



Copula Regression Spline Sample Selection Models: The R Package **SemiParSampleSel**

Małgorzata Wojtyś
Plymouth University,
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 with both decisions, the one to use health services and the one of the 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 might be especially appealing 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 relaxation of 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 when 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 Strazzeria \(2008\)](#) 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. Finally, while a fully parametric copula approach is less flexible than semi/non-parametric approaches, it still allows the user to assess the sensitivity of results to different modeling assumptions.

Some of the methods described above are implemented in popular software packages like SAS ([SAS Institute Inc. 2011](#)), Stata ([StataCorp. 2011](#)) and R ([R Core Team 2016](#)). For example, the conventional Heckman sample selection model can be fitted in SAS using `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 available 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 2015](#)). `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 ro-

bust two-stage estimator of Heckman’s approach. **sampleSelection** and **SemiParBIVProbit** support binary responses for the outcome equation, with the latter allowing for non-linear covariate effects and non-Gaussian bivariate distributions. It is also 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, Wojtyś, and Wyszynski 2016) 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-specified covariate effects are assumed then, similarly to **sampleSelection**, **SemiParSampleSel** fits the classical Heckman sample selection model using the maximum likelihood approach. 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 for 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 $i = 1, \dots, n$, 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 $y_{1i} \in \{0, 1\}$ and $y_{2i} = y_{2i}^* y_{1i}$. The variable y_{1i} governs whether or not an observation of 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

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

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 currently 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 $\tau = P((y_{11}^* - y_{12}^*)(y_{21}^* - y_{22}^*) > 0) - P((y_{11}^* - y_{12}^*)(y_{21}^* - y_{22}^*) < 0)$ for independent pairs (y_{1j}^*, y_{2j}^*) , $j = 1, 2$, that are copies of (y_1^*, y_2^*) . Testing the null hypothesis of absence of selection bias is an important issue. 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 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 $P(y_{1i} = 0)$, which is equivalent to $F_{1i}(0)$. In the second case, the joint likelihood can be expressed, using the multiplication rule, as $P(y_{1i}^* > 0)f_{2|1,i}(y_{2i}|y_{1i}^* > 0)$, 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 $\frac{1}{P(y_{1i}^* > 0)} \frac{\partial}{\partial y_2} (F_{2i}(y_2) - F_i(0, y_2))|_{y_2 \rightarrow y_{2i}}$, we obtain the log-likelihood

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

Using (1), we then have

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

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

Copula	$C(u, v; \theta)$	Parameter space
Normal	$\Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \theta)$	$\theta \in [-1, 1]$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\theta \in (0, \infty)$
Joe	$1 - \left[(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta \right]^{1/\theta}$	$\theta \in (1, \infty)$
Frank	$-\theta^{-1} \log \left[1 + (e^{-\theta u} - 1)(e^{-\theta v} - 1)/(e^{-\theta} - 1) \right]$	$\theta \in \mathbb{R} \setminus \{0\}$
Gumbel	$\exp \left\{ - \left[(-\log u)^\theta + (-\log v)^\theta \right]^{1/\theta} \right\}$	$\theta \in [1, \infty)$
FGM	$uv [1 + \theta(1-u)(1-v)]$	$\theta \in [-1, 1]$
AMH	$uv / [1 - \theta(1-u)(1-v)]$	$\theta \in [-1, 1]$

Table 1: Families of copulas implemented in **SemiParSampleSel**, with corresponding parameter range of the association parameter θ . $\Phi_2(\cdot, \cdot; \theta)$ denotes the cumulative distribution function of a standard bivariate normal distribution with correlation coefficient θ .

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

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

and that of the outcome equation as

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

where vector $\mathbf{u}_{1i}^\top = (1, u_{12i}, \dots, u_{1P_1i})$ is the i th row of $\mathbf{U}_1 = (\mathbf{u}_{11}, \dots, \mathbf{u}_{1n})^\top$, 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} . Similarly, $\mathbf{u}_{2i}^\top = (1, u_{22i}, \dots, u_{2P_2i})$ is the i th row vector of the $n_s \times P_2$ model matrix $\mathbf{U}_2 = (\mathbf{u}_{21}, \dots, \mathbf{u}_{2n_s})^\top$, 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 $v = 1, 2$, $k_v = 1, \dots, K_v$ (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., $s_k(z_{ki}) = \sum_{j=1}^{J_k} \beta_{kj} b_{kj}(z_{ki}) = \boldsymbol{\beta}_k^\top \mathbf{B}_k(z_{ki})$, 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., $\mathbf{B}_k(z_{ki}) = \{b_{k1}(z_{ki}), b_{k2}(z_{ki}), \dots, b_{kJ_k}(z_{ki})\}^\top$, 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). Our implementation also supports varying coefficients' models, obtained by multiplying one or more smooth terms by some predictor(s), smooth functions of two or more (e.g., spatial) covariates, random effect and Markov random field smooth functions, to name but a few (Wood 2006). These cases follow a similar construction as described above. 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 $\eta_{vi} = \mathbf{u}_{vi}^\top \boldsymbol{\alpha}_v + \mathbf{B}_{vi}^\top \boldsymbol{\beta}_v$, where $\mathbf{B}_{vi}^\top = \{\mathbf{B}_{v1}(z_{v1i})^\top, \dots, \mathbf{B}_{vK_v}(z_{vK_v i})^\top\}$ and $\boldsymbol{\beta}_v^\top = (\beta_{v1}^\top, \dots, \beta_{vK_v}^\top)$, 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 $\boldsymbol{\delta}^\top = (\boldsymbol{\delta}_1^\top, \boldsymbol{\delta}_2^\top, \sigma, \theta)$ and $\boldsymbol{\delta}_v^\top = (\boldsymbol{\alpha}_v^\top, \boldsymbol{\beta}_v^\top)$, 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 $\iint \left(\frac{\partial^2 s_{12}}{\partial z_1^2} \right)^2 + 2 \left(\frac{\partial^2 s_{12}}{\partial z_1 \partial z_2} \right)^2 + \left(\frac{\partial^2 s_{12}}{\partial z_2^2} \right)^2 dz_1 dz_2$, 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 $\boldsymbol{\beta}^\top \mathbf{S}_\lambda \boldsymbol{\beta}$ where $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)$, $\mathbf{S}_\lambda = \sum_{v=1}^2 \sum_{k_v=1}^{K_v} \lambda_{vk_v} \mathbf{S}_{vk_v}$ 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 $\boldsymbol{\delta}_*^\top = (\boldsymbol{\delta}_1^\top, \boldsymbol{\delta}_2^\top, \sigma^*, \theta^*) \in \mathbb{R}^p$, where p is the total number of parameters. Therefore, the function to maximize is

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

Given a parameter vector value for $\hat{\boldsymbol{\lambda}}^\top = (\hat{\lambda}_{1k_1}, \dots, \hat{\lambda}_{1K_1}, \hat{\lambda}_{2k_2}, \dots, \hat{\lambda}_{2K_2})$, 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).

Consider a minimization problem and 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_{a+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 shrunk 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 the trust region approach, the search for m_{a+1} is always a solution to the trust region problem; if the function at the proposed m_{a+1} is not finite or not better than the value at m_a , then the proposal is rejected and the trust region shrunk. 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

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

where $\|\cdot\|$ denotes the Euclidean norm and $r^{[a]}$ represents the radius of the trust region. \mathbf{S}_λ^* 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 $\boldsymbol{\mathcal{H}}_{r,h} = \partial^2 \ell(\boldsymbol{\delta}_*) / \partial \boldsymbol{\delta}_r \partial \boldsymbol{\delta}_h^\top$, $r, h = 1, \dots, 4$, 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, $\check{\ell}_p(\boldsymbol{\delta}_*^{[a]})$ 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 shrunk 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 like 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

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

where $\mathbf{z} = \mathbf{z}_i$ is the 4-dimensional vector $\mathbf{z}_i = \mathbf{X}_i \boldsymbol{\delta}_*^{[a]} + \mathbf{W}_i^{-1} \mathbf{d}_i$, $\mathbf{d}_i = \{\partial \ell(\boldsymbol{\delta}_*)_i / \partial \eta_{1i}, \partial \ell(\boldsymbol{\delta}_*)_i / \partial \eta_{2i}, \partial \ell(\boldsymbol{\delta}_*)_i / \partial \eta_{3i}, \partial \ell(\boldsymbol{\delta}_*)_i / \partial \eta_{4i}\}^\top$, $\eta_{3i} = \sigma^*$, $\eta_{4i} = \theta^*$, \mathbf{W}_i is a 4×4 matrix with (r, h) th element $(\mathbf{W}_i)_{rh} = -\partial^2 \ell(\boldsymbol{\delta}_*)_i / \partial \eta_{ri} \partial \eta_{hi}$, $r, h = 1, \dots, 4$, $\mathbf{X}_i = \text{diag} \left\{ (\mathbf{u}_{1i}^\top, \mathbf{B}_{1i}^\top), (\mathbf{u}_{2i}^\top, \mathbf{B}_{2i}^\top), 1, 1 \right\}$, $n_* = 4n$, $\mathbf{A}_\lambda = \mathbf{X}(\mathbf{X}^\top \mathbf{W} \mathbf{X} + \mathbf{S}_\lambda^*)^{-1} \mathbf{X}^\top \mathbf{W}$ is the hat matrix, and $\text{tr}(\mathbf{A}_\lambda)$ 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.

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

$$\boldsymbol{\delta}_* | \mathbf{y} \rightsquigarrow \mathcal{N}(\hat{\boldsymbol{\delta}}_*, \mathbf{V}_{\boldsymbol{\delta}_*}), \quad (7)$$

where \mathbf{y} refers to the response vectors, $\hat{\boldsymbol{\delta}}_*$ is an estimate of $\boldsymbol{\delta}_*$ and $\mathbf{V}_{\boldsymbol{\delta}_*} = (-\mathcal{H} + \mathbf{S}_\lambda^*)^{-1}$. 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

$$\hat{s}_k(z_{ki}) \dot{\sim} \mathcal{N}(s_k(z_{ki}), \mathbf{B}_k(z_{ki})^\top \mathbf{V}_{\delta_{*k}} \mathbf{B}_k(z_{ki})), \quad (8)$$

where $\mathbf{V}_{\delta_{*k}}$ is the submatrix of \mathbf{V}_{δ_*} 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) can be conveniently obtained by simulation from the posterior distribution of δ_* . 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 $\hat{\mathbf{s}}_k = \mathbf{B}_k(\mathbf{z}_k) \hat{\beta}_k$, where $\mathbf{B}_k(\mathbf{z}_k)$ denotes a full column rank matrix and $\mathbf{z}_k = (z_{k1}, z_{k2}, \dots, z_{kn})^\top$, and $\mathbf{V}_{\mathbf{s}_k} = \mathbf{B}_k(\mathbf{z}_k) \mathbf{V}_{\delta_{*k}} \mathbf{B}_k(\mathbf{z}_k)^\top$. It is then possible to obtain approximate p values for testing smooth components for equality to zero based on

$$T_{r_k} = \hat{\mathbf{s}}_k^\top \mathbf{V}_{\mathbf{s}_k}^{r_k-} \hat{\mathbf{s}}_k \dot{\sim} \chi_{r_k}^2,$$

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](#))

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

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 $\mathbf{U}_k \Lambda_k \mathbf{U}_k^\top$, 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, $AIC = -2\ell(\hat{\delta}_*) + 2\text{edf}$ and $BIC = -2\ell(\hat{\delta}_*) + \log(n)\text{edf}$, where the log-likelihood is evaluated at the penalized parameter estimates and $\text{edf} = \text{tr}(\hat{\mathbf{A}}_{\hat{\lambda}})$.

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 \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 $\theta \rightarrow \infty$ 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 the log-likelihood (2) may be approximately equal to 1, hence producing $-\text{Inf}$. 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 \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 <https://CRAN.R-project.org/package=SemiParSampleSel>. The package depends on **copula** (Yan 2007), **mgcv** (Wood 2006), **mvtnorm** (Genz and Bretz 2009) and imports functions from packages **magic** (Hankin 2005), **trust** (Geyer 2013), **VGAM** (Yee 2014) and **Matrix** (Bates and Maechler 2014). 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(list(formula.eq1, formula.eq2), data = list(),
  BivD = "N", margins = c("N", "N"), infl.fac = 1, ...)
```

The first argument is a list of `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 **infl.fac** 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; **infl.fac** = 1.4 typically achieves this and was 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", "CO", "JO", "FGM", "F", "AMH" and "GO" 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. \(2016\)](#).

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 plotting, **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** method for 'gam' objects in **mgcv** to which the reader is referred for full details.
- **aver(x, sig.lev = 0.05, ...)**. This function takes an object fitted with the function **SemiParSampleSel()** and calculates the overall estimated average correct for non-random sample selection, with corresponding confidence interval obtained using the delta method.
- **summary(object, n.sim = 1000, s.meth = "svd", prob.lev = 0.05, ...)**. This function produces some summaries of an object fitted with **SemiParSampleSel()**. **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 and θ , 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.
- **conv.check(x)** which produces some diagnostic information about the fitting procedure for an object fitted with **SemiParSampleSel()**.

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. The simulation study was performed using version 1.1 of the package **SemiParSampleSel**. 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

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

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 $s_{11}(z_{1i}) = -0.7 \{4z_{1i} + 2.5z_{1i}^2 + 0.7 \sin(5z_{1i}) + \cos(7.5z_{1i})\}$, $s_{12}(z_{2i}) = -0.4 \{-0.3 - 1.6z_{2i} + \sin(5z_{2i})\}$, and $s_{21}(z_{1i}) = 0.6 \{\exp(z_{1i}) + \sin(2.9z_{1i})\}$. 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.

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 $\sqrt{\frac{1}{200} \sum_{b=1}^{200} \{\hat{s}(z_{1b}) - s(z_{1b})\}^2}$,

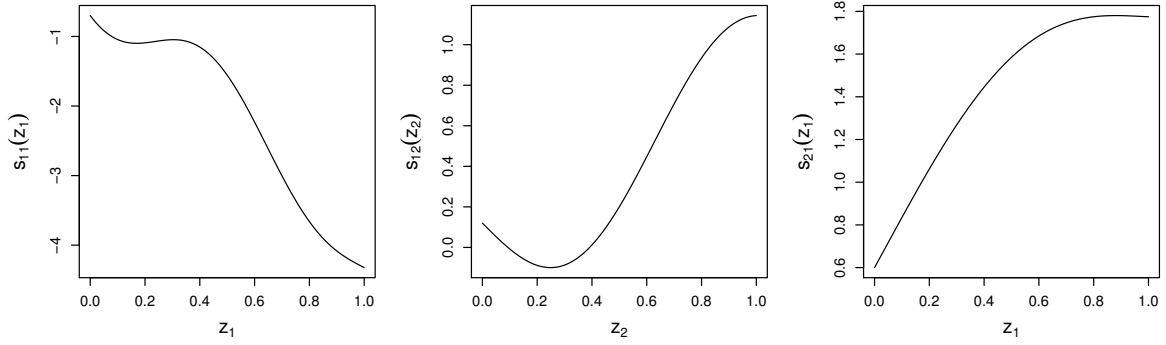


Figure 1: The test functions used in the simulation studies.

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 $\tau = 0.1$, and Tables 7 and 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 $\tau \rightarrow 0$. 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. 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 15 and 16 in Appendix B, where the experiment was repeated for samples of size $n = 3000$ and $n = 5000$ pertaining to bivariate normal distribution. There we can also

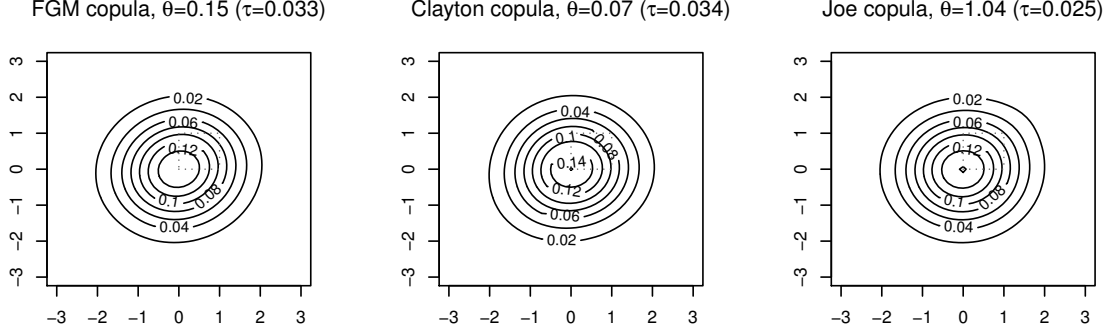


Figure 2: Contour plots of FGM, Clayton and Joe copulas with normal margins for small values of the dependence parameter.

observe consistency of the estimators when the right copula is chosen. In the case of a wrong copula the estimators are inconsistent.

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 given at the beginning of 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 5b) and Figure 6a).

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 associated then 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 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 $\theta = 0$ for normal, Frank, FGM and AMH copulas and $\theta = 1$ for Gumbel copula (note that Clayton and Joe copulas do not allow independence). 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 (9), the null rejection probabilities of the test for absence of selectivity bias have 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 test would always reject the null hypothesis in these cases. For the case of a lack of the sample selection bias ($\tau = 0$), the results of the Monte Carlo simulations based on 250 repetitions and sample sizes $n = 1000, 3000, 5000$ are presented in Table 11(a).

For $n = 1000$ the test using Frank, normal and Gumbel copulas suffers from high rejection frequency, whereas that using FGM has low rejection frequency. The rejection probabilities for the test using AMH copula are close to the nominal values. As n increases the null rejection probabilities converge to the nominal values; AMH and Frank copulas perform best achieving probabilities of the test that are very close to the theoretical value for $n = 5000$. The poorest performance occurs in the case of the Gumbel copula for which the rejection probabilities converge to the theoretical values at a much slower rate, as the sample size increases. Note, however, that the null hypothesis $\tau = 0$ involves the boundary value of τ allowed under the Gumbel copula which implies that the testing is a difficult issue in this case.

Tables 11(b) and (c) present the null rejection probabilities for the sample selection bias test for data without exclusion restriction, generated using (10) and (11), respectively. In both cases, a negative influence of lack of the exclusion restriction can be observed as the values of probabilities are larger than those presented in Table 11(a) where the exclusion restriction is used. Moreover, the effect of lack of exclusion restriction is more severe for data generated using (11) where the variable z_1 enters both, the selection and the outcome equations, in the same functional form. However, in Tables 11(b) and (c) the same tendency regarding the comparison between different copulas can be observed with FGM, AMH and Frank having rejection probabilities close to the nominal values and Gumbel copula displaying the worst performance.

A study of power of the test for sample selection bias has also been conducted for the copulas where the null rejection probabilities were reasonable (FGM, AMH and Frank). Results are reported in Table 12. Using Frank copula leads to the most powerful tests. A poor performance can be observed when using the FGM and AMH copulas with FGM copula performing the worst. Those are the copulas allowing very limited scope for τ ($\tau \in [-2/9, 2/9]$ for FGM and $\tau \in [-0.1817, 1/3]$ for AMH) which makes them perform poorly when a strong dependence holds.

Tables 13 and 14 present the power of the test for sample selection bias in the absence of an exclusion restriction. Overall, the power of the test is smaller than when the exclusion restriction is present with the best performance observed, as before, when using the Frank copula and the worst when using the FGM copula.

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.

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

Variable	Definition
lnmeddol	Log of the medical expenses of the individual (<i>outcome variable</i>).
binexp	Binary variable indicating whether the medical expenses are positive (<i>selection variable</i>).
logc	Log of the coinsurance rate (<i>coins</i>) plus 1.
idp	Binary variable for individual deductible plans.
pi	Participation incentive payment.
fmde	Is 0 if <i>idp</i> = 1, and $\log[\max\{1, \text{maximum expenditure offer}/(0.01 \cdot \text{coins})\}]$ 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.

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 `infl.fac = 1.4` to obtain smoother models.

