Generalised endogenous switching regression models and multiple imputation with applications in health economics

Generalised endogenous switching regression models and multiple imputation with applications in health economics

Jose Maria Camarena Brenes

A thesis submitted for the degree of Doctor of Philosophy

Department of Economics, Mathematics and Statistics

Birkbeck, University of London

September, 2023

Declaration

I hereby declare that, except where specific reference is made to the work of others, the contents of this thesis are my own and have not been submitted in whole or in part for consideration for any other degree or qualification at Birkbeck, University of London, or any other university.

At the time of writing, a working paper based on the contents of Chapters 2 and 3 was made available online through the Birkbeck repository (BIROn) and was subject to a claim of plagiarism by Professor Giampiero Marra and Professor Rosalba Radice. The issue was regulated by Professor Ron Smith and he can confirm that the claim was resolved and suitably comprehensive references to their work have been included in the text.

The contents of Chapter 4 are based on a joint work published in the following paper: *"Gomes, M, Radice, R, Camarena Brenes, J, Marra, G. - Copula selection models for non-Gaussian outcomes that are missing not at random. Statistics in Medicine. 2019; 38: 480-496".* I participated in the draft and publication of the paper. Specifically, my contribution to the paper was the MI approach presented in the chapter, completed under the supervision of Professor Rosalba Radice and Professor Giampiero Marra.

Abstract

In this thesis, we present extensions of the endogenous switching regression (ESR) model with an application in health economics; and an approach to multiple imputation (MI) for variables assumed to be missing not a random.

We first specify a semi-parametric ESR model where the predictors are represented using penalized regression splines, while retaining the distributional assumptions of the classical approach. We then present copula-based ESR models where the bivariate joint distributions in the model are specified using copula functions and their univariate components are specified in terms of parametric distributions. Parameter estimation and inference utilise a well-established penalized likelihood framework. We investigate insurance uptake to cover out-of-pocket expenses of prescription drugs in over 65 years olds from the United States. Our findings using the semi-parametric approach reveal evidence of self-selection into insurance and that some of the determinant factors of expenditures exhibit varying degrees of non-linear associations. An assessment of the dependence structures using the copula-based approach suggests that large values of out-of-pocket expenditures are accompanied by higher chances of having supplementary insurance however, low expenditures do not necessarily imply lower chances of having extra insurance. These features cannot be adequately captured using the classical model specification.

We also present a MI approach that obtains plausible imputed values for a variable assumed to be missing not a random and not restricted to be Gaussian. The approach is derived from a copula-based specification of the sample selection model. We re-examine the non-randomised component of the REFLUX study to evaluate the effect of surgery on patient's health status under several modelling assumptions. We find that estimates of the effect of surgery are significant, regardless of the modelling approach. Estimates obtained using MI are very similar to those based on the copula model and, in some instances, they have slightly smaller standard errors.

Acknowledgements

First of all, I would like to express my gratitude to my supervisors, Professor Walter Beckert and Dr. Swati Chandna, for their continuous support, patience, and encouragement throughout this journey. In addition, to Professor Ron Smith for his support and for being an inspiration to many. I also thank Professor Rosalba Radice and Professor Giampiero Marra for their initial supervision and for introducing me to their research, to sample selection models, and to their R package (GJRM). Furthermore, I thank Dr. Seonaidh Cotton for granting me access to use data from the REFLUX study used in Chapter 4 and Dr. Manuel Gomes for providing the data set.

I would also like to acknowledge my colleagues and friends at Birkbeck: Federico Corrado, Rubens Morita, Carlo Piccari, Charisios Grivas, and Angela Aguele for all the moments we have shared together. Lastly, I would like to thank my partner, Lukasz Longosz, for his understanding and encouragement during all these years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Statistical analyses of non-randomised studies subject to sample selection are common in research. The problem of sample selection appears when the data to be analysed consist of a non-random sample of the population under study which, if not accounted for, leads to biased inferences.

Selected samples may appear as a consequence of self-selection or of non-random selection and their analysis usually requires to model the selection mechanism explicitly. For instance, individuals may self-select into a treatment or join a programme in order to obtain some benefit from doing so based on characteristics that are observed as well as unobserved by the researcher. In this situation, the assignment mechanism becomes non-ignorable and the treatment is usually referred to as endogenous[1]. Non-random selection arises, for example, in situations where the researcher observes a response of interest only for individuals that take part in a study and participation is driven by some observed and unobserved individual characteristics. The missing data literature refers to the outcome of interest as being missing not at random.

The most popular models that correct for the bias arising from the analysis of selected samples are sample selection (SS; Heckman, 1974, 1976, 1979), dummy endogenous variable (DEV; Heckman, 1978), and endogenous switching regression (ESR; Roy, 1951).

The ESR model was proposed by Roy (1951) to study the factors that influence individuals when choosing between hunting or fishing as a profession, and can be understood as a generalization of SS and DEV models. In SS models, the outcome of interest is partially observed and the selection mechanism determines whether the response is observed or missing. In DEV models, the outcome of interest is fully observed and the selection or assignment mechanism determines the

[1] In a broad sense, the concept of endogeneity in a statistical model is attributed to describing situations where an explanatory variable is correlated with the random component of the model, for example, due to omitted variables that are unaccounted for, predictors measured with errors, or simultaneity (Wooldridge, 2010, p. 54).

level of the treatment. In contrast, in ESR models the researcher observes the outcome of interest at both levels of the treatment however, unlike DEV models, ESR allows for the effects of all the covariates that explain the outcome to vary according to the treatment variable, and not just the intercept. These models were originally specified using a system of linear regression equations where the error terms are assumed to be multivariate normally distributed. Throughout this thesis, we will refer to models using this specification as the classical approaches.

Generalizations and extensions of the aforementioned models stem from criticisms to their lack of robustness to distributional and/or functional form misspecification. Most of these extensions have been proposed in the context of SS and EDV models, using non- or semi-parametric methods, modelling the joint distribution using flexible multivariate distributions, or using copula functions (see, for example, Pigini, 2015; Puhani, 2000; Vella, 1998, for reviews).

The contents of this thesis revolve around some aspects of ESR and SS models: extensions of the ESR model and their application to data from the health economics literature; and the inclusion of flexible SS models into the multiple imputation framework (Rubin, 1977, 1978, 1987) to deal with variables assumed to be missing not at random. Specifically, the objectives of this thesis are:

(i) Present extensions of the ESR model that relax the functional form specification of the predictors and the distributional assumptions of the classical approach.

(ii) Investigate the effects of insurance uptake, and other socio-economic and health-related characteristics, on out-of-pocket expenditures for prescribed drugs in individuals that are over 65 years old in the United Stated using data from the Medical Expenditure Panel Survey (MEPS).

(iii) Present a multiple imputation method, derived from a copula-based specification of the SS model, that obtains plausible values for a partially observed continuous variable assumed to be missing not at random and not restricted to be Gaussian.

To relax the functional form specification of the predictors, the ESR model is formulated using a penalized regression spline framework where the deterministic components are defined using a semi-parametric GAM-style specification (Generalized Additive Models; Hastie & Tibshirani, 1990; Wood, 2017). The distributional assumptions are relaxed by embedding the model into a distributional regression framework and specifying the stochastic model components in terms of parametric bivariate copula functions and their univariate margins. The marginal distributions are specified using the GAMLSS framework (Generalized Additive Models for Location, Scale,

and Shape; Rigby & Stasinopoulos, 2005), where each distribution parameter can be modelled using a flexible additive predictor of explanatory variables. Parameter estimation utilises a well-established penalized likelihood framework and inference follows from the Bayesian interpretation of the penalization process (Marra et al., 2017; Marra & Radice, 2019; Wojtyś et al., 2018).

We investigate insurance uptake in individuals over 65 years of age to cover for out-of-pocket prescribed drugs expenditures first by using the semi-parametric model under the classical distributional assumptions and then using the copula-based approach. The semi-parametric framework intends to capture and correct for self-selection into insurance while accounting for several continuous covariates such as age, income, and number of chronic conditions, which may have a non-linear association with out-of-pocket expenditures. Using the copula-based models, we also intent to assess whether the response of interest is best described by several parametric distributions that are used to model healthcare expenditures (Deb & Norton, 2018; Manning et al., 2005; Manning & Mullahy, 2001; Mullahy, 2009), and whether the dependence structure in the data can be best captured by using several copula functions other than the Gaussian.

The multiple imputation approach draws plausible values of a variable subject to missingness from a density derived from a copula-based specification of the SS model. The procedure allows to impute values that are suspected to be missing not at random under several distributional assumptions and constitutes a valuable method that can be incorporated into a fully conditional specification strategy to multiple imputation. We re-examine the non-randomised component of the REFLUX study (Grant et al., 2008, 2013; Gomes et al., 2019, 2020) in order to evaluate the effect of surgery on health status among individuals with gastro-oesophageal reflux disease, using several assumptions about the missingness mechanism and the distribution of the response.

Using simulations, we find that the semi-parametric and copula-based extensions of the ESR model and the multiple imputation approach obtain parameter estimates that are near their true values and are less variable as the sample size increases. In particular, the semi-parametric approach appears to mitigate the effects of residual confounding that result from not modelling the continuous covariates flexibly (Benedetti & Abrahamowicz, 2004; Radice et al., 2016; Slama & Werwatz, 2005); the copula-based models provide reasonable results under mild model misspecification; and the multiple imputation approach performs similarly to the copula-based SS model that is based on, and appears to yield parameter estimates that are slightly less variable under mild model misspecification.

In the application to insurance uptake using the MEPS data, we find that adopting the semi-

parametric or copula-based modelling approaches provide more reliable estimates and reveal characteristics of the data that cannot be captured using the classical models. Specifically, relaxing the functional form of the deterministic model components results in parameter estimates that are, overall, more precise (in terms of obtaining smaller standard errors and narrower confidence intervals) than those obtained using the classical approach. Furthermore, the semi-parametric model reveals covariate effects with different degrees of non-linearity that cannot be captured using the classical model. Using the copula-based approach, we find that the dependence structure between insurance status and expenditures is not symmetric. The analysis suggests that large values of expenditures are associated with higher chances of obtaining supplementary insurance however, low expenditures do not necessarily imply lower chances of having insurance. This is also a feature that cannot be captured using the classical distributional assumptions. In the analysis of the REFLUX study, we find that estimates of the effect of surgery are significant, regardless of the modelling approach. The MI estimates for the effect of surgery and other model parameters are very similar to those obtained using the copula-based SS model and, in some instances, they have slightly smaller standard errors.

The remainder of this thesis is structured as follows: Chapter 2 presents a semi-parametric ESR model. We describe two approaches to parameter estimation that follow from a well-established penalized regression framework. We also discuss the main inferential results and conduct a simulation study. In an application, we study insurance uptake of individuals over 65 years old in the US to cover for out-of-pocket expenditures of prescription drugs. Chapter 3 presents copula-based ESR models for continuous responses. The models are specified using copula functions where their univariate marginals are represented using the GAMLSS framework. Parameter estimation and inference follow from the methods described in Chapter 2. We perform a simulation study to evaluate the empirical properties of the approach. We then revisit the application on insurance uptake in order to make an assessment of the dependence structures in the data, and evaluate the univariate marginal distributions of out-of-pocket expenditures using several parametric distributions. In Chapter 4, we first provide an overview of the main terminology used in the missing data literature and the role of the missing data mechanism in statistical modelling. We also review the specification of the classical SS model, a flexible copula-based extension, and the multiple imputation framework. We then present a multiple imputation approach for variables assumed to be missing not at random and assess its performance in a simulation study. Lastly, we use data from the REFLUX study to re-evaluate the robustness of the results obtained under different modelling

assumptions about the missing data mechanism and the distribution of the response.

Chapter 2

A semi-parametric endogenous switching regression model with an application in health economics

This chapter presents a semi-parametric endogenous switching regression model that extends the classical modelling approach by introducing flexible covariate effects in the model predictors. Specifically, we use a GAM-style specification of the predictors, where associations between the continuous covariates and the response are represented via penalized regression splines. Parameter estimation utilises a well-established penalized regression framework and inference follows from the Bayesian interpretation of the smoothing process. In an application, we investigate insurance uptake to cover out-of-pocket expenses related to prescription drugs in individuals over 65 years old from the United States. Our findings reveal evidence of individual self-selection into insurance and that some of the determinant factors of out-of-pocket expenditures exhibit varying degrees of non-linearity, which are not adequately captured by the classical approach.

2.1 Introduction

A typical question of interest in applied research is to examine the effect of a binary variable (or treatment) on an outcome of interest, and to determine whether the effects of other covariates differ across treatment levels.

In regression analysis of experimental studies, treatment assignment to individuals or statistical units is known and controlled by the researcher. It is common to assume that this mechanism

is unconfounded, also known as ignorable treatment assignment, or the conditional independence assumption (Imbens & Rubin, 2015). This assumption rules out the possibility of unobserved confounders that affect both the treatment and the outcome, and allows for individuals with similar characteristics to be compared at different treatment levels. With observational data, the assignment mechanism is not generally known or controlled by the researcher since random allocation of individuals may not be possible, ethical or simply, subjects may fail to adhere to the study requirements (Stuart et al., 2009). Furthermore, the conditional independence assumption may be restrictive or untenable since individuals may self-select into treatment based on particular unobserved characteristics, for example, to gain from the expected benefits of receiving the treatment. As anticipated in Chapter 1, this situation makes the treatment assignment non-ignorable and individual characteristics associated with the response may be systematically different at both levels of the treatment, due to the correlation between the unobserved individual characteristics that influence the choice of treatment and the response. Statistical analyses that do not account for a non-ignorable assignment mechanism result in biased parameter estimates.

We consider individuals who are 65 or older from the Medical Expenditure Panel Survey (MEPS) in order to assess the effects of their insurance status, and several other socio-economic factors, on their out-of-pocket prescribed medicines expenditures[1]. The subjects considered in the analysis are enrolled in Medicare (a federal health insurance programme that covers for specific healthcare services) which, at the time of data collection, did not include prescribed drugs coverage unless they obtained some form of additional insurance (employer-sponsored or union-based) to cover against certain out-of-pocket expenses (for further details, we refer the reader to Cameron & Trivedi, 2009).

The relationships between insurance status and healthcare utilization have long been studied in economics. For instance, Arrow (1978), Phelps (1973), Newhouse & Phelps (1974), and Manning et al. (1987) discuss the plausibility of endogeneity of insurance in observational data and the role of adverse selection and moral hazard effects in health care utilization. Cameron et al. (1988) develop a formal economic model from an utilitarian perspective to address the mutual dependency between the demand for health insurance and health care use. The authors argue that individuals self-select into insurance is partly based on future expected health care consumption (which is not observed), making insurance status endogenous. Deb et al. (2006) further discuss that an individual's decision to uptake insurance is based on their observed and unobserved charac-

[1] We use the subset of the MEPS data available in Cameron & Trivedi (2009).

teristics such as their future healthcare needs, their risk aversion, health status, and several socio-economic characteristics, all of which may also affect healthcare use. In the application, individual self-selection into insurance raises concerns about the endogeneity of insurance status, since it is plausible that there are some unobserved individual characteristics, such as private information or life-style choices, that affect simultaneously prescribed drugs expenditures and obtaining supplementary insurance (Deb et al., 2006). For instance, individuals might have chosen to work in a particular industry or joined a union during their working life expecting they would benefit from the extra coverage this would provide after retirement (Cameron & Trivedi, 2009; French & Jones, 2011).

Two popular models that address the problem of endogeneity of a treatment arising from self-selection are the treatment effects or dummy endogenous variable (DEV; Heckman, 1978), and endogenous switching regression (ESR; Roy, 1951). The particular specifications of each model depend on the structure of the data at hand and the assumptions made by the researcher. Details on how these models can be motivated from an utilitarian perspective can be found, for example, in Borjas (1987), Cameron et al. (1988), and Heckman & Leamer (2007).

The ESR model has its origins on the conceptual framework proposed by Roy (1951) to study the factors that influence individuals when choosing between hunting or fishing as a profession, and can be thought of an extension of sample selection and DEV models. The model aims to (i) account for the effects of endogeneity arising from self-selection in order to obtain parameter estimates that vary at each level of the treatment; (ii) estimate the extent to which unobserved confounders influence both the treatment and the outcome; and (iii) counterfactual analysis (Mare & Winship, 1988). Applications can be found in labour economics (Lee, 1978; Sakamoto & Chen, 1991), sociology and education (Gamoran & Mare, 1989; Willis & Rosen, 1979; Mare & Winship, 1988), agriculture (Wilde & Ranney, 2000), and health economics (Deb et al., 2006) among others.

The econometric formulation of the classical ESR model requires the specification of three jointly normal regression equations: one for the selection or switching mechanism, and one for each of the two regimes individuals may enter, depending on the observed value of the binary endogenous variable (see, for instance, Maddala, 1986c). In the application, the switching mechanism models the individual's choice of obtaining supplementary insurance, whereas the regime equations model out-of-pocket prescribed drug expenditures for individuals with extra insurance and for individuals using Medicare only. Note that the literature distinguishes among several types of switching regression models depending on whether (i) the selection variable and the regimes

are correlated, (ii) the sample separation is known or unknown, or (iii) the sample separation is known but observed imperfectly (for further details, we refer the reader to Maddala, 1986a). In the context of the thesis, we observed the switching variable perfectly and we assume an association between the switching and regime variables.

The model has been extended in several ways to relax the classical distributional assumptions, for instance, Choi & Min (2009) replaced the normality assumption with a multivariate version of the S_U-normal distribution (Johnson, 1949a,b) to account for asymmetry and excess kurtosis in the distribution of the response. Smith (2005) specified the model using Archimedean copulas with non-normal continuous marginals. From a Bayesian perspective, Deb et al. (2006) proposed an extension for discrete outcomes.

The aforementioned extensions are not exempt from the consequences that may arise from functional form misspecification of the deterministic components of the model. For instance, modelling the effects of continuous covariates linearly, by categorisation, or using pre-specified fixed-order polynomials may result in residual confounding (Benedetti & Abrahamowicz, 2004; Radice et al., 2016; Slama & Werwatz, 2005), which may distort the distribution of the response and result in departures from the normality assumptions in the classical model specification (Pigini, 2015).

To that end, and based on the sample selection modelling approach of Marra & Radice (2013a), we present an extension of the classical ESR model that flexibly models the covariate effects, while maintaining the classical distributional assumptions. Specifically, non-linear associations between the continuous explanatory variables and the response are modelled using penalized regression splines via a GAM-style (Hastie & Tibshirani, 1990; Wood, 2017) specification of the model predictors; parameter estimation utilise the procedure introduced in Marra et al. (2017) and Marra & Radice (2019); and inference follows from the Bayesian interpretation of the smoothing process (Marra et al., 2017; Marra & Radice, 2019; Wahba, 1978; Silverman, 1985; Wood, 2017).

The rest of this chapter is structured as follows: Section 2.2 describes a semi-parametric ESR model, in particular, we present the model specification, the modelling assumptions, and the structure of the semi-parametric additive predictors. Section 2.3 explains in detail approaches to parameter estimation, while Section 2.4 considers the main inferential results from the penalized regression framework relevant to our context. In Section 2.6 we analyse the effects of insurance status, and other socio-economic and health related factors, on out-of-pocket prescribed medicines using data from the MEPS. Lastly, in Section 2.7 we discuss the modelling approach and its limi-

tations.

2.2 Specification of a semi-parametric ESR model

Let Y_{1i}, Y_{2i}, Y_{3i}, for $i = 1, \ldots, n$, denote three random variables generated using the following rules

$$Y_{1i} = \mathbb{1}_{Y_{1i}^*>0}(Y_{1i}^*), \qquad Y_{2i} = Y_{1i}Y_{2i}^*, \qquad Y_{3i} = (1 - Y_{1i})Y_{3i}^*. \tag{2.1}$$

The Bernoulli random variable Y_{1i} represents a treatment, switch, or switching variable (suspected to be endogenous) that is determined by the sign of the continuous latent variable Y_{1i}^* through the indicator function $\mathbb{1}_{Y_{1i}^*>0}(\cdot)$. [2] The continuous variables Y_{2i} and Y_{3i} are determined by Y_{1i} and their latent counterparts Y_{2i}^* and Y_{3i}^*, that is, when $Y_{1i} = 1$ we observe $Y_{2i} = Y_{2i}^*$ otherwise, we observe $Y_{3i} = Y_{3i}^*$. Note that Y_{2i} and Y_{3i} are never observed simultaneously and represent the two possible states where a variable of interest, say Y_i, can be observed at. Dummy zero values are generally assigned to Y_{2i} and Y_{3i} accordingly (Smith, 2003, 2005). In our context, Y_1^* is an unobserved continuous variable that captures individual's propensity to obtain supplementary insurance, beyond using Medicare only. From an utilitarian perspective, Y_1^* can be thought of as the difference in expected utility between individuals with supplementary insurance and those using Medicare only. The continuous variables Y_2^* and Y_3^* represent out-of-pocket expenditures for individuals with supplementary insurance and for individuals with Medicare only, respectively. Individual's choice on insurance uptake will determine whether we observe one or the other.

Given these observation rules, a semi-parametric ESR model can be defined using the following system of equations

$$Y_{1i}^* = \mathbf{v}_{1i}^\mathsf{T}\boldsymbol{\alpha}_1 + \sum_{\bar{p}_1=1}^{\bar{P}_1} s_{1\bar{p}_1}(w_{1\bar{p}_1 i}) + \epsilon_{1i} = \eta_{1i} + \epsilon_{1i}, \tag{2.2}$$

$$Y_{2i}^* = \mathbf{v}_{2i}^\mathsf{T}\boldsymbol{\alpha}_2 + \sum_{\bar{p}_2=1}^{\bar{P}_2} s_{2\bar{p}_2}(w_{2\bar{p}_2 i}) + \epsilon_{2i} = \eta_{2i} + \epsilon_{2i}, \tag{2.3}$$

$$Y_{3i}^* = \mathbf{v}_{3i}^\mathsf{T}\boldsymbol{\alpha}_3 + \sum_{\bar{p}_3=1}^{\bar{P}_3} s_{3\bar{p}_3}(w_{3\bar{p}_3 i}) + \epsilon_{3i} = \eta_{3i} + \epsilon_{3i}, \tag{2.4}$$

where $\mathbf{v}_{mi} = (1, v_{m2i}, \ldots, v_{mP_m i})^\mathsf{T}$ is a vector of binary and/or categorical variables (including an intercept), $\boldsymbol{\alpha}_m \in \mathbb{R}^{P_m}$ is a commensurate vector of regression coefficients, and each $s_{m\bar{p}_m}(\cdot)$

[2] An indicator function $\mathbb{1}_A(a)$ is equal to 1 when $a \in A$ and 0 otherwise.

corresponds to an unknown smooth function of the continuous covariate $w_{m\bar{p}_m}$, for $m = 1, 2, 3$, and $\bar{p}_m = 1, \ldots, \bar{P}_m$. The particular representation of the smooth functions is described in Section 2.2.1.

As in the classical ESR model, the semi-parametric specification makes fully parametric assumptions on the error terms, that is, the error terms $\epsilon_{1i}, \epsilon_{2i}$, and ϵ_{3i} are assumed to follow a trivariate normal distribution, i.e.,

$$\begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix} \right),$$

where σ_1, σ_2, and σ_3 are standard deviations, and $\rho_{12}, \rho_{13}, \rho_{23}$ denote correlation coefficients. Note that the deterministic model components, encapsulated in η_{mi}, for $m = 1, 2, 3$, correspond to the well-known specification of semi-parametric additive predictors in GAMs.

Equation (2.2) represents a model for the switching mechanism or selection process, while Equations (2.3) and (2.4) correspond to models for the two states that the outcome of interest can be observed at, and are usually known as the regimes or regime equations. In our context, Equation (2.2) models the process by which individuals decide whether to obtain supplementary insurance to cover for out-of-pockets expenditures, whereas Equations (2.3) and (2.4) model the two possible regimes for out-of-pocket expenditures, based on individual's insurance status. If $\rho_{1m} = 0$, for $m = 2, 3$, the switching regression model is called exogenous and parameters can be consistently estimated by fitting two separate regressions. However, if $\rho_{1m} \neq 0$, for $m = 2, 3$, the switching regression model is called endogenous and parameters cannot be consistently estimated using separate univariate regression models since $\mathsf{E}[\epsilon_{2i} \mid \epsilon_{1i} > -\eta_{1i}] \neq 0$ and $\mathsf{E}[\epsilon_{3i} \mid \epsilon_{1i} \leq -\eta_{1i}] \neq 0$. Further distinctions can be made depending on whether Y_1 is unobserved, perfectly observed, or observed with error (for details, see, for example Maddala, 1986a). The endogenous switching regression model accounts for dependence between the error terms in the switching mechanism and each regime through the correlation coefficients ρ_{12} and ρ_{13}, which can be attributed to unobserved confounders that affect both, the probability of treatment assignment and the outcome of interest.

Identification of the model parameters is subject to certain restrictions. Since the latent variable Y_1^* is not observed, there is no information in the data to estimate σ_1, and the regression

coefficients in Equation (2.2) can only be identified up to a factor of $1/\sigma_1$. Without loss of generality, we can assume $\sigma_1 = 1$. Note that the latent variable specification of Equation (2.2) leads to a probit model for the observed binary variable Y_1. Similarly, because we do not observe Y_2 and Y_3 simultaneously, there is no information in the data to identify the correlation coefficient ρ_{23}, which is usually left unspecified or set to zero (Smith, 2005). When the additive predictors in Equations (2.2)-(2.4) contain more than one smooth function, their overall level is not identified and a sum-to-zero constraint is imposed on all non-linear terms, that is, $\sum_{i=1}^{n} s_{m\bar{p}_m}(w_{m\bar{p}_m i}) = 0$, for $m = 1, 2, 3$, and $\bar{p}_m = 1, \ldots, \bar{P}_m$ (Wood, 2017, pp. 174-175).

Potential collinearity problems may arise during estimation due to a complete overlap among the covariates in the predictors η_{mi}, for $m = 1, 2, 3$. In practice, it is common to assume that an exclusion restriction assumption on the covariates holds. A valid exclusion restriction requires the predictor η_{1i} to contain at least one covariate that is not included in η_{2i} and η_{3i}. Such covariate is usually known as an instrumental variable, which must be relevant to predicting Y_{1i} and conditionally independent of Y_{2i} and Y_{3i}.

2.2.1 Smooth function representation

The smooth functions of continuous covariates in the model are represented using a penalized regression spline framework (Eilers & Marx, 1996; Wahba, 1980; Wood, 2017). In essence, the approach consists of modelling the covariate effects using regression splines together with a penalty component that controls for smoothness of the fit and prevents over-fitting.

Assume that each of the smooth functions specified in Equations (2.2)-(2.4) can be written using basis functions expansions as follows

$$s_{m\bar{p}_m}(w_{m\bar{p}_m i}) = \sum_{j_{\bar{p}_m}=1}^{J_{\bar{p}_m}} b_{m\bar{p}_m j_{\bar{p}_m}}(w_{m\bar{p}_m i})\tilde{\alpha}_{m\bar{p}_m j_{\bar{p}_m}} = \mathbf{b}_{m\bar{p}_m}^{\mathsf{T}}(w_{m\bar{p}_m i})\tilde{\boldsymbol{\alpha}}_{m\bar{p}_m}, \tag{2.5}$$

where $\mathbf{b}_{m\bar{p}_m}(w_{m\bar{p}_m i}) = (b_{m\bar{p}_m 1}(w_{m\bar{p}_m i}), \ldots, b_{m\bar{p}_m J_{\bar{p}_m}}(w_{m\bar{p}_m i}))^{\mathsf{T}}$ is a vector that contains $J_{\bar{p}_m}$ basis functions evaluated at the continuous covariate $w_{m\bar{p}_m i}$, and $\tilde{\boldsymbol{\alpha}}_{m\bar{p}_m} \in \mathbb{R}^{J_{\bar{p}_m}}$ is a vector of regression coefficients, for $m = 1, 2, 3$, and $\bar{p}_m = 1, \ldots \bar{P}_m$. The number of basis functions is generally chosen to be 'large enough' to capture the effect of the particular covariate. In practice, using between 10 and 20 bases works well in many situations (Wood, 2017, pp. 242–243).

In addition, each smooth function has an associated penalty term $\lambda_{m\bar{p}_m}\mathcal{J}(s_{m\bar{p}_m})$, where $\lambda_{m\bar{p}_m} \geq 0$ is an unknown parameter that determines the smoothness of the estimated function

by controlling the influence of a quadratic roughness measure $\mathcal{J}(s_{m\bar{p}_m})$ on the fit, given by

$$\mathcal{J}(s_{m\bar{p}_m}) = \tilde{\boldsymbol{\alpha}}_{m\bar{p}_m}^{\mathsf{T}} \mathbf{S}_{m\bar{p}_m} \tilde{\boldsymbol{\alpha}}_{m\bar{p}_m}, \tag{2.6}$$

for some known $J_{m\bar{p}_m} \times J_{m\bar{p}_m}$ penalty matrix $\mathbf{S}_{m\bar{p}_m}$. Provided that the number of basis functions $J_{\bar{p}_m}$ used to model $s_{m\bar{p}_m}(w_{m\bar{p}_m i})$ is large enough, smoothness is ultimately controlled by $\lambda_{m\bar{p}_m}$ and the particular number of basis functions becomes irrelevant (Green & Silverman, 1993). When $\lambda_{m\bar{p}_m} \to \infty$, the penalty on the coefficients $\tilde{\boldsymbol{\alpha}}_{m\bar{p}_m}$ becomes notable, which results in the estimated function being closer to linear. On the other hand, when $\lambda_{m\bar{p}_m} \to 0$, the penalty term becomes negligible resulting in a wigglier function estimate. A key modelling issue in the penalized regression framework consists of choosing the optimal smoothing parameter that obtains the best description of the data based on a particular criterion (see Section 2.3 for further details).

Comprehensive surveys of the penalized regression framework and several specifications of the splines basis functions can be found in Fahrmeir et al. (2013), Green & Silverman (1993), and Wood (2017).

Given the aforementioned smooth function representation, each of the equations in the semi-parametric ESR model can be written succinctly as

$$Y_{mi}^{*} = \mathbf{x}_{mi}^{\mathsf{T}} \boldsymbol{\beta}_m + \epsilon_{mi} = \eta_{mi} + \epsilon_{mi}, \qquad m = 1, 2, 3,$$

where $\mathbf{x}_{mi} = (\mathbf{v}_{mi}^{\mathsf{T}}, \mathbf{b}_{mi}^{\mathsf{T}})^{\mathsf{T}}$ contains an overall intercept and the binary and/or categorical variables encapsulated in $\mathbf{v}_{mi} = (1, v_{m2i}, \ldots, v_{mP_m i})^{\mathsf{T}}$, and a vector of spline basis functions evaluated at the i^{th} observation of each of the continuous covariates, given by $\mathbf{b}_{mi} = \{\mathbf{b}_{m1}^{\mathsf{T}}(w_{m1i}), \ldots, \mathbf{b}_{m\bar{P}_m}^{\mathsf{T}}(w_{m\bar{P}_m i})\}^{\mathsf{T}}$. The corresponding vector of regression coefficients is $\boldsymbol{\beta}_m = (\boldsymbol{\alpha}_m^{\mathsf{T}}, \tilde{\boldsymbol{\alpha}}_m^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{p_m}$, where $\boldsymbol{\alpha}_m = (\alpha_{m_1}, \ldots, \alpha_{m_{P_m}})^{\mathsf{T}}$ and $\tilde{\boldsymbol{\alpha}}_m = (\tilde{\boldsymbol{\alpha}}_{m1}^{\mathsf{T}}, \ldots, \tilde{\boldsymbol{\alpha}}_{m\bar{P}_m}^{\mathsf{T}})^{\mathsf{T}}$ are the regression coefficients associated with the fully parametric and the non-parametric elements of the additive predictor, and $p_m = P_m + \sum_{\bar{p}_m=1}^{\bar{P}_m} J_{m\bar{p}_m}$. The overall penalty term associated with the m^{th} linear predictor is defined as $\sum_{\bar{p}_m=1}^{\bar{P}_m} \lambda_{m\bar{p}_m} \tilde{\boldsymbol{\alpha}}_{m\bar{p}_m}^{\mathsf{T}} \mathbf{S}_{m\bar{p}_m} \tilde{\boldsymbol{\alpha}}_{m\bar{p}_m}$, which can be expressed in terms of the overall vector of regression coefficients as $\boldsymbol{\beta}_m^{\mathsf{T}} \bar{\mathbf{S}}_m \boldsymbol{\beta}_m$, where $\bar{\mathbf{S}}_m = \text{diag}(\mathbf{0}_{P_m}^{\mathsf{T}}, \lambda_{m1}\mathbf{S}_{m1}, \ldots, \lambda_{m\bar{P}_m}\mathbf{S}_{m\bar{P}_m})$ and $\mathbf{0}_{P_m}^{\mathsf{T}}$ represents a vector of P_m zeroes. The smoothing parameters contained in the overall penalty term can also be encapsulated into an overall smoothing parameter vector as $\boldsymbol{\lambda}_m = (\lambda_{m1}, \ldots, \lambda_{m\bar{P}_m})^{\mathsf{T}}$. Lastly, note that each of the model additive predictors can be further written in vector-matrix form as $\boldsymbol{\eta}_m = \mathbf{X}_m \boldsymbol{\beta}_m$, where $\boldsymbol{\eta}_m = (\eta_{m1}, \ldots, \eta_{mn})^{\mathsf{T}}$ and $\mathbf{X}_m$ is the $n \times p_m$

design matrix associated with equation m (for further details see, for example, Marra & Radice, 2013a,b, 2021; Wood, 2017). The specification of the semi-parametric ESR model is completed by defining the structure of the error terms, which we assume to be jointly normal with zero mean and variance-covariance matrix already defined in page 11. Therefore, we maintain the distributional assumptions of the classical approach but extend the functional form of the predictors in order to capture non-linear effects of the continuous covariates.

Models that correct for self-selection are known to be sensitive to distributional and/or functional form misspecification. Identification also relies on the distributional assumptions being met (Pigini, 2015; Maddala, 1986b). The main criticisms of the ESR model are that the joint normality and the constant variance assumptions can be restrictive or inappropriate in empirical applications. Additionally, the association between the switching and regime variables is tied to the correlation coefficient, whose use as a measure of association has also been subject to criticism (Embrechts et al., 2002). Assuming the distributional assumptions hold, parameter estimates of models subject to self-selection are consistently estimated using a 2-step approach or maximum likelihood (see Section 2.3) however, departures from normality yield inconsistent parameter estimates (Vella, 1998; van der Klaauw & Koning, 2003). In the absence of exclusion restrictions, the maximum likelihood estimator is generally preferred over the 2-step approach since the latter obtains unreliable parameter estimates due to their dependence on the non-linearity of the inverse Mill's ratio (Puhani, 2000). A further issue relates to functional form misspecification, which may result in residual confounding, causing departures from the distributional assumptions and biasing the results (Benedetti & Abrahamowicz, 2004; Pigini, 2015; Radice et al., 2016; Slama & Werwatz, 2005). To address this issue, the semi-parametric ESR model intents to model the effects of continuous covariates flexibly, using the described penalized regression framework.

Following the penalized regression spline literature (for example, Green & Silverman, 1993; Wahba, 1990; Wood, 2017), we describe next two of the most popular combinations of spline basis functions and penalty terms used in practice, namely, penalized B-splines and thin plate splines. To simplify the notation and aid the exposition, we omit the sub-scripts m and $\bar{p}_m$ that refer to the model equation and the particular covariate within the equation, respectively.

Penalized B-splines

B-spline basis functions (DeBoor, 1978) are made up of piecewise polynomials which are joined continuously at some given points (or knots) in the support of the variable. B-splines are numer-

ically stable and are implemented in most statistical software. Given a set of k fixed equidistant knots $w_1^* < w_2^* < \ldots < w_k^*$ in the support of a continuous covariate w, the B-spline basis functions of order r are obtained as follows

$$b_j^r(w) = \frac{w - w_{j-r}^*}{w_j^* - w_{j-r}^*} b_{j-1}^{r-1}(w) + \frac{w_{j+1}^* - w}{w_{j+1}^* - w_{j+1-r}^*} b_j^{r-1}(w)$$

and

$$b_j^0(w) = \begin{cases} 1 & w_j^* \leq w < w_{j+1}^*, \\ 0 & \text{otherwise.} \end{cases} \qquad j = 1, \ldots, J-1.$$

In practice, it is common to set $r = 3$, since this obtains a representations of cubic spline basis (Eilers & Marx, 2021). Once the bases are constructed, a smooth function $s(w)$ can be represented as a linear combination of $J = k + r - 1$ B-spline basis functions and regression coefficients as given in (2.5).

B-splines can be combined with different types of roughness measures that lead to the quadratic form in the regression coefficients given in (2.6). For instance, Eilers & Marx (1996) provided the B-spline basis with a discrete roughness measure based on the differences of adjacent elements of the parameter vector $\tilde{\boldsymbol{\alpha}}$, known as P-splines. The d^{th}-order difference is defined recursively as $\Delta^d(\tilde{\alpha}_j) = \Delta^{d-1}(\tilde{\alpha}_j) - \Delta^{d-1}(\tilde{\alpha}_{j-1})$ and $\Delta^1(\tilde{\alpha}_j) = \tilde{\alpha}_j - \tilde{\alpha}_{j-1}$. The measure of roughness on the smooth function $s(w)$ is then based on the sum of the squares of differences of order d given by

$$\mathcal{J}(s) = \sum_{j=1+d}^{J} \{\Delta^d(\tilde{\alpha}_j)\}^2 = \tilde{\boldsymbol{\alpha}}^\mathsf{T} \mathbf{D}_d^\mathsf{T} \mathbf{D}_d \tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}^\mathsf{T} \mathbf{S}_d \tilde{\boldsymbol{\alpha}},$$

where $\mathbf{D}_d$ is a $(J-d) \times J$ difference matrix and $\mathbf{S}_d = \mathbf{D}_d^\mathsf{T} \mathbf{D}_d$ is a $J \times J$ penalty matrix. In practice, the order of the difference is usually set to $d = 2$, since this obtains a good balance between smoothness and fidelity to the data (Eilers & Marx, 1996). As an example, penalizing the squared second-order differences between two neighbour parameters leads to the following band diagonal

matrices

$$\mathbf{D}_2 = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix} \text{ and } \mathbf{S}_2 = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}.$$

A comprehensive review of P-splines, the mathematical properties that make them suitable for statistical modelling, and their applications can be found in Eilers & Marx (2021).

Alternatively, the roughness penalty can be based on a measure that quantifies the smoothness of a function such as the integrated squared second-order derivative of the smooth function, i.e., $\int\{s''(w)\}^2\mathrm{d}w$ (O'Sullivan, 1986). Using the smooth function representation given in (2.5), the penalty term can also be written as a quadratic form in the parameter vector $\tilde{\boldsymbol{\alpha}}$ as follows

$$\mathcal{J}(s) = \int\{s''(w)\}^2\mathrm{d}w = \int\left\{\tilde{\boldsymbol{\alpha}}^\mathsf{T}\mathbf{b}''(w)\right\}^2\mathrm{d}w = \tilde{\boldsymbol{\alpha}}^\mathsf{T}\left\{\int\mathbf{b}''(w)\mathbf{b}''(w)^\mathsf{T}\mathrm{d}w\right\}\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}^\mathsf{T}\mathbf{S}_\mathrm{s}\tilde{\boldsymbol{\alpha}}, \tag{2.7}$$

where $\mathbf{b}''(w)$ is a $J \times 1$ vector that contains the second-order derivative of the basis functions and the $(j_1, j_2)^{\text{th}}$ element of the $J \times J$ penalty matrix $\mathbf{S}_\mathrm{s}$ is given by $\int b''_{j_1}(w)b''_{j_2}(w)\mathrm{d}w$, for $j_1, j_2 = 1, \ldots J$. In fact, the roughness measure in (2.7) is commonly used with different types of basis functions, not just B-splines (see, for example, Green & Silverman, 1993).

Thin plate splines

Thin plate splines (TPS; Duchon, 1977) arise as the solution of the penalized least squares problem for finding the estimate of a smooth function $s(\boldsymbol{w})$ of $c \geq 1$ covariates in which the penalty term consists of a multivariate integrated squared d^{th}-order derivative (see, for example, Wahba, 1990; Wood, 2003, for details on the problem formulation). Compared to other choices of bases, TPS allow representing smooth functions of more than one covariate and their formulation does not depend on the choice of knot locations.

