62.8	34.8
	Gumbel	-17.6	0.233	-7	0.209	-0.7	0.047	8.6	0.108	0.093	11.6	8.8
$\tau = 0.7$	Normal	-5.8	0.117	-0.9	0.122	0.2	0.038	0.7	0.059	0.078	7.7	10.2
	Clayton	15.8	0.231	10.9	0.249	2.8	0.055	-7.4	0.136	0.101	5.5	5.5
	Joe	-19.4	0.179	-8.6	0.174	-0.4	0.041	4.4	0.064	0.082	8.1	4.3
	FGM	87.9	0.671	38	0.577	-13.6	0.138	-68.3	0.478	0.219	0	0.4
	AMH	79	0.604	36.1	0.549	-8.9	0.093	-54.2	0.379	0.204	0	0
	Frank	-6.9	0.111	-1.6	0.116	0	0.036	2.7	0.051	0.071	68.9	68.5
	Gumbel	-14.3	0.149	-5.2	0.141	0.4	0.039	5.4	0.062	0.077	9.8	11.1

Table 9: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}(z_1)$, and percentage frequency at which each copula model was selected by *AIC* and *BIC* for data simulated using a bivariate Frank copula with normal margins, when employing the normal, Clayton, Joe, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Number of simulated datasets is equal to 250. τ denotes the association between the selection and outcome equations. See Section 4 for further details.

		$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		
		Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	*AIC* (%)	*BIC* (%)
$\tau = 0.2$	Normal	10.7	0.266	6.8	0.243	1.6	0.04	-23.5	0.137	0.105	10.8	4
	Clayton	36.8	0.322	18	0.311	1.5	0.042	-74.2	0.168	0.127	26.4	47.6
	Joe	10.1	0.236	5.4	0.226	-1.8	0.038	-31.3	0.12	0.102	32	21.2
	FGM	28.4	0.284	14.1	0.277	0.6	0.033	-57.1	0.145	0.114	6.4	6.8
	AMH	31.3	0.294	15.5	0.287	1	0.038	-61.7	0.15	0.118	0.8	1.2
	Frank	23.2	0.281	12	0.271	0.9	0.036	-46.1	0.145	0.112	3.6	5.2
	Gumbel	3.4	0.227	3	0.213	-0.3	0.035	-13.5	0.115	0.099	20	14
$\tau = 0.5$	Normal	8.9	0.154	6.4	0.171	0.4	0.039	-9.1	0.089	0.091	11.7	3.3
	Clayton	59.7	0.489	29.3	0.468	0.1	0.048	-48.7	0.269	0.165	0.4	1.2
	Joe	2.9	0.176	1.7	0.175	-3.8	0.057	-9.5	0.105	0.096	22.5	21.2
	FGM	63.6	0.49	29.2	0.451	-5.8	0.064	-56.1	0.281	0.168	2.1	31.2
	AMH	59.7	0.465	28.6	0.444	-2.4	0.043	-48.6	0.248	0.161	1.2	5.4
	Frank	19.2	0.204	10.8	0.217	-0.8	0.039	-14.9	0.11	0.098	8.3	5
	Gumbel	-1.2	0.15	1.1	0.155	-0.8	0.041	-1.4	0.08	0.088	53.8	32.9
$\tau = 0.8$	Normal	-1	0.089	1.9	0.114	0.1	0.035	-0.2	0.041	0.072	17.7	26.5
	Clayton	8.2	0.141	7.8	0.179	2.8	0.047	-2.2	0.075	0.081	2.8	3.9
	Joe	-5.5	0.096	-1.1	0.108	-1	0.037	-0.4	0.044	0.074	15.5	8.8
	FGM	103	0.785	45	0.682	-16	0.162	-72.2	0.578	0.244	0	0
	AMH	93.1	0.71	42.9	0.65	-11	0.113	-58.7	0.469	0.227	0	0.6
	Frank	1.8	0.092	3.3	0.121	-0.1	0.034	-1	0.045	0.073	2.8	9.9
	Gumbel	-4.2	0.091	0.1	0.107	0	0.035	1.9	0.04	0.073	61.3	50.3