```

R> out_N <- SemiParSampleSel(list(SE, OE), data = ND, infl.fac = 1.4)
R> out_C <- SemiParSampleSel(list(SE, OE), data = ND, BivD = "CO",
+   infl.fac = 1.4)
R> out_J <- SemiParSampleSel(list(SE, OE), data = ND, BivD = "JO",
+   infl.fac = 1.4)
R> out_FGM <- SemiParSampleSel(list(SE, OE), data = ND, BivD = "FGM",
+   infl.fac = 1.4)
R> out_F <- SemiParSampleSel(list(SE, OE), data = ND, BivD = "F",
+   infl.fac = 1.4)
R> out_AMH <- SemiParSampleSel(list(SE, OE), data = ND, BivD = "AMH",
+   infl.fac = 1.4)
R> out_G <- SemiParSampleSel(list(SE, OE), data = ND, BivD = "GO",
+   infl.fac = 1.4)

```

Given the superior performance of *AIC* on *BIC* shown in the simulation study, we use *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] 20294.87
```

```
R> AIC_C
```

```
[1] 20293.91
```

```
R> AIC_J
```

```
[1] 20336.22
```

```
R> AIC_FGM
```

```
[1] 20289.21
```

```
R> AIC_F
```

```
[1] 20280.62
```

```
R> AIC_AMH
```

```
[1] 20280.12
```

```
R> AIC_G
```

```
[1] 20293.86
```

We choose the AMH copula model as the Frank and the AMH copulas appear to be the two preferable models. Before looking at the results, we check that the algorithm has found a solution.

```
R> conv.check(out_AMH)
```

```
Largest absolute gradient value: 2.775396e-10
Observed information matrix is positive definite
Eigenvalue range: [0.5487841,1035141]
```

```
Trust region iterations before smoothing parameter estimation: 6
Loops for smoothing parameter estimation: 7
Trust region iterations within smoothing loops: 13
```

We can now look at the results.

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

ERRORS' DISTRIBUTION: AMH Copula

SELECTION EQ.

Family: Bernoulli

Link function: probit

Formula: 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)	0.647749	0.083648	7.744	9.65e-15	***
logc	-0.091062	0.028586	-3.186	0.001445	**
idp	-0.155417	0.055458	-2.802	0.005072	**
fmde	-0.005796	0.017902	-0.324	0.746115	
physlm	0.266886	0.074524	3.581	0.000342	***
disea	0.020976	0.003731	5.622	1.88e-08	***
hlthg	0.093914	0.044434	2.114	0.034554	*
hlthf	0.233172	0.084317	2.765	0.005685	**
hlthp	0.767821	0.217573	3.529	0.000417	***
female	0.432467	0.054523	7.932	2.16e-15	***
child	0.263177	0.149627	1.759	0.078597	.
fchild	-0.422033	0.080338	-5.253	1.49e-07	***
black	-0.593539	0.054514	-10.888	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Smooth components' approximate significance:

	edf	Ref.df	Chi.sq	p-value	
s(pi)	7.038	8.035	36.353	1.58e-05	***
s(inc)	2.330	2.991	28.756	2.54e-06	***
s(fam)	4.699	5.662	8.055	0.20393	
s(educdec)	1.821	2.315	16.513	0.00048	***
s(xage)	7.258	8.285	51.701	2.99e-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

OUTCOME EQ.

Family: Gaussian

Link function: identity

Formula: 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)	3.394538	0.092205	36.815	< 2e-16	***
logc	-0.049046	0.032498	-1.509	0.131253	
idp	-0.142146	0.062644	-2.269	0.023263	*
fmde	-0.025045	0.019453	-1.287	0.197939	
physlm	0.303793	0.071523	4.247	2.16e-05	***
disea	0.025533	0.003653	6.990	2.75e-12	***
hlthg	0.186160	0.049426	3.766	0.000166	***
hlthf	0.429723	0.090264	4.761	1.93e-06	***
hlthp	0.873586	0.178302	4.899	9.61e-07	***
female	0.467060	0.058727	7.953	1.82e-15	***
child	0.153816	0.177986	0.864	0.387477	
fchild	-0.481647	0.090792	-5.305	1.13e-07	***
black	-0.379393	0.067857	-5.591	2.26e-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Smooth components' approximate significance:

	edf	Ref.df	Chi.sq	p-value	
s(pi)	1.524	1.859	1.175	0.52019	
s(inc)	2.838	3.629	24.932	4.41e-05	***
s(fam)	1.073	1.142	9.844	0.00246	**
s(educdec)	1.041	1.081	0.647	0.44540	
s(xage)	7.727	8.586	61.502	5.95e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

n = 5574 n.sel = 4281 sigma = 1.511(1.472,1.553)

theta = 0.96(0.904,0.982) total edf = 65.349

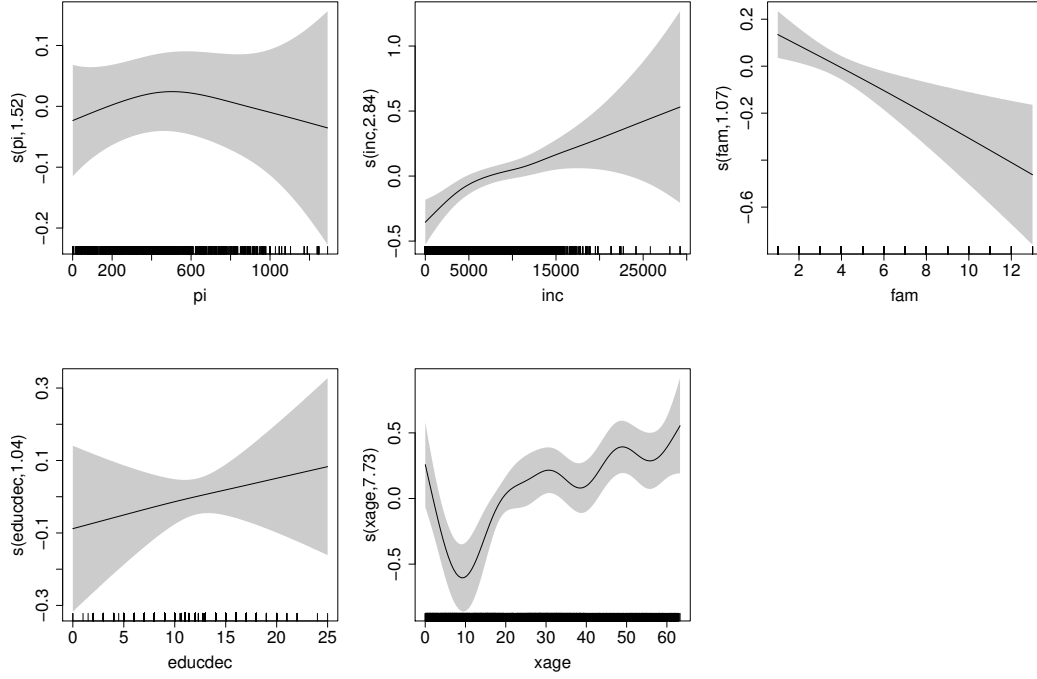


Figure 3: Smooth function estimates and 95% confidence bands obtained applying the AMH copula regression spline sample selection model on the RAND RHIE dataset described in Section 4.

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 variable `logc` seems 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 1.51 and is significantly different from zero. The estimate for θ is positive and statistically 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 65.349.

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

```
R> plot(out_AMH, 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

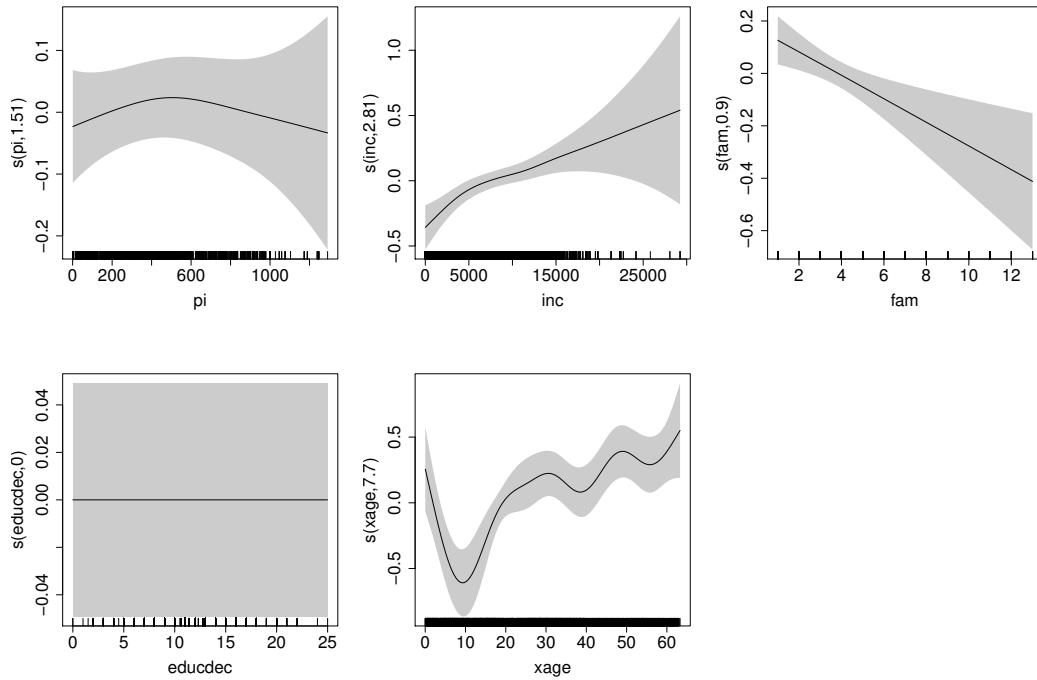


Figure 4: Smooth function estimates and 95% confidence bands obtained applying the AMH copula model with shrinkage option on the RAND RHIE dataset.

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 AMH 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_AMH_s <- SemiParSampleSel(list(SE_s, OE_s), data = ND, BivD = "AMH",
+   infl.fac = 1.4)
R> plot(out_AMH_s, eq = 2, pages = 1, scale = 0, shade = TRUE,
+   seWithMean = TRUE, cex.axis = 1.6, cex.lab = 1.6)
```

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

Finally, we use `est.aver()` to calculate the overall estimated average from the fitted copula sample selection, with corresponding confidence interval obtained using the delta method.

```
R> aver(out_AMH, sig.lev = 0.05)
```

Estimated average with 95% confidence interval:

```
3.68 (3.50,3.86)
```

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. 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.

In order to address the issue of testing for the absence of sample selection bias, an approach based on a confidence interval for the Kendall's τ association coefficient has been explored. The empirical study has indicated that this approach performs well when using the Frank copula. However, for other copulas a significantly poorer small sample performance of the test has been observed. As every copula function has different characteristics and poses different issues while testing, each of them requires a separate study. It is also unclear whether a unique test that performs equally well for a wide range of copulas can be designed. Thus a further detailed study of the problem of testing for the sample selection bias in such a general context will be another direction of future research.

The reader is cautioned that the class of models presented here is not intended to be exhaustive; as with the majority of 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 should 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 and Poisson 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. Finally, a next release of the package will allow the user to model σ and θ as functions of linear predictors like those defined in Section 2.1; the theoretical and computational framework remains essentially unchanged.

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 J (1993). “Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism.” *Econometrics*, **58**, 3–29. doi:[10.1016/0304-4076\(93\)90111-h](https://doi.org/10.1016/0304-4076(93)90111-h).
- Andrews D, Schafgans M (1998). “Semiparametric Estimation of the Intercept of a Sample Selection Model.” *Review of Economic Studies*, **65**, 497–517. doi:[10.1111/1467-937x.00055](https://doi.org/10.1111/1467-937x.00055).
- Bates D, Maechler M (2014). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.1-2, URL <https://CRAN.R-project.org/package=Matrix>.
- Bhat C, 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. doi:[10.1016/j.trb.2009.02.001](https://doi.org/10.1016/j.trb.2009.02.001).
- Brechmann E, Schepsmeier U (2013). “Modeling Dependence with C- And D-Vine Copulas: The R Package **CDVine**.” *Journal of Statistical Software*, **52**(3), 1–27. doi:[10.18637/jss.v052.i03](https://doi.org/10.18637/jss.v052.i03).
- Cameron AC, Trivedi PK (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. doi:[10.1016/j.jeconom.2009.10.022](https://doi.org/10.1016/j.jeconom.2009.10.022).
- Chib S, Greenberg E, Jeliazkov I (2009). “Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection.” *Journal of Computational and Graphical Statistics*, **18**, 321–348. doi:[10.1198/jcgs.2009.07070](https://doi.org/10.1198/jcgs.2009.07070).
- Das M, Newey W, Vella F (2003). “Nonparametric Estimation of Sample Selection Models.” *Review of Economic Studies*, **70**, 33–58. doi:[10.1111/1467-937x.00236](https://doi.org/10.1111/1467-937x.00236).
- Ding P (2014). “Bayesian Robust Inference of Sample Selection Using Selection-Models.” *Journal of Multivariate Analysis*, **124**, 451–464. doi:[10.1016/j.jmva.2013.11.014](https://doi.org/10.1016/j.jmva.2013.11.014).

