# Copula regression spline models for binary outcomes

Rosalba Radice*
Department of Economics, Mathematics and Statistics
Birkbeck, London, U.K.
Malet Street, London WC1E 7HX, U.K.

Giampiero Marra
Department of Statistical Science
University College London
Gower Street, London WC1E 6BT, U.K.

Małgorzata Wojtyś
School of Computing, Electronics and Mathematics
University of Plymouth
Drake Circus, Plymouth PL4 8AA, U.K.

May 27, 2015

**Abstract**

We introduce a framework for estimating the effect that a binary treatment has on a binary outcome in the presence of unobserved confounding. The methodology is applied to a case study which uses data from the Medical Expenditure Panel Survey and whose aim is to estimate the effect of private health insurance on health care utilization. Unobserved confounding arises when variables which are associated with both treatment and outcome are not available (in economics this issue is known as endogeneity). Also, treatment and outcome may exhibit a dependence which cannot be modeled using a linear measure of association, and observed confounders may have a non-linear impact on the treatment and outcome variables. The problem of unobserved confounding is addressed using a two-equation structural latent variable framework, where one equation essentially describes a binary outcome as a function of a binary treatment whereas the other equation determines whether the treatment is received. Non-linear dependence between treatment and outcome is dealt with by using copula functions, whereas covariate-response relationships are flexibly modeled using a spline approach. Related model fitting and inferential procedures are developed, and asymptotic arguments presented.

**Key Words:** Bivariate binary outcomes; Copula; Endogeneity; Penalized regression spline; Simultaneous equation estimation; Unobserved confounding.

## 1 Introduction

Quantifying the effect of a non-randomly assigned treatment on an outcome is a challenging task in observational studies. An approach to calculate such an effect is to match subjects on the

*r.radice.bbk@ac.uk

basis of observed features or the so-called propensity score, and then compute the treatment effect as the difference between the observed responses of the matched subjects corresponding to the levels of the treatment (e.g., Heckman et al., 1997; Rosenbaum & Rubin, 1983). However, this method is only valid when the unobserved variables that influence the treatment are independent of the outcome, conditional on the covariates in the model. We consider the situation in which the researcher is interested in estimating the effect of a binary treatment on a binary outcome in the presence of unobserved confounders (i.e., unknown or not readily quantifiable variables associated with both treatment and outcome). In economics, this problem is commonly framed in terms of a regression model from which important regressors have been omitted and hence become a part of the model's error term. In this context, the treatment is termed exogenous if it is not associated with the error term after conditioning on the observed confounders, and endogenous otherwise. We address this issue by specifying a simultaneous model for treatment and outcome; this route has been previously taken by several scholars (e.g., Chib & Hamilton, 2002; Greene, 2012; Heckman, 1978; Maddala, 1983; Marra & Radice, 2011a). Other approaches are available to account for unobserved confounding; see the detailed review of Clarke & Windmeijer (2012).

To fix ideas, let us consider a case study which uses data from the Medical Expenditure Panel Survey (MEPS) and whose goal is to estimate the effect of having private health insurance on the probability of using health care services. Private health insurance status, which is an important determinant of the use of health care services, is a potentially endogenous variable. This is because unobserved variables, such as allergy and risk aversiveness, are likely to influence both health service utilization and private insurance decision. Sometimes the effect of private health insurance can be interpreted as adverse selection or moral hazard (e.g., Buchmueller et al., 2005). Adverse selection occurs when individuals with a greater demand for medical care, because of poor health for instance, are expected to have a greater demand for insurance. Moral hazard refers to the tendency of people to be more inclined to seek health services, and doctors to be more inclined to refer them when all costs are covered. The matter is further complicated by the fact that the effects of observed confounders, such as age and education, may be complex since they embody productivity and life-cycle effects that are likely to influence private health insurance and health care utilization non-linearly. If these relationships are mismodeled then the effect of insurance on the probability of using health care services may be biased (e.g., Marra & Radice, 2011a). Moreover, insurance status and health care utilization may exhibit a non-Gaussian association (Winkelmann, 2012).

Unobserved confounding can be controlled for by using the recursive bivariate probit model (Heckman, 1978). This model controls for unobserved confounding by using a two-equation structural latent variable framework, where one equation essentially describes a binary outcome (e.g., health care utilization) as a function of a binary treatment (e.g., insurance coverage) whereas the other equation determines whether the treatment is received. The model is completed by assuming that the latent errors of the two equations follow a standard bivariate Gaussian distribution with correlation $\theta$; $\theta \neq 0$ suggests that unobserved confounding is present, hence joint estimation of the two equations is required. Some applications in economics and bio-statistics are provided

by Goldman et al. (2001), Jones et al. (2006), Gitto et al. (2006), Latif (2009), Kawatkar & Nichol (2009) and Li & Jensen (2011). The limitations of this model are, however, the inability to deal effectively with non-linear covariate effects and non-Gaussian dependencies between the treatment and outcome equations. To model flexibly covariate-response relationships, Chib & Greenberg (2007) and Marra & Radice (2011a) introduced Bayesian and likelihood estimation methods based on penalized splines, respectively. To deal with the problem of non-Gaussian dependence between treatment and outcome, Winkelmann (2012) discussed a modification of the recursive bivariate probit that maintains the Gaussian assumption for the marginal distributions of the two equations while introducing non-Gaussian dependence between them using the Frank and Clayton copulas.

The contribution of this article is twofold, one methodological and the other practical. First, we extend the procedures discussed in Marra & Radice (2011a) and Winkelmann (2012) to make it possible to deal simultaneously with unobserved confounding, non-linear covariate effects and non-Gaussian dependencies between treatment and outcome. In particular, we generalize the penalized likelihood estimation approach based on the assumption of bivariate normality presented in Marra & Radice (2011a) by allowing for non-Gaussian dependencies between the two model equations; this is achieved by employing some classic copulas, such as Clayton, Frank, Gumbel and Joe, and the rotated versions of Clayton, Gumbel and Joe. We also provide some theoretical argumentation related to the asymptotic behavior of the proposed estimator and the ensuing formula to calculate the treatment effect. Second, we implement the methods discussed in this article in the `R` package `SemiParBIVProbit` (Marra & Radice, 2015). This can be particularly attractive to practitioners who wish to fit such models. Swihart et al. (2014) and Genest et al. (2013) have also adopted the copula paradigm to model multiple binary outcomes. One of the main contributions of the former article is to establish the connection between existing marginalized multilevel models and copulas. The work by Genest et al. (2013) discusses models for vectors of binary outcomes in the which the marginal distributions depend on covariates through logistic regressions and the dependence structure is modeled through meta-elliptical copulas. Our approach does not deal with multivariate binary outcomes, although it can be extended to this context. However, as opposed to Swihart et al. (2014) and Genest et al. (2013), the proposed methodology can account for non-linear covariate effects, and more importantly can mitigate the issue of endogeneity.

The rest of the paper is organized as follows. Section 2 mainly discusses the model structure, parameter estimation, confidence intervals and variable selection. Section 3 applies the proposed methodology to the MEPS data mentioned above, whereas Section 4 discusses the limitations of the proposed framework and concludes with some future extensions. The online supplementary material includes some of the details required to calculate the asymptotic variance of the treatment effect, details on the structure of the score vector and Hessian matrix used in the algorithm, asymptotic considerations related to the proposed estimator and the ensuing formula to calculate the treatment effect, and the results of a simulation study.

# 2 Methods

## 2.1 Model definition

The focus is on a pair of random variables $(y_{1i}, y_{2i})$, for $i = 1, \ldots, n$, where $y_{vi} \in \{0, 1\}$, $v$ can take values 1 and 2, and $n$ represents the sample size. Variable $y_{1i}$ refers to the treatment and $y_{2i}$ to the outcome. The observed $y_{vi}$ is determined by a latent continuous variable $y^*_{vi}$ such that $y_{vi} = \mathbf{1}(y^*_{vi} > 0)$, where **1** is the classic indicator function. We assume that $y^*_{vi} \sim \mathcal{N}(\eta_{vi}, 1)$ where $\eta_{vi} \in \mathbb{R}$ is a linear predictor defined in the next section for $v = 1, 2$. The probability of event $(y_{1i} = 1, y_{2i} = 1)$ can be defined by using the copula representation (Sklar, 1959, 1973)

$$\mathbb{P}(y_{1i} = 1, y_{2i} = 1) = \mathcal{C}(\mathbb{P}(y_{1i} = 1), \mathbb{P}(y_{2i} = 1); \theta),$$

where $\mathbb{P}(y_{vi} = 1) = \Phi(\eta_{vi})$, $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard univariate Gaussian distribution, $\mathcal{C}$ is a two-place copula function and $\theta$ is an association parameter measuring the dependence between the two marginals $\mathbb{P}(y_{1i} = 1)$ and $\mathbb{P}(y_{2i} = 1)$. In other words, the joint distribution is expressed in terms of marginal distributions and a function $\mathcal{C}$ that binds them together. A substantial advantage of the copula approach is that the marginal distributions may come from different families. Note that the marginal cdfs are conditioned on covariates (see the definition of $\eta_{vi}$ in the next section), but for notational convenience we have suppressed this when expressing the marginal distributions. Some of the copulas considered are Clayton, Frank, Gaussian, Gumbel, and Joe as well as the rotated versions of Clayton, Gumbel and Joe. Rotation by 180 degrees leads to the survival copula ($\mathcal{C}_{180}$), while rotation by 90 ($\mathcal{C}_{90}$) and 270 degrees ($\mathcal{C}_{270}$) allows for negative dependence which is not possible with the non-rotated and survival versions. The copulas considered here are displayed in Figure 1. The counter-clockwise rotated versions can be obtained using (e.g., Brechmann & Schepsmeier, 2013)

$$\begin{aligned}
\mathcal{C}_{90}(u_i, v_i) &= v_i - \mathcal{C}(1 - u_i, v_i), \\
\mathcal{C}_{180}(u_i, v_i) &= u_i + v_i - 1 + \mathcal{C}(1 - u_i, 1 - v_i), \\
\mathcal{C}_{270}(u_i, v_i) &= u_i - \mathcal{C}(u_i, 1 - v_i),
\end{aligned}$$

where $u_i = \mathbb{P}(y_{1i} = 1)$ and $v_i = \mathbb{P}(y_{2i} = 1)$. The ranges of $\theta$ for the copulas rotated by 90 and 270 degrees are on a negative scale; e.g., for Gumbel rotated by 90 and 270 degrees $\theta$ has to be smaller than $-1$. For full details on copulas and their properties see, for instance, Nelsen (2006).

The log-likelihood function for the recursive bivariate probit model can be expressed as

$$\ell = \sum_{i=1}^{n} \left\{ y_{1i} y_{2i} \log p_{11i} + y_{1i}(1 - y_{2i}) \log p_{10i} + (1 - y_{1i}) y_{2i} \log p_{01i} + (1 - y_{1i})(1 - y_{2i}) \log p_{00i} \right\},$$

where $p_{11} = \mathbb{P}(y_{1i} = 1, y_{2i} = 1)$, $p_{10i} = \mathbb{P}(y_{1i} = 1, y_{2i} = 0) = \mathbb{P}(y_{1i} = 1) - \mathbb{P}(y_{1i} = 1, y_{2i} = 1)$, $p_{01i} = \mathbb{P}(y_{1i} = 0, y_{2i} = 1) = \mathbb{P}(y_{2i} = 1) - \mathbb{P}(y_{1i} = 1, y_{2i} = 1)$ and $p_{00i} = \mathbb{P}(y_{1i} = 0, y_{2i} = 0) = 1 - [\mathbb{P}(y_{1i} = 1) + \mathbb{P}(y_{2i} = 1) - \mathbb{P}(y_{1i} = 1, y_{2i} = 1)]$.

As it can be seen from Table 1, $\theta$ may be difficult to interpret in some cases. To this end, the well known Kendall's $\tau \in [-1, 1]$ can be utilized. Alternatively, Tajar et al. (2001) suggest using the odds ratio and gamma measure proposed by Goodman & Kruskal (1954). These can be defined as $\zeta = p_{00}p_{11}/p_{10}p_{01}$ and $\gamma = \zeta - 1/\zeta + 1$, respectively. The odds ratio has range $\mathbb{R}$ whereas $\gamma \in [-1, 1]$.

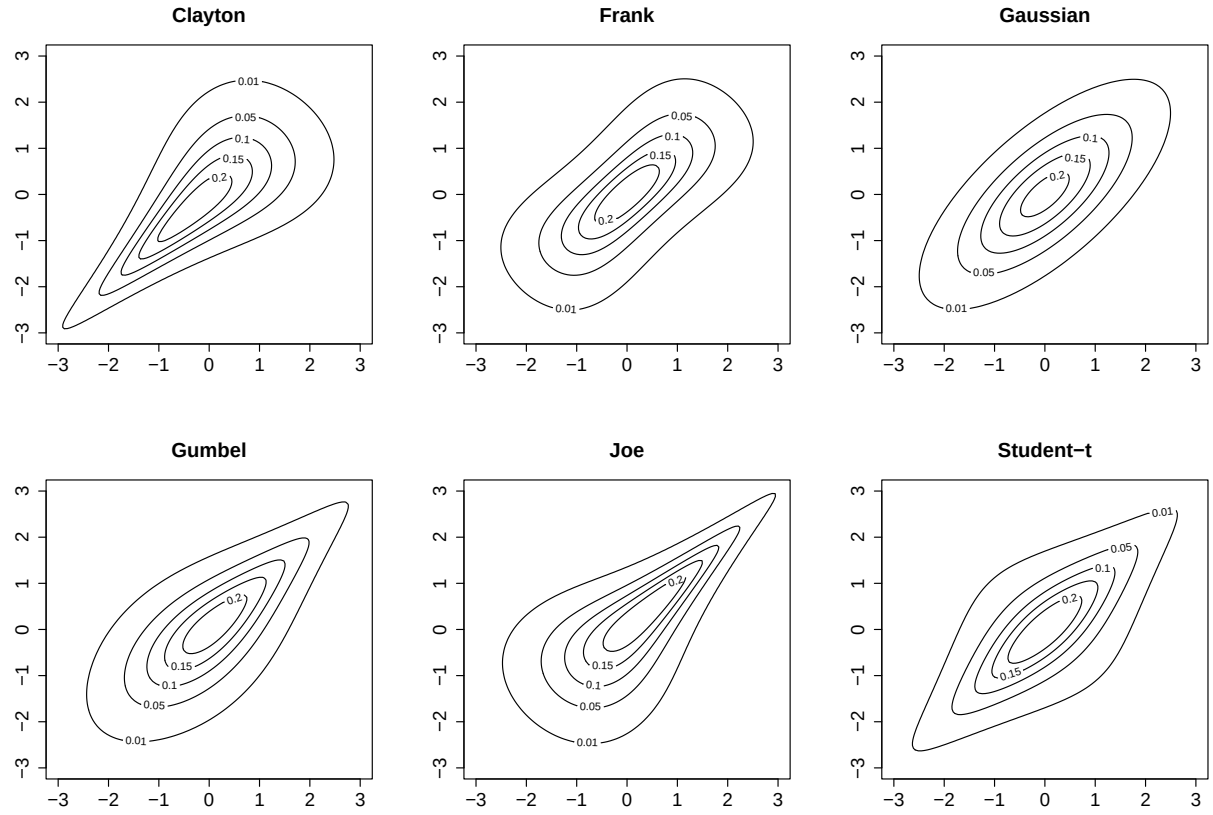


Figure 1: Contour plots of some classic copula functions with standard normal margins for data simulated using association parameters 2, 5.74, 0.71, 2, 2.86, and 0.71, respectively (these values are consistent with a medium positive correlation). The Gaussian, Student-t (here with three degrees of freedom) and Frank copulas allow for equal degrees of positive and negative dependence. Gaussian and Frank show a weaker tail dependence as compared to Student-t, and Frank exhibits a slightly stronger dependence in the middle of the distribution. Clayton is asymmetric with a strong lower tail dependence but a weaker upper tail dependence. Vice versa for the Gumbel and Joe copulas.

### 2.1.1 Linear predictor specification

The linear predictor for the treatment equation can be written as

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

whereas that for the outcome as

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

| Copula | Range of $\theta$ | $\theta_*$ |
|---|---|---|
| Clayton | $\theta \in (0, \infty)$ | $\log(\theta - \epsilon)$ |
| Frank | $\theta \in \mathbb{R} \setminus \{0\}$ | $\theta - \epsilon$ |
| Gaussian/Student-t | $\theta \in [-1, 1]$ | $\tanh^{-1}(\theta)$ |
| Gumbel | $\theta \in [1, \infty)$ | $\log(\theta - 1)$ |
| Joe | $\theta \in (1, \infty)$ | $\log(\theta - 1 - \epsilon)$ |

Table 1: Parameter range of dependence coefficient $\theta$ for some classic copula functions and transformations, $\theta_*$, of $\theta$ used in optimization. Quantity $\epsilon$ is set to the machine smallest positive floating-point number multiplied by $10^6$ and is used in some cases to ensure that the dependence parameters lie in their respective ranges.

where $\psi$ is the effect of the treatment on the outcome on the scale of the linear predictor, $\mathbf{u}_{1i}^\mathsf{T} = (1, u_{12i}, \ldots, u_{1P_1 i})$ is the $i^{th}$ row of $\mathbf{U}_1 = (\mathbf{u}_{11}, \ldots, \mathbf{u}_{1n})^\mathsf{T}$, the $n \times P_1$ model matrix containing $P_1$ parametric terms (e.g., intercept, dummy and categorical variables), $\boldsymbol{\alpha}_1$ is a coefficient vector, and the $s_{1k_1}$ are unknown smooth functions of the $K_1$ continuous covariates $z_{1k_1 i}$. Varying coefficient models can be obtained by multiplying one or more smooth terms by some predictor(s) (Hastie & Tibshirani, 1993), and smooth functions of two or more covariates can also be considered (Wood, 2006). Similarly, $\mathbf{u}_{2i}^\mathsf{T} = (1, u_{22i}, \ldots, u_{2P_2 i})$ is the $i^{th}$ row vector of the $n \times P_2$ model matrix $\mathbf{U}_2 = (\mathbf{u}_{21}, \ldots, \mathbf{u}_{2n})^\mathsf{T}$, $\boldsymbol{\alpha}_2$ is a parameter vector, and the $s_{2k_2}$ are unknown smooth terms of the $K_2$ continuous regressors $z_{2k_2 i}$. The smooth functions are subject to the centering (identifiability) constraint $\sum_{i=1}^n s_{vk_v}(z_{vk_v i}) = 0$ for $v = 1, 2$, $k_v = 1, \ldots, K_v$ (Wood, 2006).