A smooth function of several continuous covariates, say $\boldsymbol{w} = (w_1, \ldots, w_c) \in \mathbb{R}^c$, can be

represented using thin plate splines as follows

$$s(\boldsymbol{w}) = \sum_{\tilde{l}=1}^{\tilde{L}} b_{\tilde{l}}^{0}(\boldsymbol{w})\tilde{\alpha}_{\tilde{l}}^{0} + \sum_{i=1}^{n} b_{i}^{+}(\boldsymbol{w})\tilde{\alpha}_{i}^{+} \tag{2.8}$$

where $\tilde{\alpha}_{\tilde{l}}^{0}$ and $\tilde{\alpha}_{i}^{+}$ are unknown regression coefficients, $b_{\tilde{l}}^{0}(\cdot)$ are polynomials of degree up to $d-1$ that span the null space of the penalty[3], and $b_{j}^{+}(\cdot)$ are radial functions defined as

$$b_{i}^{+}(\boldsymbol{w}) = b_{i}^{+}\left(\|\boldsymbol{w}-\boldsymbol{w}_i\|\right) = \begin{cases} \frac{(-1)^{c/2+d+1}}{2^{2d-1}\pi^{c/2}(d-1)!(d-c/2)!}\|\boldsymbol{w}-\boldsymbol{w}_i\|^{2d-c}\log\left(\|\boldsymbol{w}-\boldsymbol{w}_i\|\right) & c \text{ even}, \\ \frac{\Gamma(c/2-d)}{2^{2d}\pi^{c/2}(d-1)!}\|\boldsymbol{w}-\boldsymbol{w}_i\|^{2c-d} & c \text{ odd}, \end{cases} \tag{2.9}$$

where $\|\cdot\|$ denotes the Euclidean distance, and $\boldsymbol{w}_i$ is the i^{th} data point. Note that writing $\tilde{\boldsymbol{\alpha}} = (\tilde{\boldsymbol{\alpha}}_0^{\mathsf{T}}, \tilde{\boldsymbol{\alpha}}_+^{\mathsf{T}})^{\mathsf{T}}$, such that $\tilde{\boldsymbol{\alpha}}_0 = (\tilde{\alpha}_1^0, \ldots, \tilde{\alpha}_{\tilde{L}}^0)^{\mathsf{T}}$ and $\tilde{\boldsymbol{\alpha}}_+ = (\tilde{\alpha}_1^+, \ldots, \tilde{\alpha}_n^+)^{\mathsf{T}}$; and $\mathbf{b}(\boldsymbol{w}) = (\mathbf{b}_0(\boldsymbol{w})^{\mathsf{T}}, \mathbf{b}_+(\boldsymbol{w})^{\mathsf{T}})^{\mathsf{T}}$, such that $\mathbf{b}^0(\boldsymbol{w}) = (b_1^0(\boldsymbol{w})^{\mathsf{T}}, \ldots, b_{\tilde{L}}^0(\boldsymbol{w})^{\mathsf{T}})^{\mathsf{T}}$ and $\mathbf{b}^+(\boldsymbol{w}) = (b_1^+(\boldsymbol{w})^{\mathsf{T}}, \ldots, b_n^+(\boldsymbol{w})^{\mathsf{T}})^{\mathsf{T}}$, obtains the smooth function representation given in (2.5). To ensure identifiability while estimating the regression coefficients, the constraint $\mathbf{T}^{\mathsf{T}}\tilde{\boldsymbol{\alpha}}_+ = \mathbf{0}$ is imposed on $\tilde{\boldsymbol{\alpha}}_+$, where $\mathbf{T}$ is a $n \times \tilde{L}$ matrix whose $(i,\tilde{l})^{\text{th}}$ element is given by $b_{\tilde{l}}^0(\boldsymbol{w}_i)$, for $i = 1, \ldots, n$, and $\tilde{l} = 1, \ldots, \tilde{L}$ (see Wood, 2017, p. 216 for details). Furthermore, the vector of evaluations of $s(\boldsymbol{w})$ at $\boldsymbol{w}_i$, given by $\boldsymbol{s} = (s(\boldsymbol{w}_1), \ldots, s(\boldsymbol{w}_n))^{\mathsf{T}}$, can be written as $\boldsymbol{s} = \mathbf{T}\tilde{\boldsymbol{\alpha}}_0 + \mathbf{E}\tilde{\boldsymbol{\alpha}}_+$, where the $\mathbf{E}$ is a $n \times n$ matrix with $(i,j)^{\text{th}}$ element given by $b^+\left(\|\boldsymbol{w}_i - \boldsymbol{w}_j\|\right)$.

The roughness measure on the smooth function is defined by the following d-variate integral

$$\mathcal{J}_{cd}(s) = \int_{\mathbb{R}^c} \sum_{\nu_1+\ldots+\nu_c=d} \frac{r!}{\nu_1!\ldots\nu_c!}\left(\frac{\partial^d s(\boldsymbol{w})}{\partial w_1^{\nu_1}\ldots\partial w_c^{\nu_c}}\right)^2 \mathrm{d}w_1\ldots\mathrm{d}w_c, \tag{2.10}$$

which can be understood as a generalization of the one-dimensional integrated squared second derivative that appears in Equation (2.7). Moreover, using the results shown in, for example, Green & Silverman (1993, pp. 143) or Wahba (1990, pp. 33), the roughness measure can also be written as $\tilde{\boldsymbol{\alpha}}_+^{\mathsf{T}}\mathbf{E}\tilde{\boldsymbol{\alpha}}_+$.

As an example, the TPS representation of a smooth function of two predictors $s(w_1, w_2)$, together with a roughness penalty based on second order derivatives can be written as

$$s(w_1, w_2) = \tilde{\alpha}_1^0 + \tilde{\alpha}_2^0 w_1 + \tilde{\alpha}_3^0 w_2 + \sum_{i=1}^{n} \tilde{\alpha}_i^+ b_i^+(\boldsymbol{w}),$$

[3] the space of functions where the roughness measure is zero with dimension $\tilde{L} = \binom{c+d-1}{c}$.

where $b_i^+\left(\|\boldsymbol{w}-\boldsymbol{w}_i\|\right) = \frac{1}{8\pi}\|\boldsymbol{w}-\boldsymbol{w}_i\|^2\log(\|\boldsymbol{w}-\boldsymbol{w}_i\|)$, subject to $\mathbf{T}^\mathsf{T}\tilde{\boldsymbol{\alpha}}_+ = \mathbf{0}$, where $\mathbf{T}$ is a $n\times 3$ matrix whose i^{th} row is $(1, w_{1i}, w_{2i})$, for $i = 1,\dots n$. The roughness penalty in Equation (2.10) reduces to

$$\mathcal{J}_{22}(s) = \iint \left(\frac{\partial^2 s(w_1,w_2)}{\partial w_1 \partial w_1}\right)^2 + 2\left(\frac{\partial^2 s(w_1,w_2)}{\partial w_1 \partial w_2}\right)^2 + \left(\frac{\partial^2 s(w_1,w_2)}{\partial w_2 \partial w_2}\right)^2 \mathrm{d}w_1\mathrm{d}w_2.$$

The main disadvantage of using TPS in practical terms is that they are computationally costly, since they require the estimation of as many parameters as observations. Further details on TPS, their construction, and their theoretical and numerical properties can be found in Green & Silverman (1993), Wahba (1990), and Wood (2017).

Thin plate regression splines (TPRS; Wood, 2003) are a low rank approximation to TPS that retain their good mathematical properties but are computationally efficient, since the number of operations needed to determine the smooth function is significantly lower than using TPS. The idea behind the construction of TPRS consists of keeping the bases contained in the first term on the right-hand side of (2.8), while truncating to a lower dimension $\tilde{k}$, with $\tilde{L} < \tilde{k} < n$, the n radial basis functions within the second term on the right-hand side of (2.8). Furthermore, the resulting approximation has a minimal impact on the formulation of the smoothing problem, that is, on both the model fit and the penalty functional that determines the shape of the smooth function. To see this, consider the eigenvalue decomposition of the matrix $\mathbf{E}$ given by $\mathbf{E} = \mathbf{U}^\mathsf{T}\mathbf{\Lambda}\mathbf{U}$, where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues arranged in descending order of magnitude and $\mathbf{U}$ is the matrix containing the corresponding eigenvectors. Wood (2003) then shows that, for a given value $\tilde{k}$, the truncated matrix $\mathbf{E}_{\tilde{k}} = \mathbf{U}_{\tilde{k}}^\mathsf{T}\mathbf{\Lambda}_{\tilde{k}}\mathbf{U}_{\tilde{k}}$, where $\mathbf{U}_{\tilde{k}}$ contains the first $\tilde{k}$ columns of $\mathbf{U}$ and $\mathbf{\Lambda}_{\tilde{k}}$ the first $\tilde{k}$ largest eigenvalues of $\mathbf{\Lambda}$, obtains the best approximation to $\mathbf{E}$ in the sense of minimising the spectral norm $\|\mathbf{E}-\mathbf{E}_{\tilde{k}}\|_2$. His results allow to reformulate the smoothing problem in terms of the truncated matrices and to obtain a $\tilde{k}$ rank approximation to the TPS with a small approximation error. The value of $\tilde{k}$ is chosen by the analyst and should be larger than what is believed to be needed to model the data at hand. For further details, we refer the reader to Wood (2003) and Wood (2017, pp. 215–219). TPRS are the default splines basis in the popular `mgcv` (Wood, 2020) R package.

As anticipated earlier in this section, the semi-parametric ESR model intents to model the effects of continuous covariates flexibly, without setting a pre-specified assumption about the form of the effect, using the described penalized regression framework. The flexible specification of the

additive predictors addresses the problem of residual confounding, which may cause departures from the distributional assumptions and induced bias in the results (Benedetti & Abrahamowicz, 2004; Pigini, 2015; Radice et al., 2016; Slama & Werwatz, 2005). In the application, it can be argued that variables such as age, income, and chronic conditions may affect the choice of insurance and out-of-pocket expenditures with some degree of non-linearity since they encapsulate life-cycles effects (Radice et al., 2016; Winkelmann, 2012; Wojtyś et al., 2018). Our findings in Section 2.6 reveal that these variables exhibit varying degrees of non-linearity, which are not adequately captured by the classical approach.

2.3 Parameter estimation

In this section, we describe two approaches to estimate the parameters of the semi-parametric ESR model. In particular, we place into to the current context the penalized two-step formulation of the SS model presented in Marra & Radice (2013a) and the penalized maximum likelihood estimation framework of Marra et al. (2017) and Marra & Radice (2019).

Similarly to Heckman's two-step estimation approach for SS models (Heckman, 1976), Lee (1976, 1978) proposed a two-step approach for the classical ESR model in which both regime equations are augmented with correction terms and estimated separately via least squares. Marra & Radice (2013a) adapted Heckman's approach to the penalized regression context. Using the same rationale, the semi-parametric ESR model equations given in Section 2.2 can be re-written as follows

$$Y_{1i}^* = \eta_{1i} + \epsilon_{1i}, \tag{2.11}$$

$$Y_{2i}^* = \eta_{2i} + \lambda(\eta_{1i})\beta_{\lambda_2} \quad + \zeta_{2i} \;\text{ if } Y_{1i}^* > 0, \tag{2.12}$$

$$Y_{3i}^* = \eta_{3i} - \lambda(-\eta_{1i})\beta_{\lambda_3} + \zeta_{3i} \;\text{ if } Y_{1i}^* \leq 0. \tag{2.13}$$

Equation (2.11) represents the model for the switching mechanism and it has the same specification as Equation (2.2) in the semi-parametric ESR model. The terms η_{mi}, for $m = 1, 2, 3$, correspond to the additive predictors that appear in Equations (2.2)-(2.4), whereas the second terms on the right-hand side of Equations (2.12) and (2.13) represent the expressions for $\mathsf{E}[\epsilon_{2i} \mid \epsilon_{1i} > -\eta_{1i}]$ and $\mathsf{E}[\epsilon_{3i} \mid \epsilon_{1i} \leq -\eta_{1i}]$, respectively. The regression coefficients $\beta_{\lambda_m} = \rho_{1m}\sigma_m$, for $m = 2, 3$, are associated with $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$, the inverse Mill's ratio (IMR), where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density

function (pdf) and distribution function (cdf) of the standard normal. Note that the conditional variances of the latent variables are given by $\mathsf{Var}[Y_{2i}^* \mid Y_{1i}^* > 0] = \sigma_2^2 - (\rho_{12}\sigma_2)^2\lambda(\eta_{1i})[\lambda(\eta_{1i}) + \eta_{1i}]$ and $\mathsf{Var}[Y_{3i}^* \mid Y_{1i}^* \leq 0] = \sigma_3^2 - (\rho_{13}\sigma_3)^2\lambda(-\eta_{1i})[\lambda(-\eta_{1i}) - \eta_{1i}]$ (see, for instance, Greene, 2014, and references therein). Furthermore, the error terms in Equations (2.12) and (2.13), denoted as ζ_{2i} and ζ_{3i}, have zero conditional mean and are uncorrelated with the covariates encapsulated in the predictors η_{2i}, η_{3i}, and with $\lambda(\cdot)$ (see Lee, 1978, for further details).

Estimation proceeds in two steps: the first stage estimates the parameters in Equation (2.11) using a GAM with a probit link on the observed switching variable to obtain estimates of the regression coefficients and construct $\hat{\eta}_{1i} = \mathbf{x}_{1i}^\mathsf{T}\hat{\boldsymbol{\beta}}_1$; the second stage inserts $\hat{\eta}_{1i}$ into Equations (2.12) and (2.13) and fits two separate GAMs to obtain estimates of the corresponding regression coefficients. Given the expressions for the conditional variances above, estimates for σ_2 and σ_3 can be obtained using (Maddala, 1986c)

$$\hat{\sigma}_2 = \sqrt{\hat{\boldsymbol{\zeta}}_2^\mathsf{T}\hat{\boldsymbol{\zeta}}_2/n_2 + \hat{\beta}_{\lambda_2}^2\lambda(\hat{\boldsymbol{\eta}}_1)^\mathsf{T}\left[\lambda(\hat{\boldsymbol{\eta}}_1) + \hat{\boldsymbol{\eta}}_1\right]/n_2}$$
$$\hat{\sigma}_3 = \sqrt{\hat{\boldsymbol{\zeta}}_3^\mathsf{T}\hat{\boldsymbol{\zeta}}_3/n_3 + \hat{\beta}_{\lambda_3}^2\lambda(-\hat{\boldsymbol{\eta}}_1)^\mathsf{T}\left[\lambda(-\hat{\boldsymbol{\eta}}_1) - \hat{\boldsymbol{\eta}}_1\right]/n_3},$$

respectively, where $\hat{\boldsymbol{\zeta}}_m$ represents a vector of residuals from the second stage, n_m denotes the number of observations in the m^{th} model equation, and $\hat{\beta}_{\lambda_m}$ corresponds to the estimated regression coefficient associated with the IMR, for $m = 2, 3$. Furthermore, estimates of the correlation coefficients can be obtained using $\hat{\rho}_{12} = \hat{\beta}_{\lambda_2}/\hat{\sigma}_2$ and $\hat{\rho}_{13} = -\hat{\beta}_{\lambda_3}/\hat{\sigma}_3$ (see, for example, Cameron & Trivedi, 2005; Maddala, 1986c; Toomet & Henningsen, 2008a).[4]

Two-stage approaches in models subject to self-selectivity are easy to implement using available software and yield consistent estimators under the classical assumptions however, they present certain drawbacks (see, for example, Puhani (2000) for a review, and Marra & Radice (2013a,b) for the relevant aspects in the penalized regression context). For instance, estimation of Equations (2.12) and (2.13) does not take into account the variability associated with estimating the parameters in the switching equation, leading to incorrect standard errors. Corrections to obtain the appropriate standard errors are given in Heckman (1979). In the context of binary response in a semi-parametric SS model, Marra & Radice (2013b) proposed an approach to account for such variability based on posterior simulation. In Section 2.5, we investigate some of the empirical properties of the two-stage estimator however, nowadays the two-step method is generally used to obtain initial values for maximum likelihood approaches.

[4]Estimates of ρ_{1m}, for $m = 2, 3$, are not restricted to the $[-1, 1]$ range and are truncated in practice.

Assuming the model is correctly specified, consistent parameter estimates and standard errors for the binary and the continuous model components can be obtained simultaneously via maximum likelihood. Given n independent observations, the log-likelihood function of the semi-parametric ESR model can be written as follows

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_{1i} \left\{ \log \sigma_2^{-1} + \log \left[\phi \left(\frac{y_{2i} - \eta_{2i}}{\sigma_2} \right) \right] + \log \left[\Phi \left(\frac{\eta_{1i} + \rho_{12} \left(y_{2i} - \eta_{2i} \right) / \sigma_2}{\sqrt{1 - \rho_{12}^2}} \right) \right] \right\}$$
$$+ \sum_{i=1}^{n} (1 - y_{1i}) \left\{ \log \sigma_3^{-1} + \log \left[\phi \left(\frac{y_{3i} - \eta_{3i}}{\sigma_3} \right) \right] + \log \left[1 - \Phi \left(\frac{\eta_{1i} + \rho_{13} \left(y_{3i} - \eta_{3i} \right) / \sigma_3}{\sqrt{1 - \rho_{13}^2}} \right) \right] \right\}, \tag{2.14}$$

where η_{mi}, for $m = 1, 2, 3$, are the semi-parametric additive predictors defined in the previous section and the overall parameter vector is given by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T}, \boldsymbol{\beta}_3^\mathsf{T}, \sigma_2, \sigma_3, \rho_{12}, \rho_{13})^\mathsf{T} \in \mathbb{R}^p$, where $p = 4 + \sum_{m=1}^{3} p_m$ and $\boldsymbol{\beta}_m \in \mathbb{R}^{p_m}$. Note that the parameters σ_m and ρ_{1m} are transformed for optimization to keep their domain on the real line, specifically, $\sigma_m^* = \log \sigma_m$ and $\rho_{1m}^* = \tanh^{-1}(\rho_{1m}) = \frac{1}{2} \log \left(\frac{1+\rho_{1m}}{1-\rho_{1m}} \right)$, for $m = 2, 3$ (Marra & Radice, 2013a). Also note that the trivariate normality assumption given in Section 2.2 is not actually needed for estimation since the likelihood function can be derived assuming bivariate normality of the pairs $(\epsilon_{1i}, \epsilon_{2i})$ and $(\epsilon_{1i}, \epsilon_{3i})$ (Smith, 2003). The derivation of the likelihood function is given in Appendix A.1.

Simply maximising the log-likelihood function in a penalized regression context will result in over-fitting of the smooth terms and the likelihood is regularized by imposing a roughness term that penalizes the fit (Green & Silverman, 1993; Marra & Radice, 2013a; Marra et al., 2017; Wood, 2017). The penalized maximum likelihood estimate (PMLE) is the value that maximises the penalized log-likelihood, that is,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max}\, \ell_{\mathrm{p}}(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max} \left\{ \ell(\boldsymbol{\beta}) - \mathcal{P}(\boldsymbol{\beta}, \boldsymbol{\lambda}) \right\},$$

where $\ell_{\mathrm{p}}(\boldsymbol{\beta})$ represents the penalized log-likelihood and $\mathcal{P}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ denotes an overall quadratic penalty given by $\mathcal{P}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} \boldsymbol{\beta}^\mathsf{T} \mathbf{S}_\lambda \boldsymbol{\beta}$. The matrix $\mathbf{S}_\lambda$ is a block-diagonal overall penalty matrix which contains each of the penalty matrices associated with the m^{th} model equation, given in Section 2.2.1, written as $\mathbf{S}_\lambda = \mathrm{diag}(\bar{\mathbf{S}}_1, \bar{\mathbf{S}}_2, \bar{\mathbf{S}}_3, 0, 0, 0, 0)$. The smoothing parameters embedded into the matrices $\bar{\mathbf{S}}_m$, for $m = 1, 2, 3$, can also be assembled into an overall penalty vector given by $\boldsymbol{\lambda} = (\bar{\boldsymbol{\lambda}}_1^\mathsf{T}, \bar{\boldsymbol{\lambda}}_2^\mathsf{T}, \bar{\boldsymbol{\lambda}}_3^\mathsf{T})^\mathsf{T}$. For further details on setting up the penalty terms in a penalized regression context see, for example, Marra & Radice (2013a,b, 2021) and Wood (2017).

The model penalized gradient and penalized Hessian are defined as follows

$$\mathbf{g}_{\mathrm{p}}(\boldsymbol{\beta}) = \frac{\partial \ell_{\mathrm{p}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{g}(\boldsymbol{\beta}) - \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$$

and

$$\boldsymbol{\mathcal{H}}_{\mathrm{p}}(\boldsymbol{\beta}) = \frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}}} = \boldsymbol{\mathcal{H}}(\boldsymbol{\beta}) - \mathbf{S}_{\boldsymbol{\lambda}},$$

where the vector $\mathbf{g}(\boldsymbol{\beta})$ and the matrix $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta})$ correspond to the gradient and Hessian of the unpenalized log-likelihood given by

$$\mathbf{g}(\boldsymbol{\beta}) = \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}^{\mathsf{T}}, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2}^{\mathsf{T}}, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3}^{\mathsf{T}}, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \sigma_2^*}, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \sigma_3^*}, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \rho_{12}^*}, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \rho_{13}^*} \right)^{\mathsf{T}}$$

and

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^{\intercal}} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2^{\intercal}} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_3^{\intercal}} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \sigma_2^*} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \sigma_3^*} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \rho_{12}^*} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \rho_{13}^*} \\ \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^{\intercal}} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2 \partial \sigma_2^*} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2 \partial \rho_{12}^*} & \mathbf{0} \\ \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3 \partial \boldsymbol{\beta}_3^{\intercal}} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3 \partial \sigma_3^*} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3 \partial \rho_{13}^*} \\ \cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \sigma_2^* \partial \sigma_2^*} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \sigma_2^* \partial \rho_{12}^*} & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \sigma_3^* \partial \sigma_3^*} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \sigma_3^* \partial \rho_{13}^*} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \rho_{12}^* \partial \rho_{12}^*} & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \rho_{13}^* \partial \rho_{13}^*} \end{bmatrix}$$

respectively. Their analytical expressions are derived in Appendix A.2.

Estimation of the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ proceeds by using the iterative approach proposed by Marra et al. (2017) and Marra & Radice (2019), implemented in the `SemiparBIV.fit` function and several internal routines of the `GJRM` package (Marra & Radice, 2021). At iteration a, the first step holds the smoothing parameter vector fixed at $\boldsymbol{\lambda}^{[a]}$ and optimises the penalized log-likelihood using a trust-region procedure to obtain $\boldsymbol{\beta}^{[a+1]}$; the second step holds the overall parameter vector fixed at $\boldsymbol{\beta}^{[a+1]}$ and minimises a prediction error criterion to obtain $\boldsymbol{\lambda}^{[a+1]}$. These two steps are iterated until $\frac{\left|\ell(\boldsymbol{\beta}^{[a+1]}) - \ell(\boldsymbol{\beta}^{[a]})\right|}{0.1 + \left|\ell(\boldsymbol{\beta}^{[a+1]})\right|} < 10^{-7}$ is satisfied. Initial values for the overall parameter vector are obtained using the penalized two-stage method described earlier in this section. The following two subsections describe their estimation approach in detail.

Step 1: Estimation of $\boldsymbol{\beta}$

Given some initial values $\boldsymbol{\beta}^{[0]}$, the trust-region procedure is an iterative approach that obtains a sequence of guess-estimates $\boldsymbol{\beta}^{[1]}, \ldots, \boldsymbol{\beta}^{[a]}, \ldots$ that converges to an estimate for the parameter vector $\boldsymbol{\beta}$ based on the assumption that there is a 'suitable neighbourhood' or region around the current guess-estimate $\boldsymbol{\beta}^{[a]}$ in which the penalized log-likelihood is adequately represented by a quadratic approximation.

In contrast to line search methods, which first determine the search direction and then compute the step length that obtains the best improvement, the procedure first establishes an upper bound in the step length and then finds an improvement within that region. As pointed out by Braun (2014), restricting the search for the next guess-estimate this way makes the trust-region approach appropriate for optimizing non-concave functions or those that contain regions that are nearly flat. The procedure is then particularly suitable in this context since the log-likelihood function of models with self-selectivity are not globally concave in general (Toomet & Henningsen, 2008a; Pigini, 2015). Furthermore, the absence of a variable that fulfils the exclusion restriction assumption may lead to a near-flat likelihood function over a relatively wide interval around its mode (Marra & Radice, 2011, 2013a). Trust-region algorithms have been successfully applied to graphical models (Hunter et al., 2008), beta-binomial regressions (Martin et al., 2020), difference-in-differences methods (Sant'Anna & Zhao, 2020), and in several models that fit simultaneous systems of equations (for example, Gomes et al., 2019; Marra & Radice, 2013a; Wojtyś et al., 2018, among others).

To fix ideas, at iteration a the trust-region approach proceeds on the basis of the following quadratic approximation of the penalized log-likelihood around the current parameter vector $\boldsymbol{\beta}^{[a]}$

$$\mathcal{Q}^{[a]}(\mathbf{q}) = \ell_{\mathrm{p}}(\boldsymbol{\beta}^{[a]}) + \mathbf{q}^{\mathsf{T}}\mathbf{g}_{\mathrm{p}}(\boldsymbol{\beta}^{[a]}) + \frac{1}{2}\mathbf{q}^{\mathsf{T}}\boldsymbol{\mathcal{H}}_{\mathrm{p}}(\boldsymbol{\beta}^{[a]})\mathbf{q}, \tag{2.15}$$

where $\mathbf{g}_{\mathrm{p}}(\boldsymbol{\beta}^{[a]}) = \frac{\partial \ell_{\mathrm{p}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[a]}}$ and $\boldsymbol{\mathcal{H}}_{\mathrm{p}}(\boldsymbol{\beta}^{[a]}) = \frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathsf{T}}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[a]}}$ are the penalized gradient and penalized Hessian evaluated at the current guess-estimate of the parameter vector, and $\mathbf{q} = \boldsymbol{\beta} - \boldsymbol{\beta}^{[a]} \in \mathbb{R}^p$. The 'suitable neighbourhood' around $\boldsymbol{\beta}^{[a]}$ corresponds to the trust-region at iteration a, which is defined as a ball of radius $r^{[a]}$ centred at $\boldsymbol{\beta}^{[a]}$, i.e., $\mathcal{T}^{[a]} = \{\boldsymbol{\beta} \colon \|\boldsymbol{\beta} - \boldsymbol{\beta}^{[a]}\| \leq r^{[a]}\}$. The radius $r^{[a]}$ can be understood as the bound of the trial step at the current iteration, and is updated at each iteration.

The algorithm then solves the following constrained optimization sub-problem

$$\tilde{\mathbf{q}}^{[a]} = \underset{\mathbf{q}\in\mathcal{T}^{[a]}}{\arg\min} -\mathcal{Q}^{[a]}(\mathbf{q}),$$

that is, $\tilde{\mathbf{q}}^{[a]}$ is the value that minimises the quadratic approximation of the penalized negative log-likelihood within the trust-region $\mathcal{T}^{[a]}$ at the current iteration a.

The acceptance of the trial step $\tilde{\mathbf{q}}^{[a]}$ and the update of the trust-region radius are based on the evaluation of the following ratio (Nocedal & Wright, 2006)

$$\breve{r}^{[a]} = \frac{\ell_{\mathrm{p}}(\boldsymbol{\beta}^{[a]}) - \ell_{\mathrm{p}}(\boldsymbol{\beta}^{[a]} + \tilde{\mathbf{q}}^{[a]})}{\mathcal{Q}^{[a]}(\boldsymbol{\beta}^{[a]}) - \mathcal{Q}^{[a]}(\boldsymbol{\beta}^{[a]} + \tilde{\mathbf{q}}^{[a]})},$$

where the numerator reflects the reduction in the penalized log-likelihood, and the denominator shows the reduction in the penalized log-likelihood when approximated by the quadratic function given in (2.15). Values of $\breve{r}^{[a]}$ close to 1 indicate that the current best solution $\boldsymbol{\beta}^{[a+1]} = \boldsymbol{\beta}^{[a]} + \tilde{\mathbf{q}}^{[a]}$ provides a satisfactory decrease in the value of the penalized negative log-likelihood (that is, $\mathcal{Q}^{[a]}(\boldsymbol{\beta}^{[a]})$ is a good representation of $\ell_{\mathrm{p}}(\boldsymbol{\beta}^{[a]})$), the trial step is accepted, and the radius of the trust-region can be increased at the next iteration. When $\breve{r}^{[a]}$ is negative or close to zero, the approximation $\mathcal{Q}^{[a]}(\boldsymbol{\beta}^{[a]})$ is a poor representation of $\ell_{\mathrm{p}}(\boldsymbol{\beta}^{[a]})$, the trial step $\tilde{\mathbf{q}}^{[a]}$ is discarded, and the radius of the trust-region $r^{[a]}$ is reduced. In other cases, $r^{[a]}$ is not modified and the algorithm proceeds to the next iteration. Note that, for a sufficiently large number of iterations, Nocedal & Wright (2006, pp. 92) show that the bound set by the trust-region constraint becomes irrelevant and the algorithm behaves as the classic Newton-Raphson approach.

A graphical illustration of the main ideas behind trust-region methods applied to an objective function of two variables is given in Figure 2.1 (Nocedal & Wright, 2006, pp. 67). Further information and theoretical details of the approach can be found in Conn et al. (2000) and Nocedal & Wright (2006, chapter 4). The R implementation is provided by the `trust` package (Geyer, 2015).

Step 2: Estimation of $\boldsymbol{\lambda}$

As anticipated in Section 2.2.1, efficient estimation of the smoothing parameters is a pivotal step in the penalized regression framework. Without being exhaustive, and following Wood (2017), automatic smoothing parameter estimation can be approached by choosing either a likelihood- or a model selection-based criterion which can then be applied to the model itself, or to a working

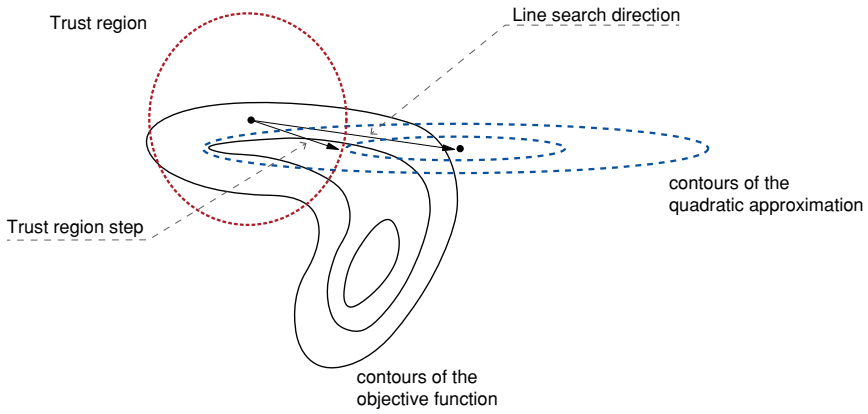

Figure 2.1: Graphical illustration of the trust-region approach applied to an objective function of two variables (black line ——). The trust-region algorithm constructs a region of trust (dotted red line ·····) around the current guess-estimate where it is believed that a quadratic approximation (dashed blue line - - -) represents well the objective function. Bounding the step-length to be contained within the trust-region avoids searching too far from the current guess-estimate and generally provides a better progress towards the minimum. Figure adapted from Nocedal & Wright (2006).

model derived from the previous estimation step. Likelihood-based methods result from taking the mixed model representation of the penalization process, where smooth functions are understood as random effects and smoothing parameters as variance components, and estimation proceeds via maximum likelihood or restricted maximum likelihood (see, for example, Ruppert et al., 2003; Wood et al., 2016). On the other hand, model selection-based approaches estimate the smoothing parameters by minimising a particular prediction error criterion that quantifies the model performance, usually based on an estimate of the mean squared error, such as generalized cross validation (GCV) or the unbiased risk estimator (UBRE; Craven & Wahba, 1978). A thorough review of these approaches can be found in Wood (2017).

In the present context, the smoothing parameter vector $\boldsymbol{\lambda}$ is estimated using an UBRE-like criterion based on a parametrization of the working model that results from the trust-region iteration using the approach of Marra et al. (2017) and Marra & Radice (2019) as described next.

First, note that at convergence of the previous step, a first order Taylor series expansion of the penalized gradient $\mathbf{g}_{\text{p}}(\boldsymbol{\beta}^{[a+1]})$ about $\boldsymbol{\beta}^{[a]}$ obtains the following expression for the next guess-estimate of the parameter vector (Marra et al., 2017; Marra & Radice, 2019)

$$\boldsymbol{\beta}^{[a+1]} = \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a]}) + \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right\}^{-1} \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a]})\right\}^{1/2} \mathbf{z}^{[a]}, \tag{2.16}$$

where

$$\mathbf{z}^{[a]} = \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a]})\right\}^{1/2} \boldsymbol{\beta}^{[a]} + \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a]})\right\}^{-1/2} \mathbf{g}(\boldsymbol{\beta}^{[a]}) \tag{2.17}$$

is a pseudo-data vector or working variable, and the matrices $\left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a]})\right\}^{1/2}$ and $\left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a]})\right\}^{-1/2}$ denote the square root of the observed information matrix and its inverse, respectively. It is tacitly assumed that the Hessian matrix is positive definite or that it can be perturbed to positive definiteness during the fitting process (see, supplementary material A of Marra et al., 2017).

Using results from likelihood theory (for example, Pawitan, 2013, p. 92) note that $\left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta})\right\}^{-1/2}\mathbf{g}(\boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix. This result implies that the pseudo-data vector is normally distributed with expectation $\mathsf{E}[\mathbf{z}] = \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta})\right\}^{1/2}\boldsymbol{\beta} = \boldsymbol{\mu}_{\mathbf{z}}$ and variance $\mathsf{Var}[\mathbf{z}] = \mathbf{I}$.

Letting $\hat{\boldsymbol{\mu}}_{\mathbf{z}} = \left\{-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}})\right\}^{1/2}\hat{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}$ be the predicted value of $\boldsymbol{\mu}_{\mathbf{z}}$, where $\mathbf{A}_{\boldsymbol{\lambda}} = \left\{-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}})\right\}^{1/2}$ $\left\{-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}}) + \mathbf{S}_{\lambda}\right\}^{-1}\left\{-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}})\right\}^{1/2}$ corresponds to an influence matrix, the smoothing parameter vector $\boldsymbol{\lambda}$ is estimated by minimising the following expected mean squared error criterion (Marra et al., 2017; Marra & Radice, 2019)

$$\mathsf{E}\left(\frac{1}{n}\left\|\boldsymbol{\mu}_{\mathbf{z}} - \hat{\boldsymbol{\mu}}_{\mathbf{z}}\right\|^2\right) = n^{-1}\mathsf{E}\left\|\mathbf{z} - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}\right\|^2 + 2n^{-1}\mathrm{tr}(\mathbf{A}_{\boldsymbol{\lambda}}) - \mathrm{n}^{-1}\bar{\mathrm{K}}, \tag{2.18}$$

where $\mathrm{tr}(\mathbf{A}_{\boldsymbol{\lambda}}) = \mathrm{tr}\left\{-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}})\left\{-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}}) + \mathbf{S}_{\lambda}\right\}^{-1}\right\}$ defines the model effective degrees of freedom (edf), and $\bar{K}$ is the total number of parameters. Note that this definition of edf can be used to obtain the Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) defined as $\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2edf$ and $\text{BIC} = -2\ell(\hat{\boldsymbol{\beta}}) + \log(n)edf$ (Hastie & Tibshirani, 1990; Marra & Radice, 2019; Wood, 2017, pp. 301).

In practical terms, given the guess-estimate $\boldsymbol{\beta}^{[a+1]}$ from Step 1, the smoothing parameters are obtained by minimizing an estimate of the expression in Equation (2.18), that is,

$$\boldsymbol{\lambda}^{[a+1]} = \underset{\boldsymbol{\lambda}}{\arg\min}\ \mathcal{V}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\lambda}}{\arg\min}\ \|\mathbf{z}^{[a+1]} - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}^{[a+1]}\|^2 + 2\mathrm{tr}(\mathbf{A}_{\boldsymbol{\lambda}}) - \bar{K}, \tag{2.19}$$

which is solved using the Newton approach proposed by Wood (2004) and implemented by the `magic` function of the `mgcv` package (Wood, 2020). As explained in Marra et al. (2017), this approach to smoothing parameter selection is efficient and computationally stable. Furthermore, minimising the mean squared error criterion $\mathcal{V}(\boldsymbol{\lambda})$ is also equivalent to minimising an approximate Akaike information criterion (Marra & Radice, 2019).

2.4 Some inferential results

We describe next some inferential results that stem from the Bayesian interpretation of the smoothing process (Marra et al., 2017; Marra & Radice, 2019; Wahba, 1978; Wood, 2017; Silverman, 1985) and can be applied to the current context. Under this view, the overall penalty term $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_{\lambda}\boldsymbol{\beta}$ corresponds to the assumption of setting a prior density on $\boldsymbol{\beta}$, that is, $f(\boldsymbol{\beta}) \propto \exp(-\frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_{\lambda}\boldsymbol{\beta})$ or, equivalently, $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{S}_{\lambda}^{-})$, where $\mathbf{S}_{\lambda}^{-}$ denotes the pseudo-inverse of $\mathbf{S}_{\lambda}$. The penalized maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ corresponds to the posterior mode and, assuming a fixed value of $\boldsymbol{\lambda}$, inference proceeds using the large sample approximation to the posterior (Wood et al., 2016)

$$\boldsymbol{\beta} \mid \boldsymbol{y} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}}) \tag{2.20}$$

where $\mathbf{V}_{\boldsymbol{\beta}} = \left\{-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}}) + \mathbf{S}_{\lambda}\right\}^{-1}$ denotes the Bayesian posterior covariance matrix for the model parameters (for further details see, for example, Wood, 2017; Wood et al., 2016, and references therein). Assessments of uncertainty and tests of hypothesis about linear and non-linear functions of the parameters are then derived based on the covariance matrix in (2.20) instead of its frequentist counterpart as we describe next.

Point-wise Bayesian credible intervals (Wahba, 1983) for smooth model components can be constructed using

$$\left(\hat{s}_{m\bar{p}_m}(w_{m\bar{p}_m i}) - z_{\varsigma/2}\sqrt{v_{m\bar{p}_m i}}, \hat{s}_{m\bar{p}_m}(w_{m\bar{p}_m i}) + z_{\varsigma/2}\sqrt{v_{m\bar{p}_m i}}\right) \qquad i = 1, \ldots, n,$$

where $\hat{s}_{m\bar{p}_m}(w_{m\bar{p}_m i}) = \mathbf{b}_{m\bar{p}_m}^{\mathsf{T}}(w_{m\bar{p}_m i})\hat{\tilde{\boldsymbol{\alpha}}}_{m\bar{p}_m}$ is the estimated smooth function, $z_{\varsigma/2}$ corresponds to the $100(1-\varsigma/2)$ percentile of the standard normal, and $v_{m\bar{p}_m i} = \mathbf{b}_{m\bar{p}_m}^{\mathsf{T}}(w_{m\bar{p}_m i})\mathbf{V}_{\tilde{\boldsymbol{\alpha}}_{m\bar{p}_m}}\mathbf{b}_{m\bar{p}_m}(w_{m\bar{p}_m i})$, where $\mathbf{V}_{\tilde{\boldsymbol{\alpha}}_{m\bar{p}_m}}$ denotes the portion of the Bayesian covariance matrix corresponding to the regression parameters in the $\bar{p}_m^{\text{th}}$ smooth term of the m^{th} model equation. Nychka (1988), Marra & Wood (2012), and Marra & Radice (2019) showed that intervals for the smooth model components constructed using the Bayesian approach have closer to nominal coverage probabilities, when interpreted 'across-the-function' rather than pointwise, than those constructed using the frequentist results. Approximate credible intervals for non-linear functions of the parameters in the model can be constructed using simulation via the following steps (Radice et al., 2016)

CI.1 draw n^* vectors $\{\boldsymbol{\beta}_i^*\}_{i=1}^{n^*}$ from the posterior distribution in (2.20),

CI.2 compute the function of the parameters of interest $\{h(\boldsymbol{\beta}_i^*)\}_{i=1}^{n^*}$,

CI.3 obtain the approximate $100(1-\varsigma)\%$ credible intervals for $h(\boldsymbol{\beta})$ using the lower $\varsigma/2$ and upper $1-\varsigma/2$ quantiles of $\{h(\boldsymbol{\beta}_i^*)\}_{i=1}^{n^*}$.

Hypothesis testing about model components being equal to zero follow from the results shown in Wood (2012), Wood (2017, pp. 304-315), and the extension of Marra (2013). For instance, letting $\boldsymbol{\alpha}_{mp_j}$ denote a sub-vector of p_j parametric components from the m^{th} model predictor, and $\mathrm{V}_{\hat{\boldsymbol{\alpha}}_{mp_j}}$ the corresponding portion of the Bayesian covariance matrix then, under the null hypothesis $\mathrm{H}_0\colon \boldsymbol{\alpha}_{mp_j} = \mathbf{0}$, it holds that approximately,

$$\hat{\boldsymbol{\alpha}}_{mp_j}^{\mathsf{T}} \mathrm{V}_{\hat{\boldsymbol{\alpha}}_{mp_j}}^{-1} \hat{\boldsymbol{\alpha}}_{mp_j} \dot{\sim} \chi^2_{p_j}$$

provided $\mathrm{V}_{\hat{\boldsymbol{\alpha}}_{mp_j}}$ is not singular. Single-parameter hypothesis can also be re-written by using the standard normal as the reference distribution for the test. Moreover, testing whether a smooth term can be included in the model, i.e. $\mathrm{H}_0\colon s_{m\bar{p}_m} = 0$, can be achieved using the following test statistic (Wood, 2012)

$$\hat{\boldsymbol{s}}_{m\bar{p}_m}^{\mathsf{T}} \mathrm{V}_{\boldsymbol{s}_{m\bar{p}_m}}^{-} \hat{\boldsymbol{s}}_{m\bar{p}_m} \dot{\sim} \chi^2_{r_{m\bar{p}_m}},$$

where $\hat{\boldsymbol{s}}_{m\bar{p}_m} = (\hat{s}_{m\bar{p}_m}(w_{m\bar{p}_m 1}), \ldots, \hat{s}_{m\bar{p}_m}(w_{m\bar{p}_m n}))^{\mathsf{T}}$, the matrix $\mathrm{V}_{\boldsymbol{s}_{m\bar{p}_m}}^{-}$ corresponds to the pseudo-inverse of $\mathrm{V}_{\boldsymbol{s}_{m\bar{p}_m}} = \tilde{\mathbf{X}}_{m\bar{p}_m} \mathbf{V}_{\boldsymbol{\beta}} \tilde{\mathbf{X}}_{m\bar{p}_m}^{\mathsf{T}}$, and $\tilde{\mathbf{X}}_{m\bar{p}_m}$ is a matrix such that $\boldsymbol{s}_{m\bar{p}_m} = \tilde{\mathbf{X}}_{m\bar{p}_m}\boldsymbol{\beta}$. The degrees of freedom are computed as follows (Marra, 2013; Wood, 2012)

$$r_{m\bar{p}_m} = \begin{cases} \lfloor edf_{m\bar{p}_m} \rfloor & \text{if } edf_{m\bar{p}_m} - \lfloor edf_{m\bar{p}_m} \rfloor < 0.05, \\ \lfloor edf_{m\bar{p}_m} \rfloor + 1 & \text{otherwise,} \end{cases}$$

where $\lfloor \cdot \rfloor$ represents the floor function and $edf_{m\bar{p}_m}$ denotes the effective degrees of freedom associated with the $\bar{p}_m$ smooth term in the m^{th} model equation. For further details on the justification and construction of the test statistic, the derivation of its distribution, and the rationale behind the calculation of $r_{m\bar{p}_m}$, see Wood (2012).