Table 10: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}(z_1)$, and percentage frequency at which each copula model was selected by *AIC* and *BIC* for data simulated using a bivariate Gumbel copula with normal margins, when employing the normal, Clayton, Joe, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Number of simulated datasets is equal to 250. τ denotes the association between the selection and outcome equations. See Section 4 for further details.

a) Normal copula



b) Clayton copula

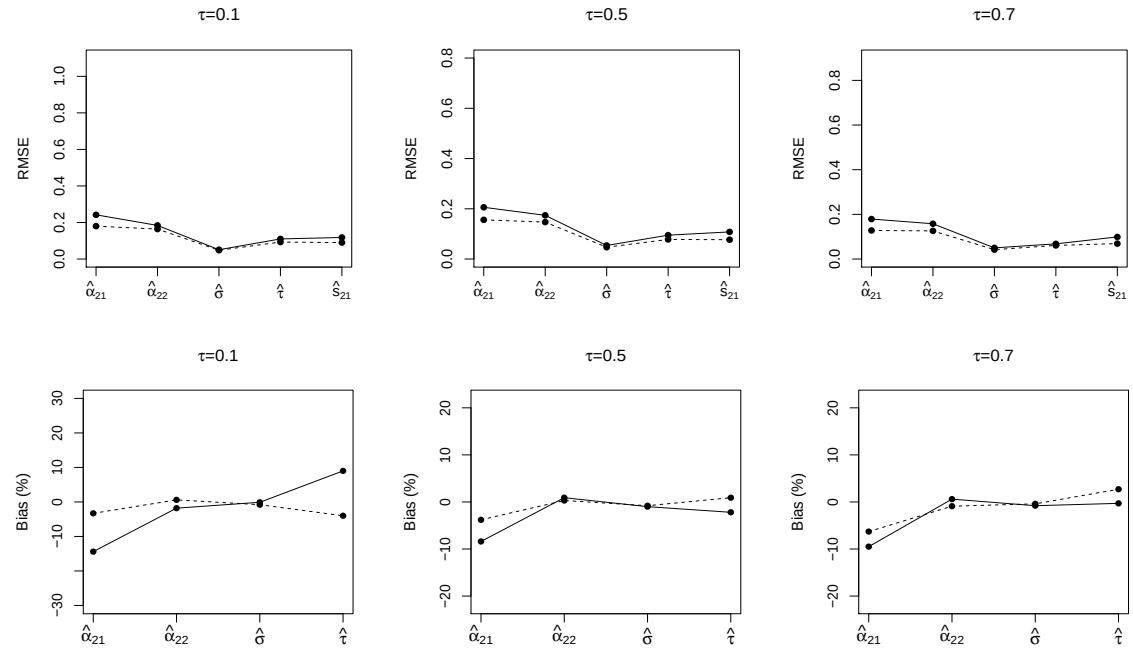
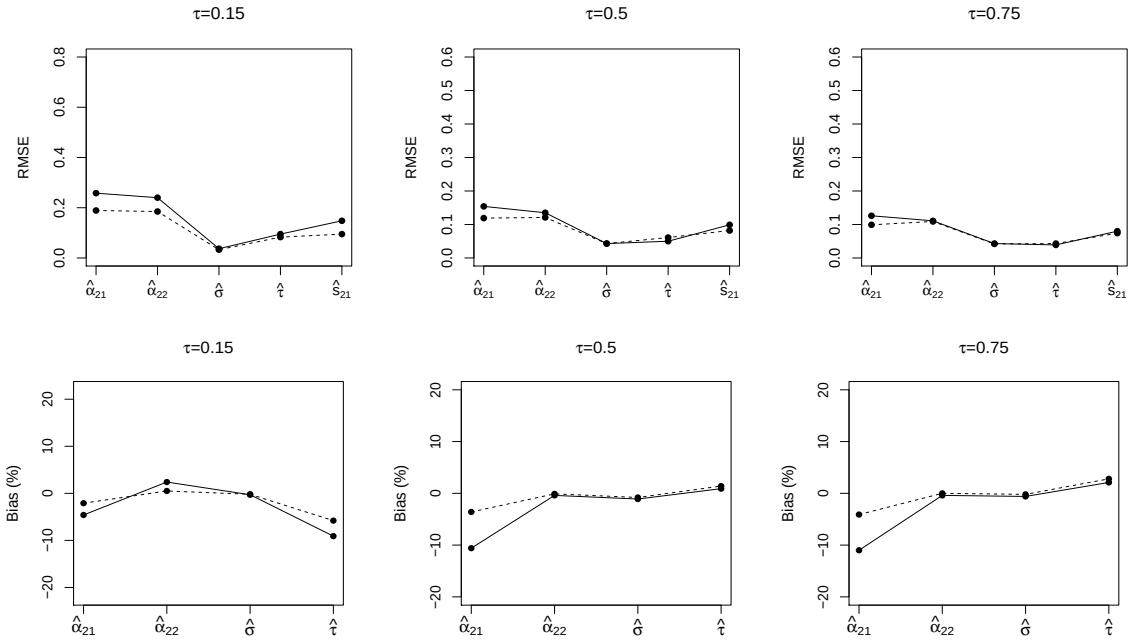