The smooth functions are represented using the regression spline approach (e.g., Ruppert et al., 2003). Specifically, $s_{vk_v}(z_{vk_v i})$ is approximated by a linear combination of known spline basis functions, $b_{vk_v j}(z_{vk_v i})$, and regression parameters, $\beta_{vk_v j}$, i.e. $s_{vk_v}(z_{vk_v i}) = \sum_{j=1}^{J_{vk_v}} \beta_{vk_v j} b_{vk_v j}(z_{vk_v i}) = \boldsymbol{\beta}_{vk_v}^\mathsf{T} \mathbf{B}_{vk_v}(z_{vk_v i})$, where $J_{vk_v}$ is the number of spline bases used to represent $s_{vk_v}(\cdot)$, $\mathbf{B}_{vk_v}(z_{vk_v i})$ is the $i^{th}$ vector of dimension $J_{vk_v}$ containing the basis functions evaluated at the observation $z_{vk_v i}$, i.e. $\mathbf{B}_{vk_v}(z_{vk_v i}) = \left\{b_{vk_v 1}(z_{vk_v i}), b_{vk_v 2}(z_{vk_v i}), \ldots, b_{vk_v J_{vk_v}}(z_{vk_v i})\right\}^\mathsf{T}$, and $\boldsymbol{\beta}_{vk_v}$ is the corresponding parameter vector. Evaluating $\mathbf{B}_{vk_v}(z_{vk_v i})$ for each $i$ yields $J_{vk_v}$ curves with different degrees of complexity which multiplied by some value of $\boldsymbol{\beta}_{vk_v}$ and then summed will give a (linear or non-linear) estimate for $s_{vk_v}(z_{vk_v})$; see Ruppert et al. (2003) for a detailed overview. Basis functions should be chosen to have convenient mathematical and numerical properties. We employ low rank thin plate regression splines (Wood, 2003), although many spline definitions (including B-splines and cubic regression splines) are supported in our implementation. Note that for one-dimensional smooth functions, the choice of spline definition does not play a crucial role in determining the shape of $\hat{s}_{vk_v}(z_{vk_v i})$ (Wood, 2006). The cases of smooth terms multiplied by some covariate(s) and of smooths of more than one variable follow a similar construction; see Wood (2006, Chapter 4) for full details. Linear predictors (1) and (2) can, therefore, be written as $\eta_{1i} = \mathbf{u}_{1i}^\mathsf{T}\boldsymbol{\alpha}_1 + \mathbf{B}_{1i}^\mathsf{T}\boldsymbol{\beta}_1$ and $\eta_{2i} = \psi y_{1i} + \mathbf{u}_{2i}^\mathsf{T}\boldsymbol{\alpha}_2 + \mathbf{B}_{2i}^\mathsf{T}\boldsymbol{\beta}_2$, where $\mathbf{B}_{vi}^\mathsf{T} = \left\{\mathbf{B}_{v1}(z_{v1i})^\mathsf{T}, \ldots, \mathbf{B}_{vK_v}(z_{vK_v i})^\mathsf{T}\right\}$ and $\boldsymbol{\beta}_v^\mathsf{T} = (\boldsymbol{\beta}_{v1}^\mathsf{T}, \ldots, \boldsymbol{\beta}_{vK_v}^\mathsf{T})$. After defining $\mathbf{X}_{1i} = (\mathbf{u}_{1i}^\mathsf{T}, \mathbf{B}_{1i}^\mathsf{T})^\mathsf{T}$ and $\mathbf{X}_{2i} = (y_{1i}, \mathbf{u}_{2i}^\mathsf{T}, \mathbf{B}_{2i}^\mathsf{T})^\mathsf{T}$, we have $\eta_{1i} = \mathbf{X}_{1i}^\mathsf{T}\boldsymbol{\delta}_1$ and $\eta_{2i} = \mathbf{X}_{2i}^\mathsf{T}\boldsymbol{\delta}_2$ where $\boldsymbol{\delta}_1^\mathsf{T} = (\boldsymbol{\alpha}_1^\mathsf{T}, \boldsymbol{\beta}_1^\mathsf{T})$ and $\boldsymbol{\delta}_2^\mathsf{T} = (\psi, \boldsymbol{\alpha}_2^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T})$. Note that the presence of a binary endogenous variable in $\eta_{2i}$ does not alter the log-likelihood function presented in the previous section; $\mathbb{P}(y_{1i}, y_{2i})$ can be written as $\mathbb{P}(y_{2i}|y_{1i})\mathbb{P}(y_{1i})$, hence its form does not change if

$\eta_{2i}$ includes $y_{1i}$.

To identify the parameters in $\eta_{2i}$, it is typically assumed that an exclusion restriction on the exogenous variables holds: the regressors in the treatment equation should contain at least one or more covariates (usually referred to as instruments) not included in the outcome equation. However, as shown for instance in Han & Vytlacil (2014), Marra & Radice (2011a) and Wilde (2000), the presence of this restriction may not be necessary.

## 2.2 Sample average treatment effect

The effect of $y_{1i}$ on the probability that $y_{2i} = 1$ is of primary interest. In other words, the aim is to investigate how the treatment changes the expected outcome. Thus, the treatment effect is given by the difference between the expected outcome with treatment and the expected outcome without treatment. Different measures of treatment effect have been proposed in the literature. Here, we focus on the average treatment effect in the specific sample at hand, rather than that in the population (SATE; Abadie et al., 2004). In our case, this can be defined as

$$\mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(y_{2i} = 1|y_{1i} = 1) - \mathbb{P}(y_{2i} = 1|y_{1i} = 0),$$

where

$$\mathbb{P}(y_{2i} = 1|y_{1i} = 1) = \frac{\mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=1)}); \theta\right)}{\Phi(\eta_{1i})},$$

$$\mathbb{P}(y_{2i} = 1|y_{1i} = 0) = \frac{\Phi(\eta_{2i}^{(y_{1i}=0)}) - \mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=0)}); \theta\right)}{1 - \Phi(\eta_{1i})},$$

the linear predictors are defined in the previous section, $\eta_{2i}^{(y_{1i}=r)}$ represents the linear predictor evaluated at $y_{1i} = r$ for $r$ equal to 1 or 0, $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \theta)$, and $\mathbf{X} = (\mathbf{x}_1|\ldots|\mathbf{x}_n)^\mathsf{T}$ where $\mathbf{x}_i$ is defined as $(\mathbf{X}_{1i}^\mathsf{T}, \mathbf{X}_{2i}^\mathsf{T})^\mathsf{T}$. $\mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})$ can be estimated using $\mathrm{SATE}(\hat{\boldsymbol{\delta}}, \mathbf{X})$, whereas a confidence interval for it can be obtained employing the delta method. Specifically, the appropriate estimator of the asymptotic variance of $\mathrm{SATE}(\hat{\boldsymbol{\delta}}, \mathbf{X})$ is

$$\left.\frac{\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})}{\partial \boldsymbol{\delta}}^\mathsf{T}\right|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} \mathbf{V}_{\boldsymbol{\delta}} \left.\frac{\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})}{\partial \boldsymbol{\delta}}\right|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}}, \tag{3}$$

where $\mathbf{V}_{\boldsymbol{\delta}}$ is the covariance matrix of $\boldsymbol{\delta}$ defined in Section 2.4 and

$$\frac{\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})}{\partial \boldsymbol{\delta}} = \left[\frac{\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})}{\partial \boldsymbol{\delta}_1}^\mathsf{T}, \frac{\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})}{\partial \boldsymbol{\delta}_2}^\mathsf{T}, \frac{\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})}{\partial \theta}\right]^\mathsf{T},$$

with elements defined in Section S.1 of the online supplementary material. Alternatively, Bayesian posterior simulation can be employed (see Section 2.4).

## 2.3 Parameter estimation

Since the range of $\theta$ is bounded in most cases, we use a proper transformation of it, $\theta_*$, and define $\boldsymbol{\delta}_*^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \theta_*)$, to ensure that in optimization $\boldsymbol{\delta}_* \in \mathbb{R}^p$, where $p$ is the total number of parameters; see Table 1 for ranges of $\theta$ and the transformations employed. Let us denote the log-likelihood for a given copula function as $\ell(\boldsymbol{\delta}_*)$. Given the flexible linear predictor structure considered here, unpenalized estimation can result in smooth term estimates that are too rough to produce practically useful results (e.g., Ruppert et al., 2003). This issue is dealt with by using a penalty term, such as $\sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\int\left\{s''_{vk_v}(z_{vk_v})\right\}^2 dz_{vk_v}$ for the one-dimensional case, which measures the second-order roughness of the smooth terms in the model. The $\lambda_{vk_v}$ are smoothing parameters controlling the trade-off between fit and smoothness and can take values in $[0,\infty)$. Since regression splines are linear in their model parameters, the overall penalty can be written as $\boldsymbol{\beta}^\mathsf{T}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$ where $\boldsymbol{\beta}^\mathsf{T} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T})$, $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\mathbf{S}_{vk_v}$ and the $\mathbf{S}_{vk_v}$ are positive semi-definite symmetric known square matrices expanded with zeros everywhere except for the elements which correspond to the coefficients of the $vk_v^{th}$ smooth term. The expressions for the $b_{vk_vj}(z_{vk_vi})$ and $\mathbf{S}_{vk_v}$ depend on the type of spline employed and we refer the reader to Ruppert et al. (2003) and Wood (2003, 2006) for these details. The function to maximize is

$$\ell_p(\boldsymbol{\delta}_*) = \ell(\boldsymbol{\delta}_*) - \frac{1}{2}\boldsymbol{\beta}^\mathsf{T}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}, \tag{4}$$

where the penalty term can be written as $\boldsymbol{\delta}_*^\mathsf{T}\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\delta}_*/2$ where $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$ is an overall penalty matrix defined as $\mathrm{diag}(\mathbf{0}_{P_1}^\mathsf{T}, \lambda_{1k_1}\mathbf{S}_{1k_1}, \ldots, \lambda_{1K_1}\mathbf{S}_{1K_1}, \mathbf{0}_{P_2}^\mathsf{T}, \lambda_{2k_2}\mathbf{S}_{2k_2}, \ldots, \lambda_{2K_2}\mathbf{S}_{2K_2}, 0)$ with $\mathbf{0}_{P_v}^\mathsf{T} = (0_{v1}, \ldots, 0_{vP_v})$.

### 2.3.1 Estimating $\delta_*$ given smoothing parameters

Given $\hat{\boldsymbol{\lambda}}^\mathsf{T} = (\hat{\lambda}_{1k_1}, \ldots, \hat{\lambda}_{1K_1}, \hat{\lambda}_{2k_2}, \ldots, \hat{\lambda}_{2K_2})$, we seek to maximize (4). To this end, we use a trust region approach which is generally more stable and faster than its line-search counterparts, particularly for functions that are, for example, non-concave and/or exhibit regions that are close to flat (Nocedal & Wright, 2006, Chapter 4). Let $a$ be an iteration index. Intuitively speaking, line search methods choose a direction to move from $m_a$ to $m_{a+1}$ and find the distance along that direction which gives the best improvement in the objective function. If the function is non-convex or has long plateaus then the optimizer may search far away from $m_a$ but still choose an $m_{a+1}$ that is close to $m_a$ (hence offering a 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 $m_{a+1}$ based on 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 shrunken if the proposed point in the region is worse/not better than the current point; the new problem with smaller region is then solved. If a point which is 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 $m_{a+1}$ is not finite or not better than the value at $m_a$ then the proposal is rejected and the trust region shrunken. Finally, a line search approach requires repeated estimation of the objective function, while trust region methods evaluate the objective function only after solving the trust region problem. Hence, trust region methods can be considerably faster when the objective function is expensive to compute. Full details can be found in (Nocedal & Wright, 2006, Chapter 4).

Let us define the penalized gradient and Hessian at iteration $a$ as $\mathbf{g}_p^{[a]} = \mathbf{g}^{[a]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}_*^{[a]}$ and $\boldsymbol{\mathcal{H}}_p^{[a]} = \boldsymbol{\mathcal{H}}^{[a]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}$, where $\mathbf{g}^{[a]}$ is made up of $\mathbf{g}_1^{[a]} = \partial\ell(\boldsymbol{\delta}_*)/\partial\boldsymbol{\delta}_1|_{\boldsymbol{\delta}_1=\boldsymbol{\delta}_1^{[a]}}$, $\mathbf{g}_2^{[a]} = \partial\ell(\boldsymbol{\delta}_*)/\partial\boldsymbol{\delta}_2|_{\boldsymbol{\delta}_2=\boldsymbol{\delta}_2^{[a]}}$ and $g_3^{[a]} = \partial\ell(\boldsymbol{\delta}_*)/\partial\theta_*|_{\theta_*=\theta_*^{[a]}}$, and the Hessian matrix has a $3\times 3$ matrix block structure with $(r,h)^{th}$ element $\boldsymbol{\mathcal{H}}_{r,h}^{[a]} = \partial^2\ell(\boldsymbol{\delta}_*)/\partial\boldsymbol{\delta}_r\partial\boldsymbol{\delta}_h^\mathsf{T}|_{\boldsymbol{\delta}_r=\boldsymbol{\delta}_r^{[a]},\boldsymbol{\delta}_h=\boldsymbol{\delta}_h^{[a]}}$, $r,h=1,\ldots,3$, where $\boldsymbol{\delta}_3=\theta_*$; details on the structure of $\mathbf{g}$ and $\boldsymbol{\mathcal{H}}$ can be found in Section S.2 of the online supplementary material. Each iteration of the trust region algorithm solves the problem

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

where $\|\cdot\|$ denotes the Euclidean norm and $r^{[a]}$ represents the radius of the trust region. At each iteration of the algorithm, $\breve{\ell}_p(\boldsymbol{\delta}_*^{[a]})$ is minimized subject to the constraint that the solution falls within a trust region with radius $r^{[a]}$. The proposed solution is then accepted or rejected and the trust region expanded or shrunken based on the ratio between the improvement in the objective function when going from $\boldsymbol{\delta}_*^{[a]}$ to $\boldsymbol{\delta}_*^{[a+1]}$ and that predicted by the quadratic approximation. The exact details of the implementation used here can be found in Geyer (2013) who also discusses numerical stability and termination criteria. Note that, near the solution, the trust region algorithm typically behaves as a classic unconstrained algorithm.

### 2.3.2 Estimating $\boldsymbol{\lambda}$ given $\boldsymbol{\delta}_*$

If the model has more than one smooth term per equation, then estimation of $\boldsymbol{\lambda}$ by direct grid search optimization of, for instance, a prediction error criterion can be computationally burdensome. It is therefore pivotal for practical modeling to estimate $\boldsymbol{\lambda}$ in an automatic way. There are many techniques for automatic multiple smoothing parameter estimation within the penalized likelihood framework; see Ruppert et al. (2003) and Wood (2006) for detailed overviews. (Note that joint estimation of $\boldsymbol{\delta}_*$ and $\boldsymbol{\lambda}$ via maximization of (4) would clearly lead to over-fitting since the highest value of $\ell_p(\boldsymbol{\delta}_*)$ would be obtained when $\hat{\boldsymbol{\lambda}}=\mathbf{0}$.)