- Dubin J, Rivers D (1989). “Selection Bias in Linear Regression, Logit and Probit Models.” *Sociological Methods and Research*, **18**, 360–390. doi:10.1177/0049124189018002006.
- Gallant R, Nychka D (1987). “Semi-Nonparametric Maximum Likelihood Estimation.” *Econometrica*, **55**, 363–390. doi:10.2307/1913241.
- Genius M, Strazzer E (2008). “Applying the Copula Approach to Sample Selection Modelling.” *Applied Economics*, **40**, 1443–1455. doi:10.1080/00036840600794348.
- 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 C (2013). *trust: Trust Region Optimization*. R package version 0.1-4, URL <https://CRAN.R-project.org/package=trust>.
- Gronau R (1974). “Wage Comparisons: A Selectivity Bias.” *Journal of Political Economy*, **82**, 1119–1143. doi:10.1086/260267.
- Hankin R (2005). “Recreational Mathematics with R: Introducing the **magic** Package.” *R News*, **5**(1), 48–51. URL <https://www.R-project.org/doc/Rnews/>.
- Hasebe T, Vijverberg W (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>.
- 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 (1990). “Varieties of Selection Bias.” *The American Economic Review*, **80**, 313–318.
- Heckman JJ (1979). “Sample Selection Bias as a Specification Error.” *Econometrica*, **47**, 153–162. doi:10.2307/1912352.
- Henningsen A (2012). *censReg: Censored Regression (Tobit) Models*. R package version 0.5-16, URL <https://CRAN.R-project.org/package=censReg>.
- Joe H (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.
- Kim Y, Gu C (2004). “Smoothing Spline Gaussian Regression: More Scalable Computation via Efficient Approximation.” *Journal of the Royal Statistical Society B*, **66**, 337–356. doi:10.1046/j.1369-7412.2003.05316.x.
- Lee D (2008). “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.” *Review of Economic Studies*, **76**(11721), 1071–1102. doi:10.1111/j.1467-937x.2009.00536.x.
- Lee LF (1982). “Some Approaches to the Correction of Selectivity Bias.” *Review of Economic Studies*, **49**, 355–372. doi:10.2307/2297361.
- Lee LF (1984). “Tests for the Bivariate Normal Distribution in Econometric Models with Selectivity.” *Econometrica*, **52**, 843–863. doi:10.2307/1911187.

- Lee LF (1994a). “Semiparametric Instrumental Variable Estimation of Simultaneous Equation Sample Selection Models.” *Journal of Econometrics*, **63**, 341–388. doi:10.1016/0304-4076(93)01571-3.
- Lee LF (1994b). “Semiparametric Two-Stage Estimation of Sample Selection Models Subject to Tobit-Type Selection Rules.” *Journal of Econometrics*, **61**, 305–344. doi:10.1016/0304-4076(94)90088-4.
- Lewis H (1974). “Comments on Selectivity Biases in Wage Comparisons.” *Journal of Political Economy*, **82**, 1145–1155. doi:10.1086/260268.
- Marchenko Y, Genton M (2012). “A Heckman Selection- t Model.” *Journal of the American Statistical Association*, **107**, 304–317. doi:10.1080/01621459.2012.656011.
- Marra G (2013). “On p -Values for Semiparametric Bivariate Probit Models.” *Statistical Methodology*, **10**, 23–28. doi:10.1016/j.stamet.2012.05.003.
- Marra G, Radice R (2010). “Penalised Regression Splines: Theory and Application to Medical Research.” *Statistical Methods in Medical Research*, **19**, 107–125. doi:10.1177/0962280208096688.
- Marra G, Radice R (2013). “Estimation of a Regression Spline Sample Selection Model.” *Computational Statistics & Data Analysis*, **61**, 158–173. doi:10.1016/j.csda.2012.12.010.
- Marra G, Radice R (2015). **SemiParBIVProbit**: *Semiparametric Bivariate Probit Modelling*. R package version 3.3, URL <https://CRAN.R-project.org/package=SemiParBIVProbit>.
- Marra G, Radice R, Wojtyś M, Wyszynski K (2016). *Semiparametric Sample Selection Modelling with Continuous Response*. R package version 1.4, URL <https://CRAN.R-project.org/package=SemiParSampleSel>.
- Marra G, Wood SN (2011). “Practical Variable Selection for Generalized Additive Models.” *Computational Statistics & Data Analysis*, **55**, 2372–2387. doi:10.1016/j.csda.2011.02.004.
- Marra G, Wood SN (2012). “Coverage Properties of Confidence Intervals for Generalized Additive Model Components.” *Scandinavian Journal of Statistics*, **39**, 53–74. doi:10.1111/j.1467-9469.2011.00760.x.
- Montes-Rojas G (2011). “Robust Misspecification Tests for the Heckman’s Two-Step Estimator.” *Econometric Reviews*, **30**, 154–172. doi:10.1080/07474938.2011.534035.
- Nelsen RB (2006). *An Introduction to Copulas*. 2nd edition. Springer-Verlag.
- Newey W (2009). “Two-Step Series Estimation of Sample Selection Models.” *Econometrica*, **12**, S217–S229. doi:10.1111/j.1368-423x.2008.00263.x.
- Newhouse J (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 SJ (2006). *Numerical Optimization*. Springer-Verlag. doi:10.1007/b98874.
- Powell J (1994). “Estimation of Semiparametric Models.” In J Heckman, E Leamer (eds.), *Handbook of Econometrics*, pp. 5307–5368. Elsevier, Amsterdam.
- Powell J, Stock J, Stoker T (1989). “Semiparametric Estimation of Index Coefficients.” *Econometrica*, **57**, 1403–1430. doi:10.2307/1913713.
- Prieger J (2002). “A Flexible Parametric Selection Model for Non-Normal Data with Application to Health Care Usage.” *Journal of Applied Econometrics*, **17**, 367–392. doi:10.1002/jae.638.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Ruppert D, Wand MP, Carroll RJ (2003). *Semiparametric Regression*. Cambridge University Press, New York.
- SAS Institute Inc (2011). *SAS/STAT Software, Version 9.3*. SAS Institute Inc., Cary, NC.
- 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 M (2003). “Modelling Sample Selection Using Archimedean Copulas.” *Econometrics Journal*, **6**, 99–123. doi:10.1111/1368-423x.00101.
- StataCorp (2011). *Stata Data Analysis Statistical Software: Release 12*. StataCorp LP, College Station, TX. URL <http://www.stata.com/>.
- Toomet O (2012). *intReg: Interval Regression*. R package version 0.1-2, URL <https://CRAN.R-project.org/package=intReg>.
- Toomet O, Henningsen A (2008). “Sample Selection Models in R: Package **sampleSelection**.” *Journal of Statistical Software*, **27**(7), 1–23. doi:10.18637/jss.v027.i07.
- Vella F (1992). “Simple Tests for Sample Selection Bias in Censored and Discrete Choice Models.” *Journal of Applied Econometrics*, **7**, 413–421. doi:10.1002/jae.3950070407.
- Wiesenfarth M, Kneib T (2010). “Bayesian Geoadditive Sample Selection Models.” *Journal of the Royal Statistical Society C*, **59**, 381–404. doi:10.1111/j.1467-9876.2009.00698.x.
- Wood S (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.
- Wood SN (2013). “On p -Values for Smooth Components of an Extended Generalized Additive Model.” *Biometrika*, **100**, 221–228. doi:10.1093/biomet/ass048.
- Yan J (2007). “Enjoy the Joy of Copulas: With a Package **copula**.” *Journal of Statistical Software*, **21**(4), 1–21. doi:10.18637/jss.v021.i04.
- Yee T (2014). *VGAM: Vector Generalized Linear and Additive Models*. R package version 0.9-6, URL <https://CRAN.R-project.org/package=VGAM>.

- Zhelonkin M, Genton M, Ronchetti E (2013). **ssmrob**: *Robust Estimation and Inference in Sample Selection Models*. R package version 0.3, URL <https://CRAN.R-project.org/package=ssmrob>.
- Zimmer D, Trivedi P (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. doi:[10.1198/073500105000000153](https://doi.org/10.1198/073500105000000153).

A. Analytical expressions for \mathbf{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 $\eta_{vi} = \mathbf{X}_{vi}\boldsymbol{\delta}_v$, $\mathbf{X}_{vi} = (\mathbf{u}_{vi}^\top, \mathbf{B}_{vi}^\top)$, 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

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\top} &= \sum_{i=1}^n \left\{ (y_{1i} - 1)F_1^{-1}(F_1^{-1}\phi(-\eta_{1i}) - \eta_{1i}) + y_{1i}P_i \right\} \phi(-\eta_{1i})\mathbf{X}_{1i}^\top \mathbf{X}_{1i}, \\ \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\top} &= \sum_{i=1}^n y_{1i} A_i \phi(-\eta_{1i})\mathbf{X}_{1i}^\top \mathbf{X}_{2i}, \\ \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \sigma^*} &= \sum_{i=1}^n y_{1i} \sigma A_i \phi(-\eta_{1i})\tilde{e}_{2i}\mathbf{X}_{1i}, \\ \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \theta^*} &= \sum_{i=1}^n y_{1i} h_{14} \phi(-\eta_{1i})\mathbf{X}_{1i}, \\ \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\top} &= \sum_{i=1}^n y_{1i} (h_i E_i - \sigma^{-2})\mathbf{X}_{2i}^\top \mathbf{X}_{2i}, \\ \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2 \partial \sigma^*} &= \sum_{i=1}^n y_{1i} \sigma \left[h_i (E_i \tilde{e}_{2i} - \sigma^{-1}) - 2\sigma^{-2}\tilde{e}_{2i} \right] \mathbf{X}_{2i}, \\ \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2 \partial \theta^*} &= \sum_{i=1}^n y_{1i} B_i \mathbf{X}_{2i}, \\ \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \sigma^{*2}} &= \sum_{i=1}^n y_{1i} \sigma^2 \tilde{e}_{2i} \left[h_i (E_i \tilde{e}_{2i} - \sigma^{-1}) - 2\sigma^{-2}\tilde{e}_{2i} \right], \\ \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \sigma^* \partial \theta^*} &= \sum_{i=1}^n y_{1i} \sigma B_i \tilde{e}_{2i}, \\ \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \theta^{*2}} &= \sum_{i=1}^n y_{1i} h_{44},\end{aligned}$$

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,

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

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

For FGM,

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

For Frank,

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

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

For AMH,

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

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}, \\
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 (\log(-\log F_1))^2 + (-\log F_2)^\theta (\log(-\log F_2))^2, \\
b_i &= \frac{z}{z-1} (\theta-1) \left[\frac{1}{\theta} \frac{C_i}{u} (1 - \theta - u^{1/\theta}) + \log(-\log F_2) - \frac{C_i}{u} \right], \\
h_i &= \frac{z}{z-1} f_2 F_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_2 F_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. \\
&\quad \left. + \frac{\log u}{\theta^2} + \frac{z}{z-1} \frac{b_i}{\theta-1} \right] - \frac{h_i b_i}{z_i} - (\theta-1) \frac{z}{z-1} f_2 F_2^{-1} (\log F_2)^{-1}, \\
E_i &= \frac{\tilde{e}_{2i}}{\sigma} - \frac{h_i}{z_i} + f_2 F_2^{-1} \left\{ 1 + (\theta-1)(\log F_2)^{-2} \frac{z}{z-1} f_2 F_2^{-1} h_i^{-1} \right. \\
&\quad \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_2 F_2^{-1} p_i (\theta + u^{1/\theta}) u^{-1} (-\log F_2)^{\theta-1} (1 - \theta - u^{1/\theta})^{-1} - \frac{h_i p_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] \\
&\quad - 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. \\
&\quad \left. - \frac{\tilde{C}_i}{u} + \frac{C_i^2}{u^2} \right] + b_i - \frac{b_i^2}{z_i}.
\end{aligned}$$

B. Tables and figures

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

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{s}_{21}(z_1)$		
	Bias (%)	RMSE		Bias (%)	RMSE		Bias (%)	RMSE		Bias (%)	RMSE		RMSE	AIC (%)	BIC (%)
Joe	Normal	4.7	0.355	3.2	0.313	0.3	0.038	0.185	-43.6	0.090	0.122	13.3	16.1		
	Clayton	4.2	0.171	3.5	0.170	0.7	0.043	0.090	-34.5	0.100	0.100	52.2	50.2		
	Joe	1.8	0.209	1.5	0.202	-2.0	0.040	0.099	-40.0	0.101	0.101	2.8	7.2		
	FGM	12.4	0.274	6.4	0.249	-0.4	0.034	0.147	-74.4	0.113	0.113	8.8	8.4		
	AMH	9.3	0.272	5.4	0.248	0.5	0.042	0.146	-56.0	0.113	0.113	6.4	5.2		
Frank	Normal	9.9	0.317	5.3	0.280	-0.1	0.037	0.169	-63.3	0.117	0.117	8.4	4.4		
	Clayton	-3.7	0.242	-0.6	0.223	-1.3	0.037	0.114	-13.8	0.103	0.103	8.0	8.4		
	Joe	-4.0	0.167	-0.4	0.164	-0.7	0.048	0.083	0.3	0.084	0.084	60.4	54.0		
	FGM	33.0	0.319	17.4	0.317	0.7	0.053	0.171	-26.3	0.122	0.122	10.0	7.2		
	AMH	11.7	0.368	4.2	0.327	-5.9	0.083	0.224	-21.0	0.120	0.120	1.2	1.6		
Gumbel	Normal	58.4	0.454	25.6	0.400	-8.1	0.087	0.284	-56.4	0.161	0.161	0.4	2.0		
	Clayton	46.1	0.367	22.2	0.353	-2.6	0.048	0.199	-38.8	0.139	0.139	1.2	10.0		
	Joe	5.5	0.213	3.6	0.203	-1.5	0.053	0.120	-7.9	0.089	0.089	10.0	10.4		
	FGM	-7.0	0.228	-2.6	0.213	-2.3	0.059	0.121	0.7	0.091	0.091	16.8	14.8		
	AMH	-5.8	0.119	-1.1	0.126	-0.3	0.043	0.050	2.0	0.075	0.075	64.1	64.1		
Joe	Normal	10.7	0.169	8.2	0.196	2.3	0.052	0.082	-4.1	0.085	0.085	11.0	9.4		
	Clayton	-12.2	0.153	-5.3	0.153	-1.8	0.052	0.064	0.1	0.082	0.082	1.2	1.6		
	Joe	89.9	0.689	38.8	0.591	-14.1	0.144	0.480	-68.5	0.221	0.221	0.0	0.4		
	FGM	77.6	0.594	35.7	0.544	-8.6	0.092	0.369	-52.6	0.198	0.198	0.0	1.2		
	AMH	-2.7	0.120	0.5	0.128	-0.3	0.046	0.056	-0.6	0.074	0.074	6.5	7.8		
Frank	Normal	-10.2	0.134	-3.5	0.134	-0.5	0.045	0.057	4.0	0.077	0.077	17.1	15.5		
	Clayton														
	Joe														
	FGM														
	AMH														

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.