Figure 5: RMSEs and percentage bias of $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}$ for data generated using the a) normal and b) Clayton copulas when employing the normal and Clayton copula regression spline sample selection models, respectively. Solid line: model (10) without exclusion restriction. Dotted line: model (9) with exclusion restriction.

a) Joe copula



b) FGM copula

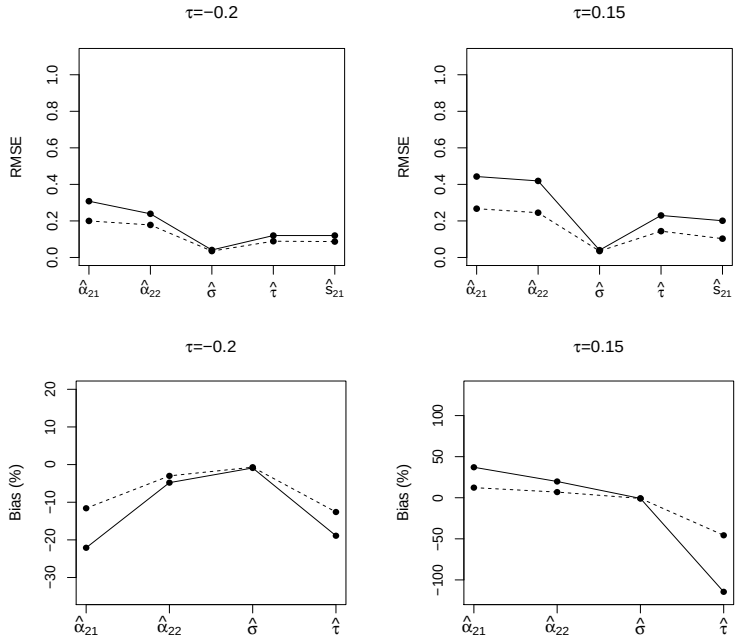
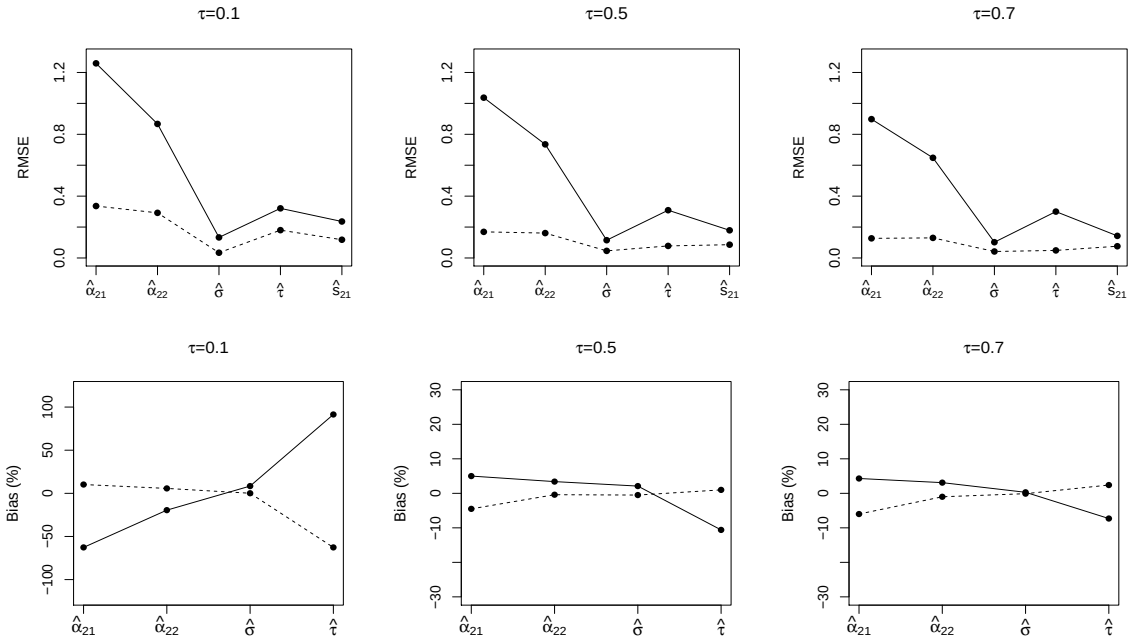