Let us define $\tilde{\mathbf{X}} = \left(\tilde{\mathbf{X}}_1|\ldots|\tilde{\mathbf{X}}_n\right)^\mathsf{T}$, where $\tilde{\mathbf{X}}_i = \text{diag}\left\{\mathbf{X}_{1i}^\mathsf{T}, \mathbf{X}_{2i}^\mathsf{T}, 1\right\}$ with $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ defined in Section 2.1.1, $\mathbf{W}^{[a]}$ as a block diagonal matrix made up of $3\times 3$ matrices $\mathbf{W}_i^{[a]}$ with $(r,h)^{th}$ element given by $-\partial^2\ell(\boldsymbol{\delta}_*)_i/\partial\eta_{ri}\partial\eta_{hi}|_{\eta_{ri}=\eta_{ri}^{[a]},\eta_{hi}=\eta_{hi}^{[a]}}$, $r,h=1,2,3$, where $\eta_{3i}=\theta_*$, and $\mathbf{d}^{[a]}$ as a vector

with $i^{th}$ element given by $\mathbf{d}_i^{[a]} = \left\{\partial\ell(\boldsymbol{\delta}_*)_i/\partial\eta_{1i}|_{\eta_{1i}=\eta_{1i}^{[a]}}, \partial\ell(\boldsymbol{\delta}_*)_i/\partial\eta_{2i}|_{\eta_{2i}=\eta_{2i}^{[a]}}, \partial\ell(\boldsymbol{\delta}_*)_i/\partial\eta_{3i}|_{\eta_{3i}=\eta_{3i}^{[a]}}\right\}^\mathsf{T}$. We then have that $\mathbf{g}_p^{[a]} = \tilde{\mathbf{X}}^\mathsf{T}\mathbf{d}^{[a]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}_*^{[a]}$ and $\boldsymbol{\mathcal{H}}_p^{[a]} = -\tilde{\mathbf{X}}^\mathsf{T}\mathbf{W}^{[a]}\tilde{\mathbf{X}} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}$. Let us use the fact that close to convergence the trust region algorithm behaves as a classic unconstrained algorithm and assume that $\boldsymbol{\delta}_*^{[a+1]}$ is a new updated guess. Applying a first order Taylor expansion to $\mathbf{g}_p^{[a+1]}$ around $\boldsymbol{\delta}_*^{[a]}$, setting the resulting expression to zero, and using the expressions above for $\mathbf{g}_p^{[a]}$ and $\boldsymbol{\mathcal{H}}_p^{[a]}$, we find that

$$\boldsymbol{\delta}_*^{[a+1]} = (\tilde{\mathbf{X}}^\mathsf{T}\mathbf{W}^{[a]}\tilde{\mathbf{X}} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}})^{-1}\tilde{\mathbf{X}}^\mathsf{T}\mathbf{W}^{[a]}\mathbf{z}^{[a]},$$

where $\mathbf{z}^{[a]} = \left(\mathbf{W}^{[a]}\right)^{-1}\mathbf{d}^{[a]} + \tilde{\mathbf{X}}\boldsymbol{\delta}_*^{[a]}$. Thus $\boldsymbol{\delta}_*^{[a+1]}$ is clearly the solution to the penalized iteratively re-weighted least squares problem

$$\underset{\boldsymbol{\delta}_*}{\arg\min}\ \|\mathbf{z}^{+,[a]} - \tilde{\mathbf{X}}^{+,[a]}\boldsymbol{\delta}_*\|^2 + \boldsymbol{\delta}_*^\mathsf{T}\tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}_*,$$

where $\mathbf{z}^{+,[a]} = \sqrt{\mathbf{W}}^{[a]}\mathbf{z}^{[a]}$ and $\tilde{\mathbf{X}}^{+,[a]} = \sqrt{\mathbf{W}}^{[a]}\tilde{\mathbf{X}}$. In the derivation above, $\mathbf{W}^{[a]}$ can also be taken to be the expectation of minus the second derivatives of the log-likelihood with respect to the linear predictors.

From standard likelihood theory, $\boldsymbol{\epsilon} = \sqrt{\mathbf{W}}\mathbf{W}^{-1}\mathbf{d}$ has mean $\mathbf{0}$ and covariance (identity) matrix $\mathbb{I}$, and $\mathbf{z}^+ = \mathbb{E}(\mathbf{z}^+) + \boldsymbol{\epsilon}$, where $\mathbb{E}(\mathbf{z}^+) = \boldsymbol{\mu}_{\mathbf{z}^+} = \sqrt{\mathbf{W}}\tilde{\mathbf{X}}\boldsymbol{\delta}_*^0$, $\boldsymbol{\delta}_*^0$ is the true parameter vector and $\mathbb{V}(\mathbf{z}^+) = \mathbb{V}(\boldsymbol{\epsilon}) = \mathbb{I}$. The predicted vector value for $\mathbf{z}^+$ is given by $\hat{\boldsymbol{\mu}}_{\mathbf{z}^+} = \mathbf{A}_{\hat{\boldsymbol{\lambda}}}\mathbf{z}^+$, where $\mathbf{A}_{\boldsymbol{\lambda}} = \sqrt{\mathbf{W}}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\mathsf{T}\mathbf{W}\tilde{\mathbf{X}} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}})^{-1}\tilde{\mathbf{X}}^\mathsf{T}\sqrt{\mathbf{W}}$ (known as influence matrix). Following the argumentation in Wood (2006, Chapter 4), $\mathbf{z}^+$ will be normally distributed in the large sample limit. Now, the smoothing parameters have to be estimated and since the estimated smooth functions should be as close as possible to the respective true functions, it makes sense to estimate $\boldsymbol{\lambda}$ so that $\hat{\boldsymbol{\mu}}_{\mathbf{z}^+}$ is as close as possible to $\boldsymbol{\mu}_{\mathbf{z}^+}$. To this end, we employ the expected mean squared error of the model, which in this case is

$$\begin{aligned}
\mathbb{E}\left(\|\boldsymbol{\mu}_{\mathbf{z}^+} - \hat{\boldsymbol{\mu}}_{\mathbf{z}^+}\|^2/\check{n}\right) &= \mathbb{E}\left(\|\mathbf{z}^+ - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}^+ - \boldsymbol{\epsilon}\|^2\right)/\check{n} \\
&= \mathbb{E}\left(\|\mathbf{z}^+ - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}^+\|^2\right)/\check{n} + \mathbb{E}\left(-\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\mu}_{\mathbf{z}^+} + 2\boldsymbol{\epsilon}^\mathsf{T}\mathbf{A}_{\boldsymbol{\lambda}}\boldsymbol{\mu}_{\mathbf{z}^+} + 2\boldsymbol{\epsilon}^\mathsf{T}\mathbf{A}_{\boldsymbol{\lambda}}\boldsymbol{\epsilon}\right)/\check{n} \\
&= \mathbb{E}\left(\|\mathbf{z}^+ - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}^+\|^2\right)/\check{n} - 1 + 2\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}})/\check{n},
\end{aligned}$$

where $\check{n} = 3n$ and $\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}})$ represents the effective degrees of freedom ($edf$) of the penalized model. The smoothing parameter vector can be estimated by minimizing an estimate of the expectation above, that is

$$\mathcal{V}(\boldsymbol{\lambda}) = \|\mathbf{z}^+ - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}^+\|^2/\check{n} - 1 + 2\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}})/\check{n}. \tag{5}$$

This is equivalent to the expression of the Un-Biased Risk Estimator reported, for instance, in Wood (2006, Chapter 4) as well as to the Akaike information criterion ($AIC$) after dropping irrelevant constant. The latter equivalence can essentially be seen by noticing that the first term on the right hand side of (5) is a quadratic approximation to $-2\ell(\hat{\boldsymbol{\delta}}_*)$ to within an additive constant.

In practice, given $\boldsymbol{\delta}_*^{[a+1]}$, we solve the problem

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \ \mathcal{V}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \|\mathbf{z}^{+,[a+1]} - \mathbf{A}_{\boldsymbol{\lambda}}^{[a+1]}\mathbf{z}^{+,[a+1]}\|^2/\check{n} - 1 + 2\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}}^{[a+1]})/\check{n} \tag{6}$$

using the automatic approach by Wood (2004), which is based on Newton's method and can evaluate in an efficient and stable way the components in $\mathcal{V}(\boldsymbol{\lambda})$ and their first and second derivatives with respect to $\log(\boldsymbol{\lambda})$ (since the smoothing parameters can only take positive values). Broadly speaking, this is achieved using a series of pivoted QR and singular value decompositions which make the evaluation of the quantities involving $\mathbf{A}_{\boldsymbol{\lambda}}^{[a+1]}$, for new trial values of $\boldsymbol{\lambda}$, cheap and derivative calculations efficient and stable; see Wood (2004) for full details.

### 2.3.3 Sketch of algorithm

The two steps, detailed in Sections 2.3.1 and 2.3.2, are iterated in a "performance iteration" fashion (Gu, 2002) until the algorithm satisfies the stopping criterion $\max \left| \boldsymbol{\delta}_*{}^{[a]} - \boldsymbol{\delta}_*{}^{[a+1]} \right| < 10^{-6}$. The steps can be summarized as follows:

**step 1** For a given parameter vector value $\boldsymbol{\delta}_*^{[a]}$ and holding the smoothing parameter vector fixed at $\boldsymbol{\lambda}^{[a]}$, find an estimate of $\boldsymbol{\delta}_*$:

$$\boldsymbol{\delta}_*^{[a+1]} = \arg\min_{\mathbf{p}} \ \breve{\ell}_p(\boldsymbol{\delta}_*^{[a]}) + \boldsymbol{\delta}_*^{[a]}.$$

**step 2** Construct the working linear model quantities needed in (6) using $\boldsymbol{\delta}_*^{[a+1]}$ and find an estimate of $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \ \mathcal{V}(\boldsymbol{\lambda}).$$

A slight modification of $\mathcal{V}(\boldsymbol{\lambda})$ is worth mentioning. If the estimated smoothing parameters yield curve estimates that are deemed to be too rough and smoother functions are desired then the trace of the influence matrix can be increased by a factor $> 1$. Kim & Gu (2004) found, in a different context, that using as inflation factor of $1.4$ corrects the tendency to over-fitting of prediction error criteria.

The asymptotic behavior of the proposed estimator and the ensuing formula to calculate the treatment effect is detailed in Section S.3 of the online supplementary material.

## 2.4 Confidence intervals and variable selection

At convergence, the covariance matrix of $\hat{\boldsymbol{\delta}}_*$ can be written as $\mathbf{V}_{\hat{\boldsymbol{\delta}}_*} = -\boldsymbol{\mathcal{H}}_p^{-1}\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{H}}_p^{-1}$. However, the alternative Bayesian result $\mathbf{V}_{\boldsymbol{\delta}_*} = -\boldsymbol{\mathcal{H}}_p^{-1}$ can be employed as well. For smooth functions, at finite sample sizes $\mathbf{V}_{\boldsymbol{\delta}_*}$ can produce intervals with close to nominal 'across-the-function' *frequentist* coverage probabilities (Marra & Wood, 2012). This is because the Bayesian covariance matrix includes both a bias and variance component in a frequentist sense, a feature that is not shared by $\mathbf{V}_{\hat{\boldsymbol{\delta}}_*}$. Note that for unpenalized model components $\mathbf{V}_{\boldsymbol{\delta}_*}$ and $\mathbf{V}_{\hat{\boldsymbol{\delta}}_*}$ are equivalent. Recall that

in (3) $\mathbf{V}_{\boldsymbol{\delta}}$ rather than $\mathbf{V}_{\boldsymbol{\delta}_*}$ is needed. This can be easily obtained by using $\theta$ in place of $\theta_*$ when constructing the covariance matrix.

Point-wise confidence intervals for $\hat{s}_{vk_v}(z_{vk_vi})$ can be obtained using $\mathcal{N}(s_{vk_v}(z_{vk_vi}), \mathbf{B}_{vk_v}(z_{vk_vi})^\mathsf{T}\mathbf{V}_{\boldsymbol{\delta}_{*vk_v}}\mathbf{B}_{vk_v}(z_{vk_vi}))$, where $\mathbf{V}_{\boldsymbol{\delta}_{*vk_v}}$ is the sub-matrix of $\mathbf{V}_{\boldsymbol{\delta}_*}$ that corresponds to the regression spline parameters associated with $\hat{s}_{vk_v}(z_{vk_vi})$. Intervals for non-linear functions of the model coefficients (e.g., $\theta$, $\gamma$ and SATE) can be conveniently obtained by simulation from the posterior distribution of $\boldsymbol{\delta}_*$ as follows:

**step 1** Draw $n_{sim}$ random vectors from $\mathcal{N}(\hat{\boldsymbol{\delta}}_*, \hat{\mathbf{V}}_{\boldsymbol{\delta}_*})$.

**step 2** Calculate $n_{sim}$ simulated realizations of the function of interest. For instance, for a Gaussian copula $\theta = \tanh(\theta_*)$, hence $\boldsymbol{\theta}^{sim} = (\theta_1^{sim}, \theta_2^{sim}, \ldots, \theta_{n_{sim}}^{sim})$ where $\theta_i^{sim} = \tanh(\theta_{*,i}^{sim})$, $i = 1, \ldots, n_{sim}$.

**step 3** Using $\boldsymbol{\theta}^{sim}$ calculate the lower, ($\varsigma/2$), and upper, $1 - \varsigma/2$, quantiles.

Small values for $n_{sim}$ are typically tolerable. Parameter $\varsigma$ is usually set to $0.05$.

Strictly speaking, point-wise confidence intervals for smooth components are not adequate for variable selection purposes, although they are often used in practice (e.g., Ruppert et al., 2003). To test smooth components for equality to zero we use the results by Wood (2013). Let us define $\hat{\mathbf{s}}_{vk_v} = \mathbf{B}_{vk_v}(\mathbf{z}_{vk_v})\hat{\boldsymbol{\beta}}_{vk_v}$, where $\mathbf{B}_{vk_v}(\mathbf{z}_{vk_v})$ denotes a full column rank matrix, $\mathbf{z}_{vk_v} = (z_{vk_v1}, z_{vk_v2}, \ldots, z_{vk_vn})^\mathsf{T}$ and $\mathbf{V}_{\mathbf{s}_{vk_v}} = \mathbf{B}_{vk_v}(\mathbf{z}_{vk_v})\mathbf{V}_{\boldsymbol{\delta}_{*vk_v}}\mathbf{B}_{vk_v}(\mathbf{z}_{vk_v})^\mathsf{T}$. It is then possible to obtain approximate p-values for testing smooth components for equality to zero based on

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

where $\mathbf{V}_{\mathbf{s}_{vk_v}}^{r_{vk_v}-}$ is the rank $r_{vk_v}$ Moore-Penrose pseudo-inverse of $\mathbf{V}_{\mathbf{s}_{vk_v}}$, which is employed to deal with possible rank deficiencies. Parameter $r_{vk_v}$ is selected using the notion of $edf$ used in (6). Because $edf$ is not an integer, it can be rounded as follows (Wood, 2013)

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

which proved effective in semiparametric bivariate probit models (Marra, 2013). Alternatively, variable selection can be achieved by adopting a single penalty shrinkage approach as described in Marra & Radice (2011a) and Marra & Wood (2011).

# 3 Analysis of health care utilization data

The analysis presented in this section was performed in the R environment (R Development Core Team, 2015) using the package `SemiParBIVProbit` (Marra & Radice, 2015) which implements the methodology discussed in this article.

| Variable | Definition |
|---|---|
| *Outcome* | |
| `visits.hosp` | =1 at least one visit to hospital outpatient departments |
| *Treatment* | |
| `private` | =1 private health insurance |
| *Demographic-socioeconomic* | |
| `age` | age in years |
| `gender` | =1 male |
| `race` | =1 white, =2 black, =3 native American, =4 others |
| `education` | years of education |
| `income` | income (000's) |
| `region` | =1 northeast, =2 mid-west, =3 south, =4 west |
| *Health-related* | |
| `health` | =1 excellent, =2 very good, =3 good, =4 fair,=5 poor |
| `bmi` | body mass index |
| `diabetes` | =1 diabetic |
| `hypertension` | =1 hypertensive on |
| `hyperlipidemia` | =1 hyperlipidemic |
| `limitation` | =1 health limits physical activity |

Table 2: Description of the outcome and treatment variables, and observed confounders.

## 3.1 Data

We used a data-set from the 2012 MEPS (`http://www.meps.ahrq.gov/`) which includes information on demographics, individual health status, health care utilization and private health insurance coverage. We excluded individuals younger than 18 years old given their different overall health profiles and expected usage patterns as compared to those of older individuals. Individuals who were older than 64 years old were also excluded since the availability of Medicare obviates the primary insurance decision for almost all US citizens. Individuals that did not have a complete set of socioeconomic and demographic control variables were excluded from the sample (e.g., missing values for education or income). After exclusions, the final data-set contains 10950 observations. Table 2 summarizes the variables used in the analysis. The choice of these variables was motivated largely by the findings reported in previous related studies (e.g., Shane & Trivedi, 2012, and references therein).

## 3.2 Models

Following previous work on the subject (e.g., Holly et al., 1998; Shane & Trivedi, 2012), the equations for private health insurance and health care utilization were specified, in R notation, as

```
treat.eq <- private ~ as.factor(health) + as.factor(race) +
as.factor(region) + limitation + gender + diabetes +
hypertension + hyperlipidemia + s(bmi) + s(income) +
s(age) + s(education)

out.eq <- visits.hosp ~ private + as.factor(health) + as.factor(race) +
```

```
as.factor(region) + limitation + gender + diabetes +
hypertension + hyperlipidemia + s(bmi) + s(income) +
s(age) + s(education)
```

where `as.factor` coerces its argument to a factor and the `s()` symbols refer to the unknown smooth functions described in Section 2.1.1. The smooth components were represented using penalized thin plate regression splines with basis dimensions equal to 20 and penalties based on second order derivatives (Wood, 2006). In cross-sectional studies, 20 bases typically suffice to represent well smooth functions, although sensitivity analysis using more spline bases is advisable when the effective degrees of freedom of the smooth components are close to the number of bases used. We also used two alternative spline definitions (i.e., B-splines with second order difference penalties and cubic regression splines with second order penalties); the resulting estimated curves did not change significantly as compared to those obtained using thin plate splines. The non-linear specification for `bmi`, `income`, `age` and `education` arises from the fact that these covariate embody productivity and life-cycle effects that are likely to affect the treatment and outcome non-linearly. In fact, in related studies, Holly et al. (1998) considered a model for health care utilization that contains linear and quadratic terms in `bmi`, `income`, `age` and `education` whereas Marra & Radice (2011b) specified a model containing smooth functions of them. Considering all copulas discussed in Section 2.1, and including the case in which the outcome equation is estimated alone (this will be referred to as Independent), we fitted 19 copula models. Based on the $AIC$ and Bayesian information criterion ($BIC$) reported in Table 3 the preferred models are the Gaussian, Gumbel$_0$, Clayton$_{180}$ and Joe$_0$. After applying the Vuong (Vuong, 1989) and Clarke (Clarke & Windmeijer, 2012) tests to the four models, it emerged that the Vuong test can not discriminate among the models whereas the Clarke test favors Gumbel$_0$ over the others.