The aforementioned inferential results are implemented in several routines from the `GJRM` (Marra & Radice, 2021) and `mgcv` (Wood, 2020) packages.

2.5 Simulation study

In this section, we perform several Monte Carlo experiments in order to (i) investigate the empirical properties of the semi-parametric approaches to estimation and compare them to those obtained using their classical parametric counterparts, (ii) assess the consequences of not accounting for non-linear effects flexibly, and (iii) evaluate the robustness of the results in the absence of the exclusion restriction assumption.

We generate the data using the observation rules given in (2.1) and the following specification of the switching mechanism and regime equations

$$
\begin{aligned}
y_{1i}^{*} &= \beta_{10} + \beta_{11} v_i + s_{11}(w_{1i}) + s_{12}(w_{2i}) + \epsilon_{1i}, \\
y_{2i}^{*} &= \beta_{20} + \beta_{21} v_i + s_{21}(w_{1i}) + \epsilon_{2i}, \\
y_{3i}^{*} &= \beta_{30} + \beta_{31} v_i + s_{31}(w_{1i}) + \epsilon_{3i},
\end{aligned}
$$

where β_{10} is set to either -3.05, -1.4, or -0.15 in order to assign, approximately, 20%, 40%, or 60% of observations to the first regime, while the values of $\beta_{11}, \beta_{20}, \beta_{21}, \beta_{30}$ and β_{31} are set to 1.2, 2.2, 1.3, 1.5, and 2.1, respectively. Following Marra & Radice (2011) and Marra & Radice (2013a), the covariates entering each additive predictor are obtained by first drawing $(v_i^{**}, w_{1i}^{**}, w_{2i}^{**})_{i=1}^{n}$ from a trivariate normal distribution with zero mean and correlation coefficients equal to 0.5, and then transforming them into uniformly distributed variables on the unit interval, using the cdf of the standard normal distribution, to obtain $(v_i^{*}, w_{1i}, w_{2i})_{i=1}^{n}$. Furthermore, the variable v_i^{*} is transformed into a binary variable using the following rule $v_i = \mathbb{1}(v_i^{*} > 0.5), \forall i$. Their procedure obtains a binary and two continuous on $(0, 1)$ variables with correlation coefficients approximately equal to 0.5. Note that the variable w_2 is used to fulfil the exclusion restriction assumption. We further set $\sigma_2 = \sigma_3 = 1$, while the correlation coefficients between the selection equation and the two regimes, ρ_{12} and ρ_{13}, are set to different combinations of $\{0.2, 0.4, 0.6, 0.8\}$. Lastly, the non-linear associations in the model equations are given by

$$
\begin{aligned}
s_{11}(w_1) &= 1 - w_1^3 - 2\exp(-180 w_1^2) - 2.3\sin(4.9 w_1), \\
s_{12}(w_2) &= -0.2(-0.3 - 1.3 w_2 + \cos(5 w_2)), \\
s_{21}(w_1) &= w_1 + \exp[-32(w_1 - 0.5)^2], \\
s_{31}(w_1) &= 0.3 + w_1 + \exp[-30(w_1 - 0.35)^2],
\end{aligned}
$$

whose graphs are shown in Figure 2.2. The data generating process corresponds to an endogenous switching regression model where the endogenous switching variable is observed without error. Note that the simulation experiments and the construction of the data generating process are similar to those that appear in Marra & Radice (2011) and Chib & Greenberg (2007) in the context of endogenous dummy variable models; and in Marra & Radice (2013a) in the context of sample selection models.

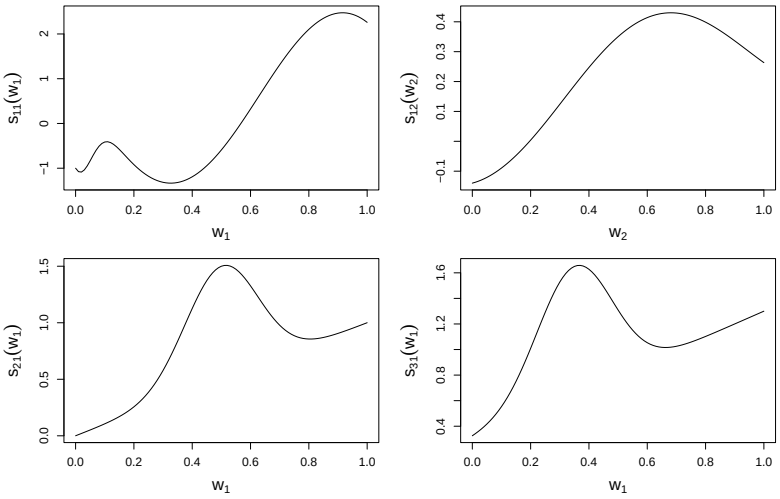

Figure 2.2: Graphs of the smooth functions representing non-linear effects used in the simulation study. Top left: $s_{11}(w_1) = 1 - w_1^3 - 2\exp(-180w_1^2) - 2.3\sin(4.9w_1)$. Top right: $s_{12}(w_2) = -0.2(-0.3 - 1.3w_2 + \cos(5w_2)$. Bottom left: $s_{21}(w_1) = w_1 + \exp[-32(w_1 - 0.5)^2]$. Bottom right: $s_{31}(w_1) = 0.3 + w_1 + \exp[-30(w_1 - 0.35)^2]$.

We perform $N = 200$ repetitions[5] for all different combinations of the values assigned to β_{10}, ρ_{12}, and ρ_{13} with sample sizes of $n = \{3000, 5000, 10000\}$, with and without the exclusion restriction assumption. We then estimate the parameters using: the semi-parametric approaches described in Section 2.3 via penalized maximum likelihood and penalized two-stage estimation, denoted as SML and S2S; and the classical maximum likelihood and two-stage methods, denoted as CML and 2S. The CML and 2S approaches are implemented in the `sampleSelection` package (Toomet & Henningsen, 2008b), where we have used third-order polynomials to model non-linear terms.

We summarise next a subset of the simulation experiments. In particular, we report the results for scenarios where the switching mechanism assigns approximately 40% of observations to the

[5]The number of repetitions is based on previous simulation studies from the literature (see, for example, Marra & Radice, 2011, 2013a). A more principled approach to choose the number of replications can be considered using the guidelines from Morris et al. (2019) and Boos & Stefanski (2013). For instance, focusing on the bias as a performance measure of the estimation method; using an initial run of $N_0 = 50$ repetitions to compute the preliminary estimates of the standard deviations for the estimated parameters of interest (say ρ_{12} and ρ_{13}) for the simulation scenario where their true values are $\rho_{12} = 0.4$ and $\rho_{13} = 0.6$ and the sample size is $n = 10000$; and aiming for a Monte Carlo standard error of 0.003 yields $N \approx 135$ and $N \approx 282$ repetitions.

first regime ($\beta_{10} = -1.4$), the parameter ρ_{12} takes values in $\{0.4, 0.6, 0.8\}$ and the value of ρ_{13} is fixed to 0.6. Figures 2.3 and 2.5 show boxplots of the estimates of β_{m1}, σ_m, and ρ_{1m}, for $m = 2, 3$, for experiments with and without an exclusion restriction. Figures 2.4 and 2.6 show the true smooth functions and the average mean effect estimates (solid lines) of $s_{m1}(w_1)$, for $m = 2, 3$, together with the 5% and 95% point-wise quantiles (shaded areas), for experiments with and without an exclusion restriction. For clarity, we only show the results of the estimated smooth effects obtained using the semi-parametric methods since the classical approaches fail to capture non-linearities adequately. In addition, Tables 2.1 and 2.2 summarise the subset of results in terms of relative bias and root mean squared error (RMSE) for the aforementioned model components.[6] The main findings are as follows:

- Overall, the boxplots show that estimates obtained using the semi-parametric methods are close to their true values and are less variable as the sample size increases for experiments with and without an exclusion restriction. Except for a few instances, the semi-parametric approaches tend to outperform their parametric counterparts in terms of relative bias and RMSE.

- In terms of the regression coefficients, β_{21} and β_{31}, both semi-parametric approaches perform better than their classical counterparts in scenarios with and without an exclusion restriction. The best results in term of bias and RMSE are given by the SML method. Moreover, the classical approaches appear to underestimate β_{21} and overestimate β_{31} slightly. This may be due to residual confounding since the effect of the continuous variable w_1 is not flexibly modelled in the CML and 2S approaches. The S2S method appears to obtain similar, and in several occasions less biased, results than the CML. In experiments without an exclusion restriction, estimates from the CML, S2S, and 2S are generally more biased, in particular those for β_{21} with relatively low values of ρ_{12}.

- In the case of the standard deviations, σ_2 and σ_3, and for experiments with and without an exclusion restriction, both maximum likelihood approaches deliver less biased estimates than their two-steps counterparts. The SML approach obtains the best overall results, except in experiments with a relatively low correlation coefficient between the selection equation

[6] The bias and RMSE of $\hat{s}_{m1}(w_1), m = 2, 3$, are obtained by evaluating the true and estimated smooth functions on a 200 points grid over $(0, 1)$ using $\text{bias}\{\hat{s}_{m1}(w_1)\} = \frac{1}{200}\sum_{i=1}^{200}\left|\frac{1}{200}\sum_{j=1}^{200}\hat{s}_{m1,j}(w_{1i}) - s_{m1}(w_{1i})\right|$ and $\text{RMSE}\{\hat{s}_{m1}(w_1)\} = \frac{1}{200}\sum_{i=1}^{200}\sqrt{\frac{1}{200}\sum_{j=1}^{200}[\hat{s}_{m1,j}(w_{1i}) - s_{m1}(w_{1i})]^2}$, for $m = 2, 3$. (Marra & Radice, 2013a, 2019; Wiesenfarth & Kneib, 2010).

and the first regime ($\rho_{12} = 0.4$), where the CML method performs slightly better in terms of precision. Nevertheless, the CML approach appears to underestimate σ_2 but overestimate σ_3, slightly. Lastly, the S2S method tends to perform better than its classical counterpart in terms of relative bias, where the latter seems to underestimate the true values of both parameters as the correlation values increase. However, the 2S approach tends to deliver better RMSE results.

- With respect to the correlation coefficients, ρ_{12} and ρ_{13}, the SML method obtains the best results overall. Furthermore, the S2S approach generally performs better than both classical approaches in terms of relative bias. Both CML and 2S methods appear to slightly underestimate the value of ρ_{12}. Omitting the exclusion restriction does not seem to affect much the estimates obtained using the semi-parametric approaches however, they are more biased when obtained using the parametric methods.

- With regard to the smooth components, $s_{21}(w_1)$ and $s_{31}(w_1)$, Figures 2.4 and 2.6 show that non-linear terms are appropriately recovered by the semi-parametric methods and that the SML outperforms the S2S in terms of relative bias and RMSE. As expected, the classical approaches fail to capture the true non-linear effects (not shown in the plots for clarity) and the estimates are severely biased.

In summary, the simulation experiments suggest that (i) the semi-parametric approaches have competitive empirical properties when compared with their parametric counterparts; (ii) the importance of flexibly modelling non-linear effects and the detrimental effect that they may have in the results; and (iii) the effect of the exclusion restriction on parameter estimates, in particular, for the 2S framework. Similar results have also been reported in Chib & Greenberg (2007) and Marra & Radice (2011, 2013a).

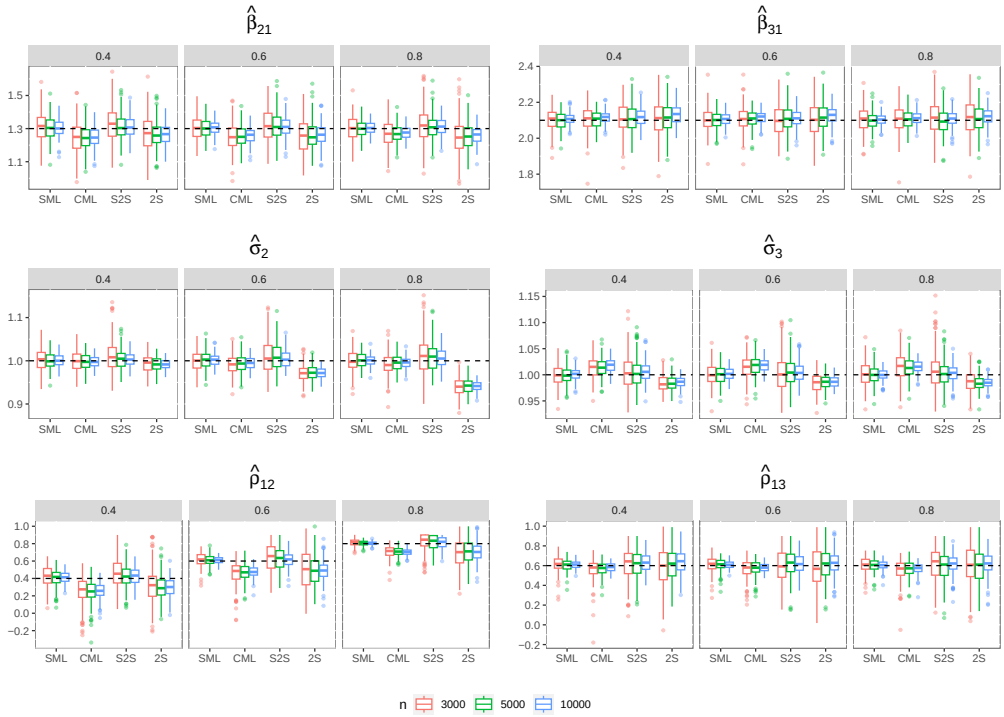

Figure 2.3: Boxplots of the estimates of the regression parameters, β_{21}, β_{31}, standard deviations σ_2, σ_3, and correlations coefficients, ρ_{12}, ρ_{13}, for experiments with an exclusion restriction where 40% of the observations are allocated to the first regime, ρ_{12} takes values in $\{0.4, 0.6, 0.8\}$, ρ_{13} is fixed at 0.6, with sample sizes of $n \in \{3000, 5000, 10000\}$. The real value of each parameter is indicated by a dashed line in each sub-plot. SML and S2S represent the semi-parametric penalized maximum likelihood and penalized two-step approaches, whereas CML and 2S denote the maximum likelihood and the two-stage classical approaches.

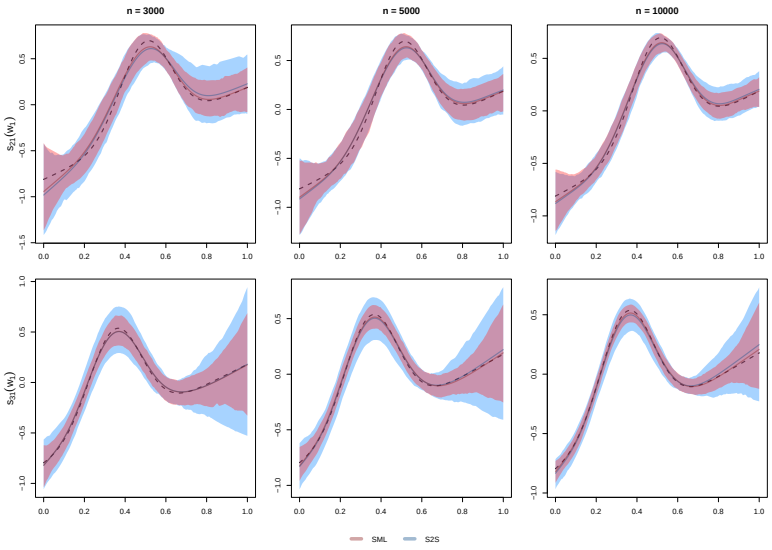

Figure 2.4: Mean estimates of $s_{21}(w_1)$ (top row) and $s_{31}(w_1)$ (bottom row) obtained using the SML (semi-parametric maximum likelihood, in red) and S2S (semi-parametric two-step, in blue) approaches for experiments with an exclusion restriction with sample sizes of $n = 3000$ (left column), $n = 5000$ (middle column), and $n = 10000$ (right column). The switching mechanism allocates 40% of the observations to the first regime, and the values of ρ_{12} and ρ_{13} are fixed to 0.4 and 0.6, respectively. The solid lines represent the average of the estimated smooth effects, while the shaded areas contain the 95% point-wise inner quantiles. The true functions are represented by dashed lines.

		Relative bias			RMSE			Relative bias			RMSE		
ρ_{12}	method | n	3000	5000	10000	3000	5000	10000	3000	5000	10000	3000	5000	10000
		$\hat{\beta}_{21}$						$\hat{\beta}_{31}$					
0.4	SML	**0.0093**	**0.0023**	**0.0022**	**0.0869**	**0.0699**	**0.0489**	**0.0028**	**-0.0006**	**0.0037**	**0.0612**	**0.0478**	**0.0309**
	CML	-0.0376	-0.0431	-0.0396	0.1063	0.0942	0.0744	0.0033	0.0025	0.0082	0.0704	0.0496	0.0364
	S2S	0.0234	0.0076	0.0079	0.1069	0.0801	0.0602	0.0038	0.0028	0.0083	0.0869	0.0745	0.0507
	2S	-0.0188	-0.0282	-0.0260	0.1094	0.0890	0.0696	0.0042	0.0070	0.0152	0.0973	0.0796	0.0615
0.6	SML	**0.0024**	**0.0038**	**0.0053**	**0.0721**	**0.0579**	**0.0426**	0.0015	**0.0004**	**0.0029**	**0.0627**	**0.0457**	**0.0345**
	CML	-0.0397	-0.0337	-0.0306	0.0977	0.0775	0.0625	0.0033	0.0037	0.0076	0.0684	0.0480	0.0390
	S2S	0.0138	0.0118	0.0072	0.1014	0.0853	0.0589	**-0.0004**	0.0035	0.0058	0.0923	0.0734	0.0495
	2S	-0.0327	-0.0323	-0.0297	0.1141	0.0961	0.0737	0.0023	0.0069	0.0130	0.1008	0.0782	0.0578
0.8	SML	**0.0011**	**-0.0001**	**0.0026**	**0.0701**	**0.0515**	**0.0363**	**0.0030**	**-0.0010**	**0.0017**	**0.0660**	**0.0466**	**0.0316**
	CML	-0.0240	-0.0222	-0.0194	0.0788	0.0613	0.0462	0.0040	0.0019	0.0056	0.0723	0.0508	0.0379
	S2S	0.0137	0.0089	0.0085	0.1058	0.0777	0.0546	0.0065	-0.0020	0.0040	0.0939	0.0717	0.0456
	2S	-0.0393	-0.0335	-0.0322	0.1187	0.0927	0.0679	0.0091	0.0040	0.0103	0.1006	0.0767	0.0545
		$\hat{\sigma}_2$						$\hat{\sigma}_3$					
0.4	SML	0.0019	**0.0004**	**0.0005**	0.0263	0.0191	0.0140	**0.0003**	**-0.0011**	**0.0016**	**0.0201**	**0.0160**	**0.0116**
	CML	**-0.0010**	-0.0021	-0.0025	**0.0241**	**0.0177**	**0.0132**	0.0141	0.0139	0.0180	0.0249	0.0219	0.0220
	S2S	0.0105	0.0036	0.0032	0.0376	0.0240	0.0162	0.0043	0.0041	0.0065	0.0322	0.0274	0.0199
	2S	-0.0066	-0.0077	-0.0084	0.0225	0.0178	0.0139	-0.0156	-0.0157	-0.0136	0.0228	0.0212	0.0167
0.6	SML	**-0.0001**	**0.0021**	**0.0018**	**0.0238**	**0.0199**	**0.0145**	**0.0008**	**0.0008**	**0.0021**	**0.0206**	**0.0173**	**0.0118**
	CML	-0.0080	-0.0059	-0.0054	0.0248	0.0212	0.0154	0.0149	0.0168	0.0181	0.0267	0.0249	0.0219
	S2S	0.0099	0.0088	0.0039	0.0416	0.0322	0.0205	0.0028	0.0058	0.0056	0.0349	0.0277	0.0204
	2S	-0.0286	-0.0275	-0.0283	0.0346	0.0320	0.0306	-0.0157	-0.0140	-0.0136	0.0236	0.0198	0.0170
0.8	SML	**0.0001**	**0.0010**	**0.0006**	**0.0241**	**0.0187**	**0.0126**	**0.0015**	**-0.0005**	**0.0000**	**0.0221**	**0.0176**	**0.0108**
	CML	-0.0073	-0.0046	-0.0049	0.0267	0.0209	0.0142	0.0155	0.0143	0.0157	0.0281	0.0238	0.0197
	S2S	0.0124	0.0087	0.0056	0.0465	0.0343	0.0236	0.0075	0.0009	0.0029	0.0362	0.0258	0.0169
	2S	-0.0599	-0.0579	-0.0587	0.0636	0.0601	0.0597	-0.0131	-0.0159	-0.0151	0.0230	0.0215	0.0177
		$\hat{\rho}_{12}$						$\hat{\rho}_{13}$					
0.4	SML	**0.0633**	**0.0211**	**0.0231**	**0.1202**	**0.0953**	**0.0656**	0.0200	**0.0061**	**0.0158**	**0.0751**	**0.0583**	**0.0392**
	CML	-0.3522	-0.3853	-0.3654	0.2118	0.1990	0.1704	-0.0608	-0.0547	-0.0306	0.1083	0.0743	0.0492
	S2S	0.1572	0.0553	0.0637	0.1900	0.1331	0.0968	**0.0190**	0.0270	0.0525	0.1693	0.1420	0.1021
	2S	-0.1985	-0.2640	-0.2533	0.2197	0.1799	0.1467	-0.0237	0.0186	0.0699	0.2116	0.1687	0.1275
0.6	SML	**0.0251**	**0.0153**	**0.0170**	**0.0774**	**0.0598**	**0.0417**	0.0211	**0.0153**	**0.0142**	**0.0779**	**0.0605**	**0.0399**
	CML	-0.2260	-0.2210	-0.2104	0.1897	0.1607	0.1407	-0.0504	-0.0439	-0.0362	0.0983	0.0747	0.0511
	S2S	0.0644	0.0471	0.0250	0.1733	0.1347	0.0903	**-0.0101**	0.0330	0.0349	0.1739	0.1460	0.1019
	2S	-0.1700	-0.1934	-0.1842	0.2415	0.2026	0.1563	-0.0290	0.0210	0.0520	0.2136	0.1736	0.1234
0.8	SML	**0.0162**	**0.0067**	**0.0037**	**0.0397**	**0.0284**	**0.0215**	**0.0141**	**0.0061**	**0.0061**	**0.0773**	**0.0633**	**0.0434**
	CML	-0.1174	-0.1177	-0.1200	0.1145	0.1063	0.1020	-0.0709	-0.0606	-0.0519	0.1117	0.0860	0.0613
	S2S	0.0190	0.0132	0.0170	0.1013	0.0842	0.0691	0.0243	-0.0073	0.0233	0.1671	0.1412	0.0994
	2S	-0.1356	-0.1157	-0.1216	0.2063	0.1741	0.1422	0.0077	-0.0011	0.0325	0.2076	0.1706	0.1188
		$\hat{s}_{21}(w_1)$						$\hat{s}_{31}(w_1)$					
0.4	SML	**0.0394**	**0.0305**	**0.0234**	**0.1367**	**0.1117**	**0.0815**	**0.0150**	**0.0127**	**0.0094**	**0.1115**	**0.0843**	**0.0631**
	CML	0.2715	0.2686	0.2669	0.3176	0.3032	0.2861	0.1776	0.1744	0.1658	0.2226	0.1967	0.1802
	S2S	0.0572	0.0387	0.0314	0.1739	0.1305	0.0961	0.0140	0.0139	0.0221	0.1568	0.1306	0.096
	2S	0.2457	0.2479	0.2478	0.3170	0.2940	0.2763	0.1910	0.1785	0.1657	0.2765	0.2336	0.2014
0.6	SML	**0.0398**	**0.0325**	**0.0178**	**0.1209**	**0.0931**	**0.0666**	0.0175	**0.0077**	**0.0105**	**0.1172**	**0.0873**	**0.0611**
	CML	0.2662	0.2608	0.2585	0.3119	0.2870	0.2744	0.1690	0.1680	0.1618	0.2165	0.1946	0.1753
	S2S	0.0546	0.0440	0.0220	0.1695	0.1308	0.0909	**0.0172**	0.0138	0.0171	0.1729	0.1354	0.0944
	2S	0.2588	0.2610	0.2597	0.3324	0.3087	0.2863	0.1886	0.1776	0.1663	0.2790	0.2381	0.1988
0.8	SML	**0.0338**	**0.0262**	**0.0191**	**0.1024**	**0.0812**	**0.0604**	**0.0129**	**0.0114**	0.0092	**0.1151**	**0.0910**	**0.0602**
	CML	0.2463	0.2467	0.2442	0.2805	0.2686	0.2576	0.1668	0.1648	0.1664	0.2157	0.1934	0.1814
	S2S	0.0499	0.0388	0.0282	0.1513	0.1175	0.0867	0.0185	0.0126	**0.0089**	0.1656	0.1308	0.0888
	2S	0.2707	0.2658	0.2659	0.3306	0.3082	0.2904	0.1761	0.1794	0.1755	0.2621	0.2381	0.2075

Table 2.1: Relative bias and RMSE for $\hat{\beta}_{12}, \hat{\beta}_{13}, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\rho}_{12}, \hat{\rho}_{13}$, and smooth functions estimates $\hat{s}_{21}(w_1)$ and $\hat{s}_{31}(w_1)$, for experiments with an exclusion restriction where 40% of the observations allocated to the first regime, ρ_{12} takes values in $\{0.4, 0.6, 0.8\}$, ρ_{13} is fixed at 0.6, with sample sizes $n \in \{3000, 5000, 10000\}$.The lowest relative bias and RMSE values for each combination of the correlation coefficients and the sample sizes are in bold. SML and S2S represent the semi-parametric penalized maximum likelihood and penalized two-step approaches, whereas CML and 2S are the maximum likelihood and the two-stage classical approaches. The true values of the parameters are $\beta_{21} = 1.3, \beta_{31} = 2.1, \sigma_2 = \sigma_3 = 1, \rho_{12} \in \{0.4, 0.6, 0.8\}$, and $\rho_{13} = 0.6$. The true smooth functions are $s_{21}(w_1) = w_1 + \exp[-32(w_1 - 0.5)^2]$ and $s_{31}(w_1) = 0.3 + w_1 + \exp[-30(w_1 - 0.35)^2]$.

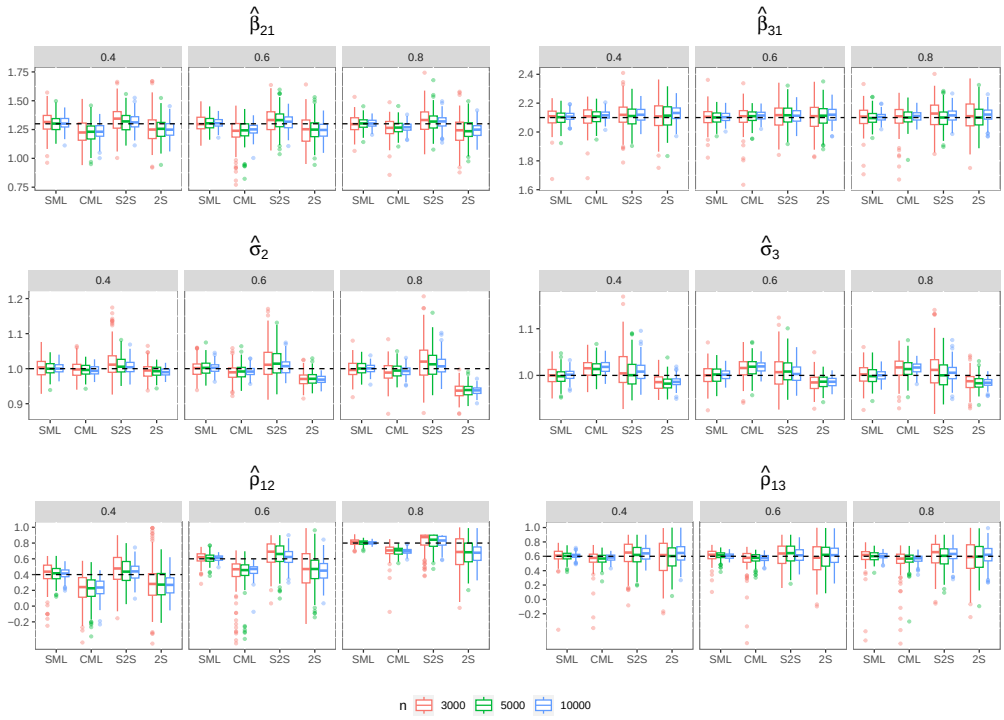

Figure 2.5: Boxplots of the estimates of the regression parameters, β_{21}, β_{31}, standard deviations σ_2, σ_3, and correlations coefficients, ρ_{12}, ρ_{13}, for experiments without an exclusion restriction where 40% of the observations are allocated to the first regime, ρ_{12} takes values in $\{0.4, 0.6, 0.8\}$, ρ_{13} is fixed at 0.6, with sample sizes of $n \in \{3000, 5000, 10000\}$. The real value of each parameter is indicated by a dashed line in each sub-plot. SML and S2S represent the semi-parametric penalized maximum likelihood and penalized two-step approaches, whereas CML and 2S denote the maximum likelihood and the two-stage classical approaches.

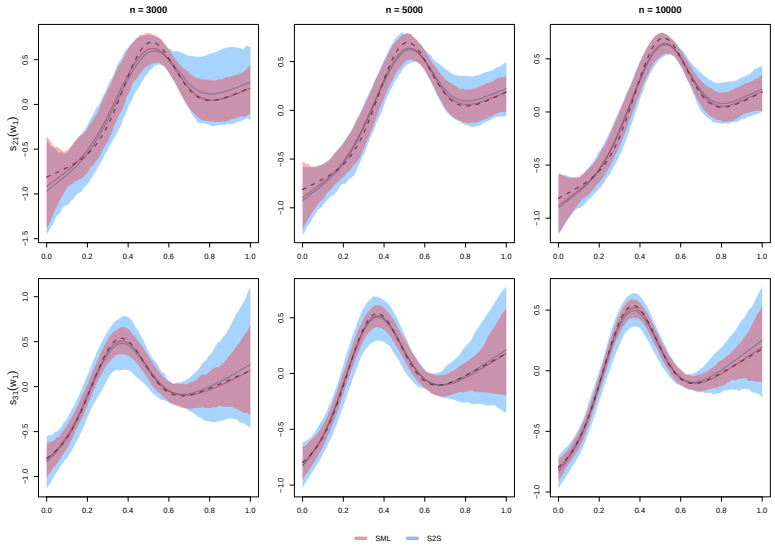

Figure 2.6: Mean estimates of $s_{21}(w_1)$ (top row) and $s_{31}(w_1)$ (bottom row) obtained using the SML (semi-parametric maximum likelihood, in red) and S2S (semi-parametric two-step, in blue) approaches for simulation experiments without an exclusion restriction and sample sizes of $n = 3000$ (left column), $n = 5000$ (middle column), and $n = 10000$ (right column). The switching mechanism allocates 40% of the observations to the first regime, and the values of ρ_{12} and ρ_{13} are fixed to 0.4 and 0.6, respectively. The solid lines represent the average of the estimated smooth effects, while the shaded areas contain the 95% point-wise inner quantiles. The true functions are represented by dashed lines.

		Relative bias			RMSE			Relative bias			RMSE		
ρ_{12}	method | n	3000	5000	10000	3000	5000	10000	3000	5000	10000	3000	5000	10000
		$\hat{\beta}_{21}$						$\hat{\beta}_{31}$					
0.4	SML	**0.0072**	**0.0000**	**0.0047**	**0.0967**	**0.0704**	**0.0534**	**0.0016**	**-0.0008**	**0.0032**	**0.0655**	**0.0483**	**0.0310**
	CML	-0.0523	-0.0596	-0.0510	0.1267	0.1186	0.0922	0.0018	0.0019	0.0076	0.0729	0.0502	0.0353
	S2S	0.0309	0.0129	0.0133	0.1221	0.0907	0.0713	0.0088	0.0027	0.0093	0.0975	0.0765	0.0539
	2S	-0.0297	-0.0385	-0.0370	0.1287	0.1092	0.0853	0.0040	0.0041	0.0141	0.1063	0.0850	0.0655
0.6	SML	**0.0027**	**0.0049**	**0.0053**	**0.0721**	**0.0595**	**0.0433**	**0.0021**	**-0.0001**	**0.0023**	**0.0654**	**0.0440**	**0.0325**
	CML	-0.0552	-0.0458	-0.0384	0.1303	0.1055	0.0719	0.0021	0.0028	0.0064	0.0793	0.0470	0.0363
	S2S	0.0248	0.0231	0.0106	0.1144	0.1010	0.0687	0.0056	0.0058	0.0052	0.0928	0.0754	0.0497
	2S	-0.0425	-0.0388	-0.0431	0.1371	0.1122	0.0904	0.0022	0.0043	0.0093	0.1001	0.0826	0.0577
0.8	SML	**-0.0002**	**0.0015**	**0.0024**	**0.0705**	**0.0516**	**0.0371**	0.0013	-0.0012	**0.0017**	**0.0736**	**0.0466**	**0.0315**
	CML	-0.0293	-0.0242	-0.0244	0.0871	0.0636	0.0513	**0.0005**	**0.0008**	0.0047	0.0817	0.0544	0.0372
	S2S	0.0221	0.0164	0.0131	0.1191	0.0864	0.0617	0.0093	-0.0009	0.0067	0.1011	0.0735	0.0482
	2S	-0.0491	-0.0424	-0.0431	0.1402	0.1071	0.0837	0.0049	0.0021	0.0088	0.1095	0.0807	0.0563
		$\hat{\sigma}_2$						$\hat{\sigma}_3$					
0.4	SML	**0.0020**	**0.0007**	**0.0010**	0.0275	0.0194	0.0150	**0.0006**	**-0.0020**	**0.0013**	**0.0200**	**0.0165**	**0.0117**
	CML	-0.0028	-0.0033	-0.0049	0.0241	0.0179	**0.0143**	0.0148	0.0133	0.0183	0.0256	0.0219	0.0224
	S2S	0.0162	0.0080	0.0056	0.0463	0.0279	0.0198	0.0118	0.0039	0.0082	0.0415	0.0297	0.0220
	2S	-0.0068	-0.0078	-0.0098	**0.0233**	**0.0176**	0.0149	-0.0146	-0.0164	-0.0135	0.0227	0.0221	0.0169
0.6	SML	**0.0002**	**0.0023**	**0.0018**	**0.0247**	**0.0209**	**0.0144**	**0.0009**	**0.0008**	**0.0017**	**0.0208**	**0.0171**	**0.0119**
	CML	-0.0109	-0.0086	-0.0086	0.0263	0.0230	0.0167	0.0154	0.0171	0.0180	0.0269	0.0252	0.0223
	S2S	0.0176	0.0163	0.0065	0.0498	0.0414	0.0245	0.0081	0.0089	0.0058	0.0386	0.0304	0.0227
	2S	-0.0299	-0.0288	-0.0307	0.0361	0.0338	0.0328	-0.0156	-0.0140	-0.0139	0.0240	0.0201	0.0178
0.8	SML	**-0.0007**	**0.0009**	**0.0006**	**0.0246**	**0.0197**	**0.0129**	**0.0016**	**-0.0008**	**0.0003**	**0.0224**	**0.0182**	**0.0113**
	CML	-0.0097	-0.0068	-0.0074	0.0286	0.0223	0.0155	0.0154	0.0143	0.0158	0.0284	0.0244	0.0200
	S2S	0.0206	0.0139	0.0098	0.0579	0.0416	0.0301	0.0118	0.0024	0.006	0.0409	0.0293	0.0205
	2S	-0.0630	-0.0613	-0.0619	0.0668	0.0639	0.0631	-0.0136	-0.0159	-0.0151	0.0232	0.0219	0.0179
		$\hat{\rho}_{12}$						$\hat{\rho}_{13}$					
0.4	SML	**0.0365**	**0.0163**	**0.0327**	**0.1514**	**0.0984**	**0.0730**	**0.0108**	**-0.0006**	**0.0118**	**0.1013**	**0.0606**	**0.0403**
	CML	-0.4591	-0.4829	-0.4464	0.2656	0.2530	0.2123	-0.0695	-0.0625	-0.0335	0.1324	0.0795	0.0510
	S2S	0.1893	0.1009	0.0906	0.2230	0.1617	0.1213	0.0514	0.0223	0.0621	0.1966	0.1501	0.1104
	2S	-0.2835	-0.3197	-0.3353	0.2742	0.2270	0.1889	-0.0225	-0.0103	0.0655	0.2505	0.1934	0.1462
0.6	SML	**0.0244**	**0.0132**	**0.0152**	**0.0742**	**0.0612**	**0.0400**	**0.0188**	0.0134	**0.0085**	**0.0847**	**0.0576**	**0.0404**
	CML	-0.2923	-0.2804	-0.2420	0.2644	0.2333	0.1633	-0.0688	-0.0482	-0.0434	0.1494	0.0763	0.0550
	S2S	0.1016	0.0834	0.0358	0.1808	0.1652	0.1090	0.0345	0.0551	0.0303	0.1820	0.1540	0.1116
	2S	-0.2121	-0.2268	-0.2472	0.2789	0.2429	0.1982	-0.0373	**0.0013**	0.0223	0.2253	0.1938	0.1420
0.8	SML	**0.0136**	**0.0059**	**0.0027**	**0.0388**	**0.0298**	**0.0204**	**-0.0016**	**0.0014**	**0.0070**	**0.1377**	**0.0630**	**0.0453**
	CML	-0.1282	-0.1269	-0.132	0.1342	0.1137	0.1116	-0.1004	-0.0719	-0.0574	0.1774	0.1077	0.0656
	S2S	0.0238	0.0211	0.0235	0.1147	0.0931	0.0770	0.0474	-0.0025	0.0478	0.1934	0.1568	0.1094
	2S	-0.1634	-0.1529	-0.1579	0.2577	0.2124	0.1816	-0.0308	-0.0217	0.0238	0.2455	0.1993	0.1373
		$\hat{s}_{21}(w_1)$						$\hat{s}_{31}(w_1)$					
0.4	SML	**0.0373**	**0.0309**	**0.0269**	**0.1547**	**0.1138**	**0.0872**	**0.0155**	0.0133	**0.0082**	**0.1177**	**0.0815**	**0.0593**
	CML	0.2963	0.2904	0.2876	0.3490	0.3303	0.3109	0.1716	0.1681	0.1600	0.2240	0.1896	0.1731
	S2S	0.0632	0.0464	0.0375	0.2031	0.1460	0.1129	0.0248	**0.0127**	0.0251	0.1825	0.1325	0.0992
	2S	0.2628	0.2600	0.2674	0.3517	0.3148	0.2994	0.1797	0.1756	0.1593	0.2886	0.2400	0.2017
0.6	CML	**0.0409**	**0.0338**	**0.0189**	**0.1206**	**0.0967**	**0.0686**	**0.0180**	**0.0081**	**0.0090**	**0.1136**	**0.0823**	**0.0593**
	CML	0.2894	0.2822	0.2720	0.3493	0.3230	0.2886	0.1634	0.1622	0.1568	0.2255	0.1882	0.1709
	S2S	0.0667	0.0552	0.0260	0.1841	0.1573	0.1053	0.0232	0.0235	0.0157	0.1737	0.1402	0.0979
	2S	0.2729	0.2741	0.2816	0.3550	0.3306	0.3097	0.1774	0.1736	0.1643	0.2718	0.2439	0.2038
0.8	SML	**0.0349**	**0.0270**	**0.0192**	**0.1043**	**0.0838**	**0.0592**	**0.0122**	**0.0117**	**0.0094**	**0.1359**	**0.0885**	**0.0569**
	CML	0.2499	0.2505	0.2492	0.2901	0.2737	0.2628	0.1682	0.1609	0.1605	0.2384	0.1967	0.175
	S2S	0.0610	0.0458	0.0332	0.1697	0.1311	0.0966	0.0294	0.0119	0.0180	0.1833	0.1391	0.0934
	2S	0.2854	0.2834	0.2836	0.3540	0.3318	0.3132	0.1778	0.1748	0.1683	0.2845	0.2437	0.2077

Table 2.2: Relative bias and RMSE for $\hat{\beta}_{12}, \hat{\beta}_{13}, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\rho}_{12}, \hat{\rho}_{13}$, and smooth functions estimates $\hat{s}_{21}(w_1)$ and $\hat{s}_{31}(w_1)$, for experiments without an exclusion restriction where 40% of the observations allocated to the first regime, ρ_{12} takes values in $\{0.4, 0.6, 0.8\}$, ρ_{13} is fixed at 0.6, and sample sizes $n \in \{3000, 5000, 10000\}$.The lowest relative bias and RMSE values for each combination of the correlation coefficients and the sample sizes are in bold. SML and S2S represent the semi-parametric penalized maximum likelihood and penalized two-step approaches, whereas CML and 2S are the maximum likelihood and the two-stage classical approaches. The true values of the parameters are $\beta_{21} = 1.3, \beta_{31} = 2.1, \sigma_2 = \sigma_3 = 1, \rho_{12} \in \{0.4, 0.6, 0.8\}$, and $\rho_{13} = 0.6$. The true smooth functions are $s_{21}(w_1) = w_1 + \exp[-32(w_1 - 0.5)^2]$ and $s_{31}(w_1) = 0.3 + w_1 + \exp[-30(w_1 - 0.35)^2]$.