Figure 6: RMSEs and percentage bias of $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}$ for data simulated using the a) Joe and b) FGM copulas when employing the Joe and FGM copula regression spline sample selection models, respectively. Solid line: model (10) without exclusion restriction. Dotted line: model (9) with exclusion restriction.

a) Normal copula



b) Clayton copula

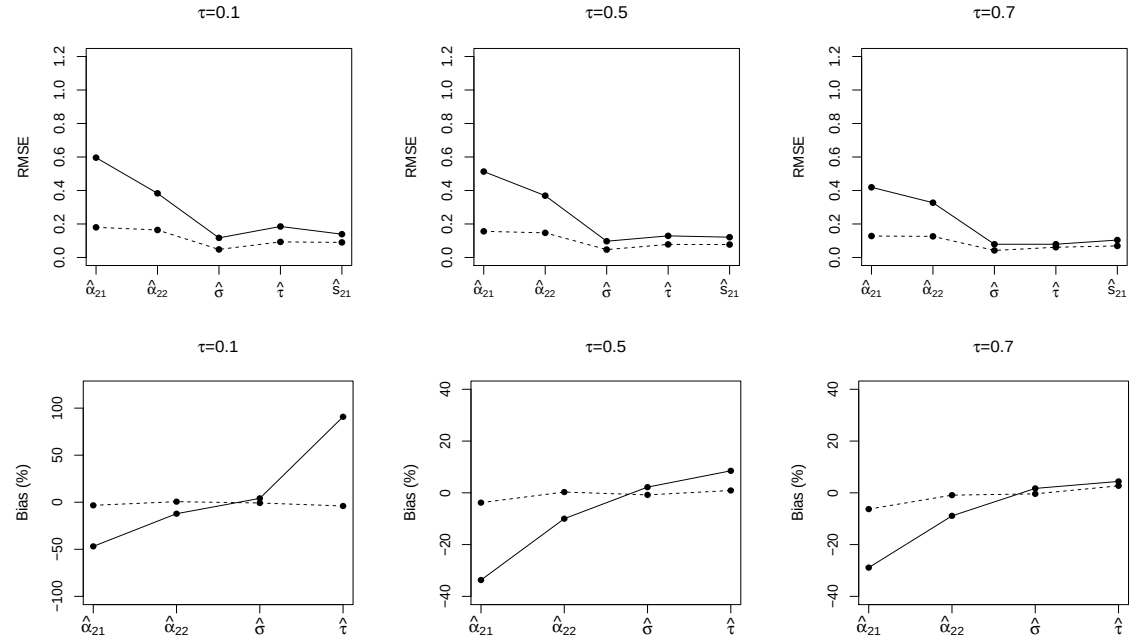