## 3.3 Empirical results

### 3.3.1 Measure of dependence

We start off by commenting on the results for the dependence measures of all models fitted (see Table 3). These represent the association between the unobserved confounders after controlling for observed confounders. Overall, the models without $AIC/BIC$ support, which account for a negative dependence, indicate absence of association between the two equations with intervals which either span all plausible (negative) values for $\gamma/\tau$ or collapse to their point estimates. This behavior is typically observed when the data are inconsistent with the restrictions on the range of the dependence parameter, case in which model misspecification should be strongly suspected (e.g., Trivedi & Zimmer, 2005). The models with $AIC/BIC$ support, which account for a positive dependence, do not exhibit such a behavior and suggest a low association. Interestingly, the small yet significant dependence parameters obtained for Gumbel$_0$ indicates that there exists some positive association between the unstructured terms of the model equations for private health insurance and hospital utilization which is most likely due to the presence of unobserved confounders. This positive relationship suggests that individuals with private health coverage are more likely to

use health care services as compared to those without coverage.

### 3.3.2 SATE of private health insurance

The estimated SATE (in $\%$) and confidence interval (CI) for all fitted copula models are reported in Table 3. The Table also reports the estimated SATE for the case in which the unobserved confounding issue is not taken into account (Independent). Several points are worth noting.

- The chosen models (Gaussian, $\text{Gumbel}_0$, $\text{Clayton}_{180}$ and $\text{Joe}_0$, which account for a positive dependence) show similar point estimates with overlapping CIs. The models that account for a negative dependence (which have no $AIC/BIC$ support) exhibit estimates that are systematically smaller than those produced by the preferred models (and that produced by the Independent model). As pointed out in the previous section, the negative dependence models have estimated dependence parameters that are on the boundary of their parameter spaces, hence suggesting that these models are not supported by the data.
- If the presence of unobserved confounders is not accounted for then the estimated SATE is smaller ($4.11\%$) than that obtained using the chosen models which can control for this issue (around $4.56\%$). Based on these estimates the direction of the bias appears to be downward. This result seems counter-intuitive in the sense that if we assume that possible confounders are allergy and risk aversiveness, then an upward bias should be expected (individuals with a greater demand for medical care are expected to have a greater demand for insurance). The explanation behind this apparent contradiction is that employer-provided insurance is generally limited to full-time workers and is positively related to the worker's income. The empirical evidence indicates that workers who are in poorer health are less likely to obtain employer-sponsored coverage (e.g., Buchmueller et al., 2005).
- Using the Gaussian copula the estimated SATE is $4.61\%$, which does not really differ from those obtained using the other supported copula models. This is most likely due to the low association observed. When $\gamma/\tau \to 0$ the copula models converge to the normal product distribution, case in which all copulas entail very similar distributions. As shown in simulation (see Section S.4 of the online supplementary material), larger differences are likely to be observed when the association between the treatment and outcome equations is stronger. In such a scenario, different copulas would entail different distributions (as shown in Figure 1), hence the use of the appropriate copula model can make a difference.

### 3.3.3 Parametric components

We report the estimated effects for the $\text{Gumbel}_0$ copula model. Similar results where obtained using the other preferred models (these are available upon request).

Most of these effects have the expected signs. Regarding `gender`, females are slightly more likely of being hospitalized than males. This may be explained by a higher demand for medical services among women during their reproductive years (e.g., Sindelar, 1982). As for `race`, there is not a significant difference between whites and nonwhites in terms of purchasing private health

| Copula | $\widehat{\text{SATE}}$ (95% CIs) | $\hat{\gamma}$ (95% CIs) | $\hat{\tau}$ (95% CIs) | $AIC$ | $BIC$ |
|---|---|---|---|---|---|
| Independent | 4.11 (0.75,7.48) | - | - | 17628.02 | 18116.06 |
| Gaussian | 4.61 (3.15,6.06) | 0.39 (0.03,0.64) | 0.13 (0.003,0.25) | 17621.97 | 18070.06 |
| Student-$t_3$ | 4.81 (3.26,6.36) | 0.61 (0.38,0.74) | 0.34 (0.22,0.45) | 17640.29 | 18085.16 |
| Student-$t_6$ | 4.56 (2.95,6.16) | 0.48 (0.15,0.71) | 0.21 (0.08,0.35) | 17628.08 | 18075.64 |
| Student-$t_9$ | 4.53 (2.95,6.10) | 0.44 (0.12,0.74) | 0.18 (0.03,0.30) | 17624.51 | 18071.92 |
| Student-$t_{12}$ | 4.53 (2.98,6.08) | 0.40 (0.07,0.71) | 0.16 (0.04,0.29) | 17623.01 | 18070.49 |
| Frank | 4.30 (2.92,5.69) | 0.29 (0.00,0.55) | 0.13 (0.001,0.25) | 17622.32 | 18070.02 |
| $\text{Clayton}_0$ | 3.98 (2.62,5.35) | 0.11 (0.01,0.73) | 0.03 (0.003,0.27) | 17624.37 | 18075.35 |
| $\text{Clayton}_{90}$ | 3.97 (2.44,5.49) | 0 (-1,0) | 0 (-1,0) | 17670.23 | 18263.54 |
| $\text{Clayton}_{180}$ | 4.52 (3.08,5.96) | 0.17 (0.064,0.45) | 0.09 (0.03,0.24) | 17622.57 | 18072.41 |
| $\text{Clayton}_{270}$ | 3.98 (2.21,5.76) | 0 (-1,0) | 0 (-1,0) | 17624.94 | 18081.31 |
| $\text{Gumbel}_0$ | 4.62 (3.17,6.08) | 0.29 (0.09,0.63) | 0.13 (0.05,0.31) | 17621.05 | 18069.71 |
| $\text{Gumbel}_{90}$ | 3.96 (2.42,5.51) | 0 (0,0) | 0 (0,0) | 17672.95 | 18261.34 |
| $\text{Gumbel}_{180}$ | 4.02 (2.64,5.40) | 0.10 (0.01,0.64) | 0.03 (0.001,0.34) | 17624.42 | 18076.04 |
| $\text{Gumbel}_{270}$ | 3.96 (2.19,5.74) | 0 (0,0) | 0 (0,0) | 17664.94 | 18280.32 |
| $\text{Joe}_0$ | 4.50 (3.06,5.95) | 0.16 (0.04,0.48) | 0.09 (0.03,0.26) | 17622.66 | 18072.62 |
| $\text{Joe}_{90}$ | 3.96 (2.58,5.34) | 0 (-1,0) | 0 (-1,0) | 17670.94 | 18287.37 |
| $\text{Joe}_{180}$ | 3.97 (2.62,5.33) | 0.04 (0.00,0.80) | 0.01 (0,0.54) | 17624.76 | 18079.31 |
| $\text{Joe}_{270}$ | 3.96 (2.26,5.67) | 0 (-1,0) | 0 (-1,0) | 17669.94 | 18291.33 |

Table 3: Estimated SATE (in %), gamma measure $\gamma$, Kendall's $\tau$, $AIC$ and $BIC$ obtained using different copula models for the 2012 MEPS data. 95% confidence intervals for the SATE have been obtained using the delta method detailed in Section 2.2, and those for $\gamma$ and $\tau$ using Bayesian posterior simulation as described in Section 2.4. For the Independent model the information criteria have been calculated assuming that the treatment and outcome equations are not associated.

insurance but there is some difference in terms of being hospitalized; black individuals seem to be less likely to use health care services as compared to whites. This is consistent with the findings by Shane & Trivedi (2012). Regarding `region`, residents of the Midwest are more likely to have a private insurance and to use health care services as compared to those of the Northeast. Individuals' evaluation of their `health` states is a potential predictor of health care utilization. Those who are in good health are less likely to access health care services. In the same vein, those who expect themselves to be in good health have little to gain from insurance while those who are in poor health are more likely to purchase health insurance. The results for the hospital utilization equation support this hypothesis indicating that the less healthy individuals are, the more likely they are to be admitted into hospitals. The positive relationship between self-assessed health and insurance purchase is counter-intuitive to the hypothesis of moral hazard and adverse selection. However, such finding is not unusual and has been obtained in several previous studies (see Srivastava & Zhao, 2008, and references therein). The more objective measures of health status (i.e., `diabetes`, `hypertension` and `hyperlipidemia`) suggest that medical need is an important determinant of hospital utilization and insurance purchase.

### 3.3.4 Non-parametric components

Figures 2 and 3 report the smooth function estimates for the treatment and outcome equations (and associated intervals) when applying the $\text{Gumbel}_0$ model on the MEPS data. The estimated smooth

| | Treatment Eq. | | Outcome Eq. | |
|---|---|---|---|---|
| Variable | Parameter estimate | Std. error | Parameter estimate | Std. error |
| gender | -0.02 | 0.03 | -0.37 | 0.03 |
| race=2 | -0.00 | 0.04 | -0.08 | 0.04 |
| race=3 | 0.04 | 0.15 | 0.35 | 0.16 |
| race=4 | -0.04 | 0.05 | -0.17 | 0.07 |
| region=2 | 0.24 | 0.05 | 0.16 | 0.06 |
| region=3 | 0.06 | 0.04 | -0.22 | 0.05 |
| region=4 | 0.01 | 0.04 | -0.37 | 0.06 |
| health=2 | 0.04 | 0.04 | 0.10 | 0.05 |
| health=3 | -0.11 | 0.04 | 0.33 | 0.05 |
| health=4 | -0.27 | 0.06 | 0.48 | 0.07 |
| health=5 | -0.39 | 0.09 | 0.67 | 0.10 |
| diabetes | 0.12 | 0.06 | 0.06 | 0.06 |
| hypertension | 0.09 | 0.04 | 0.09 | 0.04 |
| hyperlipidemia | 0.17 | 0.04 | 0.9 | 0.04 |
| limitation | 0.05 | 0.06 | -0.49 | 0.06 |

Table 4: Estimated coefficients and standard errors of the parametric components of the Gumbel$_0$ model.

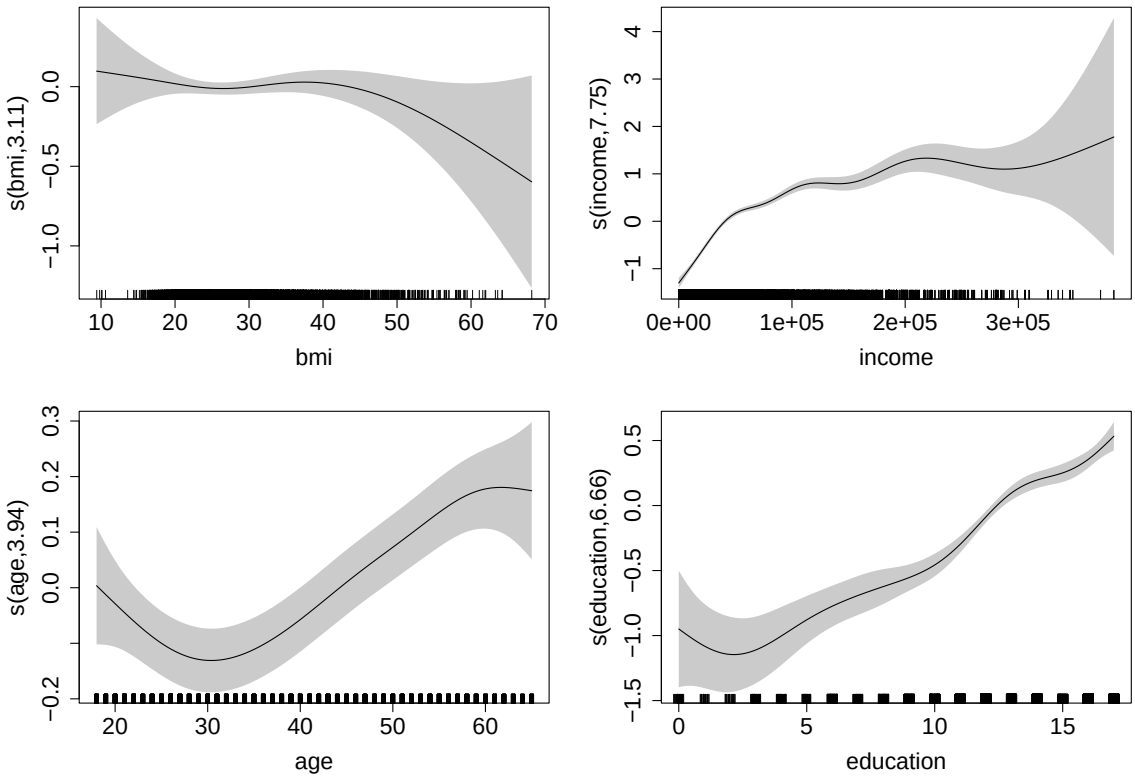


Figure 2: Smooth function estimates and associated $95\%$ point-wise confidence intervals in the treatment equation obtained by applying the Gumbel$_0$ regression spline model on the 2012 MEPS data. Results are plotted on the scale of the linear predictor. The jittered rug plot, at the bottom of each graph, shows the covariate values. The numbers in brackets in the y-axis captions are the effective degrees of freedom of the smooth curves. P-values for the smooth terms of bmi, income, age and education are $0.271$, $< 0.000$, $< 0.000$ and $< 0.000$, respectively.

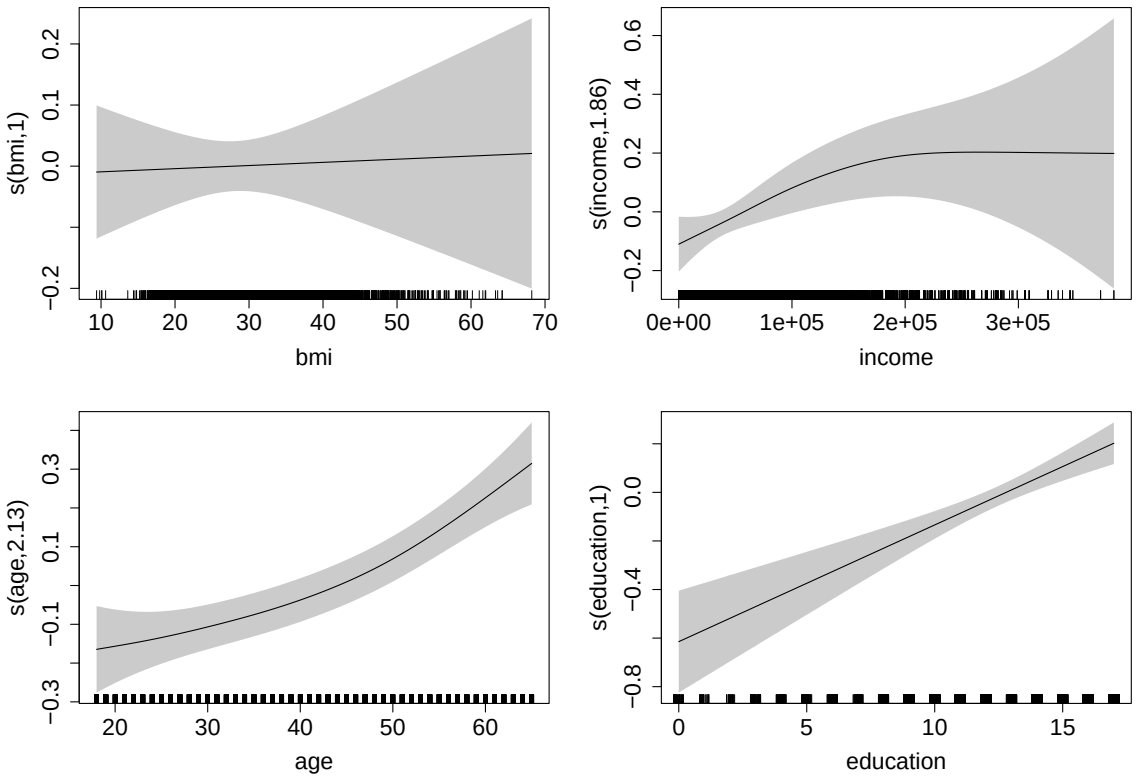


Figure 3: Smooth function estimates and associated 95% point-wise confidence intervals in the outcome equation obtained by applying the Gumbel$_0$ regression spline model on the 2012 MEPS data. Results are plotted on the scale of the linear predictor. P-values for the smooth terms of `bmi`, `income`, `age` and `education` are $0.849$, $0.01$, $< 0.000$ and $< 0.000$, respectively.

functions obtained using the other copula models (not reported here but available upon request) were similar.