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{\sigma}_{21}(z_1)$			AIC (%)	BIC (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	RMSE			
$\tau = 0.1$	Normal	15.1	0.406	7.1	0.343	-1.5	0.040	-96.0	0.231	0.133	8.8	10.4					
	Clayton	-6.3	0.187	-1.0	0.161	-0.3	0.043	4.6	0.095	0.092	45.2	39.6					
	Joe	5.2	0.203	2.7	0.192	-3.6	0.051	-60.2	0.111	0.099	2.8	5.6					
	FGM	22.0	0.338	10.0	0.293	-2.1	0.040	-125.8	0.199	0.124	12.4	9.6					
	AMH	14.8	0.325	7.4	0.278	-0.9	0.044	-87.9	0.187	0.120	9.6	6.0					
	Frank	23.3	0.384	10.6	0.330	-1.8	0.041	-131.5	0.226	0.132	10.4	14.0					
$\tau = 0.5$	Gumbel	1.3	0.227	1.2	0.206	-3.2	0.047	-41.4	0.121	0.101	10.8	14.8					
	Normal	-14.5	0.359	-6.6	0.307	-4.0	0.064	1.8	0.202	0.105	6.1	5.7					
	Clayton	-3.8	0.161	0.0	0.146	-0.9	0.045	0.4	0.082	0.081	91.5	70.3					
	Joe	10.0	0.539	2.3	0.464	-7.8	0.110	-28.8	0.326	0.166	0.0	0.4					
	FGM	56.0	0.491	22.8	0.406	-12.1	0.125	-66.5	0.360	0.159	0.0	1.6					
	AMH	38.7	0.363	17.4	0.320	-6.8	0.076	-41.0	0.243	0.128	1.6	17.5					
$\tau = 0.7$	Frank	7.7	0.531	2.7	0.449	-4.8	0.075	-21.4	0.331	0.139	0.8	3.7					
	Gumbel	-17.5	0.399	-8.5	0.349	-4.7	0.081	2.4	0.216	0.124	0.0	0.8					
	Normal	-15.8	0.213	-6.7	0.193	-2.5	0.049	3.4	0.096	0.082	4.7	8.1					
	Clayton	-4.6	0.129	-0.3	0.127	-0.9	0.038	1.7	0.058	0.073	92.4	84.7					
	Joe	-25.0	0.282	-11.9	0.263	-2.5	0.059	2.2	0.120	0.103	0.0	0.0					
	FGM	79.7	0.614	33.3	0.511	-16.2	0.164	-68.8	0.483	0.201	0.0	0.0					
$\tau = 0.7$	AMH	69.6	0.537	31.0	0.476	-11.5	0.119	-52.7	0.371	0.185	0.0	1.7					
	Frank	-15.7	0.248	-6.4	0.218	-2.0	0.047	1.0	0.123	0.084	1.7	3.8					
	Gumbel	-22.5	0.227	-9.9	0.208	-1.7	0.050	7.1	0.091	0.088	1.3	1.7					

Table 5: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\sigma}_{21}(z_1)$, 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	Bias (%)	RMSE	Bias (%)	RMSE	AIC (%)	BIC (%)	BIC (%)
$\tau = 0.1$	Normal	6.8	0.226	5.9	0.22	3.1	0.047	0.106	-16.3	0.104	6.0	4.0			
	Clayton	32.8	0.288	16.9	0.291	2.8	0.046	0.138	-83.3	0.128	59.2	71.2			
	Joe	-1.1	0.183	1.4	0.181	-0.3	0.032	0.078	-7.7	0.096	18.4	9.6			
	FGM	23.6	0.243	12.9	0.252	2.4	0.041	0.112	-57.8	0.116	1.6	2.4			
	AMH	24.6	0.252	13.5	0.259	2.8	0.047	0.116	-59.7	0.118	0.4	2.4			
	Frank	19.8	0.239	11.3	0.245	2.5	0.042	0.111	-47.0	0.113	2.0	2.0			
$\tau = 0.5$	Gumbel	-5.0	0.193	0.3	0.183	1.3	0.036	0.082	7.8	0.096	12.4	8.4			
	Normal	14.6	0.182	10.2	0.208	2.3	0.045	0.093	-11.8	0.095	0.4	1.6			
	Clayton	81.0	0.639	39.1	0.606	-0.1	0.043	0.344	-65.5	0.202	0.0	0.8			
	Joe	-4.2	0.123	0.0	0.122	-0.6	0.039	0.052	1.4	0.08	85.8	70.0			
	FGM	67.6	0.521	32.3	0.495	-3.8	0.049	0.278	-55.6	0.175	0.0	0.8			
	AMH	70.2	0.541	34.0	0.521	-1.9	0.037	0.276	-54.7	0.179	0.0	0.4			
$\tau = 0.75$	Frank	21.5	0.207	13.1	0.235	1.2	0.039	0.086	-11.6	0.097	2.8	9.7			
	Gumbel	-3.4	0.122	1.5	0.123	1.8	0.043	0.055	3.2	0.081	10.9	16.6			
	Normal	4.3	0.101	5.6	0.134	1.4	0.041	0.052	-2.4	0.081	0.0	0.5			
	Clayton	28.9	0.297	18.8	0.337	4.1	0.065	0.165	-13.3	0.12	0.0	0.0			
	Joe	-3.2	0.09	0.6	0.1	-0.5	0.036	0.04	2.3	0.073	85.3	68.8			
	FGM	100.3	0.765	45.1	0.682	-13.4	0.138	0.528	-70.4	0.242	0.0	0.0			
$\tau = 0.9$	AMH	92.7	0.708	43.5	0.658	-9.2	0.097	0.44	-58.6	0.227	0.0	0.0			
	Frank	8.5	0.108	7.0	0.144	0.0	0.035	0.038	-0.3	0.077	1.4	9.2			
	Gumbel	-2.2	0.088	2.1	0.103	1.1	0.038	0.039	2.8	0.074	13.3	21.6			

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.

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{\sigma}_{21}(z_1)$			AIC (%)	BIC (%)
	Bias (%)	RMSE		Bias (%)	RMSE		Bias (%)	RMSE		Bias (%)	RMSE		Bias (%)	RMSE			
$\tau = -0.2$	Normal	2.7	0.269	2.6	0.239	0.3	0.041	0.142	0.102	9.2	10.4						
	Clayton	-65.5	0.521	-25.3	0.406	-1.5	0.040	-127.9	0.266	0.147	51.6	52.0					
	Joe	-56.1	0.451	-21.9	0.362	-2.9	0.043	-104.5	0.217	0.129	0.4	0.8					
	FGM	-11.4	0.182	-3.1	0.169	-1.1	0.034	-10.1	0.077	0.087	14.4	11.2					
	AMH	-18.2	0.207	-5.7	0.178	-0.8	0.034	-27.9	0.090	0.088	7.2	4.4					
$\tau = 0.15$	Frank	1.9	0.253	2.5	0.233	0.1	0.040	18.4	0.135	0.103	14.8	17.2					
	Gumbel	-55.8	0.445	-21.7	0.356	-2.9	0.043	-104.3	0.216	0.128	2.4	4.0					
	Normal	11.6	0.342	6.3	0.300	-0.4	0.036	-48.6	0.187	0.119	12.8	8.8					
	Clayton	14.7	0.192	8.2	0.190	0.2	0.039	-52.2	0.111	0.102	43.2	41.6					
	Joe	6.1	0.287	3.3	0.269	-2.3	0.042	-40.7	0.147	0.116	2.8	4.8					
$\tau = 0.1$	FGM	12.0	0.268	6.5	0.244	-0.9	0.034	-47.7	0.148	0.108	10.4	10.4					
	AMH	13.7	0.267	7.5	0.245	-0.1	0.037	-48.0	0.146	0.108	7.2	7.6					
	Frank	8.9	0.317	5.2	0.280	-0.5	0.037	-38.6	0.173	0.114	16.4	13.2					
	Gumbel	2.4	0.283	2.1	0.259	-1.6	0.038	-27.2	0.146	0.112	7.2	13.6					

Table 7: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\sigma}_{21}(z_1)$, 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	AIC (%)	BIC (%)
0.12	Normal	2.6	0.330	2.6	0.282	0.1	0.041	0.179	10.0	10.0
	Clayton	-52.7	0.432	-19.9	0.336	-0.9	0.042	0.216	52.6	52.6
	Joe	-43.6	0.364	-16.8	0.296	-2.7	0.043	0.164	1.6	1.6
	FGM	-4.4	0.207	-0.3	0.184	-0.9	0.036	0.104	8.8	8.0
	AMH	-11.7	0.231	-3.1	0.193	-0.4	0.038	0.112	7.6	4.0
	Frank	8.1	0.270	5.0	0.239	0.0	0.040	0.153	17.3	19.7
0.1	Gumbel	-44.7	0.377	-17.2	0.304	-2.5	0.042	0.174	2.0	4.0
	Normal	11.5	0.370	6.5	0.320	-0.1	0.036	0.202	8.4	9.2
	Clayton	4.1	0.162	3.9	0.163	0.2	0.040	0.088	48.2	47.8
	Joe	4.9	0.215	3.4	0.206	-2.2	0.040	0.109	4.8	4.8
	FGM	13.6	0.294	7.4	0.264	-0.7	0.034	0.167	11.6	10.0
	AMH	12.6	0.281	7.3	0.253	0.1	0.038	0.160	7.2	4.8
0.28	Frank	10.6	0.318	6.1	0.280	-0.5	0.035	0.180	8.0	8.0
	Gumbel	0.9	0.237	1.9	0.217	-1.7	0.038	0.120	11.6	15.3
	Normal	10.2	0.466	4.7	0.390	-3.8	0.058	0.278	12.4	12.0
	Clayton	8.3	0.200	5.6	0.189	-1.2	0.050	0.121	39.0	22.9
	Joe	31.1	0.447	12.8	0.382	-7.0	0.084	0.277	0.0	3.2
	FGM	40.7	0.454	17.4	0.384	-6.3	0.070	0.295	4.8	7.2
0.5	AMH	26.3	0.399	12.5	0.339	-2.6	0.052	0.251	30.5	31.3
	Frank	29.0	0.556	12.5	0.464	-4.1	0.063	0.342	9.6	13.7
	Gumbel	14.5	0.422	6.0	0.362	-5.8	0.077	0.252	3.6	9.6

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.

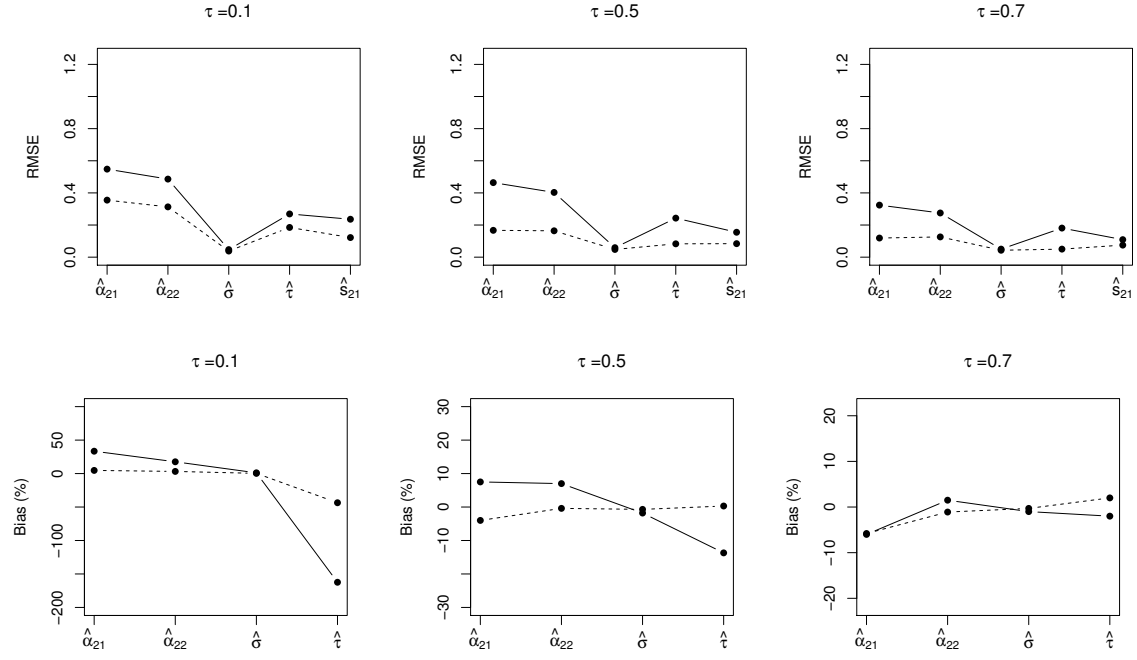
	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{\delta}_{21}(z_1)$		AIC (%)	BIC (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	RMSE				
$\tau = 0.2$	Normal	7.7	0.336	5.1	0.291	0.0	0.036	-27.6	0.184	0.117	10.8	12.4				
	Clayton	24.6	0.247	12.7	0.238	0.3	0.040	-57.5	0.145	0.115	31.2	33.6				
	Joe	4.6	0.343	2.8	0.303	-2.4	0.043	-28.3	0.173	0.123	3.6	7.2				
	FGM	13.7	0.253	7.6	0.231	-0.9	0.034	-38.0	0.141	0.107	11.6	8.8				
	AMH	17.0	0.266	9.3	0.243	0.2	0.038	-40.3	0.149	0.110	10.0	9.6				
	Frank	5.0	0.301	3.9	0.261	-0.1	0.036	-19.0	0.165	0.111	25.6	16.4				
$\tau = 0.5$	Gumbel	-1.3	0.329	0.7	0.286	-1.2	0.040	-12.2	0.169	0.120	7.2	12.0				
	Normal	-0.8	0.198	1.7	0.181	-0.6	0.043	-3.0	0.108	0.091	12.6	12.6				
	Clayton	42.0	0.371	21.7	0.366	0.0	0.046	-34.8	0.211	0.140	2.4	3.2				
	Joe	-13.4	0.317	-5.8	0.278	-2.4	0.055	0.8	0.163	0.109	5.7	5.3				
	FGM	57.0	0.443	25.6	0.398	-7.7	0.083	-56.0	0.281	0.162	0.8	6.1				
	AMH	50.6	0.400	24.3	0.382	-3.7	0.049	-43.1	0.221	0.149	4.5	12.1				
$\tau = 0.7$	Frank	-2.5	0.179	0.8	0.161	-0.5	0.040	0.2	0.096	0.084	64.8	51.4				
	Gumbel	-15.9	0.227	-5.7	0.192	-0.6	0.046	8.7	0.112	0.092	9.3	9.3				
	Normal	-4.4	0.112	0.3	0.114	0.1	0.041	1.1	0.053	0.080	7.1	10.5				
	Clayton	15.8	0.216	11.4	0.240	2.9	0.057	-6.2	0.119	0.101	7.1	6.3				
	Joe	-18.0	0.172	-7.4	0.159	-0.3	0.040	5.3	0.062	0.083	6.7	5.4				
	FGM	88.8	0.678	39.0	0.591	-13.6	0.139	-68.3	0.478	0.222	0.0	0.0				
τ	AMH	80.0	0.612	37.0	0.562	-9.0	0.095	-54.1	0.379	0.206	0.0	0.8				
	Frank	-5.2	0.105	-0.3	0.107	-0.1	0.038	2.8	0.050	0.072	65.7	64.0				
	Gumbel	-12.8	0.139	-4.0	0.124	0.4	0.040	6.1	0.059	0.079	13.4	13.0				

Table 9: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\delta}_{21}(z_1)$, 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.

	$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		BIC (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	AIC (%)	
Joe	Normal	7.2	0.257	5.2	0.236	2.2	0.044	-15.9	0.132	0.107	8.8
	Clayton	35.9	0.318	17.6	0.305	2.1	0.045	-70.7	0.163	0.134	54.0
	Joe	8.9	0.235	4.8	0.226	-1.4	0.036	-28.6	0.119	0.105	15.2
	FGM	25.5	0.261	12.9	0.254	1.1	0.034	-50.5	0.131	0.116	1.6
	AMH	26.1	0.274	13.4	0.264	2.0	0.046	-49.6	0.138	0.119	3.2
	Frank	20.6	0.268	10.8	0.256	1.4	0.038	-39.7	0.135	0.112	2.8
Joe	Gumbel	0.8	0.230	1.8	0.212	0.2	0.036	-8.2	0.113	0.100	14.4
	Normal	6.0	0.159	5.1	0.172	0.9	0.042	-6.2	0.083	0.092	16.5
	Clayton	56.7	0.472	28.1	0.455	0.7	0.048	-45.3	0.252	0.163	1.6
	Joe	0.1	0.183	0.4	0.181	-3.2	0.053	-6.5	0.098	0.092	20.5
	FGM	63.6	0.491	29.0	0.448	-5.7	0.065	-56.0	0.280	0.171	4.4
	AMH	58.5	0.457	28.1	0.436	-2.0	0.042	-46.9	0.239	0.161	2.8
Joe	Frank	15.8	0.192	9.2	0.203	-0.2	0.040	-11.5	0.095	0.096	11.2
	Gumbel	-3.7	0.152	-0.1	0.153	-0.2	0.041	1.3	0.073	0.086	43.0
	Normal	-0.4	0.099	1.9	0.124	0.2	0.039	-0.1	0.039	0.073	24.5
	Clayton	8.8	0.143	7.8	0.184	2.8	0.050	-2.2	0.065	0.083	5.6
	Joe	-4.4	0.105	-0.8	0.121	-1.0	0.042	-0.5	0.044	0.074	9.2
	FGM	103.1	0.787	44.9	0.681	-15.9	0.162	-72.2	0.578	0.245	0.0
Joe	AMH	93.2	0.712	42.8	0.649	-10.9	0.113	-58.7	0.470	0.228	0.0
	Frank	2.5	0.102	3.3	0.132	-0.1	0.040	-1.2	0.043	0.074	6.1
	Gumbel	-2.5	0.095	0.7	0.116	0.0	0.038	1.5	0.037	0.073	54.6

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.