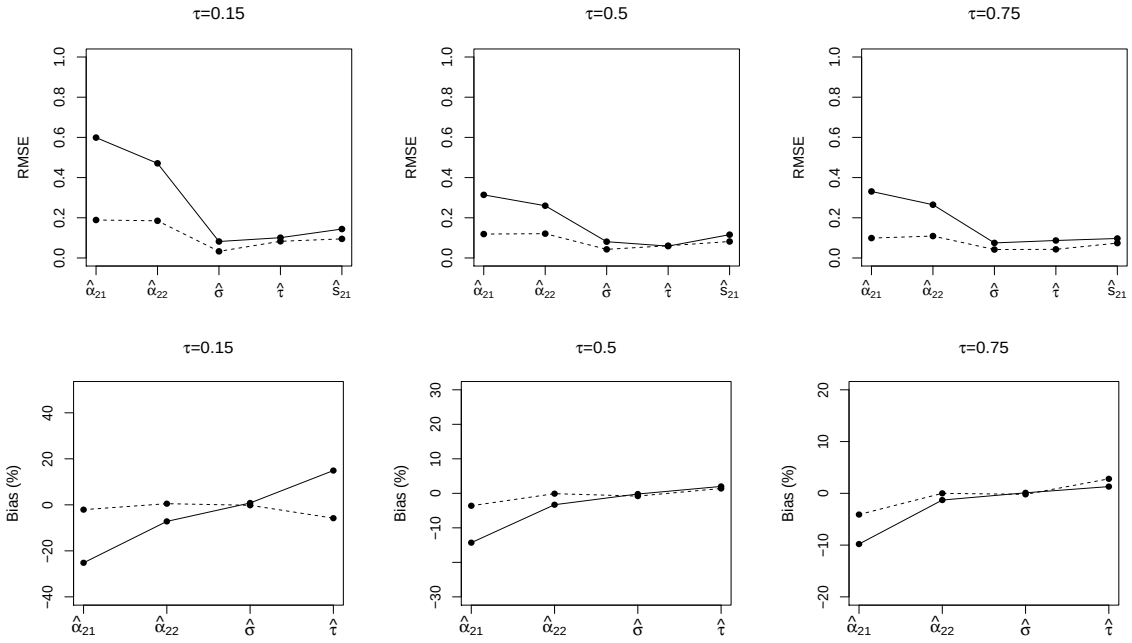


Figure 7: RMSEs and percentage bias of $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}$ for data generated using the a) normal and b) Clayton copulas when employing the normal and Clayton copula regression spline sample selection models, respectively. Solid line: model (11) without exclusion restriction. Dotted line: model (9) with exclusion restriction.