The effects of `bmi`, `income`, `age` and `education` in the treatment and outcome equations show different degrees of non-linearity. The point-wise confidence intervals of the smooth functions for `bmi` in the treatment and outcome equations contain the zero line for the whole range of the covariate values. The intervals of the smooth for `income` in the outcome equation contain the zero line for most of the covariate value range. This suggests that `bmi` is a weak predictor of private health insurance and health care utilization, and that `income` might not be an important determinant of hospital utilization. Similar conclusions can be drawn by looking at the p-values reported in the captions of Figures 2 and 3. As for the remaining variables, the estimated effects have the expected patterns. For example, `age` is a significant determinant in both equations. The probability of purchasing a private health insurance is found to increase with `age`. This is suggestive of a higher probability of private health insurance purchase as individuals become older and less likely to stay healthy (e.g., Hopkins & Kiddi, 1996). The probability of using health care services also increases with `age`. Insurance decision as well as health care utilization appear to be highly associated with `education`. Education is likely to increase individuals' awareness of health care services and the benefits of purchasing a private health insurance. Higher household income is associated with an increased probability of purchasing a private health insurance.

It is worth noting that the parametric and non-parametric estimated effects for the outcome

equation reported here should be interpreted in a qualitative way only. The actual effects can be calculated by using simulation or by adapting the formulas of Greene (2012) to the current context. This would account for the fact that the confounders appearing in the treatment equation have an indirect effect (through the endogenous variable) on the outcome and a direct effect because they also appear in the outcome equation.

# 4 Discussion

We have introduced a framework which can allow researchers to estimate the effect that a binary treatment has on a binary variable in the presence of unobserved confounding, non-linear covariate effects and non-Gaussian dependencies between the treatment and outcome equations. We have provided inferential tools for this framework and presented some argumentation related to the asymptotic behavior of the proposed penalized maximum likelihood estimator and the ensuing sample average treatment effect. We have also developed the necessary computational procedures which are incorporated in the `R` package `SemiParBIVProbit` (Marra & Radice, 2015).

Using the proposed approach, we have examined the effect of private health insurance on health care utilization using the 2012 MEPS data-set. There is a generally accepted notion that private health coverage is affected by endogeneity as it is not randomly assigned as in a controlled trial but rather is the result of individual preferences and health status, such as allergy and risk aversiveness. Also, the impacts of continuous confounders such as age and education are likely to be complex since they embody productivity and life-cycle effects that are likely to influence non-linearly private health insurance and health care utilization. Finally, insurance and health care utilization may exhibit a non-Gaussian dependence. To our knowledge, no studies have examined the impact of private health coverage accounting for endogeneity, non-linear contributions of observed confounders and non-Gaussian dependence between insurance and health care utilization, partly due to the lack of appropriate analytical and computational tools. By applying the introduced statistical framework to the 2012 MEPS data we found that not accounting for the endogeneity issue underestimates the effect of private health insurance and that some of the observed confounder effects are non-linear. We also found that the Gaussian, $\text{Gumbel}_0$, $\text{Clayton}_{180}$ and $\text{Joe}_0$ models were equally supported. This was due to the low yet significant association observed between the treatment and outcome equations, case in which the copula models entail very similar distributions. However, as shown in simulation, the use of the appropriate copula model may make a difference when the association between the two equations is strong.

Since marginal distributions other than Gaussian may be plausible in applications, we explored the possibility of modeling the margins using skew probit links derived from the standard skew-normal distribution by Azzalini (1985) as well as the power probit and reciprocal power probit links discussed by Bazan et al. (2010). We opted for these links as they include the probit link as special case and have desirable mathematical properties. The use of these approaches did not lead to SATE results different from those reported in Table 3. Moreover, the convergence of the algorithm slowed down considerably and sometimes it was not possible to find a solution. As

pointed out by Azzalini & Arellano-Valle (2013), in the simpler context of continuous outcome variables, having a parameter which regulates the distribution's skewness enjoys attractive formal properties from the probability point of view. However, a practical problem in applications is the possibility that the maximum likelihood estimate of the skewness parameter diverges. That is, the profile log-likelihood for the skewness coefficient may be flat in a non-negligible portion of situations. This issue has vanishing probability for increasing sample size, but for finite samples it occurs with non-negligible probability.

A limitation of the copulas employed in this article is that they are exchangeable (Durante, 2009; Frees & Valdez, 1998; Nelsen, 2007). In the context of our case study, this means that the probability of (not) having private health insurance conditionally to the usage (or not) of health care services is equal to the probability of using (or not) health care services knowing that a private health insurance can (not) be used. Following the approach detailed in Frees & Valdez (1998), we employed the copula $\mathcal{C}_{\kappa_1,\kappa_2}(u,v) = u^{1-\kappa_1}v^{1-\kappa_2}\mathcal{C}(u^{\kappa_1}, v^{\kappa_2}),\ 0 < \kappa_1, \kappa_2 < 1$, which has the property of including $\mathcal{C}$ as a limiting case. We encountered the same issues mentioned above, even when using a model with a small number of covariates and without smooth functions.

An interesting avenue for future research includes the use of semi- and non-parametric copula approaches. These would allow the margins and/or the copula to be estimated non-parametrically using, for instance, smoothing methods such as kernels, wavelets and orthogonal polynomials. Broadly speaking, if the specification of the model for the margins and copula is correct, then the parametric approach will outperform semi- and non-parametric methods; however, the reverse will be true under misspecification. Without any valuable prior information, semi- and non-parametric techniques should be favored as they will be more flexible in determining the shape of the underlying distribution. However, in practice, such techniques are typically limited with regard to the inclusion of a large set of covariates, may require the imposition of restrictions on the functions approximating the underlying distribution and may be computationally demanding (e.g., Deheuvels, 1981a,b; Genest et al., 1995; Tutz & Petry, 2013). While a fully parametric copula approach is less flexible than semi- and non-parametric approaches, it is computationally more feasible and it still allows the user to assess the sensitivity of results to different modeling assumptions.

Another interesting extension would be to consider trivariate system models, controlling for the endogeneity of the treatment and for non-random sample selection in the outcome (e.g., Srivastava & Zhao, 2008). Finally, a future release of `SemiParBIVProbit` will allow the user to model the copula parameter as a function of a linear predictor to allow for different degrees of endogeneity across observations; the theoretical and computational framework remains essentially unchanged.

# Acknowledgement

We would like to thank two anonymous reviewers and the Associate Editor for many suggestions which helped to clarify the contribution of the paper and improved considerably the presentation of the article.

# References


Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. W. (2004). Implementing matching estimators for average treatment effects in Stata. *Stata Journal*, 4, 290–311.

Azzalini, A. (1985). A class of distributions which includes the normal one. *Scandinavian Journal of Statistics*, 12, 171–178.

Azzalini, A. & Arellano-Valle, R. B. (2013). Maximum penalized likelihood estimation for skew-normal and skew-t distributions. *Journal of Statistical Planning and Inference*, 143, 419–433.

Barndorff-Nielsen, O. & Cox, D. (1989). *Asymptotic Techniques for Use in Statistics*. London: Chapman and Hall.

Bazan, J. L., H.Bolfarinez, & Branco, M. B. (2010). A framework for skew-probit links in binary regression. *Communications in Statistics: Theory and Methods*, 39, 678–697.

Brechmann, E. C. & Schepsmeier, U. (2013). Modeling dependence with c- and d-vine copulas: The R package CDVine. *Journal of Statistical Software*, 52(3), 1–27.

Buchmueller, T. C., Grumbach, K., Kronick, R., & Kahn, J. G. (2005). Book review: The effect of health insurance on medical care utilization and implications for insurance expansion: A review of the literature. *Medical Care Research and Review*, 62, 3–30.

Chib, S. & Greenberg, E. (2007). Semiparametric modeling and estimation of instrumental variable models. *Journal of Computational and Graphical Statistics*, 16, 86–114.

Chib, S. & Hamilton, B. H. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, 110, 67–89.

Clarke, P. S. & Windmeijer, F. (2012). Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107, 1638–1652.

Deheuvels, P. (1981a). A kolmogorov-smirnov type test for independence and multivariate samples. *Romanian Journal of Pure and Applied Mathematics*, 26, 213–226.

Deheuvels, P. (1981b). A nonparametric test for independence. *Pub. Inst. Stat. Univ. Paris*, 26, 29–50.

Durante, F. (2009). Construction of non-exchangeable bivariate distribution functions. *Statistical Papers*, 50, 383–391.

Frees, E. W. & Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2, 1–25.

Genest, C., Ghoudi, K., & Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82, 543–552.

Genest, C., Nikoloulopoulos, A. K., Rivest, L.-P., & Fortin, M. (2013). Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. *Brazilian Journal of Probability and Statistics*, 27, 265–284.

Geyer, C. J. (2013). Trust regions. `http://cran.r-project.org/web/packages/trust/vignet`

Gitto, L., Santoro, D., & Sobbrio, G. (2006). Choice of dialysis treatment and type of medical unit (private vs public), application of a recursive bivariate probit. *Health Economics*, 15, 1251–1256.

Goldman, D. P., Bhattacharya, J., McCaffrey, D. F., Duan, N., Leibowitz, A. A., Joyce, G. F., & Morton, S. C. (2001). Effect of insurance on mortality in an HIV-positive population in care. *Journal of the American Statistical Association*, 96, 883–894.

Goodman, L. A. & Kruskal, W. H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, 49, 732–764.

Greene, W. H. (2012). *Econometric Analysis*. Prentice Hall, New York.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag, London.

Han, S. & Vytlacil, E. J. (2014). Identification in a generalization of bivariate probit models with endogenous regressors. *Revise and Resubmit, The Journal of Econometrics*.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, 55, 757–796.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46, 931–959.

Heckman, J. J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64, 605–654.

Holly, A., Gardiol, L., Domenighetti, G., & Brigitte, B. (1998). An econometric model of health care utilization and health insurance in Switzerland. *European Economic Review*, 42(3-5), 513–522.

Hopkins, S. & Kiddi, M. P. (1996). The determinants of the demand for private health insurance under medicare. *Applied Economics*, 28, 1623–1632.

Jones, A. M., Koolman, X., & Doorslaer, E. V. (2006). The impact of having supplementary private health insurance on the uses of specialists. *Annales d'Economie et de Statistique*, (83/84), 251–275.

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics and Data Analysis*, 49, 169–186.

Kauermann, G., Krivobokova, T., & Fahrmeir, L. (2009). Some asymptotics results on generalized penalized spline smoothing. *Journal of Royal Statistical Society Series B*, 71, 487–503.

Kawatkar, A. A. & Nichol, M. B. (2009). Estimation of causal effects of physical activity on obesity by a recursive bivariate probit model. *Value in Health*, 12, A131–A132.

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

Latif, E. (2009). The impact of diabetes on employment in Canada. *Health Economics*, 18, 577–589.

Li, Y. & Jensen, G. A. (2011). The impact of private long-term care insurance on the use of long-term care. *Inquiry*, 48(1), 34–50.

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

Marra, G. (2013). On p-values for semiparametric bivariate probit models. *Statistical Methodology*, 10, 23–28.

Marra, G. & Radice, R. (2011a). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canadian Journal of Statistics*, 39, 259–279.

Marra, G. & Radice, R. (2011b). A flexible instrumental variable approach. *Statistical Modelling*, 11, 581–603.

Marra, G. & Radice, R. (2015). *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling*. R package version 3.3.

Marra, G. & Wood, S. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39, 53–74.

Marra, G. & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*, 55, 2372–2387.

McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman and Hall.

Nelsen, R. (2006). *An Introduction to Copulas*. New York: Springer.

Nelsen, R. B. (2007). Extremes of nonexchangeability. *Statistical Papers*, 48, 329–336.

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

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

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

Shane, D. & Trivedi, P. K. (2012). What drives differences in health care demand? the role of health insurance and selection bias. *Health, Econometrics and Data Group (HEDG) Working Papers*.

Sindelar, J. L. (1982). Differential use of medical care by sex. *The Journal of Political Economy*, 90, 1003–1019.

Sklar, A. (1959). Fonctions de répartition é n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetica*, 9, 449–460.

Srivastava, P. & Zhao, X. (2008). Impact of private health insurance on the choice of public versus private hospital services. *Health, Econometrics and Data Group (HEDG) Working Papers*.

Swihart, B. J., Caffo, B. S., & Crainiceanu, C. M. (2014). A unifying framework for marginalised random-intercept models of correlated binary outcomes. *Computational Statistics and Data Analysis*, 82, 275–295.

Tajar, A., Denuit, M., & Lambert, P. (2001). Copula-type representation for random couples with bernoulli margins. *Working Paper*.

Trivedi, P. K. & Zimmer, D. M. (2005). Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1(1), 1–111.

Tutz, G. & Petry, S. (2013). Generalized additive models with unknown link function including variable selection. *Technical Report*.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57, 307–333.

Wiesenfarth, M. & Kneib, T. (2011). Bayesian geoadditive sample selection models. *Journal of the Royal Statistical Society Series C*, 59, 381–404.

Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economics Letters*, 69, 309–312.

Winkelmann, R. (2012). Copula bivariate probit models: with an application to medical expenditures. *Health Economics*, 21, 1444–1455.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B*, 65, 95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99, 673–686.

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

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100, 221–228.

# Supplementary material to "Copula regression spline models for binary outcomes"

## S.1 Derivatives of $\mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})$ with respect to $\boldsymbol{\delta}$

The components in $\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})/\partial \boldsymbol{\delta}$ that are referred to in Section 2.2 are given below.

$$\frac{\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})}{\partial \boldsymbol{\delta}_1} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \left\{ \frac{\mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=1)}); \theta\right)}{\Phi(\eta_{1i})} - \frac{\Phi(\eta_{2i}^{(y_{1i}=0)}) - \mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=0)}); \theta\right)}{1-\Phi(\eta_{1i})} \right\}}{\partial \boldsymbol{\delta}_1}, \tag{7}$$

$$\frac{\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})}{\partial \boldsymbol{\delta}_2} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \left\{ \frac{\mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=1)}); \theta\right)}{\Phi(\eta_{1i})} - \frac{\Phi(\eta_{2i}^{(y_{1i}=0)}) - \mathcal{C}\left(, \Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=0)}); \theta\right)}{1-\Phi(\eta_{1i})} \right\}}{\partial \boldsymbol{\delta}_2}, \tag{8}$$

and

$$\frac{\partial \mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \left\{ \frac{\mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=1)}); \theta\right)}{\Phi(\eta_{1i})} - \frac{\Phi(\eta_{2i}^{(y_{1i}=0)}) - \mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=0)}); \theta\right)}{1-\Phi(\eta_{1i})} \right\}}{\partial \theta}. \tag{9}$$

The quantities inside the square brackets of (7), (8) and (9) can be written as

$$\left\{ \left. \frac{h_1 \Phi(\eta_{1i}) - \mathcal{C}(\Phi(\eta_{1i}), \Phi(\eta_{2i}); \theta)}{\Phi(\eta_{1i})^2} \right|_{y_{1i}=1} + \left. \frac{h_1 \left(1 - \Phi(\eta_{1i})\right) - \left(\Phi(\eta_{2i}) - \mathcal{C}(\Phi(\eta_{1i}), \Phi(\eta_{2i}); \theta)\right)}{\left[1 - \Phi(\eta_{1i})\right]^2} \right|_{y_{1i}=0} \right\} \phi(\eta_{1i}) \mathbf{X}_{1i},$$

$$\left. \frac{h_2 \phi(\eta_{2i}) \mathbf{X}_{2i}}{\Phi(\eta_{1i})} \right|_{y_{1i}=1} - \left. \frac{1-h_2}{1-\Phi(\eta_{1i})} \phi(\eta_{2i}) \mathbf{X}_{2i} \right|_{y_{1i}=0},$$

and

$$\left. \frac{\frac{\partial \mathcal{C}(\Phi(\eta_{1i}), \Phi(\eta_{2i}); \theta)}{\partial \theta}}{\Phi(\eta_{1i})} \right|_{y_{1i}=1} + \left. \frac{\frac{\partial \mathcal{C}(\Phi(\eta_{1i}), \Phi(\eta_{2i}); \theta)}{\partial \theta}}{1-\Phi(\eta_{1i})} \right|_{y_{1i}=0},$$

where $h_v = \partial \mathcal{C}(\Phi(\eta_{1i}), \Phi(\eta_{2i}); \theta)/\partial \Phi(\eta_{vi})$, $v$ can take values 1 and 2, $\phi(\cdot)$ is the density function of the standard univariate Gaussian distribution, and all the other quantities are defined in Section 2.

# S.2 Gradient and Hessian of $\boldsymbol{\delta}_*$

Recall that $\mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=1)}); \theta\right) = \mathbb{P}(y_{1i} = 1, y_{2i} = 1)$ and the probabilities for the other three events defined in Section 2.1. Also, recall from Section 2.3 that $\boldsymbol{\delta}_*^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \theta_*)$. The quantities $\mathbf{g}$ and $\boldsymbol{\mathcal{H}}$ that are referred to in Section 2.3.1 are given below.