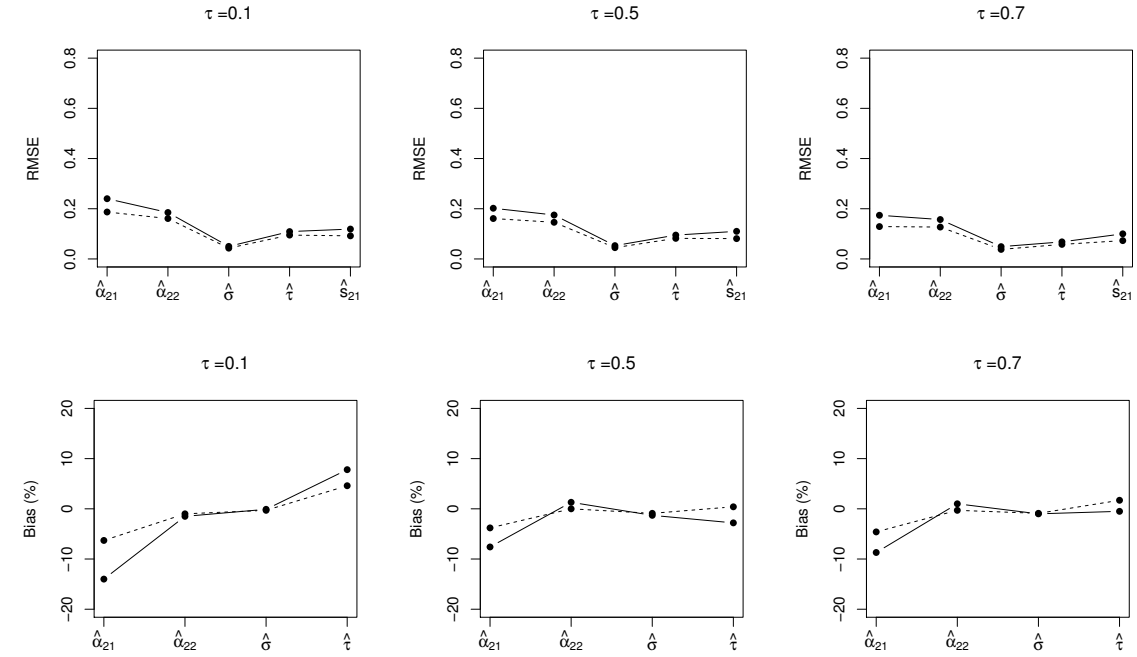
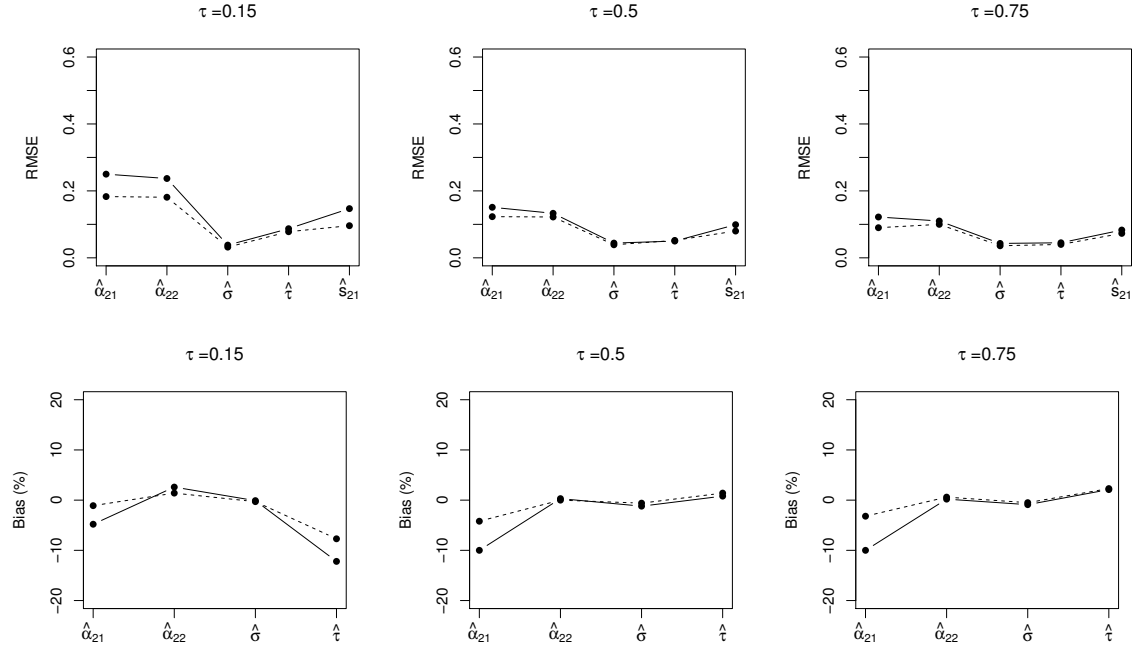


Figure 5: RMSEs and percentage bias of $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and \hat{s}_{21} for data generated using the (a) normal and (b) Clayton copulas when employing the normal and Clayton copula regression spline sample selection models, respectively. Solid line: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

(a) Joe copula.



(b) FGM copula.

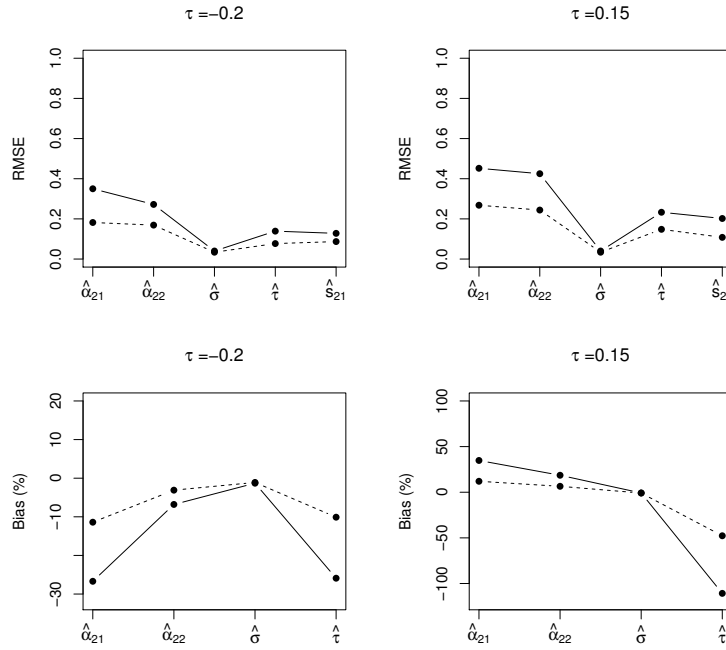
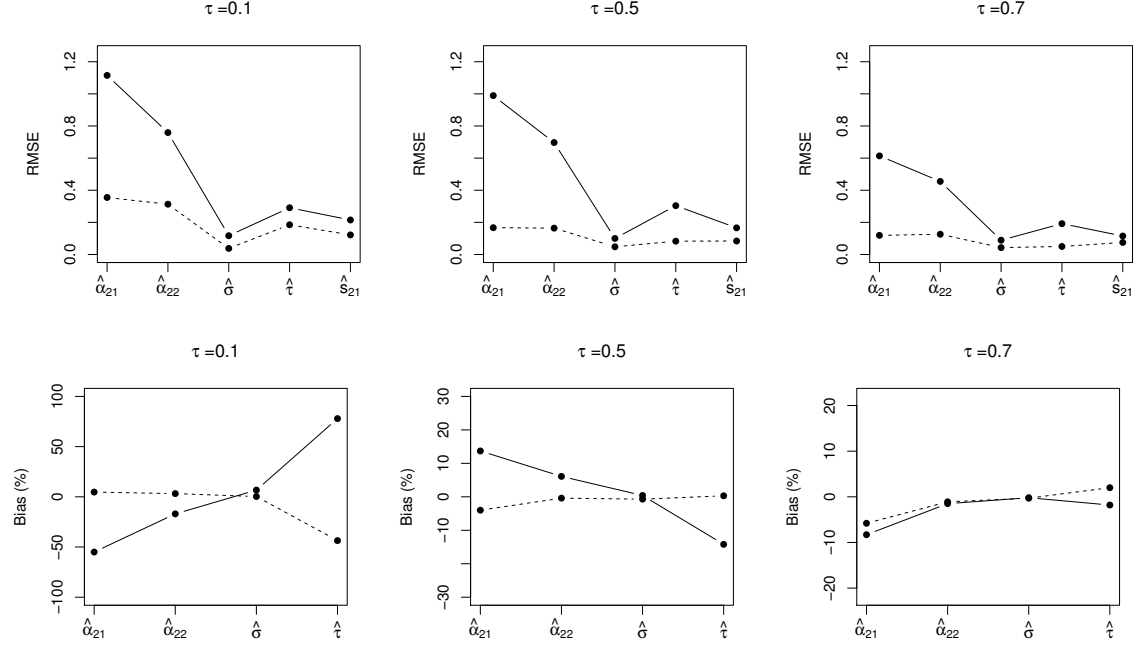


Figure 6: RMSEs and percentage bias of $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and \hat{s}_{21} for data simulated using the (a) Joe and (b) FGM copulas when employing the Joe and FGM copula regression spline sample selection models, respectively. Solid line: model (10)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

(a) Normal copula.



(b) Clayton copula.

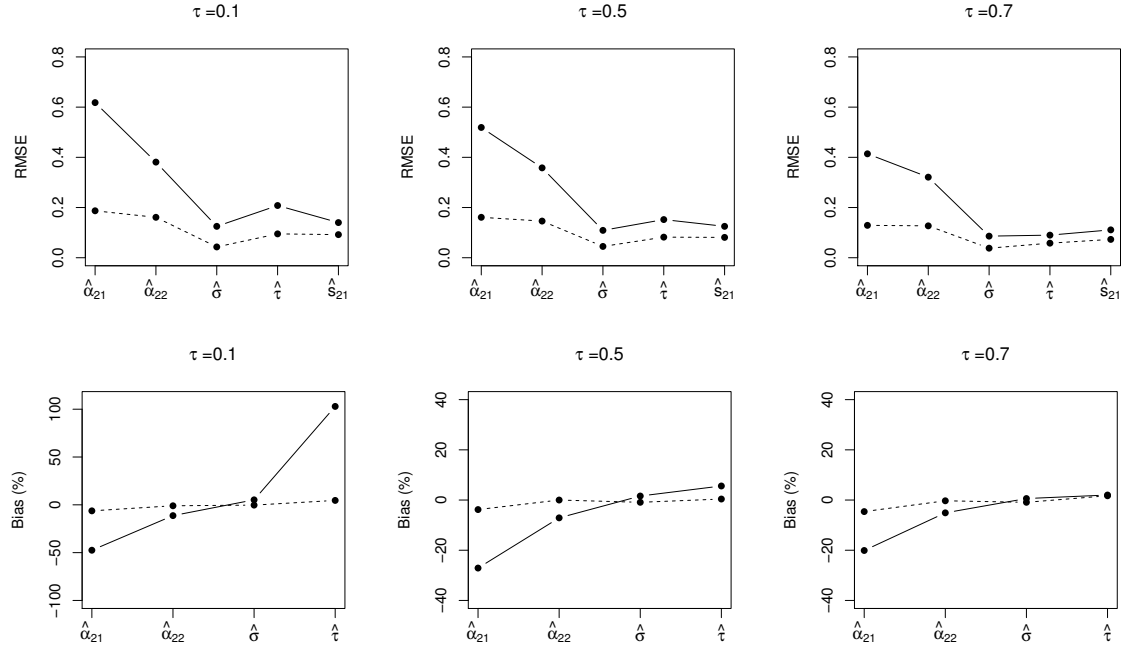
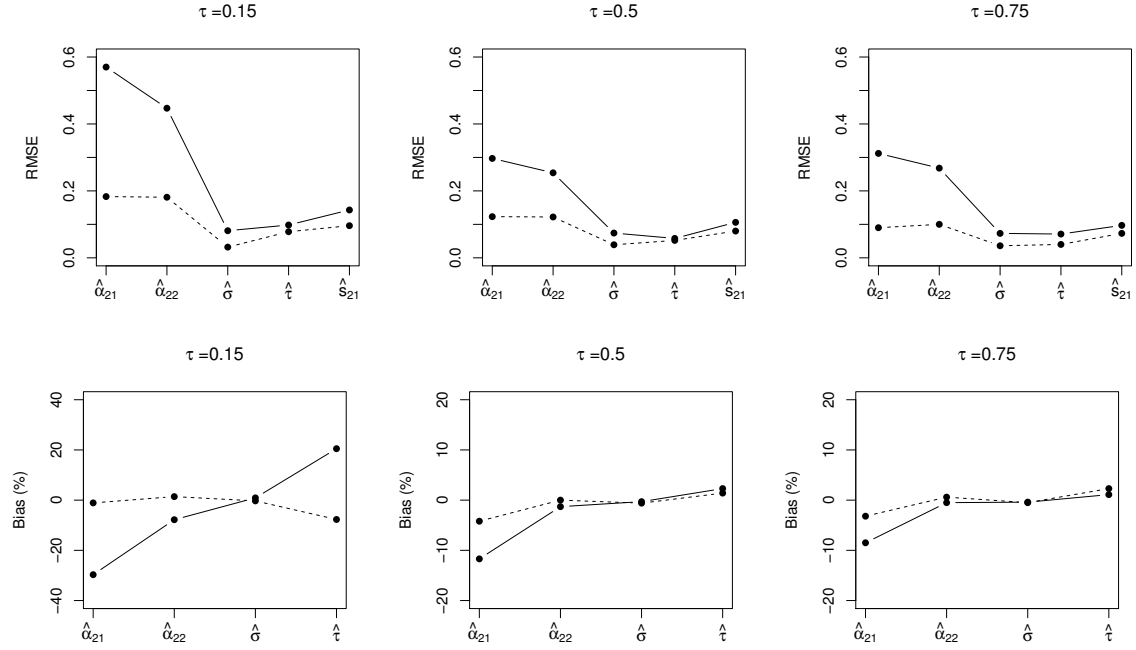


Figure 7: RMSEs and percentage bias of $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and \hat{s}_{21} for data generated using the (a) normal and (b) Clayton copulas when employing the normal and Clayton copula regression spline sample selection models, respectively. Solid line: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

(a) Joe copula.



(b) FGM copula.

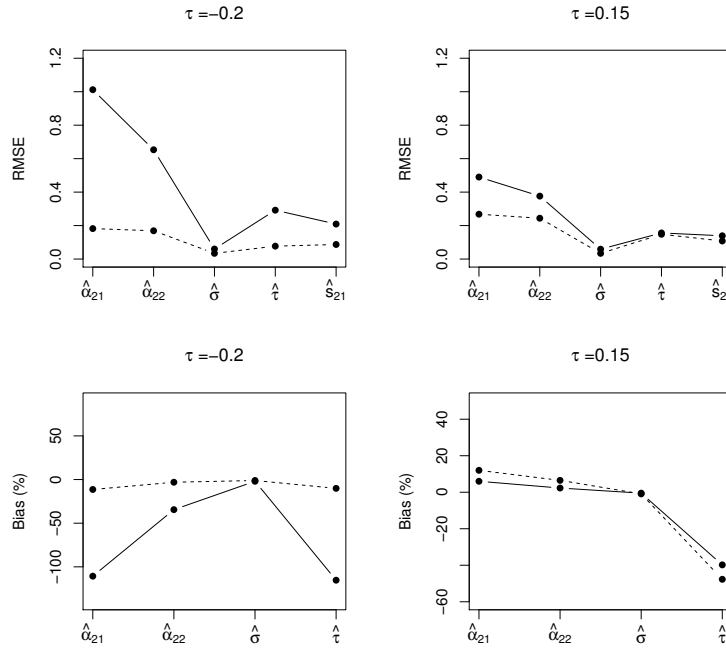


Figure 8: RMSEs and percentage bias of $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and \hat{s}_{21} for data simulated using the (a) Joe and (b) FGM copulas when employing the Joe and FGM copula regression spline sample selection models, respectively. Solid line: model (11)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

(a) Random samples generated using (9) and a normal product distribution ($\tau = 0$).

n	$\alpha(\%)$	Copula				
		Normal	FGM	AMH	Frank	Gumbel
1000	1	12	0	3	10	45
	5	19	0	6	14	46
	10	24	0	8	20	46
3000	1	5	0	2	4	38
	5	12	1	5	9	39
	10	16	6	9	16	39
5000	1	2	0	1	1	37
	5	8	4	4	6	37
	10	14	8	11	12	37

(b) Random samples generated using (10) and a normal product distribution ($\tau = 0$).

n	$\alpha(\%)$	Copula				
		Normal	FGM	AMH	Frank	Gumbel
1000	1	20	0	1	10	42
	5	33	0	2	20	43
	10	40	0	4	26	43
3000	1	32	0	1	12	37
	5	47	4	7	23	38
	10	53	12	16	32	38
5000	1	25	1	1	7	36
	5	36	6	6	15	36
	10	43	13	17	20	36

(c) Random samples generated using (11) and a normal product distribution ($\tau = 0$).

n	$\alpha(\%)$	Copula				
		Normal	FGM	AMH	Frank	Gumbel
1000	1	21	0	4	12	81
	5	29	0	9	19	81
	10	37	0	15	27	81
3000	1	23	0	12	8	63
	5	35	1	20	15	63
	10	42	4	27	19	63
5000	1	21	0	23	6	57
	5	31	1	26	12	57
	10	40	8	33	22	58

Table 11: Null rejection probabilities (%) for testing $H_0 : \tau = 0$ based on $(100 - \alpha)\%$ confidence intervals for τ , when fitting normal, FGM, AMH, Frank and Gumbel copula regression spline sample selection models.