a) Joe copula



b) FGM copula

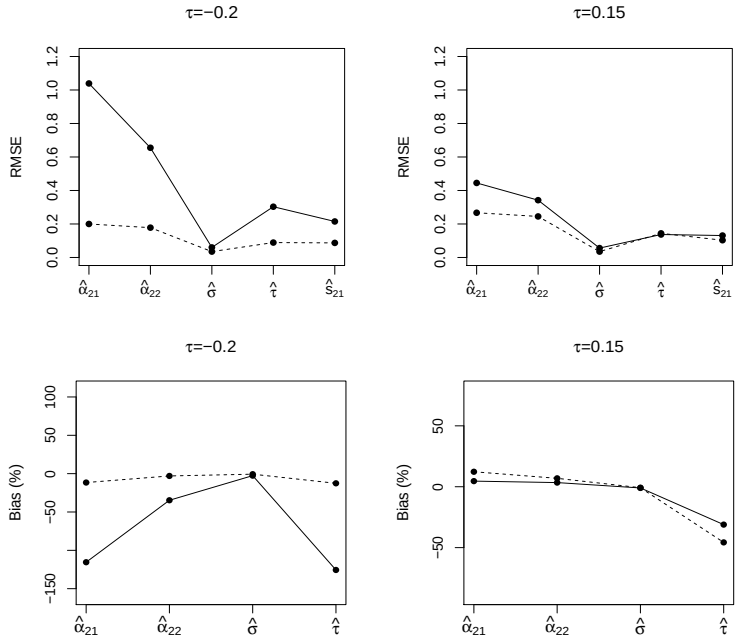


Figure 8: RMSEs and percentage bias of $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{s}_{21}$ for data simulated using the a) Joe and b) FGM copulas when employing the Joe and FGM copula regression spline sample selection models, respectively. Solid line: model (11) without exclusion restriction. Dotted line: model (9) with exclusion restriction.

n	$\alpha(\%)$	Copula				
		Normal	FGM	AMH	Frank	Gumbel
1000	1	14	0	2	9	29
	5	23	0	5	16	30
	10	28	0	9	23	31
3000	1	4	0	2	3	22
	5	11	1	4	9	27
	10	17	8	10	17	29
5000	1	3	0	1	2	13
	5	10	4	5	7	20
	10	14	8	12	13	24

Table 11: Empirical sizes (%) for testing the null hypothesis $H_0 : \tau = 0$, when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Random samples were generated using the equation (9) and a normal product distribution.

n	$\alpha(\%)$	Copula				
		Normal	FGM	AMH	Frank	Gumbel
1000	1	20	0	0	12	34
	5	33	0	2	20	35
	10	40	0	2	24	36
3000	1	11	0	0	4	27
	5	20	2	2	12	30
	10	28	7	3	19	30
5000	1	8	1	2	3	18
	5	16	4	3	9	22
	10	22	9	6	12	25

Table 12: Empirical sizes (%) for testing the null hypothesis $H_0 : \tau = 0$, when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Random samples were generated using the equation (10) and a normal product distribution.

n	$\alpha(\%)$	Copula				
		Normal	FGM	AMH	Frank	Gumbel
1000	1	20	0	3	13	40
	5	30	0	12	23	41
	10	38	0	17	29	41
3000	1	12	0	0	6	31
	5	18	3	2	13	35
	10	25	10	4	22	39
5000	1	9	1	3	5	22
	5	15	6	5	12	29
	10	20	8	7	15	32

Table 13: Empirical sizes (%) for testing the null hypothesis $H_0 : \tau = 0$, when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Random samples were generated using the equation (11) and a normal product distribution.

Affiliation:

Małgorzata Wojtyś
School of Computing and Mathematics
University of Plymouth
PL4 8AA Plymouth, United Kingdom
E-mail: malgorzata.wojtys@plymouth.ac.uk
URL: http://www.plymouth.ac.uk/staff/mwojtys
and
Faculty of Mathematics and Information Science
Warsaw University of Technology
ul. Koszykowa 75, 00-662 Warszawa

a) data generated using the bivariate normal copula

Copula	$\tau_1 = 0.1$	$\tau_2 = 0.5$	$\tau_3 = 0.7$
Normal	0.26	1.00	1.00
FGM	0.00	0.00	0.00
AMH	0.14	0.47	0.21
Frank	0.23	0.97	1.00
Gumbel	0.66	0.99	1.00

b) data generated using the bivariate Clayton copula

Copula	$\tau_1 = 0.1$	$\tau_2 = 0.5$	$\tau_3 = 0.7$
Normal	0.28	0.95	1.00
FGM	0.00	0.00	0.00
AMH	0.14	0.13	0.08
Frank	0.25	0.94	1.00
Gumbel	0.38	0.94	1.00

c) data generated using the bivariate Joe copula

Copula	$\tau_1 = 0.15$	$\tau_2 = 0.5$	$\tau_3 = 0.75$
Normal	0.23	0.99	1.00
FGM	0.00	0.00	0.00
AMH	0.05	0.99	0.95
Frank	0.11	1.00	1.00
Gumbel	0.95	1.00	1.00