## Gradient

$$\begin{aligned}
\mathbf{g}_1 = \frac{\partial \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1} = \sum_{i=1}^{n} \Bigg\{ & y_{1i}y_{2i} \frac{1}{\mathbb{P}(y_{1i} = 1, y_{2i} = 1)} \frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 1)}{\partial \boldsymbol{\delta}_1} \\
& + y_{1i}(1 - y_{2i}) \frac{1}{\mathbb{P}(y_{1i} = 1, y_{2i} = 0)} \frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 0)}{\partial \boldsymbol{\delta}_1} \\
& + (1 - y_{1i})y_{2i} \frac{1}{\mathbb{P}(y_{1i} = 0, y_{2i} = 1)} \frac{\partial \mathbb{P}(y_{1i} = 0, y_{2i} = 1)}{\partial \boldsymbol{\delta}_1} \\
& + (1 - y_{1i})(1 - y_{2i}) \frac{1}{\mathbb{P}(y_{1i} = 0, y_{2i} = 0)} \frac{\partial \mathbb{P}(y_{1i} = 0, y_{2i} = 0)}{\partial \boldsymbol{\delta}_1} \Bigg\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 1)}{\partial \boldsymbol{\delta}_1} &= h_1 \phi(\eta_{1i}) \mathbf{X}_{1i}, \\
\frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 0)}{\partial \boldsymbol{\delta}_1} &= \phi(\eta_{1i}) \mathbf{X}_{1i} - \frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 1)}{\partial \boldsymbol{\delta}_1}, \\
\frac{\partial \mathbb{P}(y_{1i} = 0, y_{2i} = 1)}{\partial \boldsymbol{\delta}_1} &= -\frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 1)}{\partial \boldsymbol{\delta}_1}, \\
\frac{\partial \mathbb{P}(y_{1i} = 0, y_{2i} = 0)}{\partial \boldsymbol{\delta}_1} &= -\frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 0)}{\partial \boldsymbol{\delta}_1}.
\end{aligned}$$

$$\begin{aligned}
\mathbf{g}_2 = \frac{\partial \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2} = \sum_{i=1}^{n} \Bigg\{ & y_{1i}y_{2i} \frac{1}{\mathbb{P}(y_{1i} = 1, y_{2i} = 1)} \frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 1)}{\partial \boldsymbol{\delta}_2} \\
& + y_{1i}(1 - y_{2i}) \frac{1}{\mathbb{P}(y_{1i} = 1, y_{2i} = 0)} \frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 0)}{\partial \boldsymbol{\delta}_2} \\
& + (1 - y_{1i})y_{2i} \frac{1}{\mathbb{P}(y_{1i} = 0, y_{2i} = 1)} \frac{\partial \mathbb{P}(y_{1i} = 0, y_{2i} = 1)}{\partial \boldsymbol{\delta}_2} \\
& + (1 - y_{1i})(1 - y_{2i}) \frac{1}{\mathbb{P}(y_{1i} = 0, y_{2i} = 0)} \frac{\partial \mathbb{P}(y_{1i} = 0, y_{2i} = 0)}{\partial \boldsymbol{\delta}_2} \Bigg\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 1)}{\partial \boldsymbol{\delta}_2} &= h_2 \phi(\eta_{2i}) \mathbf{X}_{2i}, \\
\frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 0)}{\partial \boldsymbol{\delta}_2} &= -\frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 1)}{\partial \boldsymbol{\delta}_2}, \\
\frac{\partial \mathbb{P}(y_{1i} = 0, y_{2i} = 1)}{\partial \boldsymbol{\delta}_2} &= \phi(\eta_{2i}) \mathbf{X}_{2i} - \frac{\partial \mathbb{P}(y_{1i} = 1, y_{2i} = 1)}{\partial \boldsymbol{\delta}_2}, \\
\frac{\partial \mathbb{P}(y_{1i} = 0, y_{2i} = 0)}{\partial \boldsymbol{\delta}_2} &= -\frac{\partial \mathbb{P}(y_{1i} = 0, y_{2i} = 1)}{\partial \boldsymbol{\delta}_1}.
\end{aligned}$$

$$
\begin{aligned}
g_3 = \frac{\partial \ell(\boldsymbol{\delta}_*)}{\partial \theta_*} = \sum_{i=1}^{n} \Bigg\{ & y_{1i} y_{2i} \frac{1}{\mathbb{P}(y_{1i}=1, y_{2i}=1)} \frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \theta} \\
& + y_{1i}(1-y_{2i}) \frac{1}{\mathbb{P}(y_{1i}=1, y_{2i}=0)} \frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \theta} \\
& + (1-y_{1i}) y_{2i} \frac{1}{\mathbb{P}(y_{1i}=0, y_{2i}=1)} \frac{\partial \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \theta} \\
& + (1-y_{1i})(1-y_{2i}) \frac{1}{\mathbb{P}(y_{1i}=0, y_{2i}=0)} \frac{\partial \mathbb{P}(y_{1i}=0, y_{2i}=0)}{\partial \theta} \Bigg\} \frac{\partial \theta}{\partial \theta_*},
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \theta} &= \frac{\partial \mathcal{C}(\Phi(\eta_{1i}), \Phi(\eta_{2i}); \theta)}{\partial \theta}, \\
\frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \theta} &= -\frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \theta}, \\
\frac{\partial \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \theta} &= -\frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \theta}, \\
\frac{\partial \mathbb{P}(y_{1i}=0, y_{2i}=0)}{\partial \theta} &= \frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \theta},
\end{aligned}
$$

and $\partial\theta/\partial\theta_*$ can be obtained using the transformations in Table 1.

## Hessian

$$
\begin{aligned}
\boldsymbol{\mathcal{H}}_{1,1} = \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\intercal} = \sum_{i=1}^{n} \Bigg\{ & y_{1i} y_{2i} \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\intercal} \mathbb{P}(y_{1i}=1, y_{2i}=1) - \left[\frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_1}\right]^2}{\left[\mathbb{P}(y_{1i}=1, y_{2i}=1)\right]^2} \\
& + y_{1i}(1-y_{2i}) \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\intercal} \mathbb{P}(y_{1i}=1, y_{2i}=0) - \left[\frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \boldsymbol{\delta}_1}\right]^2}{\left[\mathbb{P}(y_{1i}=1, y_{2i}=0)\right]^2} \\
& + (1-y_{1i}) y_{2i} \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\intercal} \mathbb{P}(y_{1i}=0, y_{2i}=1) - \left[\frac{\partial \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \boldsymbol{\delta}_1}\right]^2}{\left[\mathbb{P}(y_{1i}=0, y_{2i}=1)\right]^2} \\
& + (1-y_{1i})(1-y_{2i}) \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\intercal} \mathbb{P}(y_{1i}=0, y_{2i}=0) - \left[\frac{\partial \mathbb{P}(y_{1i}=0, y_{2i}=0)}{\partial \boldsymbol{\delta}_1}\right]^2}{\left[\mathbb{P}(y_{1i}=0, y_{2i}=0)\right]^2} \Bigg\},
\end{aligned}
$$

where

$$\begin{aligned}
\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\mathsf{T}} &= \frac{\partial h_1}{\partial \boldsymbol{\delta}_1}\phi(\eta_{1i})\mathbf{X}_{1i} - h_1\phi(\eta_{1i})\eta_{1i}\mathbf{X}_{1i}^\mathsf{T}\mathbf{X}_{1i}, \\
\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\mathsf{T}} &= -\phi(\eta_{1i})\eta_{1i}\mathbf{X}_{1i}^\mathsf{T}\mathbf{X}_{1i} - \frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\mathsf{T}}, \\
\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\mathsf{T}} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\mathsf{T}}, \\
\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\mathsf{T}} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_1^\mathsf{T}}.
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\mathcal{H}}_{2,2} = \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}} = \sum_{i=1}^{n} \Bigg\{ & y_{1i}y_{2i} \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}} \mathbb{P}(y_{1i}=1, y_{2i}=1) - \left[\frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_2}\right]^2}{\left[\mathbb{P}(y_{1i}=1, y_{2i}=1)\right]^2} \\
& + y_{1i}(1-y_{2i}) \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}} \mathbb{P}(y_{1i}=1, y_{2i}=0) - \left[\frac{\partial \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \boldsymbol{\delta}_2}\right]^2}{\left[\mathbb{P}(y_{1i}=1, y_{2i}=0)\right]^2} \\
& + (1-y_{1i})y_{2i} \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}} \mathbb{P}(y_{1i}=0, y_{2i}=1) - \left[\frac{\partial \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \boldsymbol{\delta}_2}\right]^2}{\left[\mathbb{P}(y_{1i}=0, y_{2i}=1)\right]^2} \\
& + (1-y_{1i})(1-y_{2i}) \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=0)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}} \mathbb{P}(y_{1i}=0, y_{2i}=0) - \left[\frac{\partial \mathbb{P}(y_{1i}=0, y_{2i}=0)}{\partial \boldsymbol{\delta}_2}\right]^2}{\left[\mathbb{P}(y_{1i}=0, y_{2i}=0)\right]^2} \Bigg\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}} &= \frac{\partial h_2}{\partial \boldsymbol{\delta}_2}\phi(\eta_{2i})\mathbf{X}_{2i} - h_2\phi(\eta_{2i})\eta_{2i}\mathbf{X}_{2i}^\mathsf{T}\mathbf{X}_{2i}, \\
\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}}, \\
\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}} &= -\phi(\eta_{2i})\eta_{2i}\mathbf{X}_{2i}^\mathsf{T}\mathbf{X}_{2i} - \frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}}, \\
\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=0)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\mathsf{T}}.
\end{aligned}$$

$$\mathcal{H}_{3,3} = \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \theta_*^2} = \sum_{i=1}^{n} \left\{ y_{1i} y_{2i} \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \theta_*^2} \mathbb{P}(y_{1i}=1, y_{2i}=1) - \left[\frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \theta_*}\right]^2}{\left[\mathbb{P}(y_{1i}=1, y_{2i}=1)\right]^2} \right.$$
$$+ y_{1i}(1-y_{2i}) \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \theta_*^2} \mathbb{P}(y_{1i}=1, y_{2i}=0) - \left[\frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \theta_*}\right]^2}{\left[\mathbb{P}(y_{1i}=1, y_{2i}=0)\right]^2}$$
$$+ (1-y_{1i}) y_{2i} \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \theta_*^2} \mathbb{P}(y_{1i}=0, y_{2i}=1) - \left[\frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \theta_*}\right]^2}{\left[\mathbb{P}(y_{1i}=0, y_{2i}=1)\right]^2}$$
$$\left. + (1-y_{1i})(1-y_{2i}) \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \theta_*^2} \mathbb{P}(y_{1i}=0, y_{2i}=0) - \left[\frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \theta_*}\right]^2}{\left[\mathbb{P}(y_{1i}=0, y_{2i}=0)\right]^2} \right\},$$

where

$$\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \theta_*^2} = \frac{\partial^2 \mathcal{C}(\Phi(\eta_{1i}), \Phi(\eta_{2i}); \theta)}{\partial \theta_*^2},$$
$$\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=0)}{\partial \theta_*^2} = -\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \theta_*^2},$$
$$\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=1)}{\partial \theta_*^2} = -\frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \theta_*^2},$$
$$\frac{\partial^2 \mathbb{P}(y_{1i}=0, y_{2i}=0)}{\partial \theta_*^2} = \frac{\partial^2 \mathbb{P}(y_{1i}=1, y_{2i}=1)}{\partial \theta_*^2}.$$

$$\boldsymbol{\mathcal{H}}_{1,2} = \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal} = \sum_{i=1}^{n} \left\{ y_{1i} y_{2i} \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal} \mathbb{P}(y_{1i}=1, y_{2i}=1) - \frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1} \frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_2}}{\left[\mathbb{P}(y_{1i}=1, y_{2i}=1)\right]^2} \right.$$
$$+ y_{1i}(1-y_{2i}) \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal} \mathbb{P}(y_{1i}=1, y_{2i}=0) - \frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_1} \frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_2}}{\left[\mathbb{P}(y_{1i}=1, y_{2i}=0)\right]^2}$$
$$+ (1-y_{1i}) y_{2i} \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal} \mathbb{P}(y_{1i}=0, y_{2i}=1) - \frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_1} \frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_2}}{\left[\mathbb{P}(y_{1i}=0, y_{2i}=1)\right]^2}$$
$$\left. + (1-y_{1i})(1-y_{2i}) \frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal} \mathbb{P}(y_{1i}=0, y_{2i}=0) - \frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_1} \frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_2}}{\left[\mathbb{P}(y_{1i}=0, y_{2i}=0)\right]^2} \right\},$$

where

$$\begin{aligned}
\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal} &= \frac{\partial h_1}{\partial \boldsymbol{\delta}_2}\phi(\eta_{1i})\mathbf{X}_{1i},\\
\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal},\\
\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal},\\
\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal} &= \frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \boldsymbol{\delta}_2^\intercal}.
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\mathcal{H}}_{1,3} = \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_1 \partial \theta_*} = \sum_{i=1}^{n} \Bigg\{ & y_{1i}y_{2i}\frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \theta_*}\mathbb{P}(y_{1i}=1,y_{2i}=1) - \frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1}\frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \theta_*}}{\left[\mathbb{P}(y_{1i}=1,y_{2i}=1)\right]^2}\\
& + y_{1i}(1-y_{2i})\frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \theta_*}\mathbb{P}(y_{1i}=1,y_{2i}=0) - \frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_1}\frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \theta_*}}{\left[\mathbb{P}(y_{1i}=1,y_{2i}=0)\right]^2}\\
& + (1-y_{1i})y_{2i}\frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \theta_*}\mathbb{P}(y_{1i}=0,y_{2i}=1) - \frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_1}\frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \theta_*}}{\left[\mathbb{P}(y_{1i}=0,y_{2i}=1)\right]^2}\\
& + (1-y_{1i})(1-y_{2i})\frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \partial \theta_*}\mathbb{P}(y_{1i}=0,y_{2i}=0) - \frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_1}\frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \theta_*}}{\left[\mathbb{P}(y_{1i}=0,y_{2i}=0)\right]^2} \Bigg\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \theta_*} &= \frac{\partial h_1}{\partial \theta_*}\phi(\eta_{1i})\mathbf{X}_{1i},\\
\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \theta_*} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \theta_*},\\
\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \theta_*} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \theta_*},\\
\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_1 \partial \theta_*} &= \frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_1 \partial \theta_*}.
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\mathcal{H}}_{2,3} = \frac{\partial^2 \ell(\boldsymbol{\delta}_*)}{\partial \boldsymbol{\delta}_2 \partial \theta_*} = \sum_{i=1}^{n} \Bigg\{ & y_{1i}y_{2i}\frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \theta_*}\mathbb{P}(y_{1i}=1,y_{2i}=1) - \frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_2}\frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \theta_*}}{\left[\mathbb{P}(y_{1i}=1,y_{2i}=1)\right]^2}\\
& + y_{1i}(1-y_{2i})\frac{\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_2 \partial \theta_*}\mathbb{P}(y_{1i}=1,y_{2i}=0) - \frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_2}\frac{\partial \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \theta_*}}{\left[\mathbb{P}(y_{1i}=1,y_{2i}=0)\right]^2}\\
& + (1-y_{1i})y_{2i}\frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \theta_*}\mathbb{P}(y_{1i}=0,y_{2i}=1) - \frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_2}\frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \theta_*}}{\left[\mathbb{P}(y_{1i}=0,y_{2i}=1)\right]^2}\\
& + (1-y_{1i})(1-y_{2i})\frac{\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_2 \partial \partial \theta_*}\mathbb{P}(y_{1i}=0,y_{2i}=0) - \frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_2}\frac{\partial \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \theta_*}}{\left[\mathbb{P}(y_{1i}=0,y_{2i}=0)\right]^2} \Bigg\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \theta_*} &= \frac{\partial h_2}{\partial \theta_*}\phi(\eta_{2i})\mathbf{X}_{2i},\\
\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=0)}{\partial \boldsymbol{\delta}_2 \partial \theta_*} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \theta_*},\\
\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \theta_*} &= -\frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \theta_*},\\
\frac{\partial^2 \mathbb{P}(y_{1i}=0,y_{2i}=0)}{\partial \boldsymbol{\delta}_2 \partial \theta_*} &= \frac{\partial^2 \mathbb{P}(y_{1i}=1,y_{2i}=1)}{\partial \boldsymbol{\delta}_2 \partial \theta_*}.
\end{aligned}$$

The expressions for $\partial \mathcal{C}(\Phi(\eta_{1i}),\Phi(\eta_{2i});\theta)/\partial\theta_*$, $\partial^2\mathcal{C}(\Phi(\eta_{1i}),\Phi(\eta_{2i});\theta)/\partial\theta_*^2$, $h_v$, $\partial h_1/\partial\boldsymbol{\delta}_v$, $\partial h_v/\partial\theta_*$, and $\partial h_2/\partial\boldsymbol{\delta}_2$ for all the copulas considered in this paper are implemented in `SemiParBIVProbit` (Marra & Radice, 2015). These have been derived analytically and verified using numerical derivatives.