(a) Data generated using the bivariate FGM copula.

n	$\alpha(\%)$	$\tau = -0.2$	$\tau = 0.15$
1000	1	0	0
	5	0	0
	10	0	0
3000	1	0	0
	5	8	10
	10	19	27
5000	1	4	7
	5	22	46
	10	34	66

(b) Data generated using the bivariate AMH copula.

n	$\alpha(\%)$	$\tau = -0.12$	$\tau = 0.1$	$\tau = 0.28$
1000	1	1	4	34
	5	2	13	49
	10	4	19	52
3000	1	0	14	87
	5	15	27	92
	10	32	38	93
5000	1	9	20	97
	5	49	36	98
	10	70	44	100

(c) Data generated using the bivariate Frank copula.

n	$\alpha(\%)$	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.7$
1000	1	28	99	100
	5	46	99	100
	10	54	99	100
3000	1	57	100	100
	5	76	100	100
	10	83	100	100
5000	1	78	100	100
	5	90	100	100
	10	91	100	100

Table 12: Null rejection probabilities (%) for testing the null hypothesis $H_0 : \tau = 0$ based on $(100 - \alpha)\%$ confidence intervals for τ , when fitting (a) FGM, (b) AMH, (c) Frank copula regression spline sample selection models. Random samples were generated using (9), with exclusion restriction.

(a) Data generated using the bivariate FGM copula.

n	$\alpha(\%)$	$\tau = -0.2$	$\tau = 0.15$
1000	1	0	0
	5	0	0
	10	0	0
3000	1	0	0
	5	9	4
	10	20	17
5000	1	2	2
	5	20	27
	10	32	46

(b) Data generated using the bivariate AMH copula.

n	$\alpha(\%)$	$\tau = -0.12$	$\tau = 0.1$	$\tau = 0.28$
1000	1	0	0	8
	5	2	2	19
	10	6	5	23
3000	1	2	4	23
	5	13	15	28
	10	32	24	36
5000	1	8	6	31
	5	40	22	39
	10	59	37	49

(c) Data generated using the bivariate Frank copula.

n	$\alpha(\%)$	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.7$
1000	1	20	97	100
	5	40	99	100
	10	47	100	100
3000	1	35	100	100
	5	53	100	100
	10	65	100	100
5000	1	58	100	100
	5	72	100	100
	10	83	100	100

Table 13: Null rejection probabilities (%) for testing the null hypothesis $H_0 : \tau = 0$ based on $(100 - \alpha)\%$ confidence intervals for τ , when fitting (a) FGM, (b) AMH, (c) Frank copula regression spline sample selection models. Random samples were generated using (10), without exclusion restriction.

(a) Data generated using the bivariate FGM copula.

n	$\alpha(\%)$	$\tau = -0.2$	$\tau = 0.15$
1000	1	0	0
	5	0	0
	10	0	0
3000	1	0	0
	5	1	4
	10	6	8
5000	1	1	2
	5	5	18
	10	15	36

(b) Data generated using the bivariate AMH copula.

n	$\alpha(\%)$	$\tau = -0.12$	$\tau = 0.1$	$\tau = 0.28$
1000	1	4	3	9
	5	12	13	21
	10	22	19	30
3000	1	20	15	33
	5	32	31	42
	10	38	38	48
5000	1	33	16	40
	5	46	32	48
	10	50	42	54

(c) Data generated using the bivariate Frank copula.

n	$\alpha(\%)$	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.7$
1000	1	22	78	94
	5	38	86	97
	10	51	92	98
3000	1	40	95	100
	5	59	97	100
	10	67	99	100
5000	1	62	99	100
	5	80	100	100
	10	85	100	100

Table 14: Null rejection probabilities (%) for testing the null hypothesis $H_0 : \tau = 0$ based on $(100 - \alpha)\%$ confidence intervals for τ , when fitting (a) FGM, (b) AMH, (c) Frank copula regression spline sample selection models. Random samples were generated using (11), without exclusion restriction.

	$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{\sigma}_{21}(z_1)$		AIC (%)	BIC (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE			
$\tau = 0.1$	Normal	0.6	0.174	2.0	0.156	0.021	-19.1	0.088	0.061	18.8	15.6	
	Clayton	10.3	0.131	6.3	0.135	0.025	-52.4	0.073	0.063	46.8	46.8	
	Joe	12.7	0.137	6.7	0.140	0.024	-73.6	0.082	0.066	5.2	7.2	
	FGM	9.0	0.176	5.4	0.166	0.020	-51.8	0.099	0.068	6.0	4.8	
	AMH	9.1	0.172	5.6	0.162	0.022	-49.3	0.097	0.068	8.4	8.4	
$\tau = 0.5$	Frank	9.1	0.173	5.4	0.165	0.020	-51.8	0.097	0.068	4.8	4.4	
	Gumbel	8.0	0.143	4.7	0.141	0.022	-53.1	0.078	0.063	10.0	12.8	
	Normal	-3.8	0.093	0.2	0.088	0.026	0.2	0.041	0.049	86.4	87.6	
	Clayton	33.8	0.276	18.3	0.29	0.031	-26.1	0.142	0.099	2.8	2.4	
	Joe	2.9	0.226	0.8	0.2	0.065	-14.2	0.138	0.071	0.0	0.0	
$\tau = 0.7$	FGM	58.5	0.446	26.0	0.395	0.077	-55.6	0.278	0.147	0.0	0.0	
	AMH	44.8	0.344	22.2	0.339	0.026	-36.6	0.184	0.119	0.4	1.6	
	Frank	4.3	0.109	3.6	0.113	0.029	-6.8	0.061	0.054	4.0	2.4	
	Gumbel	-9.1	0.135	-3.1	0.118	0.033	2.6	0.061	0.053	6.4	6.0	
	Normal	-4.7	0.072	-0.1	0.069	0.024	0.8	0.026	0.045	91.1	93.5	
$\tau = 0.9$	Clayton	14.5	0.136	10.5	0.178	0.037	-6.8	0.062	0.063	0.4	0.8	
	Joe	-10.9	0.107	-4.3	0.097	0.031	-2.0	0.034	0.05	0.0	0.0	
	FGM	90.6	0.687	39.4	0.594	0.138	-68.3	0.478	0.214	0.0	0.0	
	AMH	78.5	0.597	36.5	0.551	0.082	-52.5	0.368	0.191	0.0	0.0	
	Frank	-1.5	0.067	1.6	0.075	0.025	-2.1	0.032	0.045	1.2	0.8	
	Gumbel	-9.3	0.095	-2.6	0.08	0.025	2.7	0.031	0.046	7.3	4.8	

Table 15: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\sigma}_{21}(z_1)$, and percentage frequency at which each copula model was selected by AIC and BIC for data simulated using a normal bivariate distribution and (9), when employing the normal, Clayton, Joe, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Sample size equals $n = 3000$.

	$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		AIC (%)	BIC (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE			
Normal Clayton Joe FGM AMH	-0.7	0.14	1.5	0.121	0.0	0.015	-15.4	0.073	0.052	29.6	26.0	
	10.5	0.114	6.5	0.121	0.7	0.02	-55.4	0.07	0.055	38.8	36.0	
	14.4	0.13	7.5	0.133	-1.1	0.019	-80.8	0.085	0.06	2.4	7.2	
	7.9	0.131	5.1	0.126	-0.3	0.015	-48.9	0.077	0.053	4.8	4.0	
	7.9	0.131	5.1	0.126	0.0	0.016	-46.8	0.078	0.054	6.0	6.8	
Frank Gumbel	7.7	0.131	5.0	0.126	-0.3	0.015	-47.8	0.076	0.054	8.4	5.2	
	9.4	0.125	5.5	0.125	-1.0	0.018	-60.4	0.076	0.057	10.0	14.8	
Normal Clayton Joe FGM AMH	-4.3	0.076	0.0	0.067	-0.1	0.02	0.0	0.031	0.039	96.8	96.0	
	34.4	0.272	18.5	0.287	1.4	0.026	-27.0	0.142	0.095	0.0	0.0	
	0.3	0.169	-0.4	0.149	-5.5	0.063	-12.9	0.109	0.056	0.0	0.0	
	58.3	0.444	26.0	0.392	-7.3	0.075	-55.6	0.278	0.143	0.0	0.0	
	44.4	0.339	22.1	0.334	-1.3	0.021	-36.5	0.183	0.114	0.0	0.0	
Frank Gumbel	3.5	0.082	3.3	0.088	-0.8	0.022	-6.6	0.05	0.044	1.2	1.2	
	-10.1	0.111	-3.5	0.093	-1.6	0.028	2.7	0.041	0.042	2.0	2.8	
Normal Clayton Joe FGM AMH	-4.7	0.06	-0.1	0.052	-0.1	0.018	0.3	0.02	0.036	95.2	94.0	
	15.3	0.131	10.8	0.173	2.5	0.031	-7.7	0.062	0.058	0.0	0.0	
	-10.9	0.097	-4.4	0.085	-1.7	0.026	-2.8	0.03	0.039	0.0	0.0	
	90.2	0.686	39.3	0.591	-13.7	0.137	-68.3	0.478	0.211	0.0	0.0	
	78.1	0.594	36.3	0.547	-8.0	0.081	-52.5	0.368	0.188	0.0	0.0	
Frank Gumbel	-1.4	0.05	1.6	0.057	0.0	0.018	-2.6	0.029	0.037	0.4	1.2	
	-9.4	0.086	-2.8	0.067	-0.3	0.019	2.1	0.025	0.036	4.4	4.8	

Table 16: 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 and (9), when employing the normal, Clayton, Joe, FGM, AMH, Frank and Gumbel copula regression spline sample selection models. Sample size equals $n = 5000$.

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{\sigma}_{21}(z_1)$			<i>AIC</i> (%)	<i>BIC</i> (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	RMSE	RMSE		
$\tau = 0.1$	Normal	33.4	0.548	17.6	0.486	1.1	0.047	-162.4	0.269	0.236	14.9	11.6					
	Clayton	2.5	0.186	5.3	0.192	0.3	0.048	-45.3	0.096	0.137	60.2	59.8					
	Joe	-5.4	0.292	1.0	0.261	-2.0	0.044	-39.1	0.115	0.154	1.6	2.4					
	FGM	26.6	0.363	14.8	0.344	-0.2	0.038	-139.9	0.193	0.187	5.6	5.2					
	AMH	24.2	0.353	14.1	0.337	0.5	0.046	-123.8	0.181	0.186	2.4	4.4					
$\tau = 0.5$	Frank	28.9	0.423	15.8	0.389	0.1	0.042	-147.6	0.223	0.195	9.6	8.0					
	Gumbel	-6.8	0.307	0.6	0.267	-1.4	0.042	-27.9	0.125	0.153	5.6	8.4					
	Normal	7.5	0.464	7.0	0.403	-1.8	0.059	-13.7	0.243	0.155	44.2	26.5					
	Clayton	49.3	0.443	26.9	0.454	-2.1	0.063	-35.5	0.217	0.260	18.9	18.5					
	Joe	18.2	0.514	9.5	0.454	-6.8	0.094	-32.6	0.271	0.201	1.6	1.2					
$\tau = 0.7$	FGM	80.0	0.686	36.9	0.619	-9.3	0.099	-70.1	0.382	0.316	3.2	8.8					
	AMH	63.8	0.553	32.2	0.535	-4.5	0.066	-48.3	0.281	0.290	4.8	21.7					
	Frank	44.8	0.692	22.4	0.607	-4.1	0.073	-41.9	0.372	0.241	12.4	13.3					
	Gumbel	0.1	0.433	2.9	0.378	-3.5	0.076	-12.2	0.211	0.165	14.9	10.0					
	Normal	-6.0	0.324	1.5	0.275	-1.0	0.050	-2.0	0.181	0.109	58.5	52.4					
$\tau = 0.9$	Clayton	19.3	0.249	14.9	0.294	-0.2	0.053	-7.2	0.108	0.188	12.1	18.5					
	Joe	-16.6	0.193	-5.0	0.166	-3.0	0.059	-6.2	0.079	0.103	0.0	0.8					
	FGM	103.5	0.805	46.6	0.717	-16.5	0.168	-70.8	0.503	0.379	0.0	0.0					
	AMH	94.1	0.731	45.4	0.696	-11.5	0.120	-54.9	0.394	0.373	0.0	2.0					
	Frank	7.2	0.415	7.2	0.361	-2.1	0.058	-9.6	0.245	0.140	6.9	12.1					
	Gumbel	-17.2	0.186	-3.9	0.152	-0.8	0.049	2.0	0.055	0.096	22.6	14.1					

Table 17: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\sigma}_{21}(z_1)$, and percentage frequency at which each copula model was selected by *AIC* and *BIC* for data simulated using a normal bivariate distribution and (10), 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 4.2 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 (%)
Joe	Normal	54.3	0.628	25.7	0.546	0.2	0.045	0.340	0.277	18.6	24.3				
	Clayton	-14.0	0.240	-1.5	0.185	-0.1	0.050	0.109	0.119	44.5	41.3				
	Joe	-5.6	0.322	0.6	0.281	-3.3	0.053	0.132	0.154	2.0	1.2				
	FGM	33.9	0.391	17.3	0.359	-1.8	0.042	0.232	0.194	5.3	3.6				
	AMH	33.2	0.370	17.4	0.346	-1.2	0.043	0.216	0.194	6.1	6.5				
Frank	Normal	39.4	0.507	19.6	0.448	-1.0	0.045	0.286	0.223	18.6	16.2				
	Clayton	-8.6	0.351	-0.5	0.295	-2.8	0.053	0.150	0.157	4.9	6.9				
	Joe	77.7	1.078	34.0	0.902	-6.0	0.086	0.645	0.387	6.5	5.6				
	FGM	-7.6	0.202	1.3	0.175	-1.3	0.053	0.095	0.110	87.5	65.7				
	AMH	15.6	0.628	7.0	0.540	-9.4	0.124	0.356	0.292	0.0	0.0				
Gumbel	Normal	109.0	0.916	47.0	0.779	-13.1	0.135	0.598	0.353	0.4	2.4				
	Clayton	94.6	0.832	42.3	0.719	-10.5	0.114	0.531	0.325	0.8	12.1				
	Joe	116.2	1.159	50.2	0.976	-8.5	0.103	0.721	0.424	4.4	12.5				
	FGM	-2.3	0.600	0.4	0.504	-6.5	0.108	0.321	0.272	0.4	1.6				
	AMH	25.5	0.880	13.1	0.742	-3.5	0.073	0.534	0.264	4.9	5.7				
Frank	Normal	-8.7	0.174	1.0	0.157	-1.0	0.049	0.068	0.100	92.7	84.6				
	Clayton	-26.5	0.428	-10.2	0.365	-4.0	0.084	0.198	0.200	0.0	0.0				
	Joe	110.6	0.918	48.4	0.794	-17.9	0.187	0.636	0.375	0.4	0.8				
	FGM	99.0	0.831	45.7	0.746	-13.9	0.144	0.539	0.346	0.0	5.3				
	AMH	58.8	1.040	27.3	0.879	-5.6	0.093	0.660	0.314	0.4	2.8				
Gumbel	Normal	-31.3	0.388	-10.8	0.313	-1.4	0.070	0.156	0.182	1.6	0.8				
	Clayton														
	Joe														
	FGM														
	AMH														

Table 18: 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 and (10), 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 4.2 for further details.