2.6 Empirical application

The Medical Expenditure Panel Survey (MEPS) is a large-scale survey, conducted by the Agency for Healthcare Research and Quality in the United States, that contains information on families and individuals such as their demographic and socio-economic characteristics, their use of healthcare, and details on their health insurance coverage (further information about the survey can be found at `https://meps.ahrq.gov`). In the analysis, we use a subset of the MEPS dataset, available in Cameron & Trivedi (2009), consisting of individuals over 65 years old, their out-of-pocket prescribed drug expenditures, and several other socio-economic and health-related factors. Subjects considered in the study are automatically enrolled in Medicare (a federal health insurance programme) which, at the time of data collection, did not cover for any prescribed medicines. As a consequence, individuals may choose to obtain supplementary private services, in the form of employer-based or union-sponsored insurance, to cover against certain out-of-pocket expenses (see Cameron & Trivedi, 2009, for further details). We focus on individuals with full information on all the variables of interest and whose out-of-pocket prescribed drug expenditures are over a hundred dollars per year. Zimmer (2013) provides the justification for concentrating on these individuals: first, models for healthcare demand in the US are usually based on the distinction of two economic processes that are assumed to be statistically independent: one that determines whether individuals spend, and one that governs how much they spend (see, for example, Pohlmeier & Ulrich, 1995). Second, from a government budget perspective, these individuals represent a priority since the US Healthcare Financing Administration reports that approximately 10% of Medicare users account for 70% of Medicare spending (see Zimmer, 2013, for further details).

The analysis aims to investigate the effects of having employer-based or union-sponsored supplementary insurance on out-of-pocket prescribed drugs expenditures under different modelling approaches while accounting for several other socio-economic and health-related factors, namely `age`, `gender`, `race`, `log(income)`, and their total number of `chronic` conditions. Information on whether an individual has any extra form of insurance is given by the binary variable `supplementary`, while the outcome of interest is the total amount of out-of-pocket expenditures on prescribed medicines (for further details, see, Cameron & Trivedi, 2009). Since healthcare expenditures variables tend to be highly skewed (Deb & Norton, 2018; Manning & Mullahy, 2001), it is common to use the logarithm of out-of-pocket expenditures, `log(expenditure)`, as the response. Figure 2.7 shows the histograms of out-of-pocket prescribed drug expenditures

(left) together with the histograms of their log transformations (right) at both levels of the insurance variable. Having supplementary insurance raises concerns of potential endogeneity, since it

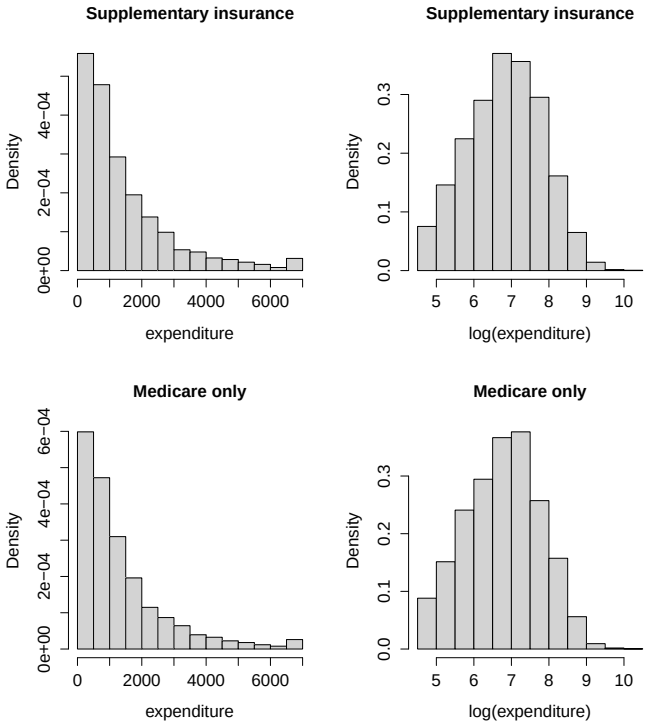

Figure 2.7: Histograms of `expenditure` (left) and `log(expenditure)` (right) by insurance status. For visual presentation only, we have top coded individuals with expenditures greater than $\$7,000$ for the plots on the left hand-side.

is plausible that there are some unobserved individual characteristics that affect simultaneously prescribed drugs expenditures and obtaining supplementary insurance (Deb et al., 2006). For example, individuals may had chosen to work in a job that would provide an extra coverage in order to benefit from having lower medical expenses on retirement (Cameron & Trivedi, 2009).

Table 2.3 contains the description and summary statistics of all the variables considered in the analysis, at sample level and by type of insurance. The proportion of individuals that have some form of supplementary health insurance is about 38%, there is a slightly higher proportion of females than males, 16% of the individuals self-identified as Black or Hispanic, and the average age is close to 75. Health-wise, the average number of chronic conditions in the sample is under 2. Regarding individuals with or without supplementary health insurance we observe that, overall, those with Medicare-only coverage have lower expenditures, are older, poorer, and have a slightly lower average number of chronic conditions. There is also a higher proportion of females and Black or Hispanic individuals covered only by Medicare compared to those who have supplementary insurance.

Variable	Description	Full sample (n = 9113)		Supplementary insurance (n = 3509)		Medicare only (n = 5604)	
		Mean	SD	Mean	SD	Mean	SD
Utilization							
`log(expenditure)`	Logarithm of drug expenditures	6.789	0.997	6.828	0.999	6.765	0.996
Insurance							
`supplementary`	=1 if supplementary health insurance	0.385	0.487	1	0	0	0
Socio-economic/health variables							
`gender`	=1 if female	0.583	0.493	0.518	0.5	0.624	0.484
`race`	=1 if individual self-identified as Black or Hispanic	0.160	0.367	0.131	0.337	0.178	0.383
`age`	Age in years (> 65)	75.083	6.645	73.778	6.523	75.900	6.591
`log(income)`	Logarithm of annual household income in thousands of dollars	2.746	0.907	2.951	0.932	2.618	0.866
`chronic`	Number of chronic conditions	1.959	1.287	1.964	1.307	1.956	1.275
Instruments							
`ssiratio`	ssiratio = $\frac{\text{social security income}}{\text{total income}}$	0.539	0.369	0.441	0.356	0.600	0.363
`multloc`	=1 if firm has multiple locations	0.060	0.238	0.098	0.297	0.037	0.189

Table 2.3: Description and summary statistics of the variables for the full sample and by type of insurance.

Cameron & Trivedi (2009) propose and discuss the validity of several potential variables to fulfil the exclusion restriction assumption based on individual and employer characteristics. For instance, the ratio of a person's social security income to their total income, denoted by `ssiratio`, and the binary indicator `multloc`, which indicates whether the firm an individual worked for has multiple locations. Recall that such variables must satisfy the conditions stated in Section 2.2, that is, `ssiratio` and/or `multloc` must be conditionally independent with respect to `log(expenditure)`, and have a high partial correlation with `supplementary` after accounting for the rest of the covariates. The authors argue the validity and relevance of `ssiratio` by assuming that income is included in each regime equation through the explanatory variable `log(income)`, and that `ssiratio` is expected to be negatively correlated with having extra insurance. The variable `multloc` is potentially a week instrument, it captures employer-based supplementary insurance but individuals in the sample are already retired. However, the literature in health economics provides several case studies, in similar settings, where `multloc` has been used as instrument, for example, Deb et al. (2006), Deb & Trivedi (2006), and Marra et al. (2020).

We proceed by specifying the following additive predictors for the selection mechanism and regime equations

$$\begin{aligned}
\eta_{1i} &= \beta_{10} + \beta_{11}\texttt{gender}_i + \beta_{12}\texttt{race}_i + \beta_{13}\texttt{multloc}_i + s_{11}(\texttt{age}_i) + s_{12}(\texttt{log(income)}_i) \\
&\quad + s_{13}(\texttt{chronic}_i) + s_{14}(\texttt{ssiratio}_i), \\
\eta_{2i} &= \beta_{20} + \beta_{21}\texttt{gender}_i + \beta_{22}\texttt{race}_i + s_{21}(\texttt{age}_i) + s_{22}(\texttt{log(income)}_i) + s_{23}(\texttt{chronic}_i), \\
\eta_{3i} &= \beta_{30} + \beta_{31}\texttt{gender}_i + \beta_{32}\texttt{race}_i + s_{31}(\texttt{age}_i) + s_{32}(\texttt{log(income)}_i) + s_{33}(\texttt{chronic}_i).
\end{aligned}$$

where $s_{m1}(\cdot), s_{m2}(\cdot)$, and $s_{m3}(\cdot)$, for $m = 1, 2, 3$, are smooth functions represented using thin plate regression splines. We then fit univariate GAMs (denoted as 1GAM) to each regime separately, the classical fully parametric (CML), and the semi-parametric (SML) ESR models. Preliminary analyses using the CML approach, where the continuous covariates were modelled using second order polynomials, failed to converge and did not provide values for the standard errors of the estimates. We settled with specifying linear effects for all continuous covariates.

Table 2.4 contains the estimates of the parametric model components at each level of the switching variable obtained using the aforementioned approaches, together with their 95% confidence/credible intervals. The effects of `gender` and `race` have the same sign, for both regimes,

Model	1GAM		CML		SML	
Estimates | Regimes	Supplementary	Medicare	Supplementary	Medicare	Supplementary	Medicare
`(Intercept)`	6.814 (6.769, 6.859)	6.800 (6.759, 6.840)	6.354 (5.962, 6.745)	7.468 (7.086, 7.850)	6.517 (6.352, 6.683)	7.215 (7.176, 7.254)
`gender:female`	0.045 (-0.016, 0.106)	-0.023 (-0.072, 0.027)	-0.018 (-0.084, 0.049)	-0.103 (-0.160, -0.047)	-0.002 (-0.071, 0.067)	-0.098 (-0.153, -0.044)
`race:Black/Hispanic`	-0.075 (-0.166, 0.016)	-0.113 (-0.176, -0.049)	-0.129 (-0.224, -0.034)	-0.180 (-0.250, -0.109)	-0.108 (-0.203, -0.014)	-0.167 (-0.236, -0.098)
`age`	(smooth term) -	(smooth term) -	-0.010 (-0.015, -0.004)	-0.013 (-0.018, -0.009)	(smooth term) -	(smooth term) -
`log(income)`	(smooth term) -	(smooth term) -	0.065 (0.024, 0.106)	0.058 (0.024, 0.093)	(smooth term) -	(smooth term) -
`chronic`	(smooth term) -	(smooth term) -	0.307 (0.282, 0.331)	0.309 (0.289, 0.329)	(smooth term) -	(smooth term) -
$\hat{\sigma}_2$	0.908 (0.885, 0.932)	- -	0.989 (0.936, 1.042)	- -	0.960 (0.929, 0.985)	- -
$\hat{\sigma}_3$	- -	0.902 (0.883, 0.920)	- -	1.016 (0.965,1.066)	- -	1.020 (0.994,1.040)
$\hat{\rho}_{12}$	- -	- -	0.458 (0.321, 0.595)	- -	0.371 (0.111, 0.507)	- -
$\hat{\rho}_{13}$	- -	- -	- -	0.588 (0.470, 0.706)	- -	0.616 (0.580,0.647)

Table 2.4: Parameter estimates and 95% confidence/credible intervals of the parametric model components obtained after fitting a univariate generalized additive model (1GAM) for each regime, the classical (CML), and the semi-parametric (SML) endogenous switching regression models.

when estimated using CML and SML but they appear to have a slightly higher magnitude when using the CML approach. This may be a result of the CML approach not modelling the effects of the continuous variables flexibly. Overall, the results indicate that females have lower out-of-pocket expenditures than men, and that Black or Hispanic individuals have lower expenditures compared to White individuals, regardless of the type of insurance. The effect of `gender` appears to be significant only for individuals in the Medicare-only regime, whereas the effect of `race` is stronger for persons without extra insurance. In particular, estimates obtained using the SML approach for those without supplementary insurance point out that females spend about 10% less than men, and that Black or Hispanic individuals spends about 17% less than White individuals. The corre-

sponding results using the 1GAM approach suggest that `gender` is not a determinant factor of `log(expenditure)`, and that `race` is significant only for individuals in the Medicare regime.

Estimates of σ_2 and σ_3 obtained via the CML and SML frameworks are very similar and suggest a higher variability in the Medicare-only regime. On the other hand, these parameters appear to be slightly under-estimated when using 1GAM models. In regard to the correlation coefficients, their estimates have the same sign but a somewhat different magnitude under the CML and SML approaches. The estimate of ρ_{12} obtained under the CML approach suggests a stronger association between `supplementary` and `log(expenditure)` than that obtained using the SML framework. The CML estimate of ρ_{13} suggest a lower strength of dependence than the one estimated via the SML model. Both correlation coefficients are significant, suggesting the presence of unobserved individual characteristics that influence the uptake of supplementary health insurance.

Parametric CML estimates of `age`, `log(income)`, and the number of `chronic` conditions suggest that `chronic` is the main determinant of `log(expenditure)` followed by `log(income)`. Both effects are positive, significant, and have a similar magnitude in both regimes. Individual's `age` has a negative, but relatively low effect on the response for both levels of insurance. Figure 2.8 (top row) shows the estimated smooth functions, and 95% credible intervals, obtained by the SML approach for the supplementary (blue) and the Medicare-only (red) regimes. The plots are centered around zero due to the identifiability constraint placed on the smooth functions (see Section 2.2). The estimated non-linear effects on `log(expenditure)` of `age`, `chronic`, and `log(income)` appear to be similar for individuals with extra insurance and for those using only Medicare. The results show that the effect of `age` on the average of `log(expenditure)` is non-linear, significant, and somewhat variable in both regimes, with an overall downward trend, which suggests that individuals appear to spend less as they get older. The effect of `chronic` shows an expected pattern and suggests that as the number of `chronic` conditions increases, the average of `log(expenditure)` increases steeply for both regimes. In terms of `log(income)`, we observe an overall downward trend for individuals with low annual income, which appears to be steeper for those without extra insurance, followed by an upward trend for individuals with higher incomes. We also observe that `log(income)` has little effect on the response for individuals with lower income levels. These results seem to be consistent with the CML estimates in particular, `chronic` has the largest effect on `log(expenditure)`, `age` has a negative but weak effect, and `log(income)` has a slightly lower impact on the response

for those with extra insurance. The estimated curves obtained using the 1GAM (not shown on the same figures for clarity) are similar to those obtained using the SML however, the p-values resulting from testing the significance of the smooth terms suggest that `age` and `log(income)` are not significantly different from zero in the supplementary regime, while `log(income)` is not significant in the Medicare regime.

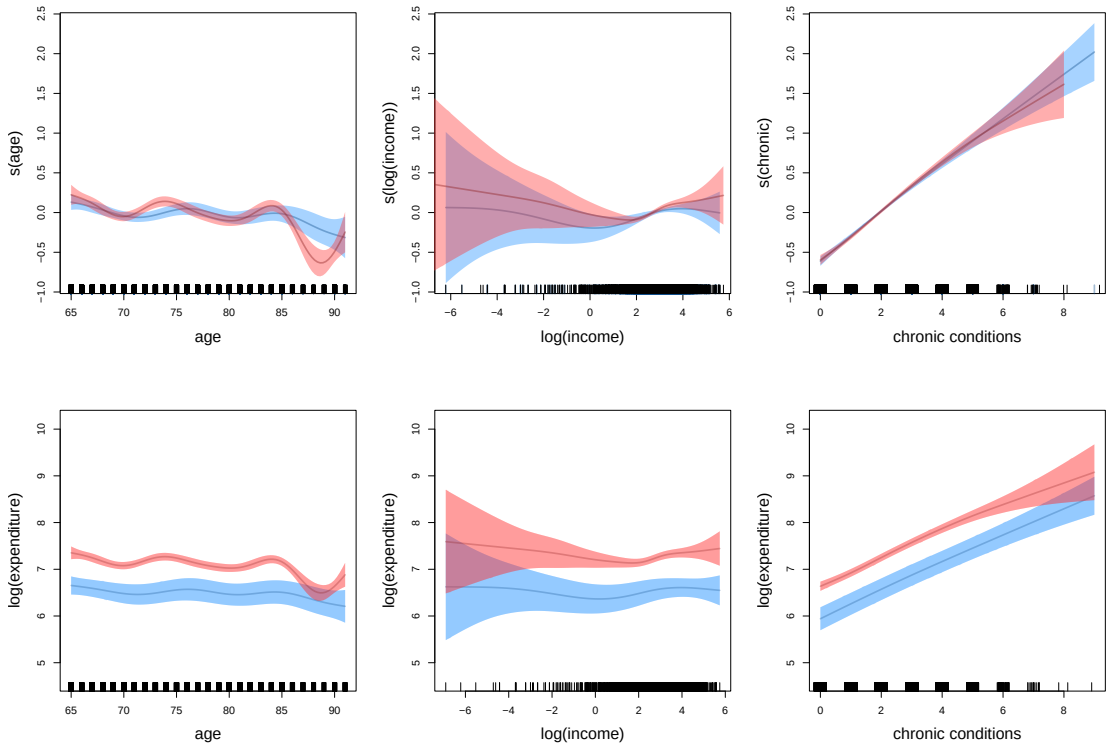

Figure 2.8: Top row: Estimated smooth functions and 95% pointwise credible intervals of `age`, `log(income)`, and `chronic` for individuals with and without supplementary insurance. The smooths estimates are vertically centered about zero due to the identifiability restrictions placed on the smooths terms. Bottom row: Estimated smooth effects and 95% pointwise credible intervals of `age`, `log(income)`, and `chronic` on `log(expenditure)` with the remaining of the covariates set at their mean/mode. Estimates in the supplementary regime are shown in blue whereas those in the Medicare regime are shown in red. The jittered rug plot in the x-axis indicates the values of the covariates in the data.

To aid further interpretation, the bottom row of Figure 2.8 shows the estimated smooth functions in the scale of `log(expenditure)` where the rest of the covariates have been set to their means or modes. In particular, they display the effect of the specific covariate on the mean `log(expenditure)` for White females where the other two continuous covariates have been set to their mean.

An estimate of the average treatment effect (ATE) of type of insurance on expenditure for a randomly chosen individual can also be obtained using the model parameter estimates (see, for example, Heckman et al., 2001). The CML ATE estimate is -1160.67. [7] In contrast the SML

[7]The `sampleSelection` package only provides the tools to obtain the conditional expectations to calculate the

estimate, together with its 95% credible interval, is $-1255.82(-1450.28, -1019.29)$. [8] The result indicates that, on average, having supplementary insurance leaves a randomly chosen individual about $\$1255$ better off in terms of prescribed drugs expenditures.

2.7 Discussion

In this chapter, we have presented a semi-parametric ESR model that relaxes the functional form specification of the deterministic model components using a well-established penalized regression spline framework. We have also described two approaches to parameter estimation and several inferential results from the penalized regression literature which are relevant to this context. The simulation study has shown that the semi-parametric ESR model provides competitive results when compared to the classical approaches, and it has also highlighted the importance of flexibly modelling the relationships between the response of interest and the predictors.

In an application, we have investigated the effects of insurance, and several other socio-economic and health-related factors, on prescribed drugs expenditures for US individuals over 65 years old. Our findings suggest individual self-selection into insurance, and that the main determinants of expenditures are the number of individual's chronic conditions, followed by race. The semi-parametric approach also reveals that the continuous covariates affect the response with different degrees of non-linearities. The estimated ATE indicates that having supplementary insurance constitutes a saving of about $\$1255$.

The modelling approach presented in this chapter still suffers from some of the criticisms introduced in Section 2.1. For instance, the joint normality, which also implies normally distributed regimes, and the constant variance assumptions may be restrictive or inappropriate in empirical applications. In Chapter 3, we explore extensions of the semi-parametric ESR model in several ways. First, the joint distributions of the switching variable and each of the regime responses are not restricted to bivariate Gaussian, and are specified using parametric bivariate copula functions. Second, the regime responses are specified using several families of parametric distributions. And third, each of the distribution parameters can be modelled using a flexible linear predictor of covariates that accounts for several types of effects.

ATE estimate but not to construct the confidence intervals.

[8] Credible intervals are constructed using the simulation steps described in Section 2.4 implemented in the `GJRM` package.

Chapter 3

Copula-based endogenous switching regression models with an application in health economics

This chapter presents copula-based endogenous switching regression models for continuous responses. In an application, we study insurance uptake in the US to cover for out-of-pocket expenditures of prescriptions drugs. The modelling approach allows for the joint distributions of the switching variable and the regimes to be specified using copula functions, whereas the univariate marginal components are specified employing parametric families of distributions. The models are embedded into a distributional regression framework, where all the distribution parameters can be defined using additive predictors that account for several types of covariate effects. Parameter estimation and inference utilise a well-established penalized likelihood framework. The results obtained from the copula-based approach suggest that large values of out-of-pocket prescribed drug expenditures are accompanied by a higher chances of having supplementary insurance however, low expenditures do not necessarily imply lower chances of having extra insurance. This feature cannot be captured under the assumptions of the classical and semi-parametric approaches.

3.1 Introduction

As pointed out in Chapters 1 and 2, the normality assumption in the context of models subject to sample selection has often been criticised since it limits the applicability of these models when the researcher wants to consider more complex distributions for the response (Klein & Kneib, 2016;

Pigini, 2015; Puhani, 2000; Smith, 2003). Furthermore, the association between the switching and regime variables is tied to the correlation coefficient, which only measures linear dependence, and whose use as a measure of association has also been subject to criticism (Embrechts et al., 2002). Although there are multivariate alternatives to the normal, they tend to be restrictive in applications, since they do not allow to specify different distributions for each of the marginals, and measures of dependence appear in the specification of the marginal distributions (Frees & Valdez, 1998; Kotz et al., 2004).

Some of the aforementioned limitations can be dealt with by using a copula-based approach. The copula approach to modelling jointly determined random variables allows to separate the specification of their marginal distributions from their dependence structure by specifying their joint distribution in terms of their marginals and a copula function that joins them together. The definition and properties of a copula function (an m-dimensional copula is a multivariate cdf with m standard uniform margins) and the work of Sklar (1959) and Patton (2006) provide the most important results for statistical modelling using copulas. Sklar shows that any multivariate distribution can be represented as a composition of a copula function and its univariate margins, while Patton extends Sklar's theorem to a situation where the univariate margins are conditional distribution functions.

As a consequence of these results, and in the context of the chapter, statistical modelling using copulas is approached in two steps: first, specifying a parametric model for each of the latent variables in the ESR model; second, choosing an appropriate copula function that links the univariate components together and captures their dependence structure. The modelling approach permits to relax the joint normality assumption between the switching and each regime variables, accommodates different types of marginal distributions, and accounts for several forms of dependence structures (implied by the copula) between the model components.

Copula-based regression models are not new in the literature. Lee (1983) developed a SS model, based on the bivariate Gaussian copula, in which the marginals could be assumed to belong to different families of distributions; Prieger (2002) proposed SS models based on the FGM copula to model hospitalization stays; and Smith (2003, 2005) introduced SS and ESR models using Archimedean copulas with non-normal continuous margins. More recently, the copula-based regression literature has focused on using flexible distributional specifications in which each of the distribution parameters can be flexibly modelled using predictors that incorporate several types of covariates effects. For example, bivariate regression models embedded into the GAM/GAMLSS

frameworks from the Bayesian (Klein & Kneib, 2016) and frequentist perspectives (Marra & Radice, 2017; Vatter & Chavez-Demoulin, 2015); and GAM- and GAMLSS-type specifications of SS models (Marra et al., 2017; Wiemann et al., 2022; Wojtyś & Marra, 2015; Wojtyś et al., 2018).

In this chapter, and based on the modelling framework of Marra et al. (2017) and Wojtyś et al. (2018), we present flexible copula-based ESR models for continuous regime responses. In an application, we study insurance uptake in the US to cover for out-of-pocket expenses of prescription drugs, and make an assessment of whether different univariate distributions and dependence structures provide some insights that cannot be captured using the classical distributional assumptions of the ESR model. The adopted framework allows for the joint distributions of the switching and each of the regime variables to be modelled using copula functions to capture their dependence structure. The distributions of the variables that govern the switching mechanism and both regimes are embedded into the GAMLSS framework (Rigby & Stasinopoulos, 2005). In particular, the distribution that models the switching mechanism is not restricted to be Gaussian, whereas the distributions of the regime variables can be modelled using several parametric continuous distribution functions. Furthermore, all the distribution parameters can be modelled using a vector of explanatory variables to accommodate various forms of covariate effects. Parameter estimation and inference follow from the well-established penalized likelihood framework of Marra et al. (2017) and Marra & Radice (2019) described in the previous chapter. At the time of writing, the models presented here have become a subset of those that appear in Marra et al. (2022), implemented in the `GJRM` package (Marra & Radice, 2022), which further allows for the distribution of the regime variables to be binary or discrete and the copula parameters to be specified as a function of covariates.

In the application, we find evidence of individual self-selection into insurance and that the association between insurance uptake and out-of-pocket expenditures for prescribed drugs is best described by the Joe and the Gumbel copula families. The dependence structures implied by the copula functions suggest that individuals with higher expected out-of-pocket expenditures are more likely to uptake supplementary insurance.

The rest of this chapter is structured as follows: Section 3.2 describes copula-based ESR models, the specification of the univariate and joint distributions functions, and the structure of the generic predictor used to model each of the univariate distribution parameters. Section 3.3 considers parameter estimation, while Section 3.4 evaluates the empirical properties of the copula-based approach via simulation. In Section 3.5, we analyse the effects of insurance status and several

socio-economic and health related characteristics on out-of-pocket prescribed drug expenditures using the copula-based specification of the ESR model. Lastly, in Section 3.6 we discuss the modelling approach and the empirical results.

3.2 Specification of copula-based endogenous switching regression models

In this section, we describe in detail the structure of copula-based ESR models. First, we revisit the observation rules that govern the ESR model, and then we show how the model likelihood function can be written in terms of the joint distribution functions that characterise the switching and each regime variables, and their univariate marginals. In Section 3.2.1 we specify the marginal distributions using the GAMLSS framework, whereas in Section 3.2.2 we specify the joint distributions in terms of their marginals and bivariate copula functions that link them together. Section 3.2.3 describes the structure of the additive predictor associated with each of the parameters of the marginal distributions.

Recall from Chapter 2 that observations from the ESR model are generated using the following rules

$$Y_{1i} = \mathbb{1}_{Y_{1i}^*>0}(Y_{1i}^*), \qquad Y_{2i} = Y_{1i}Y_{2i}^*, \qquad Y_{3i} = (1-Y_{1i})Y_{3i}^*, \quad i = 1, \ldots, n. \tag{3.1}$$

The Bernoulli random variable Y_{1i} is determined by the sign of the continuous latent variable Y_{1i}^* through the indicator function $\mathbb{1}_{Y_{1i}^*>0}(\cdot)$. The continuous random variables Y_{2i} and Y_{3i} are determined by Y_{1i} and their latent counterparts Y_{2i}^* and Y_{3i}^*, that is, when $Y_{1i} = 1$ we observe $Y_{2i} = Y_{2i}^*$ otherwise, we observe $Y_{3i} = Y_{3i}^*$.

Assume that each of the latent variables in the model can be described by parametric families of distributions, conditional on covariates, and denote their pdfs and cdfs as $f_m(y_m^* \mid \boldsymbol{\theta}_m)$ and $F_m(y_m^* \mid \boldsymbol{\theta}_m)$, respectively, where $\boldsymbol{\theta}_m \in \mathbb{R}^{\tilde{n}_m}$ represents a vector of $\tilde{n}_m$ distribution parameters, for $m = 1, 2, 3$. Let us also denote the joint cdf of the pair of latent random variables (Y_1^*, Y_m^*) as $F_{1m}(y_1^*, y_m^* \mid \boldsymbol{\theta}_{1m})$, where $\boldsymbol{\theta}_{1m} \in \mathbb{R}^{\tilde{n}_{1m}}$ represents a vector of $\tilde{n}_{1m}$ of distribution parameters, for $m = 2, 3$.

Given a random sample of observations, and omitting the distribution parameters and covariates for simplicity, the likelihood function of the ESR model can be written in a generic form as a

function of the univariate and bivariate distributions of the latent variables in (3.1). Conditioning on the switching variable, note that when $y_{1i} = 1$ we observe y_{2i} and the contribution to the likelihood corresponds to $f(y_{2i} \mid y_{1i}^* > 0)\mathbb{P}(Y_{1i}^* > 0)$. Otherwise, we observe y_{3i} and the contribution to the likelihood is $f(y_{3i} \mid y_{1i}^* \leq 0)\mathbb{P}(Y_{1i}^* \leq 0)$, that is,

$$\prod_{i=1}^{n} \left\{ f(y_{2i} \mid y_{1i}^* > 0)\mathbb{P}(Y_{1i}^* > 0) \right\}^{y_{1i}} \left\{ f(y_{3i} \mid y_{1i}^* \leq 0)\mathbb{P}(Y_{1i}^* \leq 0) \right\}^{1-y_{1i}}, \tag{3.2}$$

where $f(y_{2i} \mid y_{1i}^* > 0)$ and $f(y_{3i} \mid y_{1i}^* \leq 0)$ denote the conditional densities of Y_{2i}^* and Y_{3i}^* given $Y_{1i}^* > 0$ and $Y_{1i}^* \leq 0$, respectively. The conditional densities can be written as

$$\begin{aligned} f(y_{2i} \mid y_{1i}^* > 0) = f(y_{2i} \mid y_{1i} = 1) &= \frac{\partial F(y_{2i} \mid y_{1i} = 1)}{\partial y_{2i}} \\ &= \frac{\partial}{\partial y_{2i}} \left[\frac{F_{12}(1, y_{2i})}{1 - F_1(0)} \right] \\ &= \frac{1}{1 - F_1(0)} \frac{\partial}{\partial y_{2i}} \left[F_2(y_{2i}) - F_{12}(0, y_{2i}) \right] \\ &= \frac{1}{1 - F_1(0)} \left[f_2(y_{2i}) - \frac{\partial F_{12}(0, y_{2i})}{\partial y_{2i}} \right], \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} f(y_{3i} \mid y_{1i}^* \leq 0) = f(y_{3i} \mid y_{1i} = 0) &= \frac{\partial F\left(y_{3i} \mid y_{1i} = 0\right)}{\partial y_{3i}} \\ &= \frac{\partial}{\partial y_{3i}} \left[\frac{F_{13}\left(0, y_{3i}\right)}{F_1(0)} \right] \\ &= \frac{1}{F_1(0)} \frac{\partial F_{13}\left(0, y_{3i}\right)}{\partial y_{3i}}. \end{aligned} \tag{3.4}$$

Substituting expressions (3.3) and (3.4) back into (3.2) obtains (Smith, 2003, 2005)

$$\prod_{i=1}^{n} \left\{ f_2(y_{2i}) - \frac{\partial F_{12}(0, y_{2i})}{\partial y_{2i}} \right\}^{y_{1i}} \left\{ \frac{\partial F_{13}(0, y_{3i})}{\partial y_{3i}} \right\}^{1-y_{1i}}. \tag{3.5}$$

The form of the likelihood function implies that it is not necessary to specify the joint distribution of $(Y_{1i}^*, Y_{2i}^*, Y_{3i}^*)$ in order to estimate the model parameters (Smith, 2005). Furthermore, the joint cdf of the pair (Y_{2i}^*, Y_{3i}^*) does not appear in the likelihood function and any plausible parameter(s) measuring their dependence cannot be identified. Note that, assuming the joint distributions F_{12} and F_{13} are bivariate normal, the likelihood function given in (3.5) is equivalent to the likelihood function of the classical ESR (see, for example, Smith, 2005).

3.2.1 Specification of the marginal distributions: F_1, F_2 and F_3

We now specify the marginal distribution for each of the latent variables using the GAMLSS framework. This approach assumes that each of the conditional distributions in the model belongs to a parametric family whose parameters represent several distributional characteristics such as location, scale, and shape. Furthermore, each of the distribution parameters can be modelled as a function of covariates.

The model for the switching mechanism is specified as follows

$$Y_{1i}^* \sim F_1(\boldsymbol{\theta}_{1i}), \qquad i = 1, \ldots, n,$$

such that $\boldsymbol{\theta}_{1i} = (\mu_{1i}, \sigma_{1i})^\mathsf{T}$, where μ_{1i} and σ_{1i} are location and scale parameters, respectively. The location parameter is associated with an additive predictor of covariates and regression coefficients as follows

$$\mu_{1i} = \eta_i^{\mu_1}\left(\mathbf{x}_i^{\mu_1}, \boldsymbol{\beta}_{\mu_1}\right) = \beta_0^{\mu_1} + \sum_{p_{\mu_1}=1}^{P_{\mu_1}} h_{p_{\mu_1}}^{\mu_1}\left(\tilde{\mathbf{x}}_{p_{\mu_1} i}^{\mu_1}, \boldsymbol{\beta}_{p_{\mu_1}}^{\mu_1}\right), \tag{3.6}$$

where $\mathbf{x}_i^{\mu_1}$ is a vector of explanatory variables chosen to model μ_1, $\boldsymbol{\beta}_{\mu_1} = \left(\beta_0^{\mu_1}, \boldsymbol{\beta}_1^{\mu_1\mathsf{T}}, \ldots, \boldsymbol{\beta}_{P_{\mu_1}}^{\mu_1\mathsf{T}}\right)^\mathsf{T}$ is a vector of regression coefficients, $\beta_0^{\mu_1}$ is an overall intercept, and each $h_{p_{\mu_1}}^{\mu_1}(\cdot)$ denotes particular effects of a sub-vector of covariates $\tilde{\mathbf{x}}_{p_{\mu_1} i}^{\mu_1}$ contained in $\mathbf{x}_i^{\mu_1}$, for $p_{\mu_1} = 1, \ldots, P_{\mu_1}$. The specific structure of the predictor will be described in Subsection 3.2.3. The scale parameter σ_{1i} is set to 1 for the usual identification purposes.

Note that the model for the switching mechanism corresponds to a latent variable representation of a binary model and F_1 is usually chosen to be the normal, logistic, or Gumbel distributions since they yield the typical probit, logit, or complementary log-log link functions to model the probability of success of the observed variable Y_{1i}.

In terms of the models for the regime variables Y_{2i}^* and Y_{3i}^*, we assume that their true distributions belong to parametric families with up to three parameters, that is,

$$Y_{mi}^* \sim F_m(\boldsymbol{\theta}_{mi}), \;\; m = 2, 3, \; i = 1, \ldots, n,$$

such that $\boldsymbol{\theta}_{mi} = (\mu_{mi}, \sigma_{mi}, \nu_{mi})^\mathsf{T}$, where the parameters μ_{mi}, σ_{mi}, and ν_{mi} (often representing

location, scale, and shape) are associated with additive predictors of covariates as follows

$$
\begin{aligned}
g_{\mu_m}(\mu_{mi}) &= \eta_i^{\mu_m}\left(\mathbf{x}_i^{\mu_m}, \boldsymbol{\beta}_{\mu_m}\right) = \beta_0^{\mu_m} + \sum_{p_{\mu m}=1}^{P_{\mu m}} h_{p_{\mu_m}}^{\mu_m}\left(\tilde{\mathbf{x}}_{p_{\mu_m} i}^{\mu_m}, \boldsymbol{\beta}_{p_{\mu_m}}^{\mu_m}\right), \\
g_{\sigma_m}(\sigma_{mi}) &= \eta_i^{\sigma_m}\left(\mathbf{x}_i^{\sigma_m}, \boldsymbol{\beta}_{\sigma_m}\right) = \beta_0^{\sigma_m} + \sum_{p_{\sigma m}=1}^{P_{\sigma m}} h_{p_{\sigma_m}}^{\sigma_m}\left(\tilde{\mathbf{x}}_{p_{\sigma_m} i}^{\sigma_m}, \boldsymbol{\beta}_{p_{\sigma_m}}^{\sigma_m}\right), \\
g_{\nu_m}(\nu_{mi}) &= \eta_i^{\nu_m}\left(\mathbf{x}_i^{\nu_m}, \boldsymbol{\beta}_{\nu_m}\right) = \beta_0^{\nu_m} + \sum_{p_{\nu m}=1}^{P_{\nu m}} h_{p_{\nu_m}}^{\nu_m}\left(\tilde{\mathbf{x}}_{p_{\nu_m} i}^{\nu_m}, \boldsymbol{\beta}_{p_{\nu_m}}^{\nu_m}\right).
\end{aligned}
\tag{3.7}
$$

The functions $g_{\mu_m}(\cdot)$, $g_{\sigma_m}(\cdot)$, and $g_{\nu_m}(\cdot)$ are known, monotonic, and differentiable link functions that maintain the restrictions on the range of the distribution parameters. For instance, when μ_m is restricted to the positive real numbers, one can choose a log link, i.e., $g_{\mu_m}(\mu_{mi}) = \log(\mu_{mi}) = \eta_i^{\mu_m}$. The rest of the components in (3.7) are defined similarly to those that appear in (3.6). Subsection 3.2.3 describes in more detail the particular structure of these predictors.

3.2.2 Specification of the joint distributions: F_{12} and F_{13}

We specify next the joint distributions of the pairs (Y_1^*, Y_m^*), for $m = 2, 3$, using bivariate parametric copula functions. The copula approach to modelling jointly determined random variables allows to separate the univariate distributional components from the dependence structure by specifying the joint distribution of the random variables in terms of their marginals, and a copula function that joins them together. In our context, the copula approach relaxes the joint normality assumption between the switching and each regime variables, accommodates different types of marginal distributions, and accounts for several forms of dependence structures between the model components.

A 2-dimensional copula is a bivariate cdf with standard uniform margins, i.e., a function $\mathcal{C}\colon [0,1]^2 \to [0,1]$ defined by

$$
\mathcal{C}(u_1, u_2) = \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2), \text{ where } U_1, U_2 \sim \mathcal{U}(0,1),
$$

and satisfying the following conditions: (i) $\mathcal{C}(0,u) = \mathcal{C}(u,0) = 0, \forall u \in [0,1]$; (ii) $\mathcal{C}(1,u) = \mathcal{C}(u,1) = u, \forall u \in [0,1]$; (iii) $\mathcal{C}$ is $2-$increasing. Conditions (i) and (ii) are called the boundary conditions of the copula, whereas condition (iii) is the rectangle inequality.

The work of Sklar (1959) and Patton (2006) provide the most important results for statistical modelling with copulas. The former obtains a representation of a multivariate distribution function

as a composition of a copula and its univariate margins, while the latter extends Sklar's theorem to a situation where the univariate margins are conditional distribution functions. In the current context, the aforementioned results allow to specify the joint cdfs F_{12} and F_{13} in terms of copula functions linking the marginal distributions of the latent switching and regime variables, that is,

$$F_{1m}(y_{1i}^*, y_{mi}^* \mid \boldsymbol{\theta}_{1mi}) = \mathcal{C}_{1m}\left(F_1(y_{1i}^* \mid \boldsymbol{\theta}_{1i}), F_m(y_{mi}^* \mid \boldsymbol{\theta}_{mi}) \mid \delta_{1m}\right), \quad m = 2, 3,\ i = 1, \ldots, n,$$

where the vector $\boldsymbol{\theta}_{1mi} = (\boldsymbol{\theta}_{1i}^\mathsf{T}, \boldsymbol{\theta}_{mi}^\mathsf{T}, \delta_{1m})^\mathsf{T}$ encapsulates the distribution and copula parameters, and $\mathcal{C}_{1m}(\cdot, \cdot)$ is a parametric bivariate copula function that captures the dependence between the margins $F_1(\cdot)$, and $F_m(\cdot)$ through the association parameter δ_{1m}.