# S.3 Asymptotic considerations

In this section, we present some argumentation related to the asymptotic behavior of the penalized maximum likelihood estimator defined as

$$\hat{\boldsymbol{\delta}}_* = \arg\max_{\boldsymbol{\delta}_*} \ell_p(\boldsymbol{\delta}_*),$$

where $\ell_p(\boldsymbol{\delta}_*)$ is given in (4), $\hat{\boldsymbol{\delta}}_* = (\hat{\boldsymbol{\delta}}_1^{\mathsf{T}}, \hat{\boldsymbol{\delta}}_2^{\mathsf{T}}, \hat{\theta}_*)^{\mathsf{T}}$, and the behavior of the ensuing SATE estimator constructed in Section 2.2. Note that $\widehat{\mathrm{SATE}} = \mathrm{SATE}(\hat{\boldsymbol{\delta}}, \mathbf{X})$ is based on $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\delta}}_1^{\mathsf{T}}, \hat{\boldsymbol{\delta}}_2^{\mathsf{T}}, \hat{\theta})^{\mathsf{T}}$ where $\hat{\theta} = \hat{\theta}(\hat{\theta}_*)$ is a proper inverse transformation of parameter $\hat{\theta}_*$ found as result of maximizing the penalized likelihood. We consider the situation in which the spline bases approximating the smooth components $\{b_{vk_vj},\, j = 1,\ldots,J_{vk_v}, k_v = 1,\ldots,K_v, v = 1,2\}$ are of a fixed high dimension, i.e. the $J_{vk_v}$ are fixed. Note that the unknown smooth functions $\{s_{vk_v}, k_v = 1,\ldots,K_v, v = 1,2\}$ may not have an exact representation as linear combinations of given basis functions and consequently the unknown functions and parameters may not be asymptotically identified by their estimators as the sample size grows. However, the case of fixed basis dimensions is of relevance as in practice these have to be fixed and assuming that these are of a high dimension, it is possible to assume heuristically that the approximation bias is negligible compared to estimation variability (e.g., Kauermann, 2005). In this scenario, the method provides estimates which tend in probability to quantities best approximating the unknown functions and parameters in terms of Kullback-Leibler measure. (Recall that the Kullback-Leibler distance between two density functions $f$ and $g$ is defined as $\mathrm{KL}(f||g) = \int_{-\infty}^{\infty} f\log(f/g)$ if $f$ is absolutely continuous with respect to $g$ and 0 otherwise.) Let $L^t$ be the likelihood function for the true model which, in our case, contains the true smooth functions appearing in the linear predictors $\eta_1$ and $\eta_2$ given in (1) and (2) and true value of $\theta_*$ of a given copula, and let $\ell^t$ be the corresponding log-likelihood. Then the Kullback-Leibler distance between the likelihood $L^t$ in the true model and the likelihood $L(\boldsymbol{\delta}_*)$ in the model where

$\sum_{k_1=1}^{K_1} s_{1k_1}(z_{1k_1i})$ and $\sum_{k_2=1}^{K_2} s_{1k_2}(z_{2k_2i})$ are replaced with their spline approximations $\mathbf{B}_{1i}^{\mathsf{T}}\boldsymbol{\beta}_1$ and $\mathbf{B}_{2i}^{\mathsf{T}}\boldsymbol{\beta}_2$ is equal to

$$\mathrm{KL}(L^t||L(\boldsymbol{\delta}_*)) = \mathbb{E}\left(\ell^t - \ell(\boldsymbol{\delta}_*)\right),$$

where the expectation is taken with respect to the true model distribution and $\boldsymbol{\delta}_* = (\boldsymbol{\delta}_1^{\mathsf{T}}, \boldsymbol{\delta}_2^{\mathsf{T}}, \theta_*)^{\mathsf{T}}$. Define parameter vector $\boldsymbol{\delta}_*^0 = (\boldsymbol{\delta}_1^{0\mathsf{T}}, \boldsymbol{\delta}_2^{0\mathsf{T}}, \theta_*^0)^{\mathsf{T}}$ as the minimizer of the above distance, that is

$$\boldsymbol{\delta}_*^0 = \arg\min_{\boldsymbol{\delta}_*} \mathrm{KL}\left(L^t||L(\boldsymbol{\delta}_*)\right),$$

and consequently let $\boldsymbol{\delta}^0 = (\boldsymbol{\delta}_1^{0\mathsf{T}}, \boldsymbol{\delta}_2^{0\mathsf{T}}, \theta^0)^{\mathsf{T}}$ where $\theta^0 = \theta^0(\theta_*^0)$ is a proper transformation of $\theta_*^0$. It follows that $\boldsymbol{\delta}_*^0$ is the maximizer of the expected unpenalized log-likelihood $\ell(\cdot)$ and as a consequence $\mathbb{E}\mathbf{g}(\boldsymbol{\delta}_*^0) = \mathbf{0}$. Remind that $\mathbf{g}(\boldsymbol{\delta}_*)$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*)$ denote the gradient vector and Hessian matrix of $\ell(\cdot)$ calculated at a point $\boldsymbol{\delta}_*$ and let $\mathbf{g}_p(\boldsymbol{\delta}_*) = \mathbf{g}(\boldsymbol{\delta}_*) - \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\delta}_*$ and $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_*) = \boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*) - \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$ be the penalized versions of them. Below, we define classic conditions related to the score vector, Hessian and Fisher information matrix as well as the penalty matrix (see, e.g., Kauermann (2005) who used similar assumptions in the context of survival models). The assumptions are

(A1) $\mathbf{g}(\boldsymbol{\delta}_*^0) = O_P(n^{1/2})$,
(A2) $\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0) = O(n)$,
(A3) $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0) - \mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0) = O_P(n^{1/2})$,
(A4) $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}} = o\left(n^{1/2}\right)$, where $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$ is defined in Section 2.3.

Conditions (A1) and (A3) are the assumptions of $n^{1/2}$ asymptotics (e.g., Barndorff-Nielsen & Cox, 1989). Note that, as the $n$ observations are assumed to be independent, $\mathbf{g}(\boldsymbol{\delta}_*^0)$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0)$ are made up of sums of independent random variables. Assumptions (A1) and (A3) imply that, given the model, the average values $\frac{1}{n}\mathbf{g}(\boldsymbol{\delta}_*^0)$ and $\frac{1}{n}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0)$ over the random sample converge in probability to their expected values at the rate $n^{-1/2}$. Condition (A4) can be equivalently formulated as $\lambda_{vk_v} = o\left(n^{1/2}\right)$ for $k_v = 1, \ldots, K_v$, $v = 1, 2$, assuming that the matrices $\mathbf{S}_{vk_v}$ are asymptotically bounded. This assumption is rather weak as it allows the smoothing parameters to grow as the sample size increases, at a rate smaller than $n^{1/2}$. In fact, the sequence $\hat{\boldsymbol{\lambda}}$ based on the mean squared error criterion described in subsection 2.3.2 is bounded in probability (e.g., Kauermann, 2005).

**Theorem 1.** Under conditions (A1)-(A4) we have

$$\hat{\boldsymbol{\delta}}_* - \boldsymbol{\delta}_*^0 = O_P(n^{-1/2}) \text{ as } n \to \infty.$$

**Remark 1.** Note that the above theorem states the consistency and its rate for the vector of parameters $\hat{\boldsymbol{\delta}}_* = (\hat{\boldsymbol{\delta}}_1^{\mathsf{T}}, \hat{\boldsymbol{\delta}}_2^{\mathsf{T}}, \hat{\theta}_*)^{\mathsf{T}}$ which includes the transformed dependence parameter $\theta_*$ used in optimization (see Section 2.3 and Table 1). However, if we assume that the inverse transformation $\theta_* \mapsto \theta$ is differentiable then by using the mean value theorem we immediately obtain that the above result holds also for the vector of coefficients $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\delta}}_1^{\mathsf{T}}, \hat{\boldsymbol{\delta}}_2^{\mathsf{T}}, \hat{\theta})^{\mathsf{T}}$ which contains the copula dependence parameter on the original scale.

Let $\mathrm{SATE}^0$ be equal to

$$\mathrm{SATE}^0 = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{P}^0(y_{2i}=1|y_{1i}=1) - \mathbb{P}^0(y_{2i}=1|y_{1i}=0)\right\},$$

where

$$\begin{aligned}\mathbb{P}^0(y_{2i}=1|y_{1i}=1) &= \frac{\mathcal{C}\left(\Phi(\eta_{2i}^{0\,(y_{1i}=1)}),\Phi(\eta_{1i}^0);\theta^0\right)}{\Phi(\eta_{1i}^0)},\\ \mathbb{P}^0(y_{2i}=1|y_{1i}=0) &= \frac{\Phi(\eta_{2i}^{0\,(y_{1i}=0)}) - \mathcal{C}\left(\Phi(\eta_{2i}^{0\,(y_{1i}=0)}),\Phi(\eta_{1i}^0);\theta^0\right)}{1-\Phi(\eta_{1i}^0)},\end{aligned} \tag{10}$$

$\eta_{1i}^0 = \mathbf{X}_{1i}^\mathsf{T}\boldsymbol{\delta}_1^0$ and $\eta_{2i}^0 = \mathbf{X}_{2i}^\mathsf{T}\boldsymbol{\delta}_2^0$ and $\theta^0 = \theta^0(\theta_*^0)$ is the appropriate transformation of parameter $\theta_*^0$. In order to prove consistency for the estimator of the $\mathrm{SATE}$ we introduce the additional assumption that the

(A5) probabilities (10) are differentiable as functions of $\boldsymbol{\delta}$ and their gradients are bounded in the neighborhood of $\boldsymbol{\delta}^0$, uniformly for all $\mathbf{x}_i = (\mathbf{X}_{1i}^\mathsf{T}, \mathbf{X}_{2i}^\mathsf{T})^\mathsf{T}$.

**Theorem 2.** If conditions (A1)-(A5) hold then

$$\widehat{\mathrm{SATE}} - \mathrm{SATE}^0 = O_P(n^{-1/2}) \text{ as } n\to\infty,$$

where $\widehat{\mathrm{SATE}} = \mathrm{SATE}(\hat{\boldsymbol{\delta}}, \mathbf{X})$ as defined in Section 2.2.

*Proof of Theorem 1.* We show that the following approximation holds

$$\hat{\boldsymbol{\delta}}_* - \boldsymbol{\delta}_*^0 \approx \left(-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0) + \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\right)^{-1}\left(\mathbf{g}(\boldsymbol{\delta}_*^0) - \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\delta}_*^0\right), \tag{11}$$

which implies the asymptotic consistency of $\hat{\boldsymbol{\delta}}_*$ at the rate $n^{-1/2}$. We adopt the argumentation used in the theory of maximum likelihood estimation (e.g., McCullagh, 1987) which involves a Taylor expansion of the score in the neighborhood of $\boldsymbol{\delta}_*^0$. A similar approach was used by Kauermann (2005) and Kauermann et al. (2009) in the context of penalized spline smoothing. For simplicity of notation, we omit all terms of order higher than 1 and assume that higher order derivatives of the log-likelihood behave in a similar manner as those defined in (A1)-(A3).
The first-order Taylor expansion of $\mathbf{g}_p(\cdot)$ around $\boldsymbol{\delta}_*^0$ implies

$$\mathbf{g}_p(\hat{\boldsymbol{\delta}}_*) = \mathbf{g}_p(\boldsymbol{\delta}_*^0) + \boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_*^0)(\hat{\boldsymbol{\delta}}_* - \boldsymbol{\delta}_*^0) + (\text{higher order terms}),$$

which, after using the fact that $\mathbf{g}_p(\hat{\boldsymbol{\delta}}_*) = \mathbf{0}$ and inverting the above series (e.g., Barndorff-Nielsen & Cox, 1989), leads to

$$\hat{\boldsymbol{\delta}}_* - \boldsymbol{\delta}_*^0 = -\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_*^0)^{-1}\left(\mathbf{g}(\boldsymbol{\delta}_*^0) - \tilde{\mathbf{S}}_\lambda\boldsymbol{\delta}_*^0\right) + \ldots$$

We then decompose $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_*^0)$ as

$$\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_*^0) = \left(\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0) - \mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0)\right) + \left(\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0) - \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\right) = \mathbf{R} - \mathbf{F}(\boldsymbol{\lambda}),$$

where $\mathbf{R} = \boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0) - \mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0)$ represents a stochastic error and $\mathbf{F}(\boldsymbol{\lambda}) = -\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0) + \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$ is the penalized Fisher information matrix. Now, let $f(\cdot) = (\cdot - \mathbf{F}(\boldsymbol{\lambda}))^{-1}$ be an auxiliary function of a matrix argument. Using the Taylor expansion of $f(\mathbf{R})$ around $f(\mathbf{0})$, we obtain

$$\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_*^0)^{-1} = -\mathbf{F}(\boldsymbol{\lambda})^{-1} - \mathbf{F}(\boldsymbol{\lambda})^{-1}\mathbf{R}(\mathbf{F}(\boldsymbol{\lambda})^{-1})^{\mathsf{T}} + \dots$$

Now, assumptions (A2)-(A4) imply

$$\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_*^0)^{-1} = -\mathbf{F}(\boldsymbol{\lambda})^{-1}\left(\mathbb{I} + \mathbf{R}\mathbf{F}(\boldsymbol{\lambda})^{-1} + \dots\right) = -\mathbf{F}(\boldsymbol{\lambda})^{-1}\left(\mathbb{I} + O_P(n^{-1/2})\right),$$

where $\mathbb{I}$ is an identity matrix. Thus

$$\hat{\boldsymbol{\delta}}_* - \boldsymbol{\delta}_*^0 = \mathbf{F}(\boldsymbol{\lambda})^{-1}\left(\mathbf{g}(\boldsymbol{\delta}_*^0) - \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\delta}_*^0\right)(\mathbb{I} + o_P(1)) + \dots, \tag{12}$$

which proves (11) and hence

$$\hat{\boldsymbol{\delta}}_* - \boldsymbol{\delta}_*^0 = O_P\left(n^{-1/2}\right) \text{ as } n \to \infty. \tag{13}$$

**Remark 2. (a)** From approximation (12), the asymptotic bias and covariance matrix of $\hat{\boldsymbol{\delta}}_*$ can be derived. Specifically,

$$\mathbf{bias}(\hat{\boldsymbol{\delta}}_*) = \mathbb{E}\left(\hat{\boldsymbol{\delta}}_* - \boldsymbol{\delta}_*^0\right) \approx -\mathbf{F}(\boldsymbol{\lambda})^{-1}\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\delta}_*^0,$$

where the property $\mathbb{E}\mathbf{g}(\boldsymbol{\delta}_*^0) = \mathbf{0}$ of $\boldsymbol{\delta}_*^0$ has been used, and

$$\mathbf{Cov}(\hat{\boldsymbol{\delta}}_*) \approx -\mathbf{F}(\boldsymbol{\lambda})^{-1}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0)\mathbf{F}(\boldsymbol{\lambda})^{-1}, \tag{14}$$

which follows from the fact that $\mathbf{Cov}(\mathbf{g}(\boldsymbol{\delta}_*^0)) = -\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0)$. In addition, conditions (A2) and (A4) imply that

$$\mathbf{bias}(\hat{\boldsymbol{\delta}}_*) = o(n^{-1/2}) \text{ and } \mathbf{Cov}(\hat{\boldsymbol{\delta}}_*) = O(n^{-1}).$$

**(b)** Assumption (A4) implies that

$$\sqrt{n}\mathbf{Cov}(\hat{\boldsymbol{\delta}}_*) \approx \left\{\frac{1}{\sqrt{n}}\mathbb{E}\left[-\boldsymbol{\mathcal{H}}(\delta_*^0)\right]\right\}^{-1}$$

and

$$\sqrt{n}\mathbf{V}_{\boldsymbol{\delta}_*} \approx \left(-\frac{1}{\sqrt{n}}\boldsymbol{\mathcal{H}}(\delta_*^0)\right)^{-1},$$

where $\mathbf{V}_{\boldsymbol{\delta}_*} = -\boldsymbol{\mathcal{H}}_p^{-1}$ is the Bayesian approximation of the covariance matrix of $\hat{\boldsymbol{\delta}}_*$ mentioned in Section 2.4. Thus, the frequentist asymptotic approximation (14) and the Bayesian result become equivalent as the sample size $n$ grows to $\infty$.

**(c)** As $\mathbf{g}(\boldsymbol{\delta}_*^0)$ is a sum of i.i.d. components, it follows that $\left(-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0)\right)^{-1/2}\mathbf{g}(\boldsymbol{\delta}_*^0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{I})$. Hence, approximation (12) also implies asymptotic normality of the normalized estimator $\hat{\boldsymbol{\delta}}_*$. The

asymptotic normality holds also for the vector of parameters $(\hat{\boldsymbol{\delta}}_1^\mathsf{T}, \hat{\boldsymbol{\delta}}_2^\mathsf{T}, \hat{\theta})^\mathsf{T}$ containing the dependence parameter $\theta$ on the original scale. However, as for some copulas parameter $\theta$ is bounded, the normal approximation may not be accurate for small sample sizes.