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{\delta}_{21}(z_1)$		AIC (%)	BIC (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	RMSE				
$\tau = 0.15$	Normal	27.5	0.380	17.2	0.379	3.4	0.054	-66.7	0.163	0.216	4.0	2.4				
	Clayton	40.4	0.332	22.6	0.365	2.7	0.047	-95.3	0.147	0.237	86.3	88.7				
	Joe	-4.8	0.250	2.6	0.237	-0.1	0.038	-12.2	0.087	0.147	4.0	2.8				
	FGM	35.6	0.325	20.5	0.354	2.7	0.047	-84.0	0.144	0.227	0.4	0.4				
	AMH	34.8	0.328	20.3	0.354	3.2	0.052	-80.5	0.145	0.227	0.4	0.4				
$\tau = 0.5$	Frank	33.4	0.339	19.6	0.359	2.9	0.048	-78.8	0.150	0.223	1.6	2.8				
	Gumbel	-6.7	0.257	2.3	0.239	1.5	0.043	0.5	0.095	0.144	3.2	2.4				
	Normal	19.8	0.272	15.4	0.305	0.5	0.047	-11.7	0.125	0.180	0.8	2.8				
	Clayton	102.8	0.807	51.2	0.786	-3.7	0.063	-68.1	0.360	0.427	1.2	3.2				
	Joe	-10.0	0.151	0.3	0.133	-1.2	0.044	0.8	0.050	0.099	77.9	57.8				
$\tau = 0.75$	FGM	81.5	0.628	40.8	0.622	-6.3	0.072	-56.2	0.283	0.350	0.0	4.8				
	AMH	85.7	0.660	43.5	0.663	-4.6	0.059	-53.7	0.272	0.372	0.0	0.0				
	Frank	28.5	0.288	18.8	0.330	-1.0	0.042	-12.4	0.111	0.210	6.0	18.9				
	Gumbel	-10.1	0.152	1.7	0.136	1.5	0.046	7.6	0.064	0.102	14.1	12.4				
	Normal	-3.8	0.206	5.2	0.198	2.0	0.051	2.6	0.115	0.105	0.0	1.8				
$\tau = 1$	Clayton	25.2	0.371	20.3	0.414	3.0	0.082	-6.2	0.187	0.222	0.0	0.0				
	Joe	-10.0	0.122	0.2	0.110	-0.9	0.043	2.1	0.045	0.083	78.2	65.9				
	FGM	114.6	0.875	53.2	0.805	-16.6	0.169	-70.6	0.531	0.425	0.0	0.0				
	AMH	111.1	0.849	54.0	0.816	-13.1	0.135	-59.0	0.444	0.433	0.0	0.0				
	Frank	5.4	0.236	8.2	0.233	-0.9	0.046	1.7	0.134	0.127	1.4	10.0				
	Gumbel	-10.2	0.123	1.6	0.114	1.5	0.047	6.0	0.060	0.085	20.5	22.3				

Table 19: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\delta}_{21}(z_1)$, 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 and (10), 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 4.2 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 (%)
Joe	Normal	0.1	0.533	4.0	0.442	2.0	0.055	0.055	18.8	0.246	0.203	20.9	0.203	20.9	30.5
	Clayton	-96.0	0.766	-34.3	0.553	0.1	0.053	0.053	-163.8	0.348	0.247	29.3	0.247	29.3	22.1
	Joe	-108.0	0.956	-41.4	0.755	-2.6	0.054	0.054	-162.8	0.377	0.314	2.4	0.314	2.4	2.0
	FGM	-26.7	0.350	-6.8	0.272	-1.3	0.040	0.040	-25.9	0.139	0.128	11.2	0.128	11.2	8.8
	AMH	-33.4	0.383	-9.2	0.286	-0.7	0.041	0.041	-44.6	0.158	0.129	3.6	0.129	3.6	3.2
	Frank	-7.5	0.447	1.2	0.376	0.8	0.050	0.050	9.9	0.211	0.175	27.7	0.175	27.7	29.3
Joe	Gumbel	-116.4	1.028	-44.4	0.803	-1.2	0.057	0.057	-183.1	0.428	0.336	4.8	0.336	4.8	4.0
	Normal	53.8	0.700	26.5	0.618	1.0	0.049	0.049	-156.1	0.344	0.287	17.0	0.287	17.0	13.4
	Clayton	14.3	0.224	10.7	0.239	-0.4	0.045	0.045	-55.5	0.122	0.147	44.1	0.147	44.1	47.4
	Joe	-7.5	0.432	0.3	0.380	-2.2	0.047	0.047	-25.5	0.166	0.205	4.9	0.205	4.9	3.6
	FGM	34.8	0.452	18.6	0.425	-0.8	0.040	0.040	-110.8	0.233	0.202	9.3	0.202	9.3	6.9
	AMH	37.2	0.439	20.0	0.421	-0.1	0.042	0.042	-110.4	0.224	0.206	11.3	0.206	11.3	12.6
Joe	Frank	28.6	0.520	16.1	0.471	-0.2	0.045	0.045	-95.2	0.261	0.207	7.3	0.207	7.3	10.9
	Gumbel	-8.9	0.437	0.1	0.376	-1.3	0.047	0.047	-14.9	0.180	0.201	6.1	0.201	6.1	5.3

Table 20: 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 and (10), 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 4.2 for further details.

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{\delta}_{21}(z_1)$			AIC (%)	BIC (%)
	Bias (%)	RMSE		Bias (%)	RMSE		Bias (%)	RMSE		Bias (%)	RMSE		RMSE	RMSE			
$\tau = -0.12$	Normal	7.4	0.515	7.1	0.441		1.9	0.053		49.9	0.248		0.219			21.5	29.6
	Clayton	-75.4	0.611	-26.1	0.436		-0.1	0.050		-173.4	0.264		0.187			36.8	31.6
	Joe	-75.2	0.697	-27.5	0.540		-2.4	0.054		-150.7	0.263		0.210			0.4	0.4
	FGM	-15.2	0.301	-2.1	0.253		-1.0	0.040		-2.1	0.123		0.129			6.9	4.0
	AMH	-22.2	0.329	-4.6	0.260		-0.3	0.042		-28.7	0.137		0.126			2.4	3.2
$\tau = 0.1$	Frank	0.5	0.425	4.4	0.373		0.7	0.048		40.2	0.207		0.178			25.5	25.1
	Gumbel	-80.2	0.737	-29.3	0.563		-1.8	0.054		-168.4	0.297		0.222			6.5	6.1
	Normal	47.7	0.612	24.1	0.550		1.0	0.047		-214.3	0.309		0.265			16.9	18.1
	Clayton	1.8	0.186	5.6	0.195		-0.3	0.044		-43.4	0.094		0.133			54.0	54.0
	Joe	-5.1	0.314	1.8	0.273		-2.2	0.045		-39.1	0.131		0.166			2.4	0.8
$\tau = 0.28$	FGM	27.9	0.386	15.9	0.373		-0.6	0.038		-147.0	0.205		0.189			6.5	5.2
	AMH	28.2	0.368	16.4	0.362		-0.4	0.061		-139.1	0.192		0.200			4.8	6.0
	Frank	29.1	0.464	16.5	0.429		-0.2	0.042		-151.6	0.243		0.206			10.1	11.7
	Gumbel	-5.7	0.320	1.8	0.276		-1.6	0.045		-31.5	0.140		0.161			5.2	4.0
	Normal	79.3	0.933	35.8	0.797		-3.4	0.060		-127.8	0.517		0.354			18.7	15.9
$\tau = 0.28$	Clayton	7.8	0.239	7.9	0.236		-2.2	0.058		-24.8	0.142		0.141			33.7	21.1
	Joe	24.8	0.501	12.5	0.445		-7.4	0.091		-65.0	0.278		0.250			0.0	0.0
	FGM	78.7	0.702	35.5	0.618		-6.8	0.078		-128.9	0.424		0.288			4.9	11.0
	AMH	70.1	0.655	32.6	0.582		-5.4	0.070		-111.5	0.389		0.274			16.3	23.2
	Frank	89.9	0.892	40.3	0.772		-4.8	0.067		-143.9	0.516		0.346			19.5	18.3
	Gumbel	16.5	0.508	9.4	0.440		-6.3	0.087		-51.3	0.277		0.247			6.9	10.6

Table 21: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\delta}_{21}(z_1)$, 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 and (10), 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 4.2 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 (%)
Joe	Normal	56.1	0.723	28.1	0.641	1.1	0.050	0.362	-122.5	0.362	0.293	16.1	12.9		
	Clayton	27.8	0.274	16.8	0.294	-0.3	0.042	0.159	-66.3	0.159	0.178	45.0	47.8		
	Joe	1.2	0.441	4.3	0.386	-1.9	0.048	0.186	-36.5	0.186	0.228	2.8	1.2		
	FGM	38.3	0.467	20.7	0.437	-0.6	0.038	0.246	-90.0	0.246	0.209	9.6	7.6		
	AMH	46.7	0.499	24.6	0.473	0.2	0.041	0.256	-99.7	0.256	0.232	5.6	13.7		
Frank	Normal	34.1	0.568	18.9	0.509	0.3	0.047	0.288	-81.2	0.288	0.219	15.7	10.4		
	Clayton	0.9	0.436	4.7	0.376	-0.9	0.048	0.195	-29.8	0.195	0.218	5.2	6.4		
	Joe	33.9	0.823	19.2	0.706	0.0	0.052	0.433	-35.7	0.433	0.238	13.6	7.6		
	FGM	54.8	0.482	30.3	0.496	-2.0	0.060	0.252	-41.9	0.252	0.252	8.4	10.8		
	AMH	-25.8	0.423	-8.2	0.345	-1.6	0.059	0.163	-2.9	0.163	0.189	8.0	2.8		
Gumbel	Normal	79.7	0.696	37.9	0.639	-8.4	0.090	0.396	-71.6	0.396	0.297	1.2	7.6		
	Clayton	77.0	0.678	38.3	0.641	-5.4	0.066	0.358	-61.1	0.358	0.301	3.2	23.2		
	Joe	33.0	0.770	18.6	0.663	-0.7	0.048	0.411	-33.2	0.411	0.215	56.4	43.2		
	FGM	-29.1	0.366	-8.1	0.279	0.7	0.055	0.143	8.5	0.143	0.162	9.2	4.8		
	AMH	2.9	0.546	6.5	0.469	0.5	0.051	0.306	-8.8	0.306	0.142	4.9	7.4		
Frank	Normal	21.6	0.344	17.3	0.375	1.7	0.075	0.179	-9.7	0.179	0.188	9.1	10.7		
	Clayton	-26.9	0.234	-8.4	0.173	-0.2	0.047	0.054	-0.5	0.054	0.123	12.3	8.2		
	Joe	103.9	0.819	47.8	0.744	-15.5	0.162	0.519	-72.4	0.519	0.377	0.4	0.4		
	FGM	98.6	0.778	47.8	0.741	-11.4	0.119	0.432	-59.2	0.432	0.365	0.0	4.1		
	AMH	11.0	0.602	9.4	0.521	-0.7	0.050	0.351	-12.1	0.351	0.147	62.6	61.7		
	Normal	-23.8	0.214	-5.5	0.144	1.4	0.049	0.059	4.7	0.059	0.111	10.7	7.4		

Table 22: 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 and (10), 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 4.2 for further details.

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{\delta}_{21}(z_1)$			AIC (%)	BIC (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE		
$\tau = 0.2$	Normal	41.1	0.516	22.0	0.479	2.4	0.048	-79.7	0.240	0.237	7.3	5.7					
	Clayton	46.3	0.388	24.4	0.399	1.6	0.044	-86.2	0.184	0.241	68.8	75.7					
	Joe	10.2	0.321	8.1	0.302	-0.9	0.038	-39.7	0.140	0.178	8.9	4.9					
	FGM	44.5	0.403	23.4	0.405	1.3	0.038	-84.2	0.191	0.232	1.6	0.8					
	AMH	44.5	0.404	23.6	0.407	1.8	0.046	-82.2	0.191	0.236	1.2	2.4					
	Frank	41.0	0.411	21.9	0.404	1.4	0.040	-77.7	0.195	0.227	4.5	3.6					
	Gumbel	6.3	0.321	6.9	0.295	0.5	0.039	-27.2	0.140	0.171	7.7	6.9					
	Normal	17.2	0.380	12.5	0.359	-0.5	0.051	-15.3	0.185	0.171	11.6	11.6					
	Clayton	81.1	0.653	41.0	0.643	-2.7	0.059	-55.9	0.304	0.351	2.8	4.8					
	Joe	-2.9	0.224	1.7	0.209	-3.8	0.063	-11.8	0.112	0.118	22.0	12.8					
$\tau = 0.5$	FGM	76.0	0.603	36.7	0.576	-7.0	0.077	-58.8	0.301	0.323	2.4	9.2					
	AMH	75.3	0.600	37.8	0.590	-4.2	0.059	-51.6	0.271	0.332	1.6	9.2					
	Frank	33.3	0.404	19.1	0.397	-2.5	0.053	-24.1	0.194	0.215	15.2	19.6					
	Gumbel	-6.8	0.208	1.4	0.192	-0.8	0.051	-1.1	0.090	0.115	44.4	32.8					
	Normal	-6.5	0.118	2.6	0.131	-0.1	0.045	0.9	0.043	0.095	15.5	19.0					
	Clayton	7.5	0.176	10.7	0.235	1.7	0.054	-0.2	0.079	0.159	2.6	6.5					
	Joe	-8.6	0.122	0.2	0.120	-2.2	0.050	-3.9	0.056	0.089	9.9	11.2					
	FGM	115.5	0.880	52.4	0.791	-19.0	0.192	-72.2	0.578	0.422	0.0	0.0					
	AMH	109.8	0.838	52.8	0.799	-14.4	0.148	-59.0	0.472	0.426	0.0	0.0					
	Frank	-1.0	0.112	5.0	0.146	-1.2	0.049	-0.7	0.049	0.111	4.3	10.8					
	Gumbel	-9.0	0.124	1.1	0.123	-0.2	0.046	1.5	0.041	0.089	67.7	52.6					

Table 23: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\delta}_{21}(z_1)$, 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 and (10), 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 4.2 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 (%)
Joe	Normal	-55.0	1.115	-17.0	0.759	6.7	0.117	0.291	77.9	0.291	0.215	16.7	0.215	16.7	18.3
	Clayton	-11.2	0.456	0.9	0.332	4.5	0.109	0.153	31.6	0.153	0.134	48.8	0.134	48.8	46.3
	Joe	-94.8	1.131	-36.6	0.870	-0.7	0.106	0.193	87.3	0.193	0.197	4.9	0.197	4.9	4.5
	FGM	7.2	0.480	4.4	0.369	0.4	0.057	0.141	-42.1	0.141	0.137	8.1	0.137	8.1	6.5
	AMH	-13.0	0.510	0.2	0.360	5.3	0.100	0.172	40.1	0.172	0.140	2.0	0.140	2.0	3.3
Frank	Normal	-10.8	0.788	-1.4	0.531	3.3	0.102	0.232	-2.5	0.232	0.168	9.3	0.168	9.3	9.8
	Clayton	-121.3	1.377	-44.0	1.019	4.7	0.189	0.259	162.4	0.259	0.223	10.2	0.223	10.2	11.4
	Joe	13.7	0.989	6.1	0.697	0.4	0.100	0.304	-14.2	0.304	0.166	35.4	0.166	35.4	42.4
	FGM	114.5	1.003	46.7	0.786	-5.2	0.120	0.275	-41.6	0.275	0.209	17.7	0.209	17.7	16.5
	AMH	3.3	0.951	-3.0	0.696	-6.6	0.160	0.271	-28.7	0.271	0.193	4.1	0.193	4.1	2.5
Gumbel	Normal	157.9	1.299	56.6	0.922	-14.3	0.150	0.422	-77.9	0.422	0.246	4.1	0.246	4.1	3.7
	Clayton	136.6	1.118	52.8	0.849	-8.8	0.111	0.315	-54.9	0.315	0.227	2.1	0.227	2.1	2.9
	Joe	88.6	1.235	34.7	0.865	-3.3	0.127	0.410	-43.9	0.410	0.217	13.6	0.217	13.6	7.4
	FGM	-35.5	0.879	-14.6	0.661	0.4	0.148	0.209	-3.1	0.209	0.171	23.0	0.171	23.0	24.7
	AMH	-8.3	0.614	-1.5	0.455	-0.2	0.089	0.192	-1.8	0.192	0.115	48.3	0.115	48.3	54.7
Gumbel	Normal	80.2	0.779	38.6	0.688	-3.3	0.111	0.160	-10.8	0.160	0.170	10.6	0.170	10.6	8.9
	Clayton	-37.1	0.603	-14.9	0.469	0.3	0.119	0.153	-7.2	0.153	0.136	2.1	0.136	2.1	3.4
	Joe	195.1	1.519	70.2	1.086	-25.6	0.260	0.523	-72.9	0.523	0.288	0.4	0.288	0.4	0.4
	FGM	185.0	1.432	71.0	1.089	-19.5	0.200	0.402	-55.8	0.402	0.293	0.0	0.293	0.0	0.8
	AMH	34.1	0.745	18.5	0.564	-0.9	0.104	0.257	-9.3	0.257	0.132	4.2	0.132	4.2	7.2
	Gumbel	-33.3	0.438	-12.0	0.348	1.3	0.111	0.092	2.2	0.092	0.125	34.3	0.125	34.3	24.6

Table 24: 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 and (11), 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 4.2 for further details.