As a consequence of the copula-based specification, statistical modelling using the copula approach can be carried out in two steps: first, specify a GAMLSS model for each of the marginal cdfs (that do not need to belong to the same family of distributions); second, choose an appropriate copula function that links its univariate components together and captures their dependence structure. The literature provides a large number of parametric copula families, describing different types of dependence structures, that can be used for modelling (see, for example, Joe, 2014; Nelsen, 2006). The interpretation of the dependence parameter δ_{1m} is copula-specific and not comparable across different copulas. In practical applications, it is common to transform it to rank-based measures of dependence restricted to the $[-1, 1]$ interval such as the Kendall's tau (τ) or the Spearman's rho (ρ_s).

For further details in copula-based modelling, we refer the reader to Appendix B.1, which summarises the main results from the copula literature that are relevant to this thesis.

3.2.3 Structure of the additive predictors

Following the literature in distributional regression (see, for example, Fahrmeir et al., 2004; Klein et al., 2015; Marra & Radice, 2017) and omitting parameter-specific indices for clarity, each of the additive predictors of explanatory variables used to model the parameters of the marginal distributions in the model can be written as follows

$$\eta_i(\mathbf{x}_i, \boldsymbol{\beta}) = \beta_0 + \sum_{p=1}^{P} h_p(\tilde{\mathbf{x}}_{pi}, \boldsymbol{\beta}_p) \qquad i = 1, \ldots, n, \tag{3.8}$$

where $\beta_0 \in \mathbb{R}$ is an unknown regression coefficient denoting the overall level of the predictor, each $h_p(\cdot)$, for $p = 1, \ldots, P$, corresponds to an effect-specific function of a particular sub-vector

of covariates $\tilde{\mathbf{x}}_{pi}$ contained in $\mathbf{x}_i$, and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_P^{\mathsf{T}})^{\mathsf{T}}$ is a vector of regression coefficients. It is also assumed that each effect function can be written as

$$h_p(\tilde{\mathbf{x}}_{pi}, \boldsymbol{\beta}_p) = \sum_{j_p=1}^{J_p} b_{pj_p}(\tilde{\mathbf{x}}_{pi})\beta_{pj_p} = \mathbf{b}_p^{\mathsf{T}}(\tilde{\mathbf{x}}_{pi})\boldsymbol{\beta}_p, \qquad p = 1, \ldots, P, \tag{3.9}$$

where $\mathbf{b}_p(\tilde{\mathbf{x}}_{pi}) = [b_{p1}(\tilde{\mathbf{x}}_{1i}), \ldots, b_{pJ_p}(\tilde{\mathbf{x}}_{pi})]^{\mathsf{T}}$ is a vector of known functions evaluated at $\tilde{\mathbf{x}}_{pi}$ (whose particular form depends on the type of covariate(s) included in $\tilde{\mathbf{x}}_{pi}$), and $\boldsymbol{\beta}_p = (\beta_{p1}, \ldots, \beta_{pJ_p})^{\mathsf{T}}$ is a vector of unknown regression coefficients. The linear combination of functions and regression parameters in (3.9) allows for each effect-specific function to be written in vector-matrix notation as $\tilde{\mathbf{X}}_p\boldsymbol{\beta}_p$, where $\tilde{\mathbf{X}}_p$ is an $(n \times J_p)$-matrix whose $(i, j_p)^{\text{th}}$ element is given by $b_{pj_p}(\tilde{\mathbf{x}}_{pi})$. Furthermore, to ensure that β_0 corresponds to an overall intercept, identifiability constraints must be applied to each $h_p(\cdot)$ in the overall predictor before fitting the model (Wood, 2017, pp. 250).

In addition, each effect-specific function admits a quadratic penalty term that enforces particular properties of the effect given by $\lambda_p\boldsymbol{\beta}_p^{\mathsf{T}}\mathbf{S}_p\boldsymbol{\beta}_p$, for some unknown penalty parameter $\lambda_p \geq 0$, that needs to be estimated, and some penalty matrix $\mathbf{S}_p$, whose particular structure depends on the form of the vector $\mathbf{b}_p(\tilde{\mathbf{x}}_{pi})$.

The generic predictor described in Equation (3.8) provides the ESR model with a flexible structure since each of the distribution parameters can be modelled using several types of covariate effects, for example:

- Linear effects: when modelling parametric linear effects, the effect-specific function becomes $h_p(\tilde{\mathbf{x}}_{pi}, \boldsymbol{\beta}_p) = \tilde{\mathbf{x}}_{pi}^{\mathsf{T}}\boldsymbol{\beta}_p$, that is, a classical linear regression setup where $\tilde{\mathbf{x}}_{pi}$ is a subvector of covariates (often binary or categorical) and $\boldsymbol{\beta}_p$ a vector of regression coefficients. Linear effects are not usually penalized so the corresponding penalty matrix $\mathbf{S}_p$ is made up of zeroes (Marra et al., 2017).

- Non-linear effects: for non-parametric effects of continuous covariates, the effect-specific functions in (3.9) are represented using the penalized regression spline framework described in Section 2.2.1 of Chapter 2. That is, for a particular subset of continuous covariates, $h_p(\tilde{\mathbf{x}}_{pi}, \boldsymbol{\beta}_p) = \mathbf{b}_p^{\mathsf{T}}(\tilde{\mathbf{x}}_{pi})\boldsymbol{\beta}_p$, where $\mathbf{b}_p$ is a vector of known basis functions evaluated at the covariates and $\boldsymbol{\beta}_p$ denotes a vector of regression parameters. The smoothing parameter λ_p controls the trade-off between smoothness and goodness-of-fit of the function, and the particular structure of the penalty matrix $\mathbf{S}_p$ depends on the type of basis functions chosen to represent the effect.

- Spatial effects: when the data provide some form of discrete geographic information, for example locations or regions made up of L discrete adjacent units, the effect-specific function is written as $h_p(\tilde{\mathbf{x}}_{pi}, \boldsymbol{\beta}_p) = \tilde{\mathbf{x}}_{pi}^{\mathsf{T}}\boldsymbol{\beta}_p$, where $\tilde{\mathbf{x}}_{pi}$ is a vector whose l^{th} element is equal to 1 if observation i belongs to unit l and 0 otherwise, for $l = 1, \ldots, L$. The regression coefficient vector $\boldsymbol{\beta}_p$ represents the location/region effects. Assuming that neighbouring locations are likely to share similar effects, the penalty matrix $\mathbf{S}_p$ corresponds to an adjacency matrix whose $(i,j)^{\text{th}}$ element is defined as -1 if $i \neq j$ and locations i and j are adjacent; 0 if $i \neq j$ and locations i and j are not adjacent; and N_i if $i = j$, where N_i is the total number of neighbours for region i. For further details see, for example, Stasinopoulos et al. (2017, pp. 293-296) and references therein.

The aforementioned specification allows to write all the additive predictors in the model using a vector-matrix notation. For example, the additive predictor related to the location parameter μ_1 in (3.2.1) can be written as $\boldsymbol{\eta}_{\mu_1} = \bar{\mathbf{X}}_{\mu_1}\boldsymbol{\beta}_{\mu_1}$, where $\bar{\mathbf{X}}_{\mu_1} = (\mathbf{1}, \tilde{\mathbf{X}}_1^{\mu_1}, \ldots, \tilde{\mathbf{X}}_{P_{\mu_1}}^{\mu_1})$ is an overall design matrix, $\mathbf{1}$ denotes a vector of ones, each $\tilde{\mathbf{X}}_{p_{\mu_1}}^{\mu_1}$ represents an $(n \times J_{p_{\mu_1}})$-matrix that contains the effect-specific functions evaluated at the p^{th} subset of covariates contained in $\mathbf{x}^{\mu_1}$, for $p_{\mu_1} = 1, \ldots, P_{\mu_1}$; and $\boldsymbol{\beta}_{\mu_1} = \left(\beta_0^{\mu_1}, \boldsymbol{\beta}_1^{\mu_1\mathsf{T}}, \ldots, \boldsymbol{\beta}_{P_{\mu_1}}^{\mu_1\mathsf{T}}\right)^{\mathsf{T}}$ is the corresponding vector of unknown regression coefficients. In addition, the overall quadratic penalty associated with $\boldsymbol{\eta}_{\mu_1}$ can be written as $\boldsymbol{\beta}_{\mu_1}^{\mathsf{T}}\bar{\mathbf{S}}_{\mu_1}\boldsymbol{\beta}_{\mu_1}$ where $\bar{\mathbf{S}}_{\mu_1}$ is a block diagonal penalty matrix given by $\bar{\mathbf{S}}_{\mu_1} = \text{diag}(0, \lambda_1^{\mu_1}\mathbf{S}_1^{\mu_1}, \ldots, \lambda_{P_{\mu_1}}^{\mu_1}\mathbf{S}_{P_{\mu_1}}^{\mu_1})$. An overall penalty vector $\bar{\boldsymbol{\lambda}}_{\mu_1}$, made up of all the penalty parameters included in $\bar{\mathbf{S}}_{\mu_1}$, can be written as $\bar{\boldsymbol{\lambda}}_{\mu_1} = (\lambda_1^{\mu_1}, \ldots, \lambda_{P_{\mu_1}}^{\mu_1})^{\mathsf{T}}$.

3.3 Parameter Estimation

Given a set of n independent observations and using the results given in Section 3.2, the log-likelihood function of the copula-based ESR is given by

$$\begin{aligned}\ell(\boldsymbol{\beta}) = &\sum_{i=1}^{n} y_{1i}\left\{\log f_2(y_{2i} \mid \boldsymbol{\theta}_{2i}) + \log\left(1 - \frac{\partial\, \mathcal{C}_{12}\left(F_1(0 \mid \boldsymbol{\theta}_{1i}), F_2(y_{2i} \mid \boldsymbol{\theta}_{2i}) \mid \delta_{12}\right)}{\partial F_2(y_{2i} \mid \boldsymbol{\theta}_{2i})}\right)\right\} \\ &+ \sum_{i=1}^{n}(1 - y_{1i})\left\{\log f_3(y_{3i} \mid \boldsymbol{\theta}_{3i}) + \log\left(\frac{\partial\, \mathcal{C}_{13}\left(F_1(0 \mid \boldsymbol{\theta}_{1i}), F_3(y_{3i} \mid \boldsymbol{\theta}_{3i}) \mid \delta_{13}\right)}{\partial F_3(y_{3i} \mid \boldsymbol{\theta}_{3i})}\right)\right\}.\end{aligned} \tag{3.10}$$

The vectors of distribution parameters are defined as $\boldsymbol{\theta}_{1i} = (\mu_{1i}, \sigma_{1i})^{\mathsf{T}} = \left(g_{\mu_1}^{-1}\left(\eta_i^{\mu_1}(\boldsymbol{\beta}_{\mu_1})\right), 1\right)^{\mathsf{T}}$

and $\boldsymbol{\theta}_{mi} = (\mu_{mi}, \sigma_{mi}, \nu_{mi})^{\mathsf{T}} = \left(g_{\mu_m}^{-1}\left(\eta_i^{\mu_m}(\boldsymbol{\beta}_{\mu_m})\right), g_{\sigma_m}^{-1}\left(\eta_i^{\sigma_m}(\boldsymbol{\beta}_{\sigma_m})\right), g_{\nu_m}^{-1}\left(\eta_i^{\nu_m}(\boldsymbol{\beta}_{\nu_m})\right)\right)^{\mathsf{T}}$, for $m = 2, 3$, where the functions $g^{-1}(\cdot)$ are the inverses of the link functions defined in Subsection 3.2.1, and each η_i represents an additive predictor, as described in Subsection 3.2.3, that depends on a vector of regression coefficients. The regression coefficients together with both copula parameters are encapsulated into the overall vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mu_1}^{\mathsf{T}}, \boldsymbol{\beta}_{\mu_2}^{\mathsf{T}}, \boldsymbol{\beta}_{\mu_3}^{\mathsf{T}}, \boldsymbol{\beta}_{\sigma_2}^{\mathsf{T}}, \boldsymbol{\beta}_{\sigma_3}^{\mathsf{T}}, \boldsymbol{\beta}_{\nu_2}^{\mathsf{T}}, \boldsymbol{\beta}_{\nu_3}^{\mathsf{T}}, \delta_{12}, \delta_{13})^{\mathsf{T}} \in \mathbb{R}^p$, where $p = \sum_{\kappa=1}^{7} p_\kappa + 2$ and $p_\kappa = \dim(\boldsymbol{\beta}_\kappa)$, for $\kappa \in \{\mu_1, \mu_2, \mu_3, \sigma_2, \sigma_3, \nu_2, \nu_3\}$.

Due to the flexible structures introduced in the additive predictors, directly maximising the log-likelihood function will result in over-fitting and a penalty term is added in order to control the fit (Green & Silverman, 1993; Marra et al., 2017; Wood, 2017). The penalized maximum likelihood estimator (PMLE) is defined as

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell_{\mathrm{p}}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \left\{\ell(\boldsymbol{\beta}) - \mathcal{P}(\boldsymbol{\beta}, \boldsymbol{\lambda})\right\}, \tag{3.11}$$

where $\ell_{\mathrm{p}}(\boldsymbol{\beta})$ represents the penalized log-likelihood and $\mathcal{P}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is a quadratic penalty defined as $\mathcal{P}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_\lambda\boldsymbol{\beta}$. The matrix $\mathbf{S}_\lambda$ is block diagonal and contains the penalty matrices associated with the additive predictors related to each of the distribution parameters, that is, $\mathbf{S}_\lambda = \mathrm{diag}(\bar{\mathbf{S}}_{\mu_1}, \bar{\mathbf{S}}_{\mu_2}, \bar{\mathbf{S}}_{\mu_3}, \bar{\mathbf{S}}_{\sigma_2}, \bar{\mathbf{S}}_{\sigma_3}, \bar{\mathbf{S}}_{\nu_2}, \bar{\mathbf{S}}_{\nu_3}, 0, 0)$. The unknown penalty vectors contained in each $\bar{\mathbf{S}}$ matrix can also be collected into an overall vector as follows $\boldsymbol{\lambda} = (\bar{\boldsymbol{\lambda}}_{\mu_1}^{\mathsf{T}}, \bar{\boldsymbol{\lambda}}_{\mu_2}^{\mathsf{T}}, \bar{\boldsymbol{\lambda}}_{\mu_3}^{\mathsf{T}}, \bar{\boldsymbol{\lambda}}_{\sigma_2}^{\mathsf{T}}, \bar{\boldsymbol{\lambda}}_{\sigma_3}^{\mathsf{T}}, \bar{\boldsymbol{\lambda}}_{\nu_2}^{\mathsf{T}}, \bar{\boldsymbol{\lambda}}_{\nu_3}^{\mathsf{T}})^{\mathsf{T}}$.

Estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ proceeds by using the approach of Marra et al. (2017) and Marra & Radice (2019), already described in Chapter 2. To avoid repetition, we limit the exposition to an outline of the steps of the procedure and to defining the elements that are needed for estimation in the current context.

Step 1: At iteration a, and holding $\boldsymbol{\lambda}$ fixed, the trust-region method obtains and update of the parameter vector of the form $\boldsymbol{\beta}^{[a+1]} = \boldsymbol{\beta}^{[a]} + \mathbf{q}^{[a]}$, where $\mathbf{q}^{[a]}$ corresponds to the solution of the following constrained optimization problem

$$\mathbf{q}^{[a]} = \arg\min_{\mathbf{q} \in \mathcal{T}^{[a]}} - \left\{\ell_{\mathrm{p}}(\boldsymbol{\beta}^{[a]}) + \mathbf{q}^{\mathsf{T}}\mathbf{g}_p(\boldsymbol{\beta}^{[a]}) + \frac{1}{2}\mathbf{q}^{\mathsf{T}}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\beta}^{[a]})\mathbf{q}\right\}, \tag{3.12}$$

such that $\mathcal{T}^{[a]} = \{\mathbf{q} \in \mathbb{R}^p \colon \|\mathbf{q}\| \leq r^{[a]}\}$ represents the region of trust with radius $r^{[a]} > 0$ at iteration a. The objective function in (3.12) is a quadratic approximation of the model negative penalized log-likelihood within a suitable neighbourhood of $\boldsymbol{\beta}^{[a]}$, the vector $\mathbf{g}_p(\boldsymbol{\beta}^{[a]}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[a]}} - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\boldsymbol{\beta}^{[a]} = \mathbf{g}(\boldsymbol{\beta}^{[a]}) - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\boldsymbol{\beta}^{[a]}$ denotes the penalized gradient, and the matrix

$\boldsymbol{\mathcal{H}}_p(\boldsymbol{\beta}^{[a]}) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[a]}} - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}} = \boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a]}) - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}$ the penalized Hessian, both evaluated at the current guess-estimate of the parameter vector. The gradient and Hessian of the (unpenalized) model log-likelihood are defined as

$\mathbf{g}(\boldsymbol{\beta}) = \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1}}^\intercal, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2}}^\intercal, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3}}^\intercal, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_2}}^\intercal, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_3}}^\intercal, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_2}}^\intercal, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_3}}^\intercal, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \delta_{12}}, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \delta_{13}} \right)^\intercal$, and

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}) = \begin{bmatrix}
\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\mu_1}^\intercal} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\mu_2}^\intercal} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\mu_3}^\intercal} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\sigma_2}^\intercal} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\sigma_3}^\intercal} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\nu_2}^\intercal} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\nu_3}^\intercal} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \delta_{12}} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \delta_{13}} \\
\cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2} \partial \boldsymbol{\beta}_{\mu_2}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2} \partial \boldsymbol{\beta}_{\sigma_2}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2} \partial \boldsymbol{\beta}_{\nu_2}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2} \partial \delta_{12}} & \mathbf{0} \\
\cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3} \partial \boldsymbol{\beta}_{\mu_3}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3} \partial \boldsymbol{\beta}_{\sigma_3}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3} \partial \boldsymbol{\beta}_{\nu_3}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3} \partial \delta_{13}} \\
\cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_2} \partial \boldsymbol{\beta}_{\sigma_2}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_2} \partial \boldsymbol{\beta}_{\nu_2}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_2} \partial \delta_{12}} & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_3} \partial \boldsymbol{\beta}_{\sigma_3}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_3} \partial \boldsymbol{\beta}_{\nu_3}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_3} \partial \delta_{13}} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_2} \partial \boldsymbol{\beta}_{\nu_2}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_2} \partial \delta_{12}} & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_3} \partial \boldsymbol{\beta}_{\nu_3}^\intercal} & \mathbf{0} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_3} \partial \delta_{13}} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \delta_{12} \partial \delta_{12}} & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \delta_{13} \partial \delta_{13}}
\end{bmatrix},$$

respectively. Both $\mathbf{g}(\boldsymbol{\beta})$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta})$ have been derived analytically in Appendix B.2.

Step 2: In the second step, the parameter vector is held fixed and $\boldsymbol{\lambda}$ is estimated by minimising the following criterion

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \ \mathcal{V}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\lambda}} \ \|\mathbf{z}^{[a+1]} - \mathbf{A}_{\boldsymbol{\lambda}^{[a]}} \mathbf{z}^{[a+1]}\|^2 - \bar{K} + 2\text{tr}\left(\mathbf{A}_{\boldsymbol{\lambda}^{[a]}}\right),$$

where $\mathbf{z}^{[a+1]} = \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a+1]})\right\}^{1/2} \boldsymbol{\beta}^{[a+1]} + \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a+1]})\right\}^{-1/2} \mathbf{g}(\boldsymbol{\beta}^{[a+1]})$ corresponds to a pseudo-data vector, $\mathbf{A}_{\boldsymbol{\lambda}^{[a]}} = \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a+1]})\right\}^{1/2} \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a+1]}) + \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right\}^{-1} \left\{-\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}^{[a+1]})\right\}^{1/2}$ is the influence matrix, and $\bar{K}$ is the total number of parameters (see, Marra et al., 2017, for further details). Initial values are obtained by fitting two separate copula-based SS models, in the first instance for observations where $y_{1i} = 1$ and then for observations where $y_{1i} = 0$ using the routines provided by the `GJRM` package (Marra & Radice, 2021).

Inference proceeds from the results already described in Chapter 2, and references therein, which we do not include here to avoid repetition.

3.4 Simulation study

In this section, we perform a Monte Carlo experiment in order to (i) investigate the empirical properties of the copula-based approach and (ii) assess the effect of copula misspecification on

parameter estimates. Motivated by the empirical application presented in Section 3.5, we consider a scenario where the switching and each of the regime variables are generated using the normal and the log-normal distributions, respectively, and the copula functions that characterize their joint distributions are from the Joe and Gumbel families. We assess the results under the correct model specification and compare them with those obtained when one of the copula functions is misspecified.

The simulated data consist of a set of independent observations $(\mathbf{y}_i, \mathbf{x}_i)_{i=1}^n$, where $\mathbf{y}_i = (y_{1i}, y_{2i}, y_{3i})$ are realizations of the random variables (Y_{1i}, Y_{2i}, Y_{3i}), obtained using the observations rules given in (3.1), and $\mathbf{x}_i = (x_{1i}, x_{2i})$ consists of a binary and a continuous variable generated using the approach described in the simulation study of Chapter 2 (and references therein). The underlying latent variables are generated as follows: the model for the switching mechanism is given by $Y_{1i}^* \sim F_1(\boldsymbol{\theta}_{1i})$, where F_1 corresponds to the normal distribution with parameter vector $\boldsymbol{\theta}_{1i} = (\mu_{1i}, \sigma_{1i})^\intercal$, such that σ_1 is set to one to ensure identifiability and the additive predictor associated with μ_1 is given by

$$\eta_i^{\mu_1} = \beta_0^{\mu_1} + \beta_1^{\mu_1} x_{1i} + s_1^{\mu_1}(x_{2i}).$$

The value of $\beta_0^{\mu_1}$ is set to -1.15 in order to assign, approximately, 40% of observations to the first regime, the value of $\beta_1^{\mu_1}$ is set to 1.2, and the smooth function is given by $s_1^{\mu_1}(x_{2i}) = 1 - x_{2i}^3 - 2\exp(-180x_{2i}^2) - 2.3\sin(4.9x_{2i})$. The models for the regime variables are defined as $Y_{mi}^* \sim F_m(\boldsymbol{\theta}_{mi})$, where F_m corresponds to the log-normal distribution with parameter vector $\boldsymbol{\theta}_{mi} = (\mu_{mi}, \sigma_{mi})^\intercal$ and associated additive predictors

$$\begin{aligned}\eta_i^{\mu_m} &= \beta_0^{\mu_m} + \beta_1^{\mu_m} x_{1i} + s_1^{\mu_m}(x_{2i}),\\ \eta_i^{\sigma_m} &= \beta_0^{\sigma_m},\end{aligned}$$

for $m = 2, 3$, where the values of $\beta_0^{\mu_2}$, $\beta_1^{\mu_2}$, and $\beta_0^{\sigma_2}$ are set to $2.2, 1.3$ and 0.9, and the values of $\beta_0^{\mu_3}$, $\beta_1^{\mu_3}$, and $\beta_0^{\sigma_3}$ are set to 1.5, 2.1, and 1.1, respectively. The smooth components are defined as $s_1^{\mu_2}(x_{2i}) = x_{2i} + \exp[-32(x_{2i} - 0.5)^2]$ and $s_1^{\mu_3}(x_{2i}) = 0.3 + x_{2i} + \exp[-30(x_{2i} - 0.35)^2]$.

Lastly, the joint distributions of the pairs (Y_{1i}^*, Y_{mi}^*) are defined as follows

$$F_{1m}(y_{1i}^*, y_{mi}^* \mid \boldsymbol{\theta}_{1mi}) = \mathcal{C}_{1m}\left(F_1(y_{1i}^* \mid \boldsymbol{\theta}_{1i}), F_m(y_{mi}^* \mid \boldsymbol{\theta}_{mi}) \mid \delta_{1m}\right), \quad m = 2, 3,$$

where $\mathcal{C}_{12}$ is the Joe copula, $\mathcal{C}_{13}$ is the Gumbel copula, and the parameters δ_{12} and δ_{13} are set to values corresponding to Kendall's tau of $\tau_{12} = 0.2$ and $\tau_{13} = 0.4$, respectively. The simulation settings are similar to those that appear in Wojtyś et al. (2018) in the context of copula-based SS models.

We perform $N = 300$ repetitions[1] with sample sizes of $n = \{3000, 5000, 10000\}$ and estimate the parameters of the correctly specified model (as described above) and of two misspecified models in which $\mathcal{C}_{13}$ is assumed to belong to either the Joe or the Gaussian copula families.

Figure 3.1 shows boxplots of the estimates of $\beta_1^{\mu_m}$, $\beta_0^{\sigma_m}$, and τ_{1m}, for $m = 2, 3$, whereas Figure 3.2 shows the true smooth functions (dashed lines) and the average effect estimates (solid lines) of $s_1^{\mu_m}(x_2)$, together with the 5% and 95% point-wise quantiles (shaded areas), for $m = 2, 3$, and a sample size of $n = 10000$. In addition, Table 3.1 summarises the results in terms of relative bias and root mean squared error (RMSE). Overall we observe that, under the correct marginal and copula specification, estimates of the parametric and non-parametric components are near their true values and are less variable as the sample size increases. On the other hand, when the copula is misspecified, the parameter estimates that characterise the marginal distribution of Y_3^* appear to be slightly under- or overestimated. In particular, when the model for $\mathcal{C}_{13}$ is based on the Joe copula family β^{μ_3} and τ_{13} are slightly underestimated but estimates of $\beta_0^{\sigma_3}$ are practically unbiased. When $\mathcal{C}_{13}$ is based on the Gaussian copula, the true values of β^{μ_3} and τ_{13} are slightly overestimated whereas $\beta_0^{\sigma_3}$ is underestimated. In terms of the smooth terms, the misspecified model based on the Gaussian copula recovers the shape of true function better than the one based on the Joe copula.

3.5 Empirical application

We now revisit the empirical application from Chapter 2 and analyse the subset of the MEPS data using copula-based ESR models. Recall that the study aims to investigate the effects of the uptake of supplementary insurance on out-of-pocket prescribed drugs expenditures for individuals over 65 years old, while accounting for several other socio-economic and health-related factors. At the time of data collection, Medicare did not cover for prescribed drug expenditures and individuals may have chosen to obtain supplementary services to cover against certain out-of-pocket expenses.

[1]The number of repetitions is chosen based on the simulation study in Chapter 2 and references therein. Using the principled approach described there on the scenario for the correctly specified model with $n = 10000$; an initial run of $N_0 = 50$; and focusing on the bias obtained for the Kendall's tau parameters (τ_{12} and τ_{13}) yields $N \approx 135$ and $N \approx 282$ repetitions.

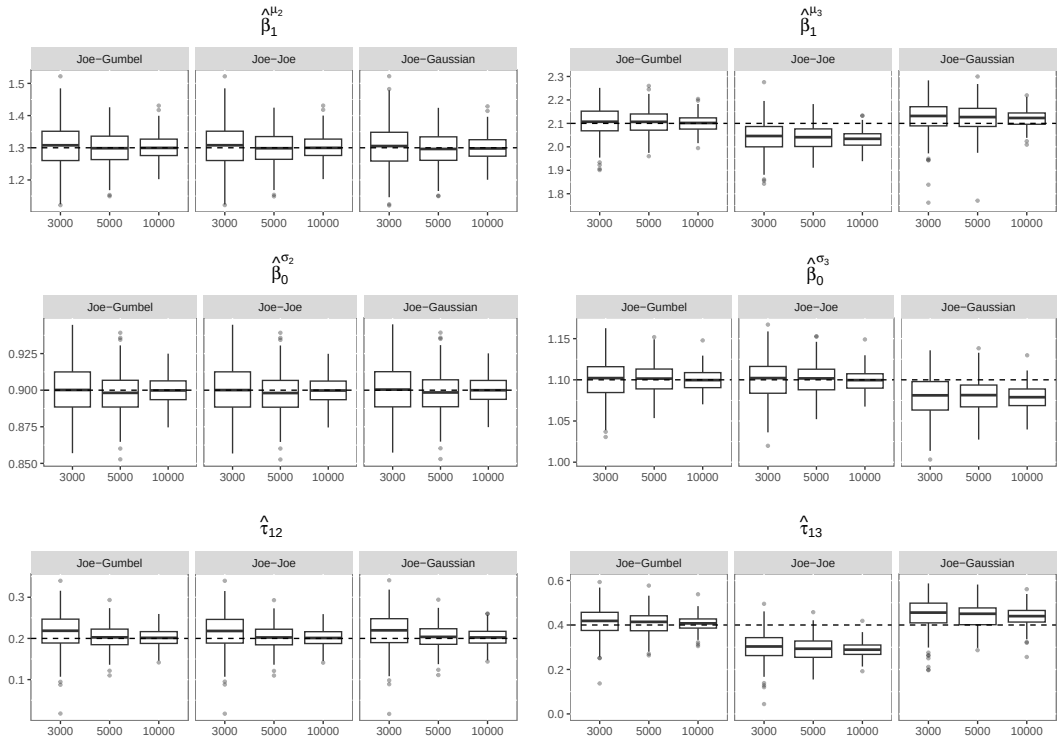

Figure 3.1: Boxplots of the estimates of $\beta_1^{\mu_2}, \beta_1^{\mu_3}, \beta_0^{\sigma_2}, \beta_0^{\sigma_3}$, and τ_{12}, τ_{13} for the correctly specified model (normal and log-normal marginals linked by the Joe and Gumbel copulas) and when one of the copulas is misspecified (normal and log-normal marginals linked by either the Joe and Joe copulas or the Joe and Gaussian copulas). The true values of the parameters are $\beta_1^{\mu_2} = 1.3, \beta_1^{\mu_3} = 2.1, \beta_0^{\sigma_2} = 0.9, \beta_0^{\sigma_3} = 1.1, \tau_{12} = 0.2$, and $\tau_{13} = 0.4$ (indicated by a dashed line in the plots). The sample size are $n \in \{3000, 5000, 10000\}$.

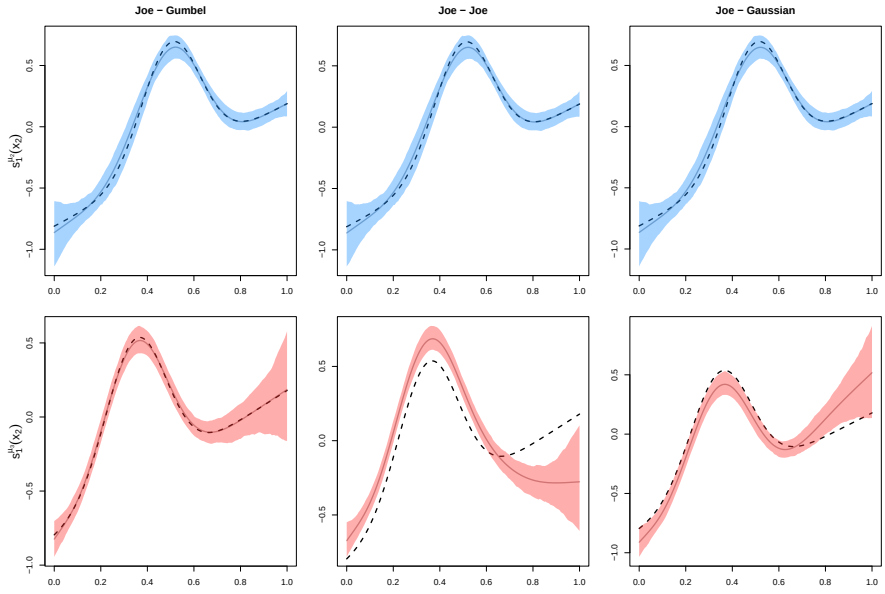

Figure 3.2: Mean estimates of $s_1^{\mu_2}(x_2)$ (top row, blue line) and $s_1^{\mu_3}(x_2)$ (bottom row, red line) for a sample size of $n = 10000$ under the correctly specified model (normal and log-normal marginals linked by the Joe and Gumbel copulas) and when one of the copulas is misspecified (normal and log-normal marginals linked by either the Joe and Joe copulas or the Joe and Gaussian copulas). The shaded areas correspond to the 5% and 95% point-wise quantiles. The true functions are represented by dashed lines.

	Relative bias			RMSE			Relative bias			RMSE		
$\mathcal{C}_{12} - \mathcal{C}_{13} \mid n$	3000	5000	10000	3000	5000	10000	3000	5000	10000	3000	5000	10000
	$\hat{\beta}_1^{\mu_2}$						$\hat{\beta}_1^{\mu_3}$					
Joe-Gumbel	0.0072	0.0003	0.0003	0.0768	0.0563	0.0388	0.0023	0.0009	-0.0009	0.0559	0.0492	0.0313
Joe-Joe	0.0074	0.0005	0.0006	0.0770	0.0563	0.0388	-0.0234	-0.0254	-0.0277	0.0742	0.0718	0.0658
Joe-Gaussian	0.0056	-0.0015	-0.0015	0.0766	0.0560	0.0388	0.0022	0.0056	0.0054	0.087	0.0654	0.0351
	$\hat{\beta}_0^{\sigma_2}$						$\hat{\beta}_0^{\sigma_3}$					
Joe-Gumbel	0.0012	-0.0016	0.0001	0.0187	0.0146	0.0102	-0.0006	0.0001	-0.0008	0.0212	0.0163	0.0122
Joe-Joe	0.0011	-0.0017	0	0.0187	0.0146	0.0102	0.0005	0.0008	-0.0004	0.0224	0.0173	0.0123
Joe-Gaussian	0.0014	-0.0014	0.0002	0.0187	0.0146	0.0102	-0.0196	-0.0180	-0.0189	0.0311	0.0264	0.0245
	$\hat{\tau}_{12}$						$\hat{\tau}_{13}$					
Joe-Gumbel	0.0972	0.0257	0.0121	0.0475	0.0311	0.0220	0.0308	0.0153	0.0094	0.0577	0.0501	0.0338
Joe-Joe	0.0950	0.0236	0.0102	0.0473	0.031	0.0219	-0.2481	-0.2673	-0.2778	0.1146	0.1181	0.1156
Joe-Gaussian	0.1019	0.0302	0.0162	0.0479	0.0312	0.0220	0.0240	0.0547	0.0674	0.1439	0.0992	0.0485
	$\hat{s}_1^{\mu_2}(x_2)$						$\hat{s}_1^{\mu_3}(x_2)$					
Joe-Gumbel	0.0498	0.0377	0.0276	0.1378	0.1043	0.0810	0.0131	0.0125	0.0091	0.1008	0.0818	0.0591
Joe-Joe	0.0496	0.0368	0.0266	0.1379	0.1040	0.0808	0.1572	0.1622	0.1619	0.1854	0.1807	0.1717
Joe-Gaussian	0.0497	0.0385	0.0289	0.1378	0.1047	0.0814	0.0710	0.0825	0.0928	0.1778	0.1446	0.1140

Table 3.1: Relative bias and RMSE for $\hat{\beta}_1^{\mu_2}, \hat{\beta}_1^{\mu_3}, \hat{\beta}_0^{\sigma_2}, \hat{\beta}_0^{\sigma_3}, \hat{\tau}_{12}, \hat{\tau}_{13}$, and smooth functions estimates $\hat{s}_1^{\mu_2}(x_2)$ and $\hat{s}_1^{\mu_3}(x_2)$, obtained under different copula specifications, with sample sizes $n \in \{3000, 5000, 10000\}$. The true copulas $\mathcal{C}_{12}$ and $\mathcal{C}_{13}$ are the Joe and the Gumbel, respectively. The misspecified models are those in which $\mathcal{C}_{13}$ corresponds to either the Joe or the Gaussian copulas. The true values of the parameters are $\beta_1^{\mu_2} = 1.3, \beta_1^{\mu_3} = 2.1, \beta_0^{\sigma_2} = 0.9, \beta_0^{\sigma_3} = 1.1, \tau_{12} = 0.2$ and $\tau_{13} = 0.4$. The true smooth functions are $s_1^{\mu_2}(x_2) = x_2 + \exp[-32(x_2 - 0.5)^2]$ and $s_1^{\mu_3}(x_2) = 0.3 + x_2 + \exp[-30(x_2 - 0.35)^2]$.

Having supplementary insurance raises then concerns of potential endogeneity, since it is plausible that there are some unobserved individual characteristics that affect simultaneously prescribed drugs expenditures and obtaining supplementary insurance. For further details, we refer the reader to Section 2.6 of Chapter 2.

As pointed out by Mullahy (2009), healthcare expenditures distributions are generally right-skewed and exhibit relatively 'heavy' upper tails. While the log transformation of the response performed in Chapter 2 enables to proceed the analysis under the classical distributional assumptions, it is plausible that the distributions of the regime variables can be better described by other parametric families. The copula-based approach allows to explore, besides the log-normal, several other distributions that have been proposed to model healthcare expenditures such as the gamma, Weibull, Dagum, and Sigh-Maddala (see, for example, Deb & Norton, 2018; Manning et al., 2005; Manning & Mullahy, 2001; Mullahy, 2009). Moreover, the bivariate normality assumption between the switching and the regimes variables is somewhat restrictive and may not reveal certain characteristics of the data. For instance, the association between the variables, measured by the correlation coefficients, is restricted to being linear and the Gaussian distribution does not capture tail dependence. The copula approach enables to inspect several modelling assumptions, such as the presence of tail dependence, in terms of the association structure implied by the copula.

We start by making an assessment of potential candidate distributions to model the switching mechanism and each of the regimes. Then, we explore several copula specifications to model the association between having supplementary insurance and prescribed-drug expenditures for each regime.

With respect to the switching mechanism, we investigate several models using the following specification

$$\texttt{supplementary}_i^* \sim F_1(\mu_{1i}, 1), \quad i = 1, \ldots, n,$$

where F_1 is either the normal, logistic, or Gumbel distribution, and the location parameter is related to the following additive predictor

$$\begin{aligned}\mu_{1i} = \eta_i^{\mu_1} = \beta_0^{\mu_1} + \beta_1^{\mu_1}\texttt{gender}_i + \beta_2^{\mu_1}\texttt{race}_i + \beta_3^{\mu_1}\texttt{multloc}_i + s_1^{\mu_1}(\texttt{age}_i) \\ + s_2^{\mu_1}(\texttt{log(income)}_i) + s_3^{\mu_1}(\texttt{chronic}_i) + s_4^{\mu_1}(\texttt{ssiratio}_i).\end{aligned}$$

The effects of the continuous covariates are modelled using smooth functions, $s_j^{\mu_1}(\cdot)$ for $j = 1, \ldots, 4$, and are represented using thin plate regression splines (see Section 2.2.1). An assessment of the results indicates that all the covariates included in the predictor are highly significant for the three different specifications. Table 3.2 shows the AIC/BIC values obtained under the different modelling options and suggest using the normal specification. A further sensitivity analysis was carried out once the rest of the ESR model components were chosen and suggested that the final results are robust to the three different specifications of the switching mechanism (not shown here).

F_1	AIC	BIC
Normal	**11279.1**	**11469.4**
Logistic	11281.6	11471.5
Gumbel	11280.9	11471.9

Table 3.2: AIC/BIC values obtained after fitting the switching mechanism using the normal, logistic and Gumbel distributions.

In terms of the models for the regime variables, we consider the aforementioned parametric families, that is, the log-normal, gamma, Weibull, Dagum, and Sigh-Maddala distributions (a table containing the expressions of the pdf, cdf, expectation and variance of these distributions can be found in Appendix B.3). We initially allow for all the covariates to enter the predictors that model the distribution parameters and proceed by using a backward selection approach, based on the AIC/BIC, to arrive at the final model specifications for each regime (see, for example,

Stasinopoulos et al., 2017; Voudouris et al., 2012, for model selection approaches in the contexts of GAMLSS). Univariate models were fitted using the `gamlss` function from the `GJRM` package (Marra & Radice, 2021). The goodness of fit of the candidate marginal distributions can be assessed using the normalised quantile residuals (Dunn & Smyth, 1996), defined as

$$\hat{r}_{mi} = \Phi^{-1}\left\{F_m(y_{mi} \mid \hat{\boldsymbol{\theta}}_{mi})\right\}, \qquad m = 2, 3,\ i = 1, \ldots, n,$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cdf of the standard normal distribution, and $\hat{\boldsymbol{\theta}}_{mi}$ is the vector of estimated distribution parameters. Letting $u_m = F_m(y \mid \boldsymbol{\theta}_m)$, it is well known that $u_m \sim \mathcal{U}(0, 1)$ and, assuming that the model is correctly specified, the quantile residuals $\hat{r}_{mi}$ are approximately standard normal. This implies that normal Q-Q plots of $\hat{r}_{mi}$ can be used as graphical tools to assist with modelling decisions and to detect lack of fit of the candidate distributions for each of the regimes.

Table 3.3 contains the AIC/BIC values corresponding to the final specifications of each of the considered modelling options for each regime, whereas Figures 3.3 and 3.4 show Q-Q plots of the normalised quantile residuals. Based on the values of the selection criteria and on an assessment of the Q-Q plots, the log-normal distribution appears to provide the best fit to the data for both regimes.

Marginal distribution	Candidate distribution	AIC	BIC
F_2	Log-normal	**57287.6**	**57390.8**
	Singh-Maddala	57388.5	57526.8
	Dagum	57424.9	57518.1
	Gamma	57575.3	57714.1
	Weibull	57661.8	57847.1
F_3	Log-normal	**90711.2**	**90829.2**
	Singh-Maddala	90865.1	90997.4
	Dagum	90914.3	91042.2
	Gamma	91159.8	91280.3
	Weibull	91308.9	91421.6

Table 3.3: AIC/BIC values corresponding to univariate GAMLSS based on the log-normal, Singh-Maddala, Dagum, gamma, and Weibull distributions for each regime. The lowest AIC/BIC values are in bold.