Table 14: Powers for testing the null hypothesis $H_0 : \tau = 0$, when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Random samples were generated using the equation (9).

a) data generated using the bivariate FGM copula

Copula	$\tau_1 = -0.2$	$\tau_2 = 0.15$
Normal	0.54	0.23
FGM	0.00	0.00
AMH	0.03	0.23
Frank	0.63	0.27
Gumbel	0.13	0.60

b) data generated using the bivariate AMH copula

Copula	$\tau_1 = -0.12$	$\tau_2 = 0.1$	$\tau_3 = 0.28$
Normal	0.45	0.29	0.65
FGM	0.00	0.00	0.00
AMH	0.04	0.17	0.54
Frank	0.48	0.20	0.60
Gumbel	0.27	0.49	0.60

c) data generated using the bivariate Gumbel copula

Copula	$\tau_1 = 0.2$	$\tau_2 = 0.5$	$\tau_3 = 0.8$
Normal	0.38	0.97	1.00
FGM	0.00	0.00	0.00
AMH	0.16	0.79	0.46
Frank	0.25	0.99	1.00
Gumbel	0.89	0.99	1.00

d) data generated using the bivariate Frank copula

Copula	$\tau_1 = 0.2$	$\tau_2 = 0.5$	$\tau_3 = 0.7$
Normal	0.41	0.99	1.00
FGM	0.00	0.00	0.00
AMH	0.43	0.87	0.85
Frank	0.53	0.99	1.00
Gumbel	0.76	0.99	1.00

Table 15: Powers for testing the null hypothesis $H_0 : \tau = 0$, when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Random samples were generated using the equation (9).

Giampiero Marra
Department of Statistical Science
University College London
WC1E 6BT London, United Kingdom
E-mail: giampiero.marra@ucl.ac.uk
URL: http://www.ucl.ac.uk/statistics/people/giampieromarra

Rosalba Radice
Department of Economics, Mathematics and Statistics
Birkbeck, University of London
WC1E 7HX London, United Kingdom
E-mail: r.radice@bbk.ac.uk
URL: http://www.ems.bbk.ac.uk/faculty/radice

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd

a) data generated using the bivariate normal copula

Copula	$\tau_1 = 0.1$	$\tau_2 = 0.5$	$\tau_3 = 0.7$
Normal	0.29	0.97	0.98
FGM	0.00	0.00	0.00
AMH	0.05	0.38	0.25
Frank	0.21	0.85	1.00
Gumbel	0.45	0.92	1.00

b) data generated using the bivariate Clayton copula

Copula	$\tau_1 = 0.1$	$\tau_2 = 0.5$	$\tau_3 = 0.7$
Normal	0.39	0.79	1.00
AMH	0.00	0.00	0.00
FGM	0.03	0.03	0.00
Frank	0.34	0.86	1.00
Gumbel	0.31	0.63	0.94

c) data generated using the bivariate Joe copula

Copula	$\tau_1 = 0.15$	$\tau_2 = 0.5$	$\tau_3 = 0.75$
Normal	0.05	0.98	1
FGM	0.00	0.00	0
AMH	0.00	0.94	1
Frank	0.02	1.00	1
Gumbel	0.89	1.00	1

Table 16: Powers for testing the null hypothesis $H_0 : \tau = 0$, when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Random samples were generated using the equation (10).