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{\sigma}_{21}(z_1)$			AIC (%)	BIC (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE		
$\tau = 0.1$	Normal	-70.1	1.301	-23.8	0.896	5.6	0.122	80.9	0.332	0.238	19.4	25.0					
	Clayton	-47.5	0.618	-11.3	0.381	5.2	0.125	102.9	0.208	0.140	40.7	41.1					
	Joe	-89.1	1.225	-34.2	0.922	-1.4	0.131	66.6	0.217	0.199	6.0	2.0					
	FGM	-5.1	0.504	-1.3	0.368	-2.3	0.059	-46.7	0.155	0.134	7.7	2.8					
	AMH	-23.9	0.547	-4.9	0.367	2.6	0.088	35.7	0.184	0.138	4.0	3.6					
$\tau = 0.5$	Frank	-32.8	0.964	-10.1	0.638	2.4	0.122	10.7	0.272	0.187	9.7	9.7					
	Gumbel	-123.8	1.409	-45.2	1.030	2.8	0.143	155.7	0.284	0.231	12.5	15.7					
	Normal	36.3	1.410	5.6	0.956	-6.7	0.150	-57.0	0.527	0.262	15.5	17.6					
	Clayton	-27.1	0.519	-7.1	0.358	1.6	0.109	5.6	0.152	0.125	68.6	61.9					
	Joe	-7.7	1.224	-11.9	0.906	-10.4	0.213	-47.1	0.382	0.235	0.8	2.1					
$\tau = 0.7$	FGM	128.5	1.096	37.3	0.664	-19.3	0.198	-100.7	0.536	0.214	1.7	0.4					
	AMH	106.9	0.963	33.1	0.602	-14.9	0.163	-77.4	0.450	0.192	1.3	2.9					
	Frank	66.5	1.287	17.7	0.814	-8.6	0.180	-70.9	0.546	0.256	4.6	4.6					
	Gumbel	-60.2	1.241	-29.5	0.940	-4.0	0.188	-17.2	0.327	0.236	7.5	10.5					
	Normal	83.3	1.631	21.3	1.088	-10.4	0.187	-62.7	0.686	0.287	8.3	11.8					
$\tau = 0.9$	Clayton	-20.1	0.414	-5.1	0.321	0.6	0.086	2.0	0.090	0.111	76.9	76.0					
	Joe	-7.7	1.291	-12.2	0.960	-8.6	0.223	-38.5	0.444	0.254	0.0	0.0					
	FGM	188.1	1.509	57.8	0.933	-27.0	0.274	-100.3	0.731	0.302	0.0	0.0					
	AMH	175.5	1.417	56.5	0.905	-23.8	0.245	-85.6	0.646	0.292	0.0	0.0					
	Frank	111.6	1.594	32.8	1.007	-11.8	0.210	-73.6	0.732	0.302	2.6	4.4					
	Gumbel	-53.4	1.052	-26.5	0.801	-2.4	0.176	-14.2	0.314	0.204	12.2	7.9					

Table 25: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\sigma}_{21}(z_1)$, 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 and (11), 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 4.2 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 (%)
Joe	Normal	-2.6	0.665	8.1	0.518	10.1	0.127	0.169	36.3	0.169	0.159	3.3	0.159	3.3	5.0
	Clayton	88.1	0.779	39.8	0.692	6.2	0.096	0.149	-76.0	0.149	0.180	77.3	0.180	77.3	79.8
	Joe	-29.7	0.570	-7.8	0.447	0.9	0.081	0.098	20.5	0.098	0.143	7.9	0.143	7.9	5.8
	FGM	79.9	0.718	36.9	0.657	5.3	0.072	0.124	-65.3	0.124	0.168	1.2	0.168	1.2	0.8
	AMH	49.9	0.597	29.0	0.570	10.4	0.140	0.130	-5.7	0.130	0.154	1.2	0.154	1.2	0.4
	Frank	67.1	0.728	32.9	0.641	6.5	0.093	0.152	-45.0	0.152	0.170	1.2	0.170	1.2	1.7
Joe	Gumbel	-53.4	1.104	-14.8	0.851	7.4	0.276	0.151	60.0	0.151	0.162	7.9	0.162	7.9	6.6
	Normal	60.8	0.840	30.9	0.685	1.4	0.088	0.209	-10.4	0.209	0.166	0.8	0.166	0.8	0.0
	Clayton	227.1	1.756	89.7	1.369	-11.4	0.135	0.418	-80.2	0.418	0.364	12.7	0.364	12.7	16.7
	Joe	-11.7	0.297	-1.3	0.254	-0.3	0.074	0.058	2.3	0.058	0.106	68.6	0.106	68.6	59.2
	FGM	182.3	1.416	74.3	1.146	-11.6	0.126	0.299	-58.3	0.299	0.288	0.4	0.288	0.4	0.8
	AMH	193.8	1.500	79.3	1.217	-10.2	0.115	0.301	-58.4	0.301	0.313	0.0	0.313	0.0	0.8
Joe	Frank	103.2	0.959	48.8	0.836	-2.3	0.074	0.191	-16.1	0.191	0.186	2.9	0.186	2.9	4.9
	Gumbel	-19.2	0.408	-1.7	0.321	4.9	0.115	0.092	13.7	0.092	0.109	14.7	0.109	14.7	17.6
	Normal	30.3	0.836	16.9	0.625	0.2	0.086	0.282	-4.1	0.282	0.130	0.5	0.130	0.5	5.4
	Clayton	113.7	1.107	54.5	0.944	-4.2	0.164	0.276	-15.0	0.276	0.232	0.5	0.232	0.5	0.0
	Joe	-8.5	0.312	-0.5	0.268	-0.4	0.073	0.071	1.1	0.071	0.097	54.1	0.097	54.1	60.5
	FGM	228.4	1.755	85.4	1.303	-28.9	0.292	0.554	-72.9	0.554	0.345	0.0	0.345	0.0	0.0
Joe	AMH	228.1	1.752	88.5	1.347	-24.8	0.253	0.477	-62.1	0.477	0.362	0.0	0.362	0.0	0.0
	Frank	64.1	0.857	31.8	0.685	-3.4	0.092	0.279	-5.8	0.279	0.141	1.6	0.141	1.6	4.9
	Gumbel	1.2	0.246	6.3	0.262	1.7	0.077	0.063	6.5	0.063	0.103	43.2	0.103	43.2	29.2

Table 26: 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 and (11), 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 4.2 for further details.

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{s}_{21}(z_1)$		<i>AIC</i> (%)	<i>BIC</i> (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	RMSE	<i>AIC</i> (%)	<i>BIC</i> (%)		
$\tau = -0.2$	Normal	-197.3	1.994	-63.9	1.312	7.9	0.141	-204.7	0.549	0.357	24.2	26.2				
	Clayton	-165.0	1.333	-47.8	0.781	5.0	0.119	-209.5	0.451	0.255	32.0	32.4				
	Joe	-244.6	2.151	-85.5	1.511	0.9	0.141	-217.4	0.493	0.368	3.7	2.5				
	FGM	-110.7	1.012	-34.5	0.653	-2.0	0.060	-115.3	0.292	0.209	5.7	7.4				
	AMH	-127.9	1.125	-37.4	0.682	2.1	0.086	-156.5	0.371	0.223	5.7	5.3				
	Frank	-129.7	1.463	-40.2	0.922	4.2	0.127	-136.0	0.434	0.268	19.7	13.1				
$\tau = 0.15$	Gumbel	-285.0	2.387	-97.7	1.635	6.0	0.164	-275.3	0.601	0.442	9.0	13.1				
	Normal	-23.4	1.271	-8.6	0.881	6.8	0.124	-7.4	0.340	0.228	18.8	14.3				
	Clayton	4.3	0.396	4.0	0.304	2.3	0.094	-17.4	0.144	0.126	36.7	40.8				
	Joe	-90.0	1.117	-36.9	0.885	-1.5	0.100	36.4	0.189	0.200	3.3	3.3				
	FGM	6.0	0.490	2.3	0.376	-0.5	0.059	-39.8	0.156	0.139	8.6	7.3				
	AMH	7.3	0.531	4.5	0.396	2.5	0.078	-21.7	0.169	0.142	3.7	3.3				
$\tau = 0.15$	Frank	-9.2	0.889	-2.6	0.599	3.7	0.110	-15.1	0.271	0.184	16.3	14.7				
	Gumbel	-123.6	1.353	-47.0	1.025	4.2	0.146	97.2	0.254	0.229	12.7	16.3				

Table 27: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ 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 and (11), 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 4.2 for further details.

	$\hat{\alpha}_{21}$		$\hat{\alpha}_{22}$		$\hat{\sigma}$		$\hat{\tau}$		$\hat{s}_{21}(z_1)$		BIC (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	AIC (%)	BIC (%)	
0.12	Normal	-163.3	1.800	-52.8	1.200	7.1	0.127	-231.2	0.479	0.315	20.9
	Clayton	-133.0	1.107	-38.2	0.659	3.8	0.106	-231.9	0.362	0.214	34.4
	Joe	-213.8	1.977	-74.9	1.404	1.1	0.148	-254.0	0.423	0.322	5.3
	FGM	-93.5	0.903	-28.7	0.592	-2.0	0.059	-130.3	0.245	0.188	4.9
	AMH	-115.3	1.053	-32.9	0.640	3.3	0.087	-198.4	0.340	0.207	10.7
	Frank	-121.3	1.387	-37.4	0.877	4.1	0.127	-174.8	0.398	0.253	9.0
0.1	Gumbel	-246.1	2.154	-84.6	1.489	5.5	0.158	-319.8	0.512	0.361	14.8
	Normal	-42.4	1.230	-13.6	0.848	6.2	0.117	37.5	0.321	0.221	17.1
	Clayton	-12.5	0.422	-0.7	0.321	2.2	0.088	13.1	0.138	0.124	44.5
	Joe	-101.5	1.207	-38.8	0.927	-1.0	0.107	89.9	0.205	0.203	2.4
	FGM	-8.6	0.508	-1.3	0.385	-0.4	0.053	-21.2	0.140	0.140	3.7
	AMH	-16.5	0.538	-1.8	0.379	3.5	0.084	28.4	0.171	0.139	5.3
0.28	Frank	-19.4	0.913	-4.9	0.607	3.6	0.108	4.8	0.264	0.188	12.7
	Gumbel	-129.0	1.344	-46.8	0.982	3.4	0.128	170.1	0.269	0.225	14.3
	Normal	-1.5	1.358	-4.5	0.953	-0.6	0.108	-44.0	0.415	0.243	23.7
	Clayton	-6.6	0.474	-0.3	0.337	-0.8	0.104	-7.9	0.176	0.123	30.7
	Joe	-51.7	1.188	-25.3	0.922	-6.1	0.162	-27.5	0.264	0.207	2.5
	FGM	62.2	0.727	17.6	0.485	-10.4	0.115	-89.6	0.321	0.158	2.1
0.5	AMH	38.3	0.638	12.6	0.441	-5.4	0.091	-53.3	0.259	0.141	5.0
	Frank	21.0	1.147	4.5	0.755	-2.2	0.129	-58.0	0.399	0.223	16.2
	Gumbel	-96.2	1.307	-39.8	0.992	-0.9	0.159	12.0	0.273	0.226	19.9

Table 28: 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 and (11), 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 4.2 for further details.

	$\hat{\alpha}_{21}$			$\hat{\alpha}_{22}$			$\hat{\sigma}$			$\hat{\tau}$			$\hat{\delta}_{21}(z_1)$			<i>AIC</i> (%)	<i>BIC</i> (%)
	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	Bias (%)	RMSE	RMSE	RMSE	RMSE		
$\tau = 0.2$	Normal	0.8	1.286	-0.2	0.886	6.6	0.122	-34.5	0.350	0.229	18.1	11.7					
	Clayton	31.0	0.478	12.4	0.365	0.9	0.087	-50.6	0.172	0.144	46.4	50.0					
	Joe	-76.5	1.011	-32.4	0.803	-2.1	0.086	7.5	0.170	0.187	3.6	2.8					
	FGM	21.9	0.508	8.0	0.387	-0.8	0.057	-49.2	0.165	0.140	4.4	3.6					
	AMH	31.2	0.581	12.6	0.428	1.8	0.073	-47.2	0.189	0.152	1.2	0.8					
	Frank	-7.5	0.868	-1.3	0.586	3.7	0.110	-13.5	0.248	0.175	14.1	15.7					
$\tau = 0.5$	Gumbel	-98.7	1.108	-38.0	0.832	2.7	0.104	46.0	0.212	0.200	12.1	15.3					
	Normal	82.2	1.665	25.3	1.130	-0.3	0.096	-59.3	0.560	0.270	22.0	14.7					
	Clayton	104.1	0.931	38.3	0.668	-7.7	0.128	-51.6	0.320	0.224	10.2	15.9					
	Joe	-52.6	1.026	-26.5	0.804	-3.2	0.136	-16.5	0.240	0.208	4.9	8.2					
	FGM	138.1	1.187	45.4	0.787	-14.0	0.149	-82.1	0.455	0.249	2.4	2.0					
	AMH	138.3	1.181	47.7	0.810	-11.2	0.125	-72.9	0.417	0.258	2.9	1.6					
$\tau = 0.7$	Frank	80.1	1.402	26.5	0.916	-1.1	0.100	-52.7	0.510	0.242	45.7	35.9					
	Gumbel	-80.2	0.908	-34.0	0.714	1.7	0.114	5.8	0.175	0.170	11.8	21.6					
	Normal	89.5	1.705	27.4	1.153	-5.3	0.122	-49.9	0.649	0.262	10.1	11.8					
	Clayton	58.0	0.729	26.7	0.577	-2.9	0.129	-13.0	0.211	0.175	23.2	21.5					
	Joe	-46.9	0.943	-24.1	0.744	-2.0	0.140	-14.2	0.260	0.182	4.8	5.7					
	FGM	202.1	1.620	66.8	1.066	-25.8	0.262	-89.7	0.661	0.333	0.0	0.4					
$\tau = 1$	AMH	198.2	1.588	68.6	1.083	-22.3	0.229	-78.2	0.595	0.341	0.9	1.8					
	Frank	108.6	1.608	36.0	1.055	-6.6	0.143	-54.4	0.662	0.265	37.7	35.5					
$\tau = 1$	Gumbel	-52.8	0.759	-23.7	0.597	0.6	0.120	-2.1	0.185	0.141	23.2	23.2					

Table 29: Percentage biases and RMSEs for $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\sigma}$, $\hat{\tau}$ and $\hat{\delta}_{21}(z_1)$, 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 and (11), 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 4.2 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 (%)
Joe	Normal	11.7	0.917	10.5	0.662		8.7	0.130		-0.6	0.227		0.172	10.1	10.5
	Clayton	87.7	0.790	38.1	0.660		4.5	0.098		-67.0	0.181		0.185	68.8	69.6
	Joe	-11.9	0.589	-3.6	0.468		-1.5	0.079		-13.2	0.112		0.148	6.5	4.9
	FGM	85.4	0.765	36.4	0.637		2.6	0.062		-69.3	0.170		0.177	0.4	0.4
	AMH	69.2	0.732	32.9	0.620		6.9	0.109		-38.0	0.164		0.168	0.8	0.4
	Frank	66.7	0.816	30.5	0.643		4.7	0.103		-47.2	0.196		0.178	3.2	4.5
Joe	Gumbel	-42.7	1.016	-12.9	0.793		4.5	0.222		25.3	0.152		0.175	10.1	9.7
	Normal	52.3	0.944	24.7	0.707		0.8	0.098		-16.8	0.256		0.163	14.0	11.5
	Clayton	196.7	1.544	76.8	1.185		-10.0	0.134		-74.8	0.399		0.312	18.1	20.2
	Joe	6.6	0.514	2.0	0.393		-5.6	0.118		-16.6	0.150		0.134	25.1	16.9
	FGM	168.3	1.315	66.2	1.026		-11.6	0.127		-63.1	0.327		0.259	0.8	2.5
	AMH	173.7	1.357	69.7	1.078		-9.2	0.112		-59.2	0.312		0.275	0.4	1.2
Joe	Frank	109.4	1.085	47.6	0.852		-3.6	0.103		-32.2	0.278		0.195	6.2	6.6
	Gumbel	-18.8	0.615	-4.5	0.485		2.3	0.159		2.6	0.109		0.129	35.4	41.2
	Normal	19.1	0.786	10.5	0.571		-1.0	0.087		-6.0	0.280		0.123	9.1	24.4
	Clayton	66.3	0.670	36.2	0.646		-1.1	0.100		-2.6	0.119		0.154	0.5	0.5
	Joe	-11.8	0.334	-2.2	0.270		-0.4	0.077		-4.5	0.095		0.100	11.5	17.2
	FGM	226.6	1.746	82.9	1.264		-30.9	0.311		-75.1	0.609		0.332	0.0	0.0
Joe	AMH	224.0	1.726	86.1	1.311		-25.7	0.261		-62.3	0.511		0.351	0.0	0.0
	Frank	50.7	0.810	26.0	0.623		-2.4	0.099		-8.9	0.290		0.133	0.5	2.9
	Gumbel	-3.8	0.326	3.0	0.316		1.6	0.097		1.8	0.041		0.106	78.5	55.0

Table 30: 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 and (11), 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 4.2 for further details.

Affiliation:

Małgorzata Wojtyś
Centre for Mathematical Sciences
Plymouth University
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

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>