The chosen models for each of the regime variables are specified as follows

$$\texttt{expenditure}^*_{mi} \sim F_2(\mu_{mi}, \sigma_{mi}), \qquad m = 2, 3, \quad i = 1, \ldots, n,$$

where F_m corresponds to the log-normal distribution and the location and scale parameters are

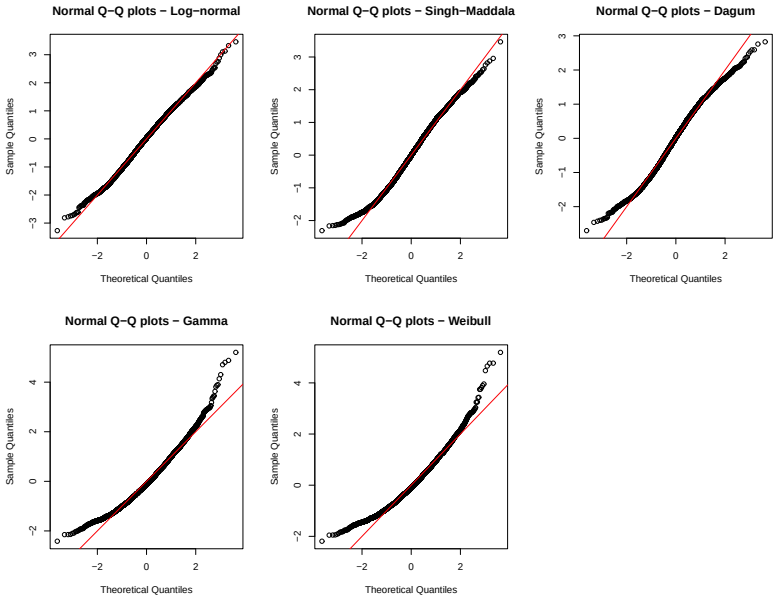

Figure 3.3: Q-Q plots of the normalised quantile residuals obtained from univariate GAMLSS for individuals with supplementary insurance based on the log-normal, Singh-Maddala, Dagum, Gamma, and Weibull distributions.

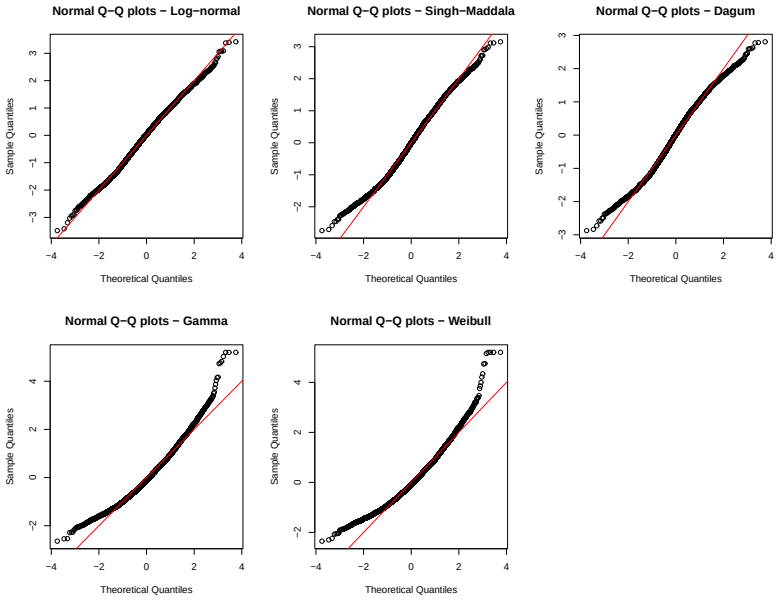

Figure 3.4: Q-Q plots of the normalised quantile residuals obtained from univariate GAMLSS for individuals without supplementary insurance based on the log-normal, Singh-Maddala, Dagum, Gamma, and Weibull distributions.

related to the following additive predictors

$$\begin{aligned}\eta_i^{\mu_m} &= \beta_0^{\mu_m} + \beta_1^{\mu_m}\texttt{gender}_i + \beta_2^{\mu_m}\texttt{race}_i + s_1^{\mu_m}(\texttt{age}_i) + s_2^{\mu_m}(\texttt{log(income)}_i) \\ &\quad + s_3^{\mu_m}(\texttt{chronic}_i), \\ \eta_i^{\sigma_m} &= \beta_0^{\sigma_m} + s_1^{\sigma_m}(\texttt{chronic}_i).\end{aligned}$$

The effects of the continuous covariates are modelled using smooth functions, $s_j^{\mu_m}(\cdot)$ for $j =$

$1, \ldots, 3$, and $s_1^{\sigma_m}(\cdot)$, which are represented using thin plate regression splines.

Given the univariate marginal models, the next stage consists of choosing candidate copula functions to specify the joint distributions $F_{1m}(\texttt{supplementary}_i^*, \texttt{expenditure}_{mi}^* \mid \boldsymbol{\theta}_{1mi})$, for $m = 2, 3$. We first perform an assessment on the strength and direction of the association parameters between the switching mechanism and each regime by fitting a model using Gaussian copulas. The results (roughly equivalent to those obtained in Chapter 2) indicate a positive association between insurance status and prescribed-drug expenditures at both levels of insurance. We then proceed by fitting several models using different combinations of copulas that allow for positive dependence, namely the Gaussian, Clayton, Gumbel, and Joe families. To choose among the candidate models, the copula-based regression literature favours the use of the AIC/BIC (Gomes et al., 2019; Zimmer, 2013) however, Brechmann & Schepsmeier (2013) also suggest the use of formal hypothesis tests for non-nested models such as those proposed by Vuong (1989) and Clarke (2007).

Table 3.4 reports the best three models in terms of AIC/BIC, the Gaussian-Gaussian specification (included for reference) and the estimated values of τ_{12} and τ_{13}. The model that obtains

$\mathcal{C}_{12}$ - $\mathcal{C}_{13}$	AIC	BIC	$\hat{\tau}_{12}$ (95% CI)	$\hat{\tau}_{13}$ (95% CI)
Joe - Gumbel	159216.3	159683.9	0.187 (0.144,0.228)	0.399 (0.321,0.475)
Joe - Gaussian	159221.9	159694.2	0.187 (0.146,0.233)	0.449 (0.373,0.524)
Joe - Joe	159222.2	159677.3	0.185 (0.145,0.231)	0.280 (0.216,0.354)
Gaussian -Gaussian	159236.0	159712.0	0.260 (0.132,0.368)	0.446 (0.362,0.521)

Table 3.4: AIC/BIC values and estimated values of τ_{12} and τ_{13} (with 95% credible intervals) obtained using different combinations of copula functions in the copula-based ESR model with marginals based on the normal and log-normal distributions. $\mathcal{C}_{12}$ and $\mathcal{C}_{13}$ denote the copula families used to model the joint distributions F_{12} and F_{13}, respectively.

the lowest AIC specifies the joint distributions F_{12} and F_{13} using the Joe and the Gumbel copulas, respectively. In contrast, the model with the lowest BIC specifies both joint distributions using the Joe copula. Performing a series of Vuong and Clarke tests to discern among the best three specifications suggest that all models provide similar fits to the data. However, as pointed out by Trivedi & Zimmer (2007), the main interest in many empirical applications lays on choosing an appropriate distribution to model the response and determining the statistical significance of association parameters, rather than on the particular form that the joint distributions F_{12} and F_{13} take. Furthermore, the dependence structures implied by the Gumbel and Joe copula families are relatively similar. They both exhibit upper tail dependence but the Gumbel has a thinner upper tail (see, for example, the contour density plots of the Gumbel and Joe copulas for different values of the Kendall's tau shown in Figure B.1 in Appendix B.1). This suggests that the interpretation of

the dependence structure implied by the models with the best AIC and BIC are relatively similar. In what follows we report the results using the Joe and Gumbel copula specifications.

Since the association parameters δ_{1m}, for $m = 2, 3$, are copula specific, we provide estimates (and 95% credible intervals) of the Kendall's tau. Both estimated values $\hat{\tau}_{12} = 0.187(0.144, 0.228)$ and $\hat{\tau}_{13} = 0.399(0.321, 0.475)$ capture individual self-selection into insurance. Similarly to the results obtained in Chapter 2, the estimated values of τ_{12} (or τ_{13}) suggest that on average, and among individuals with extra insurance (or Medicare-only individuals), unobserved characteristics that influence having supplementary coverage also influence out-of-pocket prescribed drug expenditures. The copula-based approach also allows to make an assessment of the dependence structure implied by the copula functions. The results suggest that large values of expenditures are accompanied by a higher chances of having supplementary insurance however, low expenditures do not necessarily imply lower chances of having extra insurance. This feature cannot be captured using the models presented in Chapter 2 under the standard distributional assumptions. Furthermore, these results seem to align with the literature (see, for example, Fang et al., 2008).

We now focus on the outcome of interest for each regime. As explained in Kneib et al. (2021), interpretation of the results in the context of GAMLSS models is generally difficult since we cannot directly assess the effects of the explanatory variables on the moments of the chosen distribution, the distributional parameters are subject to transformations, and the same covariate can enter different additive predictors.

Table 3.5 summarises the parametric estimates obtained by applying the 2-step approach (ESR-2-step), the classical fully parametric ESR (ESR - parametric ML), and the chosen copula-based model (ESR - Copula: Joe-Gumbel). The 2-step approach did not yield standard errors for the estimates. Estimates of the correlation coefficients for the 2-step and parametric ESR methods have been transformed to the Kendall's tau, to allow for comparison with the copula-based approach. Figure 3.5 summarises the estimates of the parametric and non-parametric model components (and 95% credible intervals) related to the distribution parameters of the univariate marginal models, for individuals with supplementary insurance (shown in blue) and for those using Medicare only (shown in red). The effects of `gender` and `race` on μ_2 and μ_3 are linear, as it is the effect of the number of `chronic` conditions on $\log \sigma_3$. Individual's `gender` is only a significant contributing factor for μ_2, whereas the effect of `race` is negative and significant for both μ_2 and μ_3. The effect of `age` shows a high degree of non-linearity and, overall, affect μ_2 and μ_3 negatively. The relationship of `log(income)` with μ_2 and μ_3 shows an overall downward

Model	ESR - 2-step		ESR - parametric ML		ESR - Copula: Joe-Gumbel	
Estimates | Regimes	Supplementary	Medicare	Supplementary	Medicare	Supplementary	Medicare
`(Intercept)`	6.369 -	7.539 -	6.354 (5.962,6.745)	7.468 (7.086,7.85)	6.597 (6.532,6.662)	7.209 (7.116,7.302)
`gender:female`	-0.04 -	-0.108 -	-0.018 (-0.084,0.049)	-0.103 (-0.16,-0.047)	-0.009 (-0.072,0.054)	-0.086 (-0.14,-0.033)
`race:Black/Hispanic`	-0.15 -	-0.184 -	-0.129 (-0.224,-0.034)	-0.18 (-0.25,-0.109)	-0.128 (-0.222,-0.034)	-0.162 (-0.23,-0.094)
`age`	-0.01 -	-0.014 -	-0.01 (-0.015,-0.004)	-0.013 (-0.018,-0.009)	(smooth term)	(smooth term)
`log(income)`	0.09 -	0.064 -	0.065 (0.024,0.106)	0.058 (0.024,0.093)	(smooth term)	(smooth term)
`chronic`	0.309 -	0.31 -	0.307 (0.282,0.331)	0.309 (0.289,0.329)	(smooth term)	(smooth term)
$\hat{\sigma}_2$	1.007 -	- -	0.989 (0.936,1.042)	- -	0.934 (0.887, 0.956)	- -
$\hat{\sigma}_3$	- -	0.979 -	- -	1.016 (0.965,1.066)	- -	1.023 (0.973,1.068)
$\hat{\tau}_{12}$	0.427 -	- -	0.303 -	- -	0.186 (0.144,0.228)	- -
$\hat{\tau}_{13}$	- -	0.455 -	- -	0.400 -	- -	0.399 (0.321,0.475)

Table 3.5: Parameter estimates and 95% confidence/credible intervals obtained using the 2-step approach (ESR - 2-step), the parametric ESR (ESR - parametric ML), and the copula-based model (ESR - Copula: Joe-Gumbel). The 2-step approach did not yield standard errors for the estimates. Smooth function estimates for the non-parametric terms of the copula-based model are reported in Figure 3.5.

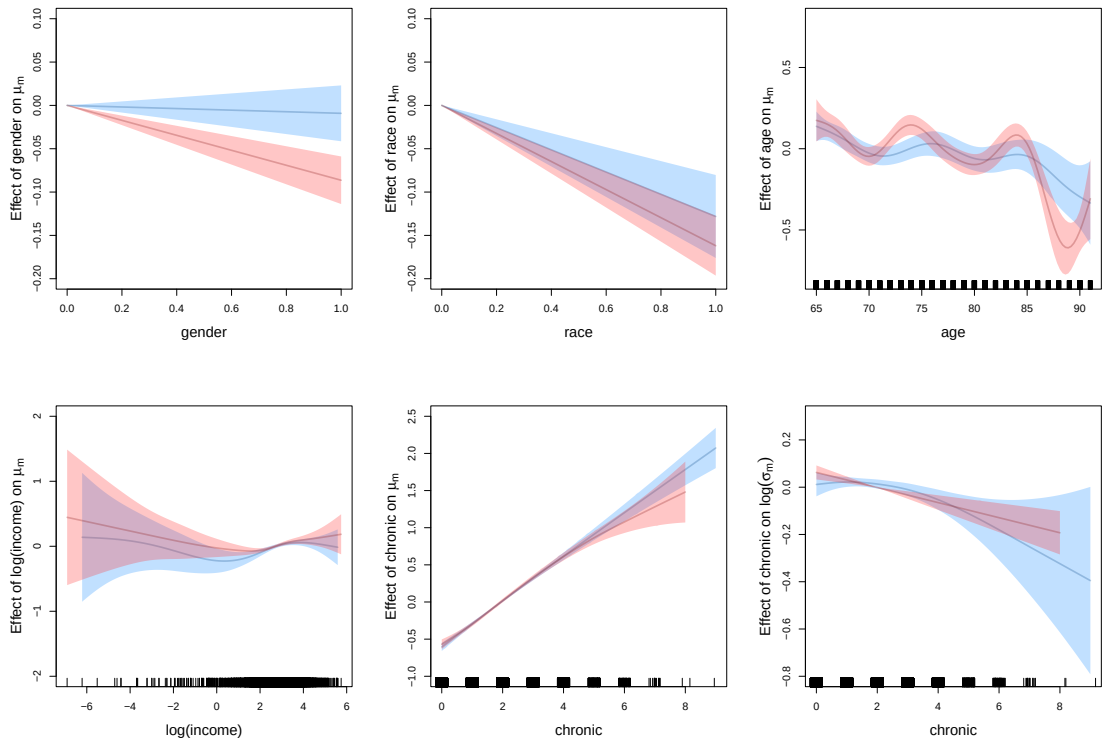

Figure 3.5: Estimated parametric and non-parametric effects, and 95% credible intervals, for μ_m and $\log(\sigma_m)$, for $m = 2, 3$. Estimates in the supplementary regime are shown in blue whereas those in the Medicare regime are shown in red. The jittered rug plot in the x-axis indicates the values of the covariates in the data.

trend for individuals with low annual income followed by an upward trend for individuals with higher incomes. Lastly, the effect of `chronic` shows that as the number of `chronic` conditions increases μ_2 and μ_3 increase but the effect is reversed for $\log \sigma_2$ and $\log \sigma_3$.

In order to provide further interpretation of the results, we follow the approach presented in Kneib et al. (2021) and Stadlmann & Kneib (2022) and show visualisations of the marginal influence of several covariates on the expected value of the response at each level of insurance. In particular, Figure 3.6 shows effect displays (Fox, 2003) that depict the influence of individual's `age`, `log(income)`, and number of `chronic` conditions on the predicted expected value of the response, for individuals with and without supplementary insurance, while keeping the rest of the covariates fixed at their mode/mean. The effects are shown in blue for individuals with supplementary insurance and in red for those with Medicare only. The shaded areas represent 95% credible intervals. Plots on the left-hand side correspond to 'average' Black/Hispanic females while those on the right-hand side correspond to 'average' White females. In terms of `age`, the expected `expenditure` appears to decrease as the 'average' individual gets older, for both levels of the insurance status and race. The effects displays of `log(income)` show that the expected values of `expenditure` for the 'average' individual follow a somewhat downward trend for low levels of income and a upward trend for higher incomes. Lastly, The effects of `chronic` on expected `expenditure` for the 'average' individual increases non-linearly with the number of `chronic` conditions, for both levels of insurance status and race.

The estimated average treatment effect of `supplementary` on `expenditure`, together with a 95% credible interval[2], for a randomly chosen individual is $-1285(-1586, -997)$. The result does not differ much from that obtained in Chapter 2 and suggests that, on average, having supplementary insurance leaves a randomly chosen individual about $\$1285$ better off in terms of prescribed drugs expenditures.

3.6 Discussion

In this chapter, we have presented copula-based ESR models that relax the distributional assumptions of the classical framework. The modelling approach allows for the joint distributions of the switching variable and each of the regime responses to be modelled using a range of parametric bivariate copula functions. Furthermore, the regime responses are embedded into the GAMLSS

[2]Credible intervals are constructed using the simulation steps described in Section 2.4 implemented in the `GJRM` package.

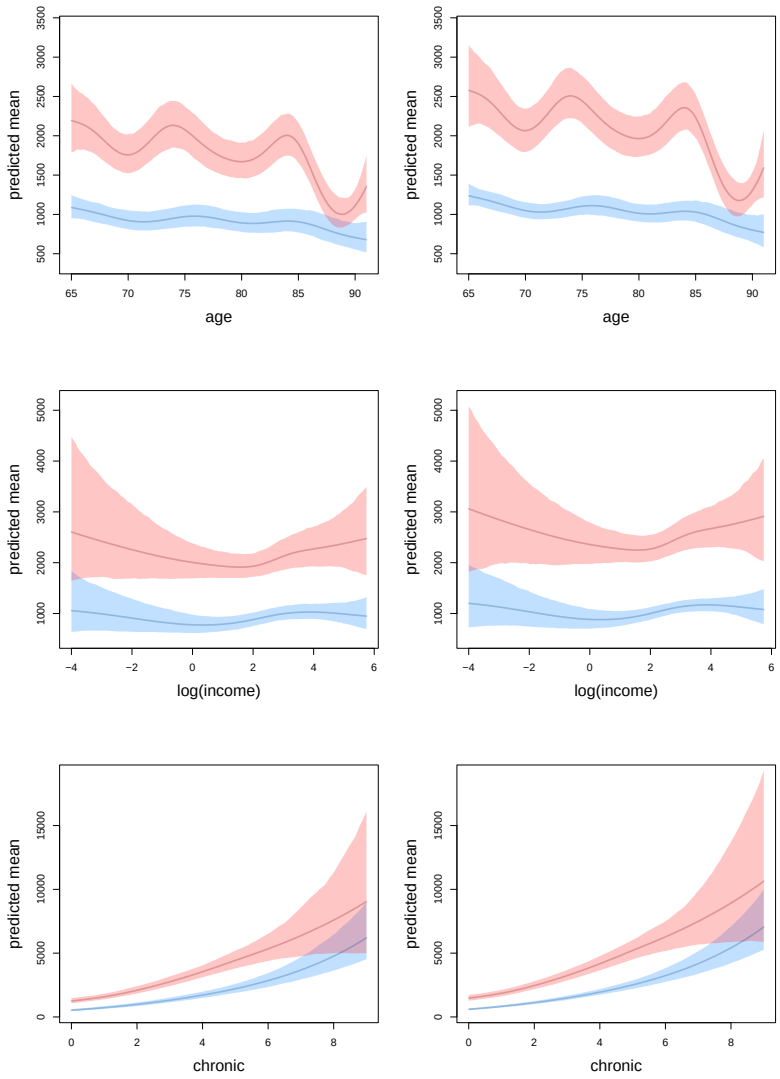

Figure 3.6: Effects of `age`, `log(income)`, and number of `chronic` conditions on the predicted expected value of the response for individuals with and without supplementary insurance. Effects on the supplementary regime are shown in blue, whereas those on the Medicare regime are shown in red. The shaded areas represent 95% credible intervals. Plots on the left-hand side correspond to 'average' Black/Hispanic females while those on the right-hand side correspond to 'average' White females.

framework, modelled using several two- and three-parameter continuous distributions, in which each distribution parameter is associated with a flexible linear predictor of covariates that accounts for linear, non-linear, and spatial effects. The approach to parameter estimation and inference is well-established in the copula regression modelling literature. We have also performed a simulation study to assess the empirical properties of the estimators and the effects of copula misspecification on the estimates.

In an application, we have studied insurance uptake of individuals over 65 to cover out-of-pocket prescription drug expenditures. Our findings suggest that, among several candidate distributions, the log-normal captures well the distribution of out-of-pocket prescribed drug expen-

ditures at both levels of insurance. Furthermore, the modelling approach captures self-selection at both levels of insurance and the dependence structures implied by the Joe and Gumbel families suggest that individuals with higher expected out-of-pocket expenditures are more likely to uptake supplementary insurance. The estimated average treatment effect indicates that having supplementary insurance constitutes a saving of about $\$1285$.

Chapter 4

A multiple imputation approach for missing not at random variables with an application in health economics

This chapter presents a multiple imputation (MI) approach that obtains plausible imputed values for a continuous variable assumed to be missing not a random and not restricted to be Gaussian. The approach is derived from a copula-based specification of the sample selection model and allows to create imputations for a partially observed variable under different assumptions about the distributions of the missingness mechanism and the variable subject to missing values. Similarly to other MI approaches in the literature, the imputation scheme can be embedded into the fully conditional specification strategy to MI.

In an application, we re-examine the non-randomised component of the REFLUX study[1] in which the response variable is missing for almost 50% of the participants and suspected to be missing not at random. The aim of the analysis is to evaluate the effect of surgery on long-term patient's health status, among individuals with gastro-oesophageal reflux disease, using different modelling strategies and assumptions about the missingness mechanism and the distribution of the response. We find that estimates of the effect of surgery are significant, regardless of the modelling approach. The MI estimates for the effect of surgery and other model parameters are very similar to those obtained using a copula-based sample selection model and, in some instances, they have slightly smaller standard errors.

[1]The REFLUX trial was funded by the NIHR Health Technology Assessment Programme (Project No 97/10/03) and was published in full in Health Technology Assessment. Volume 17, issue 22.

The contents and structure of this chapter are based on the following publication: *"Gomes, M, Radice, R, Camarena Brenes, J, Marra, G. - Copula selection models for non-Gaussian outcomes that are missing not at random. Statistics in Medicine. 2019; 38: 480-496"*.

I participated in the draft and publication of the paper. Specifically, my contributions to the paper were Section 4.4 *"MI based on the copula selection model"*, where the MI approach is presented; and appendices G, I, and J, which are related to the derivation of the copula-based likelihood model, the conditional density of the missing values, and details about the rejection algorithm used to obtain imputations. I did not contribute to the introduction, simulations, and the empirical application in the paper. The work was carried out under the supervision of Professor Rosalba Radice and Professor Giampiero Marra.

In relation to the contents of this chapter, Section 4.2 gives an overview of missing data terminology and is not included in the publication; Section 4.3 reviews sample selection models and is similar to Sections 4.1, 4.2, and 4.3 in the published paper; Section 4.4 expands on Section 4.4 from the publication. The simulation study in Section 4.5 has been carried out separately, and complements the results of the published paper. The empirical application in Section 4.6, which is the same as in the published paper, has also been carried out separately for this chapter, using the data provided by Dr Manuel Gomes, and with the permission of the Centre for Healthcare Randomised Trials unit from the University of Aberdeen.

4.1 Introduction

Missing data is a common problem encountered by researchers across different fields. For instance, situations such as non-response in surveys, failure to return self-reported questionnaires after interventions, or individual dropout in clinical studies are frequent (Grant et al., 2013; Little, 1982; Molenberghs & Kenward, 2007; Gomes et al., 2019). In the presence of partially observed variables, the challenge is to obtain valid inferences about the process that generated the data using only the observed data.

The formal framework to statistical modelling with missing data assumes that there is an underlying process that determines whether data are observed or missing and, generally, requires the specification of a joint model for the data and the missing data mechanism. Rubin (1976) defined and classified the missingness mechanism as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Without being exhaustive, the methodology

for handling missing data can be categorized into likelihood-based methods (either from a frequentist or a Bayesian point of view), weighting, and MI approaches (see, for example, Molenberghs et al., 2014). In this chapter, we present a MI approach based on the sample selection (SS) modelling framework that obtains plausible values for imputation of a partially observed continuous variable assumed to be MNAR.

As anticipated in Chapter 1, SS models (Heckman, 1974, 1976, 1979) are frequently used in situations where the researcher suspects that a partially observed outcome of interest is MNAR (see, for example Bärnighausen et al., 2011; Genius & Strazzera, 2008; Gomes et al., 2020, 2019; Sales et al., 2004). The classical specification of SS models consists of a simultaneous system of regression equations describing the selection or missingness mechanism and the substantive model of interest, where the error terms are assumed to be bivariate normal. Estimates of the model parameters are usually obtained using a two-step approach or via maximum likelihood. The classical framework has been extended in several directions, for example, based on semi- and non-parametric frameworks (Ahn & Powell, 1993; Gallant & Nychka, 1987), using alternative distributional assumptions such as the t or the skew-normal (Marchenko & Genton, 2012; Ogundimu & Hutton, 2016), or using copula functions to model the joint distribution of the selection and response variables (Smith, 2003; Wojtyś et al., 2018). In the context of missing data, a drawback of the SS framework is that the model only allows for missing values on the response and discards all the observational units with missing values in any of the explanatory variables. This may lead to biased and less precise parameter estimates (Carpenter & Smuk, 2021).

The MI framework was introduced by Rubin (1977, 1978, 1987) in the context of non-response in surveys and has become a widely used strategy to deal with missing data in several disciplines (King et al., 2001; Sterne et al., 2009; van Buuren et al., 1999). The MI approach can be understood as a way of imitating the data generating process by filling in the missing values with multiple plausible draws from appropriate imputation models in order to obtain several imputed data sets. The completed data sets are then analysed, employing the methods that would have been used in the absence of missing data, and the results are combined using what are commonly known as Rubin's rules. The main advantages of MI over other approaches are that the framework can be applied to impute several partially observed variables in the data, the analyses carried out in the completed data sets can generally be performed using standard software, and the combining Rubin's rules are generic and can be applied to estimates obtained from a wide range of analyses (see, for example, Harel & Zhou, 2007; Rubin, 1996; Zhang, 2003, for reviews on MI).

The two main strategies to MI when a data set contains several variables with missing values are referred to as the joint and the fully conditional specifications. The former models the data jointly using a multivariate distribution, usually the multivariate normal, and imputations are drawn from this distribution once the parameters have been estimated (Schafer, 1997). The latter specifies separate univariate conditional models for each variable subject to missingness given the observed (or already imputed) data, and then impute the missing values sequentially (van Buuren, 2007).

MI approaches are usually constructed under the MAR assumption however, the framework can also be employed when data is assumed to be MNAR (Gomes et al., 2019; Rubin, 1987; Schafer, 1999). For instance, Galimard et al. (2016) proposed an imputation approach based on Heckman's two-step estimation method and Ogundimu & Collins (2019) proposed imputing the missing values from an imputation model derived from a SS model based on the bivariate t distribution. In contrast, the imputation approach presented in this chapter is derived from a specification of the SS model using copula functions (Gomes et al., 2019; Smith, 2003; Wojtyś et al., 2018). The copula modelling framework allows to specify the models for the missingness mechanism and the MNAR variable using a wide range of distributions. This, in turn, permits to construct imputation schemes under different modelling assumptions about the missingness mechanism and the partially observed variable. The imputation model can be easily incorporated into a fully conditional strategy to MI to deal with missing values in several variables in the data.

In an application, we re-examine the non-randomised component of the REFLUX study (Grant et al., 2008, 2013; Gomes et al., 2019, 2020), a five year follow-up analysis of a clinical trial that evaluates the effect of using surgery, compared to continuing medication, on long-term patient's health status among individuals with gastro-oesophageal reflux disease in the UK. The response variable is constructed using participants' self-reported questionnaires and is missing for almost 50% of the individuals due to patients not returning the questionnaires at some point during the follow-up period. As pointed out by Gomes et al. (2019), the researchers in the study suspected that it was plausible for the outcome to be MNAR. The aim of the analysis is to compare the robustness of the conclusions of the study under different modelling strategies based on assumptions about the distributions of the missing mechanism and the response. We find that estimates of the effect of surgery are significant, regardless of the approach used, but their magnitude differs slightly depending on the modelling strategy. The MI estimates for the effect of surgery and other model parameters are very similar to those obtained using a copula-based SS model and, in some instances, they have slightly smaller standard errors.

The remainder of this chapter is structured as follows: Section 4.2 gives a general overview of the terminology used in the missing data literature, and the role of the mechanism governing the missing data in statistical modelling. Section 4.3 describes the classical SS model and the copula-based SS model proposed by Wojtyś et al. (2018). In Section 4.4, we present a MI approach for variables assumed to be MNAR based on the aforementioned copula SS model. Section 4.5 contains a simulation experiment to study the empirical properties of the imputation approach. Section 4.6 investigates the robustness of the conclusions from the non-randomised component of the REFLUX study under different modelling assumptions. Lastly, in Section 4.7 we discuss the MI approach and the empirical results.

4.2 An overview of missing data terminology

In this section, we provide an overview of the main terminology used in the missing data literature and some of the notation we will use in this chapter. For simplicity, we focus the exposition on likelihood-based approaches where a response variable is subject to missingness, given a fully observed vector of covariates.

Let Y_1 denote a missing data indicator such that $Y_1 = 1$ when the outcome variable of interest Y_2 is observed and $Y_1 = 0$ when missing, and let $\mathbf{x}$ correspond to a vector of fully observed covariates. In the presence of missing data, likelihood-based approaches to inference are based on the joint model for the missing data indicator and the response given the covariates with density $f(y_1, y_2 \mid \mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes a vector of parameters.

Statistical analysis with missing data is tied to a problem of identification of the parameters characterising $f(y_1, y_2 \mid \mathbf{x}, \boldsymbol{\theta})$, which can be understood by inspecting the following factorization of the joint density (Daniels & Hogan, 2008)

$$\begin{aligned} f(y_1, y_2 \mid \mathbf{x}, \boldsymbol{\theta}) = &\Big\{ f(y_2 \mid y_1 = 1, \mathbf{x}, \boldsymbol{\theta}_O) \mathbb{P}\left[Y_1 = 1 \mid \mathbf{x}\right] \Big\}^{y_1} \\ &\Big\{ f(y_2 \mid y_1 = 0, \mathbf{x}, \boldsymbol{\theta}_M) \mathbb{P}\left[Y_1 = 0 \mid \mathbf{x}\right] \Big\}^{1-y_1}, \end{aligned} \tag{4.1}$$

where $f(y_2 \mid y_1 = 1, \mathbf{x}, \boldsymbol{\theta}_O)$ and $f(y_2 \mid y_1 = 0, \mathbf{x}, \boldsymbol{\theta}_M)$ correspond to the conditional densities of the observed and missing values, characterized by the parameter vectors $\boldsymbol{\theta}_O$ and $\boldsymbol{\theta}_M$, respectively. Since we do not observed Y_2 when $Y_1 = 0$, the parameters in $f(y_2 \mid y_1 = 0, \mathbf{x}, \boldsymbol{\theta}_M)$ cannot be identified unless the researcher imposes parametric assumptions about either the distribution of the missing values given the observed, or the joint model and/or the process governing the missing

data, which are unverifiable from the data at hand (Molenberghs et al., 2014).

Given a sample of n incomplete observations $(y_{1i}, y_{2i}, \mathbf{x}_i)_{i=1}^n$, the observed data likelihood is obtained by integrating out the missing values from the joint density, that is,

$$\begin{aligned}\mathcal{L}_{obs}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \int_{\{y_2 :\, y_1=0\}} f(y_{1i}, y_{2i} \mid \mathbf{x}_i, \boldsymbol{\theta}) \mathrm{d}y_2 \\ &= \prod_{y_{1i}=1} f(y_{1i} \mid y_{2i}, \mathbf{x}_i, \boldsymbol{\theta}_1) f(y_{2i} \mid \mathbf{x}_i, \boldsymbol{\theta}_2) \\ &\quad \prod_{y_{1i}=0} \int_{\{y_2 :\, y_1=0\}} f(y_{1i} \mid y_{2i}, \mathbf{x}_i, \boldsymbol{\theta}_1) f(y_{2i} \mid \mathbf{x}_i, \boldsymbol{\theta}_2) \mathrm{d}y_2,\end{aligned} \tag{4.2}$$

where $f(y_{1i} \mid y_{2i}, \mathbf{x}_i, \boldsymbol{\theta}_1)$ corresponds to a model for the missingness mechanism, $f(y_{2i} \mid \mathbf{x}_i, \boldsymbol{\theta}_2)$ corresponds to a model for partially observed variable, and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are parameter vectors.

Inference about the parameters of interest depends on the assumptions made about the process characterising the relationship between the missing and observed data. Rubin (1976) categorised these assumptions as follows:

- Data are assumed to be MCAR if the probability of the response being observed or missing does not depend on the response itself or any other variable. Under this assumption, the model for the missingness mechanism can be written as follows $f(y_{1i} \mid y_{2i}, \mathbf{x}_i, \boldsymbol{\theta}_1) = f(y_{1i} \mid \boldsymbol{\theta}_1)$.

- Data are said to be MAR if the probability of the response being observed or missing does not depend on the missing values given the observed. That is, we write the model for the missingness mechanism as follows $f(y_{1i} \mid y_{2i}, \mathbf{x}_i, \boldsymbol{\theta}_1) = f(y_{1i} \mid \mathbf{x}_i, \boldsymbol{\theta}_1)$.

- Data are thought to be MNAR if the missingness mechanism depends on the unobserved components of Y_2, which implies that the probability of the variable of interest being observed or missing depends on the variable itself, even after accounting for other observed variables. We write the model for the MNAR mechanism as $f(y_{1i} \mid y_{2i}, \mathbf{x}_i, \boldsymbol{\theta}_1)$.

The MAR assumption implies that the observed data likelihood in (4.2) can be written as follows

$$\mathcal{L}_{obs}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_{1i} \mid \mathbf{x}_i, \boldsymbol{\theta}_1) \int_{\{y_2 :\, y_1=0\}} f(y_{2i} \mid \mathbf{x}_i, \boldsymbol{\theta}_2) \mathrm{d}y_2.$$

A consequence of assuming data is MAR is that we implicitly make an assumption about the conditional density of the missing values (given the observed data) that identifies the model parameters. MAR implies that Y_1 and Y_2 are conditionally independent and therefore the distribution

of the response is the same regardless of whether Y_2 is observed of not, i.e.,

$$f(y_{2i} \mid y_{1i} = 1, \mathbf{x}_i, \boldsymbol{\theta}_O) = f(y_{2i} \mid \mathbf{x}_i, \boldsymbol{\theta}) = f(y_{2i} \mid y_{1i} = 0, \mathbf{x}_i, \boldsymbol{\theta}_M).$$

Moreover, if in addition to MAR, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are disjoint or variation independent[2], the missingness mechanism is said to be ignorable and inference can proceed based on a model for the observed data, ignoring the process that causes the missing data (Rubin, 1976). In a regression context, the ignorability assumption implies that, under a correctly specified model, a complete-case analysis (using only the observational units with fully observed variables) yields consistent parameter estimates and valid inferences since the i^{th} contribution to the observed data likelihood for an individual with missing response is equal to 1 (see, for example, Carpenter & Kenward, 2012; Carpenter & Smuk, 2021).

When data are thought to be MNAR, the missingness mechanism is called non-ignorable and obtaining valid inferences about the parameters of interest requires specifying a joint model for the pair of variables (Y_1, Y_2), for instance, using the SS framework which we describe next.

4.3 Sample selection models

Sample selection models (SS; Heckman, 1974, 1976, 1979) are frequently used in situations where the researcher suspects that the outcome of interest is missing not at random (MNAR; see, for example Bärnighausen et al., 2011; Genius & Strazzera, 2008; Gomes et al., 2020, 2019; Sales et al., 2004). For instance, the response variable in the REFLUX study (Section 4.6) measures long term patient's health status and is missing for almost 50% of the individuals. The variable was constructed using self-reported questionnaires and researchers in the study suspected that patients in worse health were less likely to participate. This suggests that the missingness mechanism may depend on the variable that is missing and raises concerns about the plausibility of the variable being MNAR (see, Grant et al., 2008, 2013; Gomes et al., 2019, for further details).

The classical specification of SS models consists of a simultaneous system of regression equations describing the selection or missingness mechanism and the substantive model of interest, where the error terms are assumed to be bivariate normal. SS models are sensitive to distributional misspecification (Pigini, 2015) and the classical model has been extended in several directions,

[2]The parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are said to be disjoint if the parameter space of $\boldsymbol{\theta}$ is the product of the parameter spaces of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. This is also called the separability condition. From a Bayesian perspective, the condition also requires that the parameters are a priori independent. (Rubin, 1976).

for example, based on semi- and non-parametric approaches (Ahn & Powell, 1993; Gallant & Nychka, 1987), using alternative distributional assumptions (Marchenko & Genton, 2012; Ogundimu & Hutton, 2016), or using copula functions (Smith, 2003; Wojtyś et al., 2018). The copula approach provides a convenient framework to specify the distributions of the missingness/selection mechanism and the outcome of interest separately from the copula function that captures the dependence structure in the data. The framework also allows researchers to work under several distributional assumptions and assess the robustness of their results. Furthermore, fully parametric copula approaches are usually less computationally intensive than non- and semi-parametric methods (Wojtyś et al., 2018).

We describe next the specification and approaches to parameter estimation of the classical SS model and the copula-based extension proposed by Wojtyś et al. (2018). The latter constitutes the main building block for the MI approach presented in this chapter. Since the assumptions and identifiability restrictions of these models are similar to those already discussed in Chapters 2 and 3, in order to avoid repetition, we restrict our exposition to model specification and to the main concepts needed for this chapter.

Let (Y_{1i}, Y_{2i}) denote a pair of random variables generated using the following rules

$$Y_{1i} = \mathbb{1}_{Y_{1i}^*>0}(Y_{1i}^*), \qquad Y_{2i} = Y_{1i}Y_{2i}^*, \qquad i = 1, \ldots, n. \tag{4.3}$$

The Bernoulli random variable Y_{1i} represents the missing data indicator and is determined by the sign of the continuous latent variable Y_{1i}^* through the indicator function $\mathbb{1}_{Y_{1i}^*>0}(\cdot)$. The variable Y_{2i} denotes the partially observed variable of interest and is determined by Y_{1i} and its latent counterpart Y_{2i}^*, that is, when $Y_{1i} = 1$ we observe $Y_{2i} = Y_{2i}^*$ otherwise, Y_{2i} is assigned a dummy value of zero to represent that the variable is missing. Consider now that Y_{1i}^* and Y_{2i}^* can be described by parametric families of distributions, conditional on covariates, and denote their pdfs and cdfs as $f_m(y_{mi}^* \mid \boldsymbol{\theta}_{mi})$ and $F_m(y_{mi}^* \mid \boldsymbol{\theta}_{mi})$, respectively, where $\boldsymbol{\theta}_{mi} \in \mathbb{R}^{\tilde{n}_m}$ represents a vector of $\tilde{n}_m$ distribution parameters, for $m = 1, 2$. Let us also denote the joint cdf as $F(y_{1i}^*, y_{2i}^* \mid \boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i \in \mathbb{R}^{\tilde{n}}$ represents a vector of $\tilde{n}$ of distribution parameters.

The classical specification of the SS model follows from assuming that the joint distribution of (Y_{1i}^*, Y_{2i}^*) is bivariate normal, leading to univariate Gaussian specifications for the missingness mechanism and the substantive model. The model is generally specified using the following

system of equations

$$Y_{1i}^* = \eta_{1i} + \epsilon_{1i} = \mathbf{x}_{1i}^\mathsf{T}\boldsymbol{\beta}_1 + \epsilon_{1i},$$
$$Y_{2i}^* = \eta_{2i} + \epsilon_{2i} = \mathbf{x}_{2i}^\mathsf{T}\boldsymbol{\beta}_2 + \epsilon_{2i},$$

where η_{mi} denotes a linear predictor of covariates and unknown regression coefficients and ϵ_{mi} is a normally distributed error term with zero mean and variance σ_m, for $m = 1, 2$. The correlation coefficient between Y_{1i}^* and Y_{2i}^* is denoted by ρ_{12}. Estimates of the model parameters can be obtained using the Heckman's two-step approach or maximum likelihood. The two-step approach proceeds by fitting a probit model to the first equation in order to obtain estimates of $\boldsymbol{\beta}_1$ and compute the estimated linear predictor $\hat{\eta}_{1i}$. In the second step, the second equation is augmented with a correction term and parameters are estimated via least squares (further details can be found in Section 2.3 in the context of the ESR model). On the other hand, the maximum likelihood approach maximises the following log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} y_{1i} \left\{ \log \sigma_2^{-1} + \log \left[\phi \left(\frac{y_{2i} - \eta_{2i}}{\sigma_2} \right) \right] + \log \left[\Phi \left(\frac{\eta_{1i} + \rho_{12} \left(y_{2i} - \eta_{2i} \right) / \sigma_2}{\sqrt{1 - \rho_{12}^2}} \right) \right] \right\}$$
$$+ \sum_{i=1}^{n} (1 - y_{1i}) \left\{ \log \left[1 - \Phi \left(\eta_{1i} \right) \right] \right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T}, \sigma_2, \rho_{12})^\mathsf{T}$. Further details and reviews of the classical SS modelling approach can be found, for example, in Pigini (2015), Puhani (2000), and Vella (1998).