a) data generated using the bivariate FGM copula

Copula	$\tau_1 = -0.2$	$\tau_2 = 0.15$
Normal	0.70	0.34
FGM	0.00	0.00
AMH	0.04	0.11
Frank	0.56	0.34
Gumbel	0.46	0.48

b) data generated using the bivariate AMH copula

Copula	$\tau_1 = -0.12$	$\tau_2 = 0.1$	$\tau_3 = 0.28$
Normal	0.64	0.35	0.67
FGM	0.01	0.00	0.00
AMH	0.05	0.05	0.18
Frank	0.56	0.26	0.63
Gumbel	0.34	0.34	0.35

c) data generated using the bivariate Gumbel copula

Copula	$\tau_1 = 0.2$	$\tau_2 = 0.5$	$\tau_3 = 0.8$
Normal	0.15	0.94	1.00
FGM	0.00	0.00	0.00
AMH	0.03	0.77	0.67
Frank	0.09	0.98	1.00
Gumbel	0.76	1.00	1.00

d) data generated using the bivariate Frank copula

Copula	$\tau_1 = 0.2$	$\tau_2 = 0.5$	$\tau_3 = 0.7$
Normal	0.28	0.95	1.00
FGM	0.00	0.00	0.00
AMH	0.14	0.65	0.78
Frank	0.37	0.98	1.00
Gumbel	0.51	0.96	1.00

Table 17: Powers for testing the null hypothesis $H_0 : \tau = 0$, when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Random samples were generated using the equation (10).

a) data generated using the bivariate normal copula

Copula	$\tau_1 = 0.1$	$\tau_2 = 0.5$	$\tau_3 = 0.7$
Normal	0.26	0.78	1
FGM	0.00	0.00	0
AMH	0.07	0.38	0
Frank	0.12	0.78	1
Gumbel	0.90	1.00	1

b) data generated using the bivariate Clayton copula

Copula	$\tau_1 = 0.1$	$\tau_2 = 0.5$	$\tau_3 = 0.7$
Normal	0.38	0.69	0.70
FGM	0.00	0.00	0.00
AMH	0.09	0.03	0.04
Frank	0.26	0.64	0.79
Gumbel	0.75	0.82	0.91

c) data generated using the bivariate Joe copula

Copula	$\tau_1 = 0.15$	$\tau_2 = 0.5$	$\tau_3 = 0.75$
Normal	0.14	0.96	1.00
FGM	0.00	0.00	0.00
AMH	0.02	0.81	0.86
Frank	0.10	1.00	1.00
Gumbel	0.98	1.00	1.00

Table 18: Powers for testing the null hypothesis $H_0 : \tau = 0$, when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Random samples were generated using the equation (11).

a) data generated using the bivariate FGM copula

Copula	$\tau_1 = -0.2$	$\tau_2 = 0.15$
Normal	0.50	0.47
FGM	0.00	0.00
AMH	0.16	0.14
Frank	0.38	0.34
Gumbel	0.88	0.84

b) data generated using the bivariate AMH copula

Copula	$\tau_1 = -0.12$	$\tau_2 = 0.1$	$\tau_3 = 0.28$
Normal	0.52	0.38	0.57
FGM	0.00	0.00	0.00
AMH	0.16	0.21	0.23
Frank	0.43	0.39	0.57
Gumbel	0.81	0.83	0.78

c) data generated using the bivariate Gumbel copula

Copula	$\tau_1 = 0.2$	$\tau_2 = 0.5$	$\tau_3 = 0.8$
Normal	0.26	0.78	1
FGM	0.00	0.00	0
AMH	0.07	0.38	0
Frank	0.12	0.78	1
Gumbel	0.90	1.00	1

d) data generated using the bivariate Frank copula

Copula	$\tau_1 = 0.2$	$\tau_2 = 0.5$	$\tau_3 = 0.7$
Normal	0.42	0.83	0.89
FGM	0.00	0.00	0.00
AMH	0.20	0.53	0.34
Frank	0.45	0.92	0.96
Gumbel	0.82	0.97	0.98

Table 19: Powers for testing the null hypothesis $H_0 : \tau = 0$, when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Random samples were generated using the equation (11).