*Proof of Theorem 2.* Recall that $\widehat{\mathrm{SATE}} = \mathrm{SATE}(\hat{\boldsymbol{\delta}}, \mathbf{X})$ where $\mathrm{SATE}(\boldsymbol{\delta}, \mathbf{X})$ can be expressed as $\frac{1}{n}\sum_{i=1}^n \mathrm{sate}(\boldsymbol{\delta}, \mathbf{x}_i)$, with $\mathrm{sate}(\boldsymbol{\delta}, \mathbf{x}_i)$ determined by

$$\frac{\mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=1)}); \theta\right)}{\Phi(\eta_{1i})} - \frac{\Phi(\eta_{2i}^{(y_{1i}=0)}) - \mathcal{C}\left(\Phi(\eta_{1i}), \Phi(\eta_{2i}^{(y_{1i}=0)}); \theta\right)}{1 - \Phi(\eta_{1i})}.$$

The mean value theorem yields

$$\mathrm{sate}(\hat{\boldsymbol{\delta}}, \mathbf{x}_i) = \mathrm{sate}(\boldsymbol{\delta}^0, \mathbf{x}_i) + \frac{\partial}{\partial \boldsymbol{\delta}}\mathrm{sate}(\tilde{\boldsymbol{\delta}}, \mathbf{x}_i)^\mathsf{T}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0),$$

for some $\tilde{\boldsymbol{\delta}} = (1-c)\boldsymbol{\delta}^0 + c\hat{\boldsymbol{\delta}}$, $c > 0$, where $\frac{\partial}{\partial \boldsymbol{\delta}}\mathrm{sate}(\tilde{\boldsymbol{\delta}}, \mathbf{x}_i)$ is the gradient vector of $\mathrm{sate}(\cdot, \mathbf{x}_i)$ expressed as a function of $\boldsymbol{\delta}$ calculated at a point $\boldsymbol{\delta} = \tilde{\boldsymbol{\delta}}$, for $i = 1, \ldots, n$. Thus,

$$\begin{aligned}\widehat{\mathrm{SATE}} &= \frac{1}{n}\sum_{i=1}^n \mathrm{sate}(\boldsymbol{\delta}^0, \mathbf{x}_i) + \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\delta}}\mathrm{sate}(\tilde{\boldsymbol{\delta}}, \mathbf{x}_i)^\mathsf{T}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0) \\ &= \mathrm{SATE}^0 + \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\delta}}\mathrm{sate}(\tilde{\boldsymbol{\delta}}, \mathbf{x}_i)^\mathsf{T}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0) \qquad (15)\end{aligned}$$

As for the second term in (15), Schwarz's inequality implies

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\delta}}\mathrm{sate}(\tilde{\boldsymbol{\delta}}, \mathbf{x}_i)^\mathsf{T}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0) \leq \frac{1}{n}\sum_{i=1}^n \left|\left|\frac{\partial}{\partial \boldsymbol{\delta}}\mathrm{sate}(\tilde{\boldsymbol{\delta}}, \mathbf{x}_i)\right|\right| ||\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0||.$$

Given the assumption that $\frac{\partial}{\partial \boldsymbol{\delta}}\mathrm{sate}(\cdot, \mathbf{x}_i)$ is bounded in the neighborhood of $\boldsymbol{\delta}^0$ uniformly for all $\mathbf{x}_i$ and that $||\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0|| = O_P(n^{-1/2})$ proved in (13) (see also Remark 1), the assertion follows.

**Remark 3.** Expression (14) for the asymptotic covariance matrix of $\hat{\boldsymbol{\delta}}$ can be used to construct the asymptotic variance of $\widehat{\mathrm{SATE}}$ using the delta method, namely

$$\mathrm{Var}\,\widehat{\mathrm{SATE}} \approx -\frac{\partial \mathrm{SATE}(\boldsymbol{\delta}^0, \mathbf{X})}{\partial \boldsymbol{\delta}}^\mathsf{T} \mathbf{F}(\boldsymbol{\lambda})^{-1}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_*^0)\mathbf{F}(\boldsymbol{\lambda})^{-1}\frac{\partial \mathrm{SATE}(\boldsymbol{\delta}^0, \mathbf{X})}{\partial \boldsymbol{\delta}}\left(\frac{\partial \theta}{\partial \theta_*}(\theta_*^0)\right)^2,$$

which is equivalent in the limit to expression (3) given in Section 2.2, as motivated in Remark 2(b). Moreover, it follows from the delta method and Remark 2(c) that the normalized estimator $\widehat{\mathrm{SATE}}$ is asymptotically normal. Here again, it is worth noting that the normal approximation would not be accurate for relatively small sample sizes for copulas having bounded scope of $\theta$.

## S.4 Simulations

To assess the empirical effectiveness of the proposed methodology, we conducted a simulation study. Following a reviewer's suggestion, we used the findings of Section 3 and employed a smaller set of covariates and model settings to keep the study feasible. In particular, we included two binary variables and two continuous regressors in both the treatment and outcome equations with effects and covariate range values that were similar to some of those found in Sections 3.3.3 and 3.3.4. We also simulated the model errors using a Gumbel distribution with low and high dependence parameter: $\theta$ was set to 1.18 (which is what we obtained in the case study) and 7. Sample sizes were set to 5000 and 1000 and the number of replicates to 250. The models employed were Gaussian, Student-$t_3$, Frank, Clayton, Gumbel and Joe and their rotated versions. The `R` code used to simulate the data was

```
library(copula)
teta  <- 1.18 # or 7
n     <- 5000 # or 10000
n.rep <- 250
myCop <- archmCopula(family = "gumbel", dim = 2, param = teta)
bivg  <- mvdc(copula = myCop, c("norm", "norm"),
              list(list(mean = 0, sd = 1),
                   list(mean = 0, sd = 1)) )
u  <- rMvdc(n, bivg)
x1 <- runif(n,18,66)
x2 <- runif(n,10,70)
x3 <- runif(n,0,20)
x4 <- round(runif(n))
x5 <- round(runif(n))
s1 <- function(x) -0.2*sin(pi/46*x)
s2 <- function(x) -0.0004*(x+0.01*x^3)
s3 <- function(x) 0.0006*exp(0.1*x)
s4 <- function(x) 0.03*x
y1 <- ifelse(0.7 + s1(x1) + s2(x2) + 0.6*x4 - 0.4*x5 + u[,1] > 0, 1, 0)
y2 <- ifelse(-1.5 - 0.18*y1 + s3(x1) + s4(x3) - x4 + 0.75*x5 + u[,2] > 0, 1, 0)
```

The models were fitted using `SemiParBIVProbit(list(eq1,eq2), BivD=D)`, where `eq1` and `eq2` were specified according to the simulated `y1` and `y2` above, and `D` was equal to `"N"`, `"T"`, `"F"`, `"C0"`, `"C90"`, `"C180"`, `"C270"`, `"J0"`, `"J90"`, `"J180"`, `"J270"`, `"G0"`, `"G90"`, `"G180"` and `"G270"`. The sample average treatment effect (with interval obtained by posterior simulation or delta method) for each replicate and fitted model was extracted using `AT()` from the package `SemiParBIVProbit`, whereas the information criteria were obtained using `AIC()` and `BIC()`. For each model and case considered, we calculated the percentage bias and root mean squared error (RMSE) for $\widehat{\text{SATE}}$, coverage probabilities of the two types of intervals for SATE, and proportions of times that the models were selected by $AIC$ and $BIC$ over the replicates. For each replicate, we also stored the estimated smooth functions evaluated at 200

fixed values in the ranges of the respective covariates (e.g., Wiesenfarth & Kneib, 2011).

In Table 5, we have a total of four cases to which we refer to as Case 1 ($\theta = 1.18$, $n = 5000$), Case 2 ($\theta = 1.18$, $n = 10000$), Case 3 ($\theta = 7$, $n = 5000$) and Case 4 ($\theta = 7$, $n = 10000$). In all cases, the models which can only account for a negative dependence do not obviously exhibit a good performance. In Case 1, $\text{Gumbel}_0$ is outperformed by Frank and $\text{Joe}_{180}$, although the biases of these three models are negligible and the RMSEs do not differ. In Case 2, the performance of all models but $\text{Gumbel}_0$ worsens indicating that as the sample size grows the correct model tends to outperform the competing ones. In these two cases the choice of the correct copula model based on an empirical sample is extremely difficult and the information criteria are not able to discriminate between $\text{Gumbel}_0$ and some of the competing models. As explained in Section 3.3.2, in the presence of a low association the copula models entail very similar distributions, hence they can not be easily separated. In Case 3 and Case 4, the preferred model is $\text{Gumbel}_0$. In these instances, the association between the treatment and outcome equations is strong which means that it is easier to select the correct model as the different copula models entail different distributions. For instance, by comparing the Gaussian copula model (the traditional choice) to $\text{Gumbel}_0$ (the copula model used to simulate the data), the performance of the latter is superior in terms of both bias and variability. This illustrates that erroneously modeling the dependence affects the quantity of interest (SATE) in terms of bias and efficiency. The empirical coverage probabilities of the intervals for SATE calculated by posterior simulation and delta method are essentially identical and very close to the nominal $95\%$ level when the bias is negligible; as the bias increases the coverage worsens since the interval is centered on a biased point estimate. Figure 4 shows the estimated smooth functions associated to all replications for Case 3 (which seemed to be slightly more challenging than the other cases as more iterations were needed to achieve convergence). Overall, the estimated curves recover the underlying functions fairly well, with some exceptions in which the estimated functions are rougher than they should be.

| | | $n = 5000$ | | | | | | $n = 10000$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Bias (%) | RMSE | AIC (%) | BIC (%) | PS | DM | Bias (%) | RMSE | AIC (%) | BIC (%) | PS | DM |
| $\theta = 1.18$ | Gaussian | 1.3 | 0.009 | 12 | 13 | 0.94 | 0.95 | 1.5 | 0.007 | 14 | 14 | 0.95 | 0.95 |
| | Student-$t_3$ | 4.0 | 0.010 | 5 | 4 | 0.94 | 0.95 | 4.1 | 0.008 | 4 | 3 | 0.94 | 0.95 |
| | Frank | 0.3 | 0.010 | 8 | 9 | 0.95 | 0.96 | 0.5 | 0.007 | 7 | 7 | 0.94 | 0.95 |
| | $\text{Clayton}_0$ | 0.8 | 0.009 | 11 | 10 | 0.95 | 0.95 | 1.2 | 0.007 | 0 | 1 | 0.95 | 0.94 |
| | $\text{Clayton}_{90}$ | -12.7 | 0.011 | 7 | 7 | 0.82 | 0.80 | -13.4 | 0.009 | 8 | 8 | 0.80 | 0.79 |
| | $\text{Clayton}_{180}$ | -3.1 | 0.010 | 4 | 5 | 0.94 | 0.95 | -3.3 | 0.007 | 17 | 18 | 0.95 | 0.95 |
| | $\text{Clayton}_{270}$ | -13.4 | 0.011 | 3 | 4 | 0.81 | 0.80 | -13.5 | 0.009 | 7 | 7 | 0.81 | 0.80 |
| | $\text{Gumbel}_0$ | 0.4 | 0.010 | 16 | 15 | 0.95 | 0.95 | 0.4 | 0.007 | 19 | 18 | 0.96 | 0.95 |
| | $\text{Gumbel}_{90}$ | -13.1 | 0.011 | 0 | 1 | 0.81 | 0.82 | -13.5 | 0.009 | 0 | 0 | 0.78 | 0.81 |
| | $\text{Gumbel}_{180}$ | 2.2 | 0.010 | 12 | 11 | 0.96 | 0.95 | 2.7 | 0.007 | 3 | 4 | 0.95 | 0.95 |
| | $\text{Gumbel}_{270}$ | -12.8 | 0.011 | 0 | 0 | 0.81 | 0.81 | -13.5 | 0.009 | 0 | 0 | 0.80 | 0.79 |
| | $\text{Joe}_0$ | -3.7 | 0.010 | 19 | 18 | 0.94 | 0.95 | -4.0 | 0.007 | 19 | 20 | 0.95 | 0.94 |
| | $\text{Joe}_{90}$ | -13.6 | 0.012 | 2 | 1 | 0.83 | 0.83 | -13.6 | 0.009 | 1 | 0 | 0.81 | 0.80 |
| | $\text{Joe}_{180}$ | 0.2 | 0.009 | 1 | 2 | 0.95 | 0.94 | 0.7 | 0.007 | 1 | 0 | 0.95 | 0.94 |
| | $\text{Joe}_{270}$ | -12.2 | 0.011 | 0 | 0 | 0.82 | 0.80 | -13.2 | 0.009 | 0 | 0 | 0.77 | 0.80 |
| $\theta = 7$ | Gaussian | -8.2 | 0.022 | 17 | 16 | 0.91 | 0.92 | -8.3 | 0.021 | 14 | 14 | 0.91 | 0.90 |
| | Student-$t_3$ | -10.1 | 0.025 | 14 | 13 | 0.91 | 0.91 | -10.3 | 0.025 | 15 | 14 | 0.91 | 0.90 |
| | Frank | -8.3 | 0.022 | 2 | 2 | 0.91 | 0.90 | -8.3 | 0.021 | 3 | 3 | 0.90 | 0.90 |
| | $\text{Clayton}_0$ | -8.2 | 0.022 | 7 | 6 | 0.92 | 0.91 | -8.2 | 0.021 | 4 | 4 | 0.90 | 0.91 |
| | $\text{Clayton}_{90}$ | -18.0 | 0.046 | 3 | 4 | 0.71 | 0.70 | -17.9 | 0.046 | 2 | 2 | 0.70 | 0.71 |
| | $\text{Clayton}_{180}$ | -8.6 | 0.023 | 3 | 2 | 0.91 | 0.91 | -8.7 | 0.022 | 15 | 16 | 0.91 | 0.92 |
| | $\text{Clayton}_{270}$ | -18.0 | 0.046 | 3 | 3 | 0.71 | 0.72 | -18.1 | 0.046 | 2 | 1 | 0.70 | 0.72 |
| | $\text{Gumbel}_0$ | -6.4 | 0.019 | 19 | 20 | 0.92 | 0.93 | -4.5 | 0.016 | 25 | 24 | 0.93 | 0.94 |
| | $\text{Gumbel}_{90}$ | -18.0 | 0.046 | 0 | 0 | 0.70 | 0.72 | -18.0 | 0.046 | 0 | 1 | 0.70 | 0.72 |
| | $\text{Gumbel}_{180}$ | -8.2 | 0.022 | 13 | 14 | 0.91 | 0.91 | -8.2 | 0.021 | 2 | 1 | 0.90 | 0.91 |
| | $\text{Gumbel}_{270}$ | -18.0 | 0.046 | 0 | 0 | 0.71 | 0.70 | -17.9 | 0.046 | 0 | 1 | 0.71 | 0.70 |
| | $\text{Joe}_0$ | -8.6 | 0.023 | 16 | 15 | 0.92 | 0.91 | -8.7 | 0.022 | 18 | 17 | 0.92 | 0.92 |
| | $\text{Joe}_{90}$ | -18.0 | 0.046 | 0 | 0 | 0.71 | 0.73 | -18.2 | 0.046 | 0 | 1 | 0.71 | 0.72 |
| | $\text{Joe}_{180}$ | -8.2 | 0.022 | 3 | 4 | 0.90 | 0.91 | -8.1 | 0.021 | 0 | 0 | 0.92 | 0.91 |
| | $\text{Joe}_{270}$ | -18.0 | 0.046 | 0 | 1 | 0.72 | 0.71 | -18.0 | 0.046 | 0 | 0 | 1.71 | 0.70 |

Table 5: Percentage biases and RMSEs for $\widehat{\text{SATE}}$, percentage frequency at which each copula model was selected by $AIC$ and $BIC$ and empirical coverage probabilities of the intervals for SATE calculated by posterior simulation (PS) and delta method (DM). Data were simulated using a Gumbel copula with normal margins.

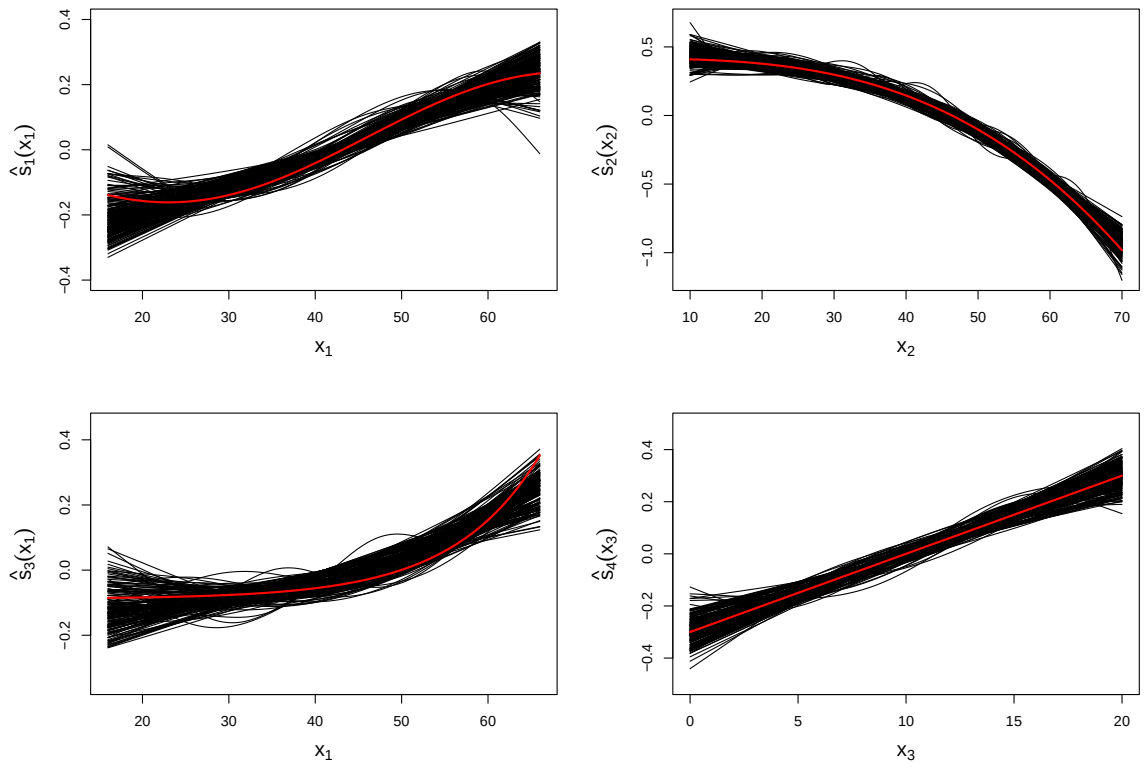


Figure 4: Estimated smooth functions for $s_1(x_1)$, $s_2(x_2)$, $s_3(x_1)$ and $s_4(x_3)$ obtained when employing the Gumbel$_0$ model for Case 3 (i.e., $\theta = 7$, $n = 5000$). Results are plotted on the scale of the linear predictors. The black lines in each plot represent the estimated smooth functions from all replications, evaluated at 200 fixed values in the range of the respective covariate. The true functions are represented by the red solid lines.