The copula-based SS framework specifies the joint cdf of the pair of latent variables (Y_{1i}^*, Y_{2i}^*) using a bivariate copula function as follows

$$F(y_{1i}^*, y_{2i}^* \mid \boldsymbol{\theta}_i) = \mathcal{C} \left(F_1(y_{1i}^* \mid \boldsymbol{\theta}_{1i}), F_2(y_{2i}^* \mid \boldsymbol{\theta}_{2i}) \mid \delta_{12} \right), \tag{4.4}$$

where $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{1i}^\mathsf{T}, \boldsymbol{\theta}_{2i}^\mathsf{T}, \delta_{12})^\mathsf{T}$ is a vector of parameters, $\mathcal{C}(\cdot, \cdot)$ is a parametric bivariate copula function, and δ_{12} corresponds to an association parameter that captures the strength of dependence between its margins and induces the outcome of interest to be MNAR. Recall that an introduction to the copula-based modelling framework is provided in Section B.1 of Appendix B. Wojtyś et al. (2018) specify the marginal distributions using the GAMLSS framework, which allows for the distribution parameters to be associated with flexible additive predictors in order to account for several types of covariate effects. For instance, the missingness mechanism is specified as

$Y_{1i}^* \sim F_1(\boldsymbol{\theta}_{1i})$ such that $\boldsymbol{\theta}_{1i} = (\mu_{1i}, \sigma_{1i})^\mathsf{T}$, where F_1 corresponds to either the normal, logistic, or Gumbel distributions. The location parameter is associated with an additive predictor as follows

$$\mu_{1i} = \eta_i^{\mu_1}\left(\mathbf{x}_i^{\mu_1}, \boldsymbol{\beta}_{\mu_1}\right) = \beta_0^{\mu_1} + \sum_{p_{\mu_1}=1}^{P_{\mu_1}} h_{p_{\mu_1}}^{\mu_1}\left(\tilde{\mathbf{x}}_{p_{\mu_1} i}^{\mu_1}, \boldsymbol{\beta}_{p_{\mu_1}}^{\mu_1}\right), \tag{4.5}$$

where $\mathbf{x}_i^{\mu_1}$ is the vector of covariates chosen to model μ_1, $\boldsymbol{\beta}_{\mu_1} = \left(\beta_0^{\mu_1}, \boldsymbol{\beta}_1^{\mu_1\mathsf{T}}, \ldots, \boldsymbol{\beta}_{P_{\mu_1}}^{\mu_1\mathsf{T}}\right)^\mathsf{T}$ is a vector of regression coefficients, and each $h_{p_{\mu_1}}^{\mu_1}\left(\tilde{\mathbf{x}}_{p_{\mu_1} i}^{\mu_1}, \boldsymbol{\beta}_{p_{\mu_1}}^{\mu_1}\right)$ denotes a function of a sub-vector of explanatory variables contained in $\mathbf{x}_i^{\mu_1}$ and regression coefficients whose particular structure depends on the type of covariate effect, for $p_{\mu_1} = 1, \ldots, P_{\mu_1}$. The scale parameter σ_{1i} is set to 1 for the usual identification purposes. Recall that each of the functions in the additive predictor is associated with a quadratic penalty term that imposes particular properties of the effect, say $\lambda_{p_{\mu_1}} \boldsymbol{\beta}_{\mu_1}^\mathsf{T} \mathbf{S}_{\lambda_{p_{\mu_1}}}^{\mu_1} \boldsymbol{\beta}_{\mu_1}$, where $\lambda_{p_{\mu_1}} > 0$ is the smoothing parameter and $\mathbf{S}_{\lambda_{p_{\mu_1}}}^{\mu_1}$ is a penalty matrix. For further details on the structure of the predictor in (4.5) we refer the reader to Wojtyś et al. (2018). On the other hand, the substantive model is specified as $Y_{2i}^* \sim F_2(\boldsymbol{\theta}_{2i})$, such that $\boldsymbol{\theta}_{2i} = (\mu_{2i}, \sigma_{2i}, \nu_{2i})^\mathsf{T}$, where F_2 can be chosen from a range of two- and three-parameter families of distributions (see, for example, Table 2 in Marra & Radice, 2017, for the definition of these distributions). Each distribution parameter is associated with a flexible additive predictor as follows

$$\begin{aligned}
g_{\mu_2}(\mu_{2i}) &= \eta_i^{\mu_2}\left(\mathbf{x}_i^{\mu_2}, \boldsymbol{\beta}_{\mu_2}\right) = \beta_0^{\mu_2} + \sum_{p_{\mu_2}=1}^{P_{\mu_2}} h_{p_{\mu_2}}^{\mu_2}\left(\tilde{\mathbf{x}}_{p_{\mu_2} i}^{\mu_2}, \boldsymbol{\beta}_{p_{\mu_2}}^{\mu_2}\right), \\
g_{\sigma_2}(\sigma_{2i}) &= \eta_i^{\sigma_2}\left(\mathbf{x}_i^{\sigma_2}, \boldsymbol{\beta}_{\sigma_2}\right) = \beta_0^{\sigma_2} + \sum_{p_{\sigma_2}=1}^{P_{\sigma_2}} h_{p_{\sigma_2}}^{\sigma_2}\left(\tilde{\mathbf{x}}_{p_{\sigma_2} i}^{\sigma_2}, \boldsymbol{\beta}_{p_{\sigma_2}}^{\sigma_2}\right), \\
g_{\nu_2}(\nu_{2i}) &= \eta_i^{\nu_2}\left(\mathbf{x}_i^{\nu_2}, \boldsymbol{\beta}_{\nu_2}\right) = \beta_0^{\nu_2} + \sum_{p_{\nu_2}=1}^{P_{\nu_2}} h_{p_{\nu_2}}^{\nu_2}\left(\tilde{\mathbf{x}}_{p_{\nu_2} i}^{\nu_2}, \boldsymbol{\beta}_{p_{\nu_2}}^{\nu_2}\right),
\end{aligned}$$

where $g_{\mu_2}(\cdot)$, $g_{\sigma_2}(\cdot)$, and $g_{\nu_2}(\cdot)$ are known, monotonic, and differentiable link functions that maintain the restrictions on the range of the distribution parameters, and the rest of the components are defined equivalently to those in Equation (4.5).

Given a random sample of observations the log-likelihood function of the copula-based SS

model can be written as follows

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} y_{1i} \log \left\{ f_2\left(y_{2i} \mid \boldsymbol{\theta}_{2i}\right) \left(1 - \frac{\partial \mathcal{C}\left(F_1(0 \mid \boldsymbol{\theta}_{1i}), F_2\left(y_{2i} \mid \boldsymbol{\theta}_{2i}\right) \mid \delta_{12}\right)}{\partial F_2\left(y_{2i} \mid \boldsymbol{\theta}_{2i}\right)}\right)\right\} + \sum_{i=1}^{n} (1 - y_{1i}) \log\left\{F_1(0 \mid \boldsymbol{\theta}_{1i})\right\},$$

where the vectors of distribution parameters are defined as $\boldsymbol{\theta}_{1i} = (\mu_{1i}, \sigma_{1i})^{\mathsf{T}} = \left(g_{\mu_1}^{-1}\left(\eta_i^{\mu_1}(\boldsymbol{\beta}_{\mu_1})\right), 1\right)^{\mathsf{T}}$ and $\boldsymbol{\theta}_{2i} = (\mu_{2i}, \sigma_{2i}, \nu_{2i})^{\mathsf{T}} = \left(g_{\mu_2}^{-1}\left(\eta_i^{\mu_2}(\boldsymbol{\beta}_{\mu_2})\right), g_{\sigma_2}^{-1}\left(\eta_i^{\sigma_2}(\boldsymbol{\beta}_{\sigma_2})\right), g_{\nu_2}^{-1}\left(\eta_i^{\nu_2}(\boldsymbol{\beta}_{\nu_2})\right)\right)^{\mathsf{T}}$, and each $g^{-1}(\cdot)$ is the inverse of the link function mapping the distribution parameter to its associated additive predictor. The overall parameter vector contains the regression coefficients together with the copula association parameter, i.e., $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\mu_1}^{\mathsf{T}}, \boldsymbol{\beta}_{\mu_2}^{\mathsf{T}}, \boldsymbol{\beta}_{\sigma_2}^{\mathsf{T}}, \boldsymbol{\beta}_{\nu_2}^{\mathsf{T}}, \delta_{12})^{\mathsf{T}}$. Estimates of the parameters are obtained by maximising the penalized log-likelihood

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_{\mathrm{p}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \left\{ \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\mathbf{S}_{\lambda}\boldsymbol{\theta} \right\},$$

where $\frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\mathbf{S}_{\lambda}\boldsymbol{\theta}$ is an overall penalty term and $\mathbf{S}_{\lambda}$ is a block-diagonal matrix that contains the penalty terms associated with the model additive predictors. Inference about the model coefficients is made from a Bayesian perspective by considering that the penalty term corresponds to the assumption of setting a prior density on $\boldsymbol{\theta}$. The penalized maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ can then be interpreted as the posterior mode and, assuming the smoothing parameters are fixed, the inferential results are based on the large sample normal approximation to the posterior (Wood et al., 2016)

$$\boldsymbol{\theta} \mid \mathbf{y} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}}\right), \tag{4.6}$$

where $\mathbf{V}_{\boldsymbol{\theta}} = \left(-\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}}) + \mathbf{S}_{\lambda}\right)^{-1}$ and $\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})$ denotes the Hessian matrix evaluated at the estimated parameter vector. For further details of the framework we refer the reader to Wojtyś et al. (2018) and Gomes et al. (2019).

4.4 Multiple imputation under MNAR

We present next a MI approach that obtains plausible values for a partially observed variable suspected to be MNAR based on the copula SS model specification described in the previous section. We start by giving a general overview of the MI process and describing the combination

rules that produce overall MI estimates, and then we derive the imputation model adapted to our context. The approach is implemented in the function `imputeSS` of the `GJRM` package.

4.4.1 The MI approach to missing data

The MI approach to missing data was introduced by Rubin (1977, 1978, 1987) in the context of non-response in surveys. In order to reduce item non-response, and to preserve the privacy of the respondents while avoiding distortions in the data, MI is generally used before releasing surveys for public access (Kennickell, 2017). For instance, the Medical Expenditure Panel Survey, the Survey of Consumer Finances, and the Consumer Expenditure Survey in the US, all contain imputed values in several of their reported variables. MI can also be used by researchers when the data to be analysed contain several partially observed variables. An introduction to MI in the context of economics can be found in Cameron & Trivedi (2005).

The MI framework consists of three steps: (i) generate $M > 1$ completed data sets by filling in the missing values with draws from an appropriate imputation model; (ii) perform the required analysis on each of the M completed data sets to obtain estimates of the model parameters; (iii) combine the results from the previous step to produce overall estimates that take into account the imputation process. Recent studies recommend using a number of imputations that is at least equal to the percentage of missing values (White et al., 2011; van Buuren, 2012; Molenberghs et al., 2014).

The main advantages of MI over other approaches that deal with missing data are that the framework can be applied to impute several partially observed variables in a data set under different assumptions about the missing data; the analyses carried out in the completed data sets can be performed using standard software; and the combining Rubin's rules are generic and can be applied to estimates obtained from a wide range of analyses (see, for example, Harel & Zhou, 2007; Rubin, 1996; Zhang, 2003, for reviews on MI).

Until recently, most MI approaches in the literature assumed missing data to be missing at random (MAR). When the researcher suspects that a partially observed outcome is missing not at random (MNAR), sample selection models (SS; Heckman, 1974, 1976, 1979) are used however, this approach only allows for missing values in the response and discards all the observational units with missing values in any of the explanatory variables. Extending the MI framework to impute variables suspected to be MNAR is an area of current research. For instance, Galimard et al. (2016) proposed an imputation approach based on Heckman's two-step estimation method

and Ogundimu & Collins (2019) proposed an imputation scheme derived from a SS model based on the bivariate t distribution.

The attractiveness of the MI approach presented in this chapter is in terms of flexibility. Although fully parametric, the method allows to construct imputation schemes under different modelling assumptions about the missingness mechanism, the distribution of the partially observed variable, and the dependence structure. The approach also allows the researcher to contrast the robustness of their results under different distributional assumptions, since it is not limited to obtain plausible draws for imputation from the Gaussian or t distributions as in the aforementioned developments. Furthermore, the approach can be embedded into a fully conditional strategy to MI to deal with several partially observed variables in a data set.

Rubin (1977, 1978, 1987) derived the MI approach from a Bayesian point of view in order to incorporate uncertainty about the missing data in the imputation process and to draw inferences that reflect the additional variability due to missing data. From a Bayesian perspective, both the parameter of interest and the missing values are thought of having a distribution given the observed data. Letting $Y_{2,obs}$ and $Y_{2,mis}$ denote the observed and missing components of Y_2 and dropping the vector of covariates for simplicity, the posterior distribution of $\boldsymbol{\theta}$ given the observed data is given by

$$\begin{aligned} f(\boldsymbol{\theta} \mid y_1, y_{2,obs}) &= \int f(\boldsymbol{\theta}, y_{2,mis} \mid y_1, y_{2,obs}) \mathrm{d}y_{2,mis} \\ &= \int f(\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}) f(y_{2,mis} \mid y_1, y_{2,obs}) \mathrm{d}y_{2,mis} \\ &= \mathsf{E}\left[f(\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}) \mid y_1, y_{2,obs} \right], \end{aligned} \tag{4.7}$$

where the first element of the integral in the second line of (4.7) corresponds to the posterior of $\boldsymbol{\theta}$ given the 'complete' data, in the sense that it would correspond to the posterior of interest had all the missing values been observed, and the second component represents the posterior predictive distribution of the missing data given the observed. Furthermore, assuming the order of integration can be exchanged, the posterior mean and variance of $\boldsymbol{\theta}$ can be written as follows (Carpenter & Kenward, 2012; Little & Rubin, 2019)

$$\mathsf{E}\left[\boldsymbol{\theta} \mid y_1, y_{2,obs} \right] = \mathsf{E}\left\{ \mathsf{E}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis} \right] \mid y_1, y_{2,obs} \right\}, \tag{4.8}$$

and

$$\begin{aligned}\mathsf{Var}\left[\boldsymbol{\theta} \mid y_1, y_{2,obs}\right] = \mathsf{E}\left\{\mathsf{Var}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}\right] \mid y_1, y_{2,obs}\right\} \\ + \mathsf{Var}\left\{\mathsf{E}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}\right] \mid y_1, y_{2,obs}\right\}.\end{aligned} \tag{4.9}$$

The main idea behind MI is that, provided a large number of draws, say $y_{2,mis}^{(1)}, \ldots y_{2,mis}^{(M)}$, can be obtained from $f(y_{2,mis} \mid y_1, y_{2,obs})$, the observed data posterior of $\boldsymbol{\theta}$ can be approximated by averaging the 'completed' data posterior $f(\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}^{(m)})$ over the repeated draws from $f(y_{2,mis} \mid y_1, y_{2,obs})$. Rubin's rules to obtain the MI estimator for $\boldsymbol{\theta}$, and the corresponding estimator for the covariance matrix, arise as approximations to the posterior mean and variance in (4.8) and (4.9). Approximating the inner expectation and variance requires assuming that, with complete data, the typical normal approximation to the posterior of $\boldsymbol{\theta}$ holds, i.e, $\boldsymbol{\theta} \mid y_2 \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimator and $\mathbf{V}_{\boldsymbol{\theta}}$ the covariance matrix evaluated at $\hat{\boldsymbol{\theta}}$. In our context, this implies that the posterior mean and variance of $\boldsymbol{\theta}$ given the completed data can be approximated by the mean and variance of the normal approximation, that is,

$$\mathsf{E}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}^{(m)}\right] \approx \hat{\boldsymbol{\theta}}^{(m)} \quad \text{and} \quad \mathsf{Var}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}^{(m)}\right] \approx \mathbf{V}_{\boldsymbol{\theta}}^{(m)},$$

where $\hat{\boldsymbol{\theta}}^{(m)}$ and $\mathbf{V}_{\boldsymbol{\theta}}^{(m)}$ are the estimators obtained from each imputed data set, for $m = 1, \ldots, M$. The MI estimator corresponds to

$$\hat{\boldsymbol{\theta}}_{\text{MI}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\theta}}^{(m)} \approx \frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}^{(m)}\right] \approx \mathsf{E}\left[\boldsymbol{\theta} \mid y_1, y_{2,obs}\right],$$

whereas the estimator for the covariance of $\hat{\boldsymbol{\theta}}_{\text{MI}}$ is

$$
\begin{aligned}
\hat{\boldsymbol{V}}_{\text{MI}} &= \hat{\boldsymbol{W}}_{\text{MI}} + \hat{\boldsymbol{B}}_{\text{MI}} \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathbf{V}_{\boldsymbol{\theta}}^{(m)} + \frac{1}{M-1}\sum_{m=1}^{M} (\hat{\boldsymbol{\theta}}^{(m)} - \hat{\boldsymbol{\theta}}_{\text{MI}})(\hat{\boldsymbol{\theta}}^{(m)} - \hat{\boldsymbol{\theta}}_{\text{MI}})^{\mathsf{T}} \\
&\approx \frac{1}{M}\sum_{m=1}^{M} \mathsf{Var}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}^{(m)}\right] \\
&+ \frac{1}{M-1}\sum_{m=1}^{M} \left(\mathsf{E}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}^{(m)}\right] - \frac{1}{M}\sum_{m'=1}^{M} \mathsf{E}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}^{(m')}\right]\right) \\
&\qquad \left(\mathsf{E}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}^{(m)}\right] - \frac{1}{M}\sum_{m'=1}^{M} \mathsf{E}\left[\boldsymbol{\theta} \mid y_{2,obs}, y_{2,mis}^{(m')}\right]\right)^{\mathsf{T}} \\
&\approx \mathsf{Var}\left[\boldsymbol{\theta} \mid y_1, y_{2,obs}\right].
\end{aligned}
$$

The variance estimator is made up of two terms $\hat{\boldsymbol{W}}_{\text{MI}}$ and $\hat{\boldsymbol{B}}_{\text{MI}}$, that correspond to the average and variance of $\boldsymbol{\theta}$ over the repeated draws from $f(y_{2,mis} \mid y_1, y_{2,obs})$, respectively, and represent the within- and between-imputation variability (Little & Rubin, 2019).

When the number of imputations is large, Rubin's rules allow to carry out inference about individual parameters the standard way, that is, confidence intervals and tests can be constructed using the normal as the reference distribution (Molenberghs et al., 2014). For a small number of imputations, Rubin adjusted the variance estimator by a factor of $1 + M^{-1}$ and inference is based on a t distribution with degrees of freedom given by $\varsigma = (M-1)(1 + r_M^{-1})^2$, where $r_M = (1 + M^{-1})\hat{\boldsymbol{B}}_{\text{MI}}^{[kk]}/\hat{\boldsymbol{W}}_{\text{MI}}^{[kk]}$, such that $\hat{\boldsymbol{B}}_{\text{MI}}^{[kk]}$ and $\hat{\boldsymbol{W}}_{\text{MI}}^{[kk]}$ are the between and within variances for the k^{th} element of $\hat{\boldsymbol{\theta}}_{\text{MI}}$. These adjustments were important at the time when the computational cost of generating a large number of imputations (and storing the completed data sets) was high however, they are less used in current research (Molenberghs et al., 2014). For further details in the MI framework we refer the reader to Rubin (1987).

The MI approach tackles the problem of missing data using similar ideas to those from the Expectation-Maximization algorithm (EM, Dempster et al., 1977) and its simulation-based extensions, such as the stochastic EM (stEM, Celeux & Diebolt, 1985), the simulated EM (sEM, Ruud, 1991), and the Monte Carlo EM (MCEM, Wei & Tanner, 1990). The aforementioned strategies all proceed by filling-in the missing values in some way, and then solving the estimation problem using the methods that would have been used in the absence of missing values (Rubin, 1991). As we have seen above, the MI framework is derived from a Bayesian perspective and computes the

observed data posterior of $\boldsymbol{\theta}$ by averaging the complete data posterior with respect to the posterior predictive distribution of the missing values given the observed. Parameter estimates and their covariance matrix are approximations to the posterior mean and posterior variance, respectively. In contrast, EM approaches obtain the maximum likelihood estimate of $\boldsymbol{\theta}$ by iteratively maximising the expected value of the log-likelihood with respect to the conditional density of the missing values given the observed, and a current approximation to the maximum likelihood estimate. The EM algorithm can be summarized as follows: at iteration j, given an estimate of the parameter of interest, say $\boldsymbol{\theta}^{(j)}$, the following two steps are iterated until some convergence criterion is satisfied (see, for example, McLachlan & Krishnan, 2008)

$$\begin{aligned}
\textbf{E-step:} \qquad Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)}) &= \int_{\{y_2 \colon y_1=0\}} \ell(\boldsymbol{\theta} \mid y_1, y_2) f(y_2 \mid y_1 = 0, \boldsymbol{\theta}^{(j)}) \mathrm{d}y_2 \\
&= \mathsf{E}\left[\ell(\boldsymbol{\theta} \mid y_1, y_2) \mid y_1, \boldsymbol{\theta}^{(j)}\right], \\
\textbf{M-step:} \qquad \boldsymbol{\theta}^{(j+1)} &= \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)}),
\end{aligned}$$

where $\ell(\boldsymbol{\theta} \mid y_1, y_2)$ denotes the log-likelihood function, and $f(y_2 \mid y_1 = 0, \boldsymbol{\theta}^{(j)})$ corresponds to the density of the missing values given the observed and the current estimate of the parameter vector. Dempster et al. (1977) showed that the value of the likelihood never decreases after each iteration and, under regularity conditions, Wu (1983) proved that the approach yields a sequence of values that converges to the maximum likelihood estimate of $\boldsymbol{\theta}$.

When the expectation in the E-step is analytically intractable, the simulation-based EM approaches replace the integral by an estimate using Monte Carlo simulation, i.e., the E-step becomes

$$\tilde{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)}) = \frac{1}{\tilde{M}} \sum_{\tilde{m}=1}^{\tilde{M}} \ell(\boldsymbol{\theta} \mid y_1, y_2^{(\tilde{m})})$$

where $y_2^{(\tilde{m})}$ represents the partially observed response variable where the missing values have been replaced by draws from $f(y_2 \mid y_1 = 0, \boldsymbol{\theta}^{(j)})$ at iteration j. The M-step consists of maximising $\tilde{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)})$ and both steps are iterated until convergence. The difference between stEM and MCEM is that the former only obtains $\tilde{M} = 1$ set of imputations per iteration, whereas the later uses a large number of draws (Nielsen, 2000b). The MCEM and sEM approaches differ in the way the random draws from $f(y_2 \mid y_1 = 0, \boldsymbol{\theta}^{(j)})$ are obtained. MCEM obtains independent realisations at each iteration, while sEM reuses the draws of values to impute in the first iteration at later iterations (Wang & Robins, 1998; Nielsen, 2000a).

The main drawback of the EM approaches is that convergence may be slow, in particular, when the missingness mechanism is non-ignorable (Little & Rubin, 2002). Moreover, the Monte Carlo step in the simulation-based approaches adds extra computational cost when $\tilde{M}$ is large. Wang & Robins (1998) and Wei & Tanner (1990) suggest to start the iterations by allocating a small number of imputations when the current estimate may be far from the true maximiser of the log-likelihood, and increase it at later iterations. In addition, EM approaches require further computation to obtain the covariance matrix of the estimates using, for example, bootstrapping or Louis' approach (Louis, 1982).

For further comparisons between these approaches we refer the reader to Nielsen (2000a, 2003), McLachlan & Krishnan (2008), Noghrehchi et al. (2021), Robins & Wang (2000), Wang & Robins (1998), and Schafer (1997).

4.4.2 The imputation model

As anticipated in the previous section, the first step of the MI process involves obtaining M random draws $y_{2,mis}^{(1)}, \ldots, y_{2,mis}^{(M)}$ from the posterior predictive distribution of the missing values given the observed, defined as

$$f(y_{2,mis} \mid y_1, y_{2,obs}) = \int f(y_{2,mis} \mid y_1, y_{2,obs}, \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid y_1, y_{2,obs}) \mathrm{d}\boldsymbol{\theta}, \tag{4.10}$$

where $f(\boldsymbol{\theta} \mid y_1, y_{2,obs})$ represents the observed data posterior distribution of $\boldsymbol{\theta}$ (or an approximation) and $f(y_{2,mis} \mid y_1, y_{2,obs}, \boldsymbol{\theta})$ is the conditional density of the missing values given the observed and $\boldsymbol{\theta}$.

Little & Rubin (2019) describe several approaches to draw plausible imputations from the frequentist and Bayesian points of view. For instance, following a frequentist approach, one can fix the value of $\boldsymbol{\theta}$ to the maximum likelihood estimate of the observed data, say $\tilde{\boldsymbol{\theta}}_{\text{MLE}}$, and obtain the random draws from $f(y_{2,mis} \mid y_1, y_{2,obs}, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_{\text{MLE}})$. This approach fails to account for the uncertainty in estimating $\boldsymbol{\theta}$ since the same value of the estimated parameter vector is used to obtain all imputations. Alternatively, following a Bayesian approach, random draws from $f(y_{2,mis} \mid y_1, y_{2,obs})$ can be obtained iteratively by first drawing $\tilde{\boldsymbol{\theta}}^{(m)}$ from $f(\boldsymbol{\theta} \mid y_1, y_{2,obs})$ and then drawing $y_{2,mis}^{(m)}$ from $f(y_{2,mis} \mid y_1, y_{2,obs}, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}^{(m)})$, for $m = 1, \ldots, M$. In practical terms, a fully Bayesian approach entails obtaining the posterior distribution of the parameters and the use of Markov chain Monte Carlo methods, such as data augmentation (Tanner & Wong, 1987). This

approach accounts for the uncertainty in estimating $\boldsymbol{\theta}$ but it may be computationally costly in complex models and convergence may be difficult to assess (Harel & Zhou, 2007). The strategy we follow in this chapter samples iteratively from the two densities within the integral in (4.10) however, instead of computing the posterior we use the large sample normal approximation to the posterior of $\boldsymbol{\theta}$ to draw $\tilde{\boldsymbol{\theta}}^{(m)}$.

Letting $\boldsymbol{\theta}$ denote the parameter vector characterising the copula-based SS model described in Section 4.3, that is, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\mathsf{T}, \boldsymbol{\theta}_2^\mathsf{T}, \delta_{12})^\mathsf{T}$ where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_{\mu_1}^\mathsf{T}, 1)^\mathsf{T}$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_{\mu_2}^\mathsf{T}, \boldsymbol{\beta}_{\sigma_2}^\mathsf{T}, \boldsymbol{\beta}_{\nu_2}^\mathsf{T})^\mathsf{T}$, realisations from the observed data posterior of $\boldsymbol{\theta}$ are obtained using the large sample result given in (4.6), i.e., $\boldsymbol{\theta} \mid \mathbf{y} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}}\right)$, where $\hat{\boldsymbol{\theta}}$ is the penalized maximum likelihood estimate and $\mathbf{V}_{\boldsymbol{\theta}}$ the Bayesian covariance matrix evaluated at $\hat{\boldsymbol{\theta}}$. The conditional distribution of the missing values corresponds to the following conditional density, derived from the copula-based representation of the joint distribution of (Y_1^*, Y_2^*) given in (4.4),

$$\begin{aligned} f(y_2 \mid y_1 = 0, \boldsymbol{\theta}) &= \frac{\partial}{\partial y_2} F(y_2 \mid y_1 = 0, \boldsymbol{\theta}) \\ &= \frac{\partial}{\partial y_2}\left[\frac{F(0, y_2 \mid \boldsymbol{\theta})}{F_1(0 \mid \boldsymbol{\theta}_1)}\right] \\ &= \frac{1}{F_1(0 \mid \boldsymbol{\theta}_1)} \frac{\partial \mathcal{C}(F_1(0 \mid \boldsymbol{\theta}_1), F_2(y_2 \mid \boldsymbol{\theta}_2) \mid \delta_{12})}{\partial F_2(y_2 \mid \boldsymbol{\theta}_2)} \frac{\partial F_2(y_2 \mid \boldsymbol{\theta}_2)}{\partial y_2}. \end{aligned} \tag{4.11}$$

This imputation scheme is appealing since it propagates the uncertainty about the model parameters and reduces the computational complexity of a fully Bayesian approach. As pointed out in Section 4.2, the parametric assumptions about the joint model using a copula function, and the parametric distributions of the missingness mechanism and the response, help to identify the parameters in the conditional distribution of the missing values.

Random draws from the target density in (4.11) are obtained using the acceptance/rejection method (see, e.g., Robert & Casella, 2005). This sampling approach requires finding a proposal or instrumental density, from which draws can be easily obtained, and finding a constant K that bounds the ratio of the target to the proposal density from above. In our context, it is reasonable to set the instrumental density to $f_2(y_2^* \mid \boldsymbol{\theta}_2)$, the density of the latent variable Y_2^* defined in Section 4.3. The value of K is computed by maximising the ratio $k(y_2) = \frac{f(y_2|y_1=0,\boldsymbol{\theta})}{f_2(y_2|\boldsymbol{\theta}_2)}$ using a trust-region approach. The acceptance/rejection algorithm then proceeds iteratively by drawing $\tilde{y}$ from $f_2(y_2^* \mid \boldsymbol{\theta}_2)$ and $u \sim \mathcal{U}(0,1)$ independently, and accepting $\tilde{y}$ if $u \leq \frac{f(\tilde{y}|y_1=0,\boldsymbol{\theta})}{K f_2(\tilde{y}|\boldsymbol{\theta}_2)}$. To avoid constraints in the optimization problem that obtains the value of K, the candidate for imputation $\tilde{y}$ is transformed using a differentiable and monotone function. For instance, when the support of

the distribution is restricted to positive numbers $\tilde{y}$ is transformed using $\breve{y} = \log \tilde{y}$. Furthermore, recall that trust-region methods require the analytical expressions for the first and second order derivatives of the objective function given by

$$\frac{dk(\tilde{y})}{d\breve{y}} = \frac{1}{F_1(0 \mid \boldsymbol{\theta}_1)} \frac{\partial^2 \mathcal{C}(F_1\left(0 \mid \boldsymbol{\theta}_1\right), F_2(\tilde{y} \mid \boldsymbol{\theta}_2))}{\partial F_2(\tilde{y} \mid \boldsymbol{\theta}_2) \partial F_2(\tilde{y} \mid \boldsymbol{\theta}_2)} f_2(\tilde{y} \mid \boldsymbol{\theta}_2) \frac{d\tilde{y}}{d\breve{y}},$$

and

$$\begin{aligned}\frac{d^2k(\tilde{y})}{d\breve{y}^2} = \frac{1}{F_1(0 \mid \boldsymbol{\theta}_1)} \Bigg[& \frac{\partial^3 \mathcal{C}(F_1\left(0 \mid \boldsymbol{\theta}_1\right), F_2(\tilde{y} \mid \boldsymbol{\theta}_2))}{\partial F_2(\tilde{y} \mid \boldsymbol{\theta}_2) \partial F_2(\tilde{y} \mid \boldsymbol{\theta}_2) \partial F_2(\tilde{y} \mid \boldsymbol{\theta}_2)} \left(f_2(\tilde{y} \mid \boldsymbol{\theta}_2)\right)^2 \\ & + \frac{\partial^2 \mathcal{C}(F_1\left(0 \mid \boldsymbol{\theta}_1\right), F_2(\tilde{y} \mid \boldsymbol{\theta}_2))}{\partial F_2(\tilde{y} \mid \boldsymbol{\theta}_2) \partial F_2(\tilde{y} \mid \boldsymbol{\theta}_2)} \frac{df_2(\tilde{y} \mid \boldsymbol{\theta}_2)}{d\tilde{y}} \Bigg] \left(\frac{d\tilde{y}}{d\breve{y}}\right)^2 + \frac{dk(\tilde{y})}{d\tilde{y}} \frac{d^2\tilde{y}}{d\breve{y}^2}.\end{aligned}$$

To summarise, the imputation scheme proceeds iteratively as follows, for $m = 1, \ldots, M$:

(i) draw $\tilde{\boldsymbol{\theta}}^{(m)} = (\tilde{\boldsymbol{\theta}}_1^{(m)\mathsf{T}}, \tilde{\boldsymbol{\theta}}_2^{(m)\mathsf{T}}, \tilde{\delta}_{12}^{(m)})^{\mathsf{T}}$ from $\mathcal{N}(\hat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$,

(ii) draw $u^{(m)} \sim \mathcal{U}(0,1)$ and $y_{2,mis}^{(m)}$ from $f_2(y_2^* \mid \boldsymbol{\theta}_2 = \tilde{\boldsymbol{\theta}}_2^{(m)})$ independently,

(iii) compute $K^{(m)} = \arg\max \frac{f(y_{2,mis}^{(m)} \mid y_1 = 0, \tilde{\boldsymbol{\theta}}^{(m)})}{f_2(y_{2,mis}^{(m)} \mid \tilde{\boldsymbol{\theta}}_2^{(m)})}$ using a trust-region method,

(iv) accept $y_{2,mis}^{(m)}$ if

$$u^{(m)} \leq \frac{f(y_{2,mis}^{(m)} \mid y_1 = 0, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}^{(m)})}{K^{(m)} f_2(y_{2,mis}^{(m)} \mid \boldsymbol{\theta}_2 = \tilde{\boldsymbol{\theta}}_2^{(m)})}.$$

Note that this approach to MI can also be easily adapted to impute missing values of a variable believed to be MAR based on univariate GAMLSS models. As we indicated in Section 4.2, the MAR assumption implies that the conditional density of the variable subject to missingness is the same, regardless the variable is observed or not, since Y_1 and Y_2 are conditionally independent given the observed data. Imputations from the posterior predictive distribution in (4.10) can be obtained by drawing iteratively from large sample normal approximation to the posterior of the parameter vector characterising the GAMLSS model, and the distribution of the partially observed variable. A similar approach to imputation based on GAMLSS models under the assumption of MAR was proposed by de Jong et al. (2016) however, they obtained draws from the observed data posterior using a parametric bootstrap approach.

4.5 Simulation study

We conduct a Monte Carlo study in order to investigate the empirical properties of the MI approach and compare the results with those obtained using a copula-based SS model. We have restricted the scope of the study since the relative performance of the copula-based and MI approaches compared to the classical SS methods has already been investigated by Gomes et al. (2019). Their simulation study was mainly focused on the performance of the approaches in terms of estimating the effect of a binary treatment variable on the response and did not include non-parametric components to capture non-linear effects.

In this study, we consider a scenario in which the selection and the substantive models are generated using the normal and gamma distributions, respectively, where the additive predictors associated with the distribution parameters also contain non-parametric terms. The joint distribution is specified using a copula function from the Gumbel family. We assess the results under the correct specification and when the copula function is misspecified using the Joe family instead.

The simulated data consist of a set of independent observations $(\mathbf{y}_i, \mathbf{x}_i)_{i=1}^n$, where $\mathbf{y}_i = (y_{1i}, y_{2i})$ are realizations of the pair or random variables (Y_{1i}, Y_{2i}) obtained using the observations rules given in (4.3), and $\mathbf{x}_i = (x_{1i}, x_{2i})$ consists of a binary and a continuous variable generated using the approach described in Chapters 2 and 3, and references therein. The model for the missingness mechanism is given by $Y_{1i}^* \sim F_1(\boldsymbol{\theta}_{1i})$, where F_1 corresponds to the normal distribution with parameters $\boldsymbol{\theta}_{1i} = (\mu_{1i}, \sigma_{1i})^\intercal$, σ_1 is set to one to ensure identifiability, and the additive predictor associated with μ_1 is given by

$$\eta_i^{\mu_1} = \beta_0^{\mu_1} + \beta_1^{\mu_1} x_{1i} + s_1^{\mu_1}(x_{2i}),$$

where $\beta_0^{\mu_1} = -0.55$, $\beta_1^{\mu_1} = 1.2$, and $s_1^{\mu_1}(x_{2i}) = 1 - x_{2i}^3 - 2\exp(-180x_{2i}^2) - 2.3\sin(4.9x_{2i})$. The missingness mechanism generates approximately 50% of missing observations in the partially observed variable. The substantive model is defined as $Y_{2i}^* \sim F_2(\boldsymbol{\theta}_{2i})$, where F_2 corresponds to the gamma distribution with parameters $\boldsymbol{\theta}_{2i} = (\mu_{2i}, \sigma_{2i})^\intercal$ and associated additive predictors

$$\begin{aligned}\eta_i^{\mu_2} &= \beta_0^{\mu_2} + \beta_1^{\mu_2} x_{1i} + s_1^{\mu_2}(x_{2i}),\\ \eta_i^{\sigma_2} &= \beta_0^{\sigma_2},\end{aligned}$$

where $\beta_0^{\mu_2} = 2.2$, $\beta_1^{\mu_2} = 1.3$, $s_1^{\mu_2}(x_{2i}) = x_{2i} + \exp(-32(x_{2i} - 0.5)^2)$, and $\beta_0^{\sigma_2} = 0.9$. The joint

distribution of (Y_{1i}^*, Y_{2i}^*), is specified using the Gumbel copula family, where the copula parameter δ_{12} has been set to values corresponding to Kendall's tau of $\tau_{12} \in \{0.2, 0.4, 0.6\}$.

We perform $N = 200$ repetitions[3] with sample sizes of $n = \{500, 1000, 3000\}$ and estimate the parameters using the copula-based SS (SS-COP) and the imputation (MI) approaches for the correctly specified model (scenario I) and for a misspecified model in which the joint distribution is assumed to belong to the Joe copula family (scenario II). In terms of the imputation approach, we set the number of imputations to $M = 100$.

Figure 4.1 shows boxplots of the estimates of $\beta_1^{\mu_2}$ and $\beta_0^{\sigma_2}$ for both scenarios, whereas Figure 4.2 shows the true smooth functions (dashed lines) and the average effect estimates (solid lines) of $s_1^{\mu_2}(x_2)$, together with the 5% and 95% point-wise quantiles (shaded areas), for a sample size of $n = 3000$ and association parameter $\tau_{12} = 0.4$. For brevity, we have omitted the plots of smooth functions for other sample sizes and strengths of association since the results are similar and do not change the conclusion of the simulation study. In addition, Table 4.1 summarises the results in terms of relative bias and root mean squared error (RMSE).

Overall, we observe that MI obtains very similar performance, in terms of bias and RMSE, to SS-COP and, in occasions, MI estimates are slightly less biased and more precise. In particular, MI estimates of $\beta_0^{\sigma_2}$ appear to be slightly less biased and yield lower RMSE than those obtained using SS-COP. Both approaches also perform relatively well when the copula is mildly misspecified. These results are consistent with those shown by Gomes et al. (2019), who studied the performance of the approaches using a wider range of MNAR settings. We conducted further experiments using different marginal distributions, copula functions, and strengths of dependence and the results were similar to those shown here.

4.6 Empirical application

The REFLUX study (Grant et al., 2008, 2013) evaluates the effect of using surgery, compared to continuing medication, on long-term patient's health status among individuals with gastro-oesophageal reflux disease in the UK. The trial comprised a randomised arm, in which 357 individuals were randomised to either surgical or medical management, and a parallel non-randomised arm, in which 453 participants were assigned to treatment based on their preferences. We focus

[3]The number of repetitions is chosen based on the simulation study in Chapter 2 and references therein. Using the principled approach described there on the scenario for the correctly specified model with $n = 3000$; an initial run of $N_0 = 50$; and focusing on the bias obtained for $\beta_1^{\mu_2}$ and $\beta_0^{\sigma_2}$, we obtained $N \approx 297$ and $N \approx 33$.

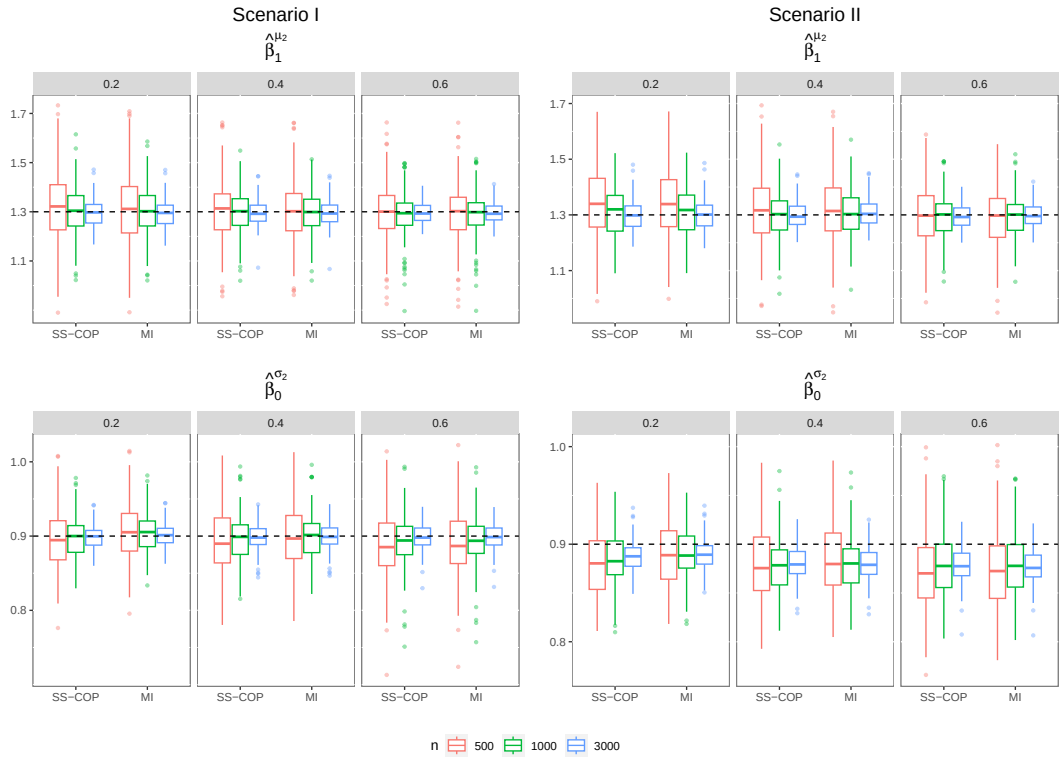

Figure 4.1: Boxplots of the estimates of $\beta_1^{\mu_2}$ and $\beta_0^{\sigma_2}$ obtained using the copula-based (SS-COP) and multiple imputation (MI) approaches for the correctly specified model (scenario I: normal-gamma marginals and Gumbel copula) and when the copula is misspecified (scenario II: normal-gamma marginals and Joe copula). The sample sizes are $n \in \{500, 1000, 3000\}$. The strength of association is given by Kendall's tau values of $\tau_{12} \in \{0.2, 0.4, 0.6\}$. The true values of the parameters are $\beta_1^{\mu_2} = 1.3$ and $\beta_0^{\sigma_2} = 0.9$ (indicated by a dashed line in the plot).

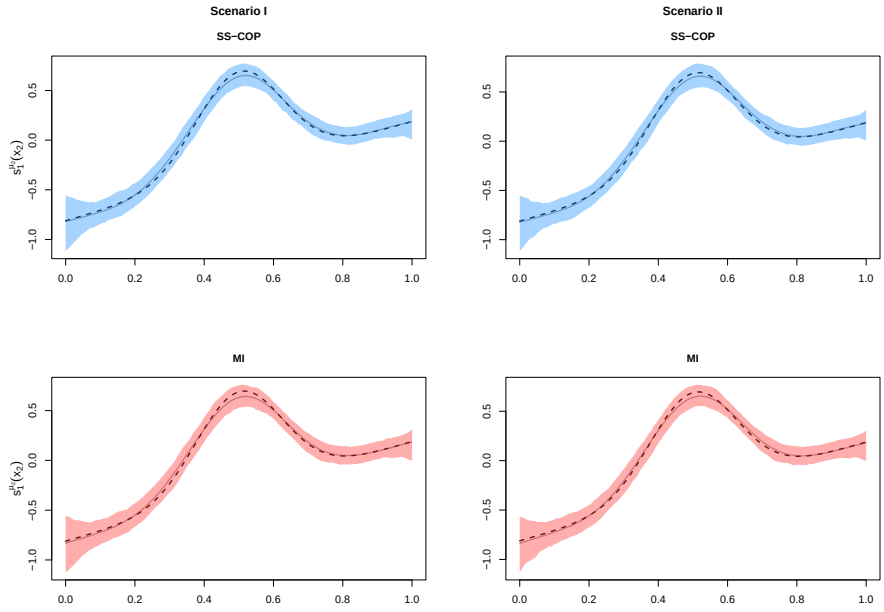

Figure 4.2: Mean estimates of $s_1^{\mu_2}(x_2)$ obtained using the copula-based (SS-COP) and multiple imputation (MI) approaches with a sample size of $n = 3000$ and a strength of association of $\tau_{12} = 0.4$, for the correctly specified model (scenario I: normal-gamma marginals and Gumbel copula) and when the copula is misspecified (scenario II: normal-gamma marginals and Joe copula). The shaded areas correspond to the 5% and 95% point-wise quantiles. The true functions are represented by dashed lines.

		Scenario I						Scenario II					
		Relative bias			RMSE			Relative bias			RMSE		
τ_{12}	method | n	500	1000	3000	500	1000	3000	500	1000	3000	500	1000	3000
		$\hat{\beta}_1^{\mu_2}$						$\hat{\beta}_1^{\mu_2}$					
0.2	SS-COP	0.0162	0.0017	-0.0040	0.1439	0.0988	0.0535	0.0303	0.0084	-0.0019	0.1387	0.0927	0.0533
	MI	0.0114	-0.0006	-0.0047	0.1425	0.0986	0.0532	0.0289	0.0073	-0.0008	0.1393	0.0932	0.0536
0.4	SS-COP	0.0064	-0.0011	-0.0037	0.1272	0.0849	0.0501	0.0122	0.0007	-0.0004	0.1247	0.0845	0.0486
	MI	0.0036	-0.0023	-0.0038	0.1272	0.0845	0.0501	0.0128	0.0040	0.0044	0.1245	0.0844	0.0486
0.6	SS-COP	0.0002	-0.0067	-0.0031	0.1215	0.0883	0.0422	-0.0018	-0.0043	-0.0040	0.1141	0.0760	0.0428
	MI	-0.0041	-0.0078	-0.0030	0.1197	0.0888	0.0424	-0.0053	-0.0046	-0.0004	0.1118	0.0753	0.0423
		$\hat{\beta}_0^{\sigma_2}$						$\hat{\beta}_0^{\sigma_2}$					
0.2	SS-COP	-0.0048	-0.0017	-0.0014	0.0419	0.0283	0.0158	-0.0212	-0.0171	-0.0141	0.0398	0.0301	0.0202
	MI	0.0070	0.0052	0.0012	0.0413	0.0275	0.0157	-0.0106	-0.0108	-0.0117	0.0355	0.0269	0.0186
0.4	SS-COP	-0.0074	-0.0037	-0.0010	0.0463	0.0300	0.0173	-0.0217	-0.0248	-0.0218	0.0446	0.0350	0.0257
	MI	-0.0008	-0.0007	-0.0001	0.0443	0.0293	0.0172	-0.0162	-0.0230	-0.0221	0.0416	0.0335	0.0258
0.6	SS-COP	-0.0160	-0.0076	-0.0015	0.0761	0.0341	0.0176	-0.0295	-0.0238	-0.0238	0.0488	0.0382	0.0278
	MI	-0.0099	-0.0067	-0.0012	0.0497	0.0333	0.0176	-0.0280	-0.0242	-0.0256	0.0473	0.0384	0.0290
		$\hat{s}_1^{\mu_2}(x_2)$						$\hat{s}_1^{\mu_2}(x_2)$					
0.2	SS-COP	0.0577	0.0437	0.0217	0.1772	0.1302	0.0753	0.0580	0.0460	0.0184	0.1764	0.1221	0.0774
	MI	0.0615	0.0453	0.0220	0.1767	0.1303	0.0753	0.0575	0.0468	0.0194	0.1734	0.1224	0.0771
0.4	SS-COP	0.0477	0.0342	0.0178	0.1609	0.1091	0.0665	0.0483	0.0336	0.0164	0.1570	0.1055	0.0639
	MI	0.0518	0.0361	0.0186	0.1595	0.1086	0.0666	0.0524	0.0345	0.0173	0.1537	0.1045	0.0635
0.6	SS-COP	0.0422	0.0301	0.0152	0.1710	0.1102	0.0576	0.0419	0.0300	0.0151	0.1345	0.0959	0.0583
	MI	0.0453	0.0316	0.0159	0.1564	0.1100	0.0574	0.0435	0.0284	0.0140	0.1295	0.0935	0.0564

Table 4.1: Relative bias and RMSE for $\hat{\beta}_1^{\mu_2}, \hat{\beta}_0^{\sigma_2}$, and smooth function estimate $\hat{s}_1^{\mu_2}(x_2)$ obtained using the copula-based (SS-COP) and multiple imputation (MI) approaches, for the correctly specified model (scenario I: normal-gamma marginals and Gumbel copula) and when the copula is misspecified (scenario II: normal-gamma marginals and Joe copula). The sample sizes are $n \in \{500, 1000, 3000\}$. The strength of association is given by Kendall's tau values of $\tau_{12} \in \{0.2, 0.4, 0.6\}$. The true values of the parameters are $\beta_1^{\mu_2} = 1.3$ and $\beta_0^{\sigma_2} = 0.9$. The true smooth function is $s_1^{\mu_2}(x_2) = x_2 + \exp[-32(x_2 - 0.5)^2]$.

our interest in the analysis of the preference arm and refer the reader to Grant et al. (2008, 2013) for further details on the randomized component.

The study gathered patient information using self-reported questionnaires before treatment, three months after treatment, and yearly, together with hospital case notes reviews. We use the variable `QALY` (5-year Quality-Adjusted life-years) as a measure of long-term patient's health, which was constructed using a combination of health related quality of life scores, obtained from the questionnaires, and estimated length of life after treatment (see Gomes et al., 2019, 2020, for further details). `QALY` has a total of 222 missing observations due to patients not returning the questionnaires at some point during the follow-up period. As pointed out by Gomes et al. (2019), the study investigators suspected that individuals in better health after treatment were more likely to be engaged with the study and return the questionnaires. This indicates that the process governing the missing values may depend on the variable that is missing, raising concerns about the possibility of `QALY` being MNAR. A further concern relates to the plausibility of the normality assumption about the response. The variable varies from -0.52 to 4.67 and the histogram in Figure 4.3 shows that its distribution is left-skewed. Gomes et al. (2019) suggested that a Gumbel

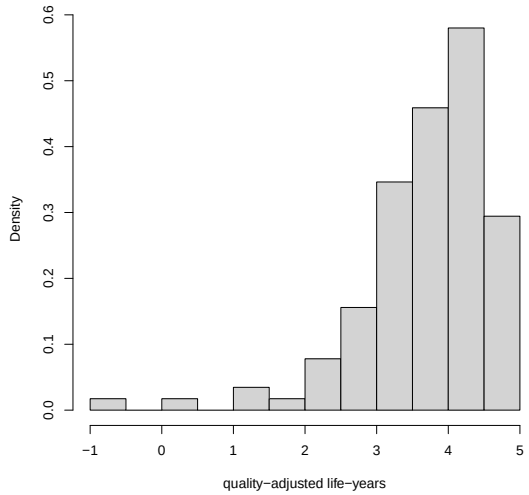

Figure 4.3: Histogram of 5-year quality-adjusted life-years (`QALY`)

distribution may provide a better fit.

Similarly to Gomes et al. (2019, 2020), the aim of the analysis is to investigate the effect of `treatment` (surgical or medical management) on `QALY` under different assumptions about the missingness mechanism, the distribution of the response, and the structure of their dependence, while accounting for several baseline patient's characteristics. Specifically, we compare the results obtained assuming the missing mechanism is ignorable, using complete-case analyses based on different distributional specifications for the response, with the results obtained under the MNAR assumption, using the SS modelling framework and the MI approach presented in Section 4.4.

Table 4.2 contains a description and summary statistics of the response, the baseline patient's characteristics, and variables that are plausible to meet the criteria for the exclusion restriction assumption that will be considered in the models under the MNAR assumption. We observe that the overall `QALY` score after five years of treatment is higher for individual that opted for surgery than for those that selected long-term medication. Furthermore, participants that chose surgery were younger and with worse health-related symptoms (judged by the lower values in most of the score variables) than patients than went through medical management. The potential variables that are considered to fulfil the exclusion restriction assumption are `csize`, which refers to the number of participants in each of the hospitals providing treatment, and `belief1` to `belief8`, which are related to patients' opinions about medicine and doctors. Gomes et al. (2019, 2020), and references therein, argue that centres with a higher number of patients may be more involved collecting self-reported questionnaires, and that patient's views about medicine are likely to influence their

Variable	Description	Medical management (n = 192)	Surgery (n = 261)
		Mean (SD)/Count (percentage)	Mean (SD)/Count (percentage)
Response			
QALY	Quality-Adjusted Life-Years	3.594 (0.834)	3.777 (0.942)
Baseline patient's characteristics			
gender	=1 if male	111 (58%)	170 (65%)
age	age in years	49.89 (11.75)	44.45 (11.97)
EQ-5D	European quality of life score	0.75 (0.22)	0.68 (0.26)
REFLUX-QoL	REFLUX quality of life score	76.11 (19.84)	55.87 (22.82)
bmi	Body mass index	27.42 (4.11)	27.72 (3.95)
heartburn	Heartburn score	73.06 (21.29)	9.13 (24.44)
gastro1	Gastro symptom score 1	59.63 (22.68)	47.05 (21.39)
gastro2	Gastro symptom score 2	82.98 (17.68)	75.84 (22.04)
nausea	Nausea symptom score	89.67 (13.61)	76.86 (19.90)
activity	Activity score	86.82 (12.97)	74.36 (16.11)
hernia	=1 if previous hiatus hernia	73 (38%)	76 (29%)
smoker	=1 if smoker	39 (20%)	71 (27%)
asthma	=1 if asthma	36 (19%)	30 (11%)
duration	duration of prescribed medication (days)	45.79 (54.10)	55.96 (68.04)
Candidate instruments			
csize	Centre size (No. patients treated)	27 (11%)	28 (9%)
belief1	Doctors use too many medicines	32 (17%)	61(23%)
belief2	People should pause treatments	42 (22%)	76 (29%)
belief3	Medicines are addictive	22 (11%)	39 (15%)
belief4	Natural remedies are safer	30 (16%)	38 (15%)
belief5	Medicines do more harm than good	4 (2%)	5 (2%)
belief6	All medicines are poisons	13 (7%)	7 (3%)
belief7	Doctors trust medicines too much	26 (14%)	51 (20%)
belief8	Doctors should spend more time with patients	69 (36%)	94 (36%)

Table 4.2: Description and summary statistics of the response, baseline individual characteristics, and available candidates to meet the exclusion restriction criteria by treatment level. The continuous covariates are summarised using the mean and standard deviation whereas the binary variables are reported using the number of patients and the corresponding percentage.

engagement with the study. However, it does not seem obvious that these variables may have a direct effect on the response given the baseline individual characteristics. Their analysis also suggest that the aforementioned variables have a relatively low association with patients returning the questionnaires, which indicate that they may be considered as 'weak' instruments.

To determine the effect of `treatment` under different modelling assumptions we consider the following approaches: assuming the missingness mechanism is ignorable, we perform complete-case analyses using the normal and the Gumbel distributions to model the response (denoted as CC-N and CC-G, respectively). Under the MNAR assumption, we use the classical SS model estimated via the two-step Heckman (2S) method and maximum likelihood (SS-ML), the copula-based SS framework (SS-COP), and the multiple imputation (MI) approach.

For all the aforementioned modelling strategies, we specify a linear predictor associated with the location parameter of the model of interest using several baseline patient's characteristics re-

ported in the REFLUX study, that is,

$$\begin{aligned}\eta_i^{\mu} = \beta_0 &+ \beta_1 \texttt{treatment}_i + \beta_2 \texttt{gender}_i + \beta_3 \texttt{age}_i + \beta_4 \texttt{EQ-5D}_i + \beta_5 \texttt{REFLUX-QoL}_i \\ &+ \beta_6 \texttt{bmi}_i + \beta_7 \texttt{heartburn}_i + \beta_8 \texttt{gastro1}_i + \beta_9 \texttt{gastro2}_i + \beta_{10} \texttt{nausea}_i \\ &+ \beta_{11} \texttt{activity}_i,\end{aligned} \tag{4.12}$$

where β_1 corresponds to the parameter of interest. In the approaches that assume QALY is MNAR, the model for the missingness mechanism contains the same covariates that appear in the substantive model together with all the candidate variables that are believed to fulfil the exclusion restriction assumption (see Gomes et al., 2019, and references therein).

In terms of the SS-COP approach, we first make an assessment using different marginal distributions to model the missingness mechanism and the response, and then investigate the dependence structure employing several copula functions. Among the different model specifications, we select the model with the lowest values of the AIC and BIC. Regarding the binary response model for the missing data mechanism, we find that a complementary log-log link (which corresponds to assuming a Gumbel distribution for the latent variable governing the missingness mechanism) delivers the best fit. As for the substantive model, the chosen distribution is the Gumbel. Lastly, in terms of their association, the Frank copula appears to deliver the best description of the dependence structure.

With regard to the MI approach, we perform $M = 100$ imputations based on the SS-COP model described above. The analysis of the completed data sets are carried out using a GAMLSS model based on the Gumbel distribution and the resulting parameter estimates are combined using the Rubin's rules described in Section 4.4.

Table 4.3 contains the estimated parameters, together with their standard errors, obtained after fitting the different models to the data. We focus our attention on the parameter of interest. The estimated effect of treatment is positive and significant across all modelling approaches providing evidence that individuals that opted for surgical management of gastro-oesophageal reflux disease had higher average quality of life than those that followed medical management. These effects are larger for models under the classical distributional assumptions (CC-N, 2S, SS-ML) compared to those that employ flexible distributional approaches (CC-G, SS-COP, MI). In terms of the approaches that assume QALY is MNAR, we observe that the estimated effects are

Estimates | models	CC-N	CC-G	2S	SS-ML	SS-COP	MI
`treatment`	0.429***	0.362***	0.427***	0.427***	0.372***	0.372***
	(0.096)	(0.072)	(0.095)	(0.095)	(0.073)	(0.075)
`gender:male`	0.262**	0.189**	0.237*	0.237*	0.143*	0.152*
	(0.095)	(0.070)	(0.096)	(0.096)	(0.073)	(0.071)
`age`	-0.005	-0.005*	-0.004	-0.004	-0.003	-0.003
	(0.004)	(0.003)	(0.005)	(0.005)	(0.003)	(0.003)
`EQ-5D`	2.037***	1.360***	2.025***	2.026***	1.372***	1.361***
	(0.215)	(0.161)	(0.216)	(0.215)	(0.174)	(0.151)
`REFLUX-QoL`	-0.007*	-0.002	-0.009**	-0.009**	-0.003	-0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
`bmi`	-0.010	-0.014	-0.010	-0.010	-0.017*	-0.017
	(0.012)	(0.010)	(0.013)	(0.012)	(0.010)	(0.009)
`heartburn`	0.005*	0.003	0.005*	0.005*	0.002	0.003
	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
`gastro1`	0.005*	0.004*	0.005*	0.005*	0.005**	0.005**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
`gastro2`	-0.002	0.001	-0.001	-0.001	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
`nausea`	0.001	0.003	0.002	0.002	0.005*	0.005*
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)
`activity`	0.005	-0.002	0.006	0.006	-0.004	-0.003
	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
`(Intercept)`	1.864***	2.983***	1.706**	1.703**	2.757***	2.741***
	(0.485)	(0.393)	(0.545)	(0.534)	(0.419)	(0.346)

Table 4.3: Parameter estimates and standard errors of the coefficients in the substantive model obtained under different modelling assumptions. CC-N: complete-case analysis assuming a normal distribution; CC-G: complete-case analysis assuming a Gumbel distribution; 2S: classical SS model using the Heckman's two-step estimation approach; SS-ML: classical SS model using maximum likelihood estimation; SS-COP: copula-based SS model; MI: multiple imputation approach. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

similar to their complete-cases counterparts, which suggest a relatively weak strength of dependence. In fact, the estimates of the correlation coefficients obtained using 2S and SS-ML, 0.07 and $0.08(-0.61, 0.77)$ respectively[4], support the assumption that `QALY` is MAR. In contrast, the estimated value of the Kendall's tau in the copula-based SS model is positive and relatively significant, $0.291(0.04, 0.48)$, providing some evidence in favour of the MNAR assumption. These results suggest that the distributional assumptions in the CC-N, 2S, and ML-SS approaches may not hold and the models appear to overestimate the effect of the `treatment`. As expected, the MI approach yields estimates that are very similar to those obtained using the copula-based model and, in some instances, provides slightly smaller standard errors of the regression coefficients.

[4]The 2S method did not provide a value for the standard error of the coefficient.

4.7 Discussion

In this chapter, we have briefly reviewed some of the main concepts that are used in the missing data literature, discussed two SS modelling frameworks that deal with MNAR outcomes, and presented a MI approach that obtains plausible draws for imputation of a continuous variable assumed to be MNAR and not restricted to be Gaussian. The MI approach stems from a copula-based specification of SS models and complements recent research that addresses MNAR variables within the MI framework (Galimard et al., 2016; Ogundimu & Collins, 2019). The flexibility of the approach, inherited from the copula-based SS models proposed by Wojtyś et al. (2018), allows researchers to contrast the robustness of their results under different assumptions about the missingness mechanism, the distribution of the partially observed variable, and the dependence structure. Furthermore, the MI scheme can be embedded into a general fully conditional specification framework, such as that provided by the `mice` package, to deal with several partially observed variables in a data set. In a simulation study, we have shown that the performance of the MI approach is comparable to that of the copula-based SS model and that it performs relatively well under mild misspecification. These results are also consistent with those obtained in Gomes et al. (2019) using a wider range of MNAR settings.

In an application, we have re-examined data from the REFLUX study to evaluate the robustness of the conclusions under different assumptions regarding the missing mechanism, the distribution of the response, and the shape of the association structure. Although the results are significant under all modelling strategies, we find some evidence suggesting that models in which the response variable is assumed to be normally distributed appear to overestimate the effect of the treatment variable. In addition, the classical SS models do not appear to capture well the dependence structure between the missingness indicator and the response. The MI approach obtains results that are very similar to the copula-based SS model.

References

Ahn, H. & Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1-2), 3–29.

Ali, M. M., Mikhail, N., & Haq, M. S. (1978). A class of bivariate distributions including the bivariate logistic. *Journal of multivariate analysis*, 8(3), 405–412.

Arrow, K. J. (1978). Uncertainty and the welfare economics of medical care. In *Uncertainty in economics* (pp. 345–375). Elsevier.

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., & Canning, D. (2011). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology*, (pp. 27–35).

Benedetti, A. & Abrahamowicz, M. (2004). Using generalized additive models to reduce residual confounding. *Statistics in Medicine*, 23(24), 3781–3801.

Boos, D. & Stefanski, L. (2013). *Essential Statistical Inference: Theory and Methods*. Springer Texts in Statistics. Springer New York.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *The American Economic Review*, 77(4), 531–553.

Braun, M. (2014). trustOptim: An R Package for Trust Region Optimization with Sparse Hessians. *Journal of Statistical Software*, 60(4), 1 – 16.

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

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

Cameron, A. & Trivedi, P. (2009). *Microeconometrics Using Stata*. Stata Press.

Cameron, A. C., Trivedi, P. K., Milne, F., & Piggott, J. (1988). A microeconometric model of the demand for health care and health insurance in australia. *The Review of Economic Studies*, 55(1), 85–106.

Carpenter, J. & Kenward, M. (2012). *Multiple Imputation and its Application*. Statistics in Practice. Wiley.

Carpenter, J. R. & Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal*, 63(5), 915–947.

Celeux, G. & Diebolt, J. (1985). The SEM: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2, 73–82.

Chib, S. & Greenberg, E. (2007). Semiparametric modeling and estimation of instrumental variable models. *Journal of Computational and Graphical Statistics*, 16(1), 86–114.

Choi, P. & Min, I. (2009). Estimating endogenous switching regression model with a flexible parametric distribution function: application to Korean housing demand. *Applied Economics*, 41(23), 3045–3055.

Clarke, K. A. (2007). A simple distribution-free test for nonnested model selection. *Political Analysis*, 15(3), 347–363.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1), 141–151.

Conn, A. R., Gould, N. I., & Toint, P. L. (2000). *Trust region methods*. SIAM.

Craven, P. & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4), 377–403.

de Jong, R., van Buuren, S., & Spiess, M. (2016). Multiple Imputation of Predictor Variables Using Generalized Additive Models. *Communications in Statistics - Simulation and Computation*, 45(3), 968–985.

Deb, P., Munkin, M. K., & Trivedi, P. K. (2006). Private Insurance, Selection, and Health Care Use. *Journal of Business & Economic Statistics*, 24(4), 403–415.

Deb, P. & Norton, E. C. (2018). Modeling Health Care Expenditures and Use. *Annual Review of Public Health*, 39(1), 489–505.

Deb, P. & Trivedi, P. K. (2006). Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: Application to health care utilization. *The Econometrics Journal*, 9(2), 307–331.

DeBoor, C. (1978). *A Practical Guide to Splines*. Springer New York.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive theory of functions of several variables* (pp. 85–100). Springer.

Dunn, P. K. & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236–244.

Eilers, P. & Marx, B. (2021). *Practical Smoothing: The Joys of P-splines*. Cambridge University Press.

Eilers, P. H. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, (pp. 89–102).

Embrechts, P., McNeil, A., & Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1, 176–223.

Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, (pp. 731–761).

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg.

Fang, H., Keane, M., & Silverman, D. (2008). Sources of Advantageous Selection: Evidence from the Medigap Insurance Market. *Journal of Political Economy*, 116(2), 303–350.

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15), 1 – 27.

Frank, M. J. (1979). On the simultaneous associativity of F(x,y) and x + y - F(x,y). *Aequationes mathematicae*, 19(1), 194–226.

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

French, E. & Jones, J. B. (2011). The effects of health insurance and self-insurance on retirement behaviour. *Econometrica*, 79(3), 693–732.

Galimard, J.-E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine*, 35(17), 2907–2920.

Gallant, A. R. & Nychka, D. W. (1987). Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica*, 55(2), 363–390.

Gamoran, A. & Mare, R. D. (1989). Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? *American journal of Sociology*, 94(5), 1146–1183.

Genius, M. & Strazzera, E. (2008). Applying the copula approach to sample selection modelling. *Applied Economics*, 40(11), 1443–1455.

Geyer, C. J. (2015). *trust: Trust Region Optimization*. R package version 0.1-7. `https://CRAN.R-project.org/package=trust`.

Gomes, M., Kenward, M. G., Grieve, R., & Carpenter, J. (2020). Estimating treatment effects under untestable assumptions with nonignorable missing data. *Statistics in Medicine*, 39(11), 1658–1674.

Gomes, M., Radice, R., Camarena Brenes, J., & Marra, G. (2019). Copula selection models for non-Gaussian outcomes that are missing not at random. *Statistics in Medicine*, 38(3), 480–496.

Grant, A., Wileman, S., Ramsay, C., Bojke, L., Epstein, D., Sculpher, M., Macran, S., Kilonzo, M., Vale, L., Francis, J., Mowat, A., Krukowski, Z., Heading, R., Thursz, M., Russell, I., Campbell, M., & Group, R. T. (2008). The effectiveness and cost-effectiveness of minimal access surgery amongst people with gastro-oesophageal reflux disease - a UK collaborative study. The REFLUX trial. *Health technology assessment*, 12(31).

Grant, A. M., Boachie, C., Cotton, S. C., Faria, R., Bojke, L., Epstein, D. M., Ramsay, C. R., Corbacho, B., Sculpher, M., Krukowski, Z. H., Heading, R. C., Campbell, M. K., & REFLUX trial group (2013). Clinical and economic evaluation of laparoscopic surgery compared with medical management for gastro-oesophageal reflux disease: 5-year follow-up of multicentre randomised trial (the REFLUX trial). *Health Technol Assess*, 17(22), 1–167.

Green, P. J. & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Greene, W. (2014). *Econometric Analysis: International Edition: Global Edition*. Pearson series in economics. Pearson Education Limited.

Gumbel, E. J. (1960). Distributions des valeurs extremes en plusiers dimensions. *Publ. Inst. Statist. Univ. Paris*, 9, 171–173.

Harel, O. & Zhou, X.-H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, 26(16), 3057–3077.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Heckman, J. (1974). Shadow Prices, Market Wages, and Labor Supply. *Econometrica*, 42(4), 679–694.

Heckman, J. & Leamer, E., Eds. (2007). *Handbook of Econometrics*, volume 6B. Elsevier, 1 edition.

Heckman, J., Tobias, J. L., & Vytlacil, E. (2001). Four parameters of interest in the evaluation of social programs. *Southern Economic Journal*, (pp. 211–223).

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4* (pp. 475–492).

Heckman, J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, 46(4), 931–959.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1), 153–161.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, 24(3), 1 – 29.

Imbens, G. & Rubin, D. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of multivariate analysis*, 46(2), 262–282.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Joe, H. (2014). *Dependence Modeling with Copulas*. CRC Press.

Johnson, N. L. (1949a). Bivariate distributions based on simple translation systems. *Biometrika*, 36(3/4), 297–304.

Johnson, N. L. (1949b). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2), 149–176.

Kennickell, A. B. (2017). Multiple imputation in the Survey of Consumer Finances. *Statistical Journal of the IAOS*, 33(1), 143–151. Publisher: IOS Press.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*, 95(1), 49–69.

Klein, N. & Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Statistics and Computing*, 26(4), 841–860.

Klein, N., Kneib, T., Klasen, S., & Lang, S. (2015). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, (pp. 569–591).

Kneib, T., Silbersdorff, A., & Säfken, B. (2021). Rage Against the Mean - A Review of Distributional Regression Approaches. *Econometrics and Statistics*.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2004). *Continuous multivariate distributions, Volume 1: Models and applications*, volume 1. John Wiley & Sons.

Lee, L.-F. (1976). *Estimation of Limited Dependent Variables by Two Stage Method*. Unpublished PhD Thesis, Department of Economics, University of Rochester.

Lee, L.-F. (1978). Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables. *International Economic Review*, 19(2), 415–433.

Lee, L.-F. (1983). Generalized Econometric Models with Selectivity. *Econometrica*, 51(2), 507–512.

Little, R. J. (1982). Models for nonresponse in sample surveys. *Journal of the American statistical Association*, 77(378), 237–250.

Little, R. J. & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Little, R. J. A. & Rubin, D. B. (2002). Author Index. In *Statistical Analysis with Missing Data* (pp. 365–369). John Wiley & Sons, Inc.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2), 226–233.

Maddala, G. (1986a). Chapter 28 disequilibrium, self-selection, and switching models. volume 3 of *Handbook of Econometrics* (pp. 1633–1688). Elsevier.

Maddala, G. (1986b). Chapter 28 disequilibrium, self-selection, and switching models. volume 3 of *Handbook of Econometrics* (pp. 1633–1688). Elsevier.

Maddala, G. (1986c). *Limited-Dependent and Qualitative Variables in Econometrics*. Econometric Society Monographs. Cambridge University Press.

Manning, W. G., Basu, A., & Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of health economics*, 24(3), 465–488.

Manning, W. G. & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20(4), 461–494.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., & Leibowitz, A. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *The American Economic Review*, 77(3), 251–277.

Marchenko, Y. V. & Genton, M. G. (2012). A Heckman Selection-t Model. *Journal of the American Statistical Association*, 107(497), 304–317.

Mare, R. D. & Winship, C. (1988). Endogenous switching regression models for the causes and effects of discrete variables. In *Long SJ Common problems in quantitative social research*. Sage Press.

Marra, G. (2013). On p-values for semiparametric bivariate probit models. *Statistical Methodology*, 10(1), 23–28.

Marra, G. & Radice, R. (2011). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canadian Journal of Statistics*, 39(2), 259–279.

Marra, G. & Radice, R. (2013a). Estimation of a regression spline sample selection model. *Computational Statistics & Data Analysis*, 61, 158–173.

Marra, G. & Radice, R. (2013b). A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electronic Journal of Statistics*, 7, 1432–1455.

Marra, G. & Radice, R. (2017). Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 112, 99–113.

Marra, G. & Radice, R. (2019). Copula Link-Based Additive Models for Right-Censored Event Time Data. *Journal of the American Statistical Association*, (pp. 1–20).

Marra, G. & Radice, R. (2021). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-4. `https://CRAN.R-project.org/package=GJRM`.

Marra, G. & Radice, R. (2022). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-6. `https://CRAN.R-project.org/package=GJRM`.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., & McGovern, M. E. (2017). A Simultaneous Equation Approach to Estimating HIV Prevalence With Nonignorable Missing Responses. *Journal of the American Statistical Association*, 112(518), 484–496.

Marra, G., Radice, R., & Zimmer, D. (2022). A Unifying Switching Regime Regression Framework with Applications in Health Economics. *https://openaccess.city.ac.uk/id/eprint/28121/*.

Marra, G., Radice, R., & Zimmer, D. M. (2020). Estimating the binary endogenous effect of insurance on doctor visits by copula-based regression additive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(4), 953–971.

Marra, G. & Wood, S. N. (2012). Coverage Properties of Confidence Intervals for Generalized Additive Model Components. *Scandinavian Journal of Statistics*, 39(1), 53–74.

Martin, B. D., Witten, D., & Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *The annals of applied statistics*, 14(1), 94–115.

McLachlan, G. J. & Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley & Sons, Inc.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., & Verbeke, G. (2014). *Handbook of Missing Data Methodology*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Molenberghs, G. & Kenward, M. (2007). *Missing data in clinical studies*. John Wiley & Sons.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102.

Mullahy, J. (2009). Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. *Medical care*, 47(7_Supplement_1), S104–S108.

Nelsen, R. (2006). *An Introduction to Copulas*. Springer Series in Statistics. Springer.

Newhouse, J. P. & Phelps, C. E. (1974). Price and income elasticities for medical care services. In *The Economics of Health and Medical Care: Proceedings of a Conference Held by the International Economic Association at Tokyo* (pp. 139–161).: Springer.

Nielsen, S. F. (2000a). On simulated EM algorithms. *Journal of Econometrics*, 96(2), 267–292.

Nielsen, S. F. (2000b). The Stochastic EM Algorithm: Estimation and Asymptotic Results. *Bernoulli*, 6(3), 457–489.

Nielsen, S. F. (2003). Proper and Improper Multiple Imputation. *International Statistical Review / Revue Internationale de Statistique*, 71(3), 593–607.

Nocedal, J. & Wright, S. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York.

Noghrehchi, F., Stoklosa, J., Penev, S., & Warton, D. I. (2021). Selecting the model for multiple imputation of missing data: Just use an IC! *Statistics in Medicine*, 40(10), 2467–2497.

Nychka, D. (1988). Bayesian Confidence Intervals for Smoothing Splines. *Journal of the American Statistical Association*, 83(404), 1134–1143.

Ogundimu, E. & Hutton, J. (2016). A Sample Selection Model with Skew-normal Distribution. *Scandinavian Journal of Statistics*, 43(1), 172–190.

Ogundimu, E. O. & Collins, G. S. (2019). A robust imputation method for missing responses and covariates in sample selection models. *Statistical Methods in Medical Research*, 28(1), 102–116.

O'Sullivan, F. (1986). A Statistical Perspective on Ill-Posed Inverse Problems. *Statistical Science*, 1(4), 502–518.

Patton, A. J. (2006). Modelling Asymmetric Exchange Rate Dependence. *International Economic Review*, 47(2), 527–556.

Pawitan, Y. (2013). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. OUP Oxford.

Phelps, C. E. (1973). *Demand for Health Insurance: A Theoretical and Empirical Investigation.* PhD thesis, The University of Chicago.

Pigini, C. (2015). Bivariate non-normality in the sample selection model. *Journal of Econometric Methods*, 4(1), 123–144.

Pohlmeier, W. & Ulrich, V. (1995). An econometric model of the two-part decisionmaking process in the demand for health care. *Journal of Human Resources*, (pp. 339–361).

Prieger, J. E. (2002). A flexible parametric selection model for non-normal data with application to health care usage. *Journal of applied econometrics*, 17(4), 367–392.

Puhani, P. A. (2000). The Heckman Correction for Sample Selection and Its Critique. *Journal of Economic Surveys*, 14(1), 53.

Radice, R., Marra, G., & Wojtyś, M. (2016). Copula regression spline models for binary outcomes. *Statistics and Computing*, 26(5), 981–995.

Rigby, R., Stasinopoulos, M., Heller, G., & De Bastiani, F. (2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. Chapman & Hall/CRC The R Series. CRC Press.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 54(3), 507–554.

Robert, C. & Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York.

Robins, J. M. & Wang, N. (2000). Inference for Imputation Estimators. *Biometrika*, 87(1), 113–124.

Roy, A. D. (1951). Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*, 3(2), 135–146.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.

Rubin, D. B. (1977). *The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys*. Technical report, U.S. Social Security Administration.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1 (pp. 20–34).: American Statistical Association.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library. Wiley.

Rubin, D. B. (1991). EM and beyond. *Psychometrika*, 56(2), 241–254.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434), 473–489.

Ruppert, D., Wand, M., & Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Ruud, P. A. (1991). Extensions of estimation methods using the EM algorithm. *Journal of Econometrics*, 49(3), 305–341.

Sakamoto, A. & Chen, M. D. (1991). Inequality and attainment in a dual labor market. *American Sociological Review*, (pp. 295–308).

Sales, A. E., Plomondon, M. E., Magid, D. J., Spertus, J. A., & Rumsfeld, J. S. (2004). Assessing response bias from missing quality of life data: the Heckman method. *Health and quality of life outcomes*, 2(1), 1–10.

Sant'Anna, P. H. & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1), 101–122.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Schafer, J. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1), 3–15.

Silverman, B. W. (1985). Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(1), 1–52.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Slama, R. & Werwatz, A. (2005). Controlling for continuous confounding factors: non-and semi-parametric approaches. *Revue d'epidemiologie et de sante publique*, 53, 65–80.

Smith, M. D. (2003). Modelling sample selection using Archimedean copulas. *Econometrics Journal*, 6(1), 99–123.

Smith, M. D. (2005). Using Copulas to Model Switching Regimes with an Application to Child Labour. *Economic Record*, 81, S47–S57.

Stadlmann, S. & Kneib, T. (2022). Interactively visualizing distributional regression models with distreg.vis. *Statistical Modelling*, 22(6), 527–545.

Stasinopoulos, M., Rigby, R., Heller, G., Voudouris, V., & De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. Chapman & Hall/CRC The R Series. CRC Press.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 339(b2393).

Stuart, E. A., Marcus, S. M., Horvitz-Lennon, M. V., Gibbons, R. D., Normand, S.-L. T., & Hendricks, B. C. (2009). Using Non-Experimental Data to Estimate Treatment Effects. *Psychiatric Annals*, 39(7), 719–728.

Tanner, M. A. & Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.

Toomet, O. & Henningsen, A. (2008a). Sample Selection Models in R: Package sampleSelection. *Journal of Statistical Software*, 27(7).

Toomet, O. & Henningsen, A. (2008b). Sample selection models in R: Package sampleSelection. *Journal of Statistical Software*, 27(7).

Trivedi, P. & Zimmer, D. (2007). *Copula Modeling: An Introduction for Practitioners*. Foundations and trends in econometrics. Now.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3), 219–242.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

van Buuren, S., Boshuizen, H., & Knook, D. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6), 681–694.

van der Klaauw, B. & Koning, R. (2003). Testing the normality assumption in the sample selection model with an application to travel demand. *Journal of Business and Economic Statistics*, 21(1), 31–42.

Vatter, T. & Chavez-Demoulin, V. (2015). Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141, 147–167.

Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources*, 33(1), 127–169.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J., & Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, 39(6), 1279–1293.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, (pp. 307–333).

Wahba, G. (1978). Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3), 364–372.

Wahba, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation theory III*, 2.

Wahba, G. (1983). Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1), 133–150.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wang, N. & Robins, J. M. (1998). Large-Sample Theory for Parametric Multiple Imputation Procedures. *Biometrika*, 85(4), 935–948.

Wei, G. C. G. & Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411), 699–704.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377–399.

Wiemann, P. F., Klein, N., & Kneib, T. (2022). Correcting for sample selection bias in Bayesian distributional regression models. *Computational Statistics & Data Analysis*, 168, 107382.

Wiesenfarth, M. & Kneib, T. (2010). Bayesian geoadditive sample selection models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(3), 381–404.

Wilde, P. E. & Ranney, C. K. (2000). The monthly food stamp cycle: shopping frequency and food intake decisions in an endogenous switching regression framework. *American Journal of Agricultural Economics*, 82(1), 200–213.

Willis, R. J. & Rosen, S. (1979). Education and self-selection. *Journal of political Economy*, 87(5, Part 2), S7–S36.

Winkelmann, R. (2012). Copula bivariate probit models: with an application to medical expenditures. *Health economics*, 21(12), 1444–1455.

Wojtyś, M. & Marra, G. (2015). Copula based generalized additive models with non-random sample selection. `https://arxiv.org/abs/1508.04070`.

Wojtyś, M., Marra, G., & Radice, R. (2018). Copula based generalized additive models for location, scale and shape with non-random sample selection. *Computational Statistics & Data Analysis*, 127, 1–14.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673–686.

Wood, S. N. (2012). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Wood, S. N. (2020). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.* R package version 1.8-33. `https://CRAN.R-project.org/package=mgcv`.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, 111(516), 1548–1563.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wu, C. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1), 95–103.

Zhang, P. (2003). Multiple Imputation: Theory and Method. *International Statistical Review / Revue Internationale de Statistique*, 71(3), 581–592.

Zimmer, D. (2013). Analysis of Mixed Outcomes in Econometrics: Applications in Health Economics. In *Analysis of Mixed Data: Methods and Applications*. CRC Press.

Appendix A

Complements to Chapter 2

A.1 Derivation of the likelihood function

The likelihood function of the ESR model can be specified as follows: when $y_{1i} = 1$ we observe y_{2i} and the contribution to the likelihood corresponds to

$$\begin{aligned}
\mathbb{P}\left[Y_{1i}^* > 0, Y_{2i}\right] &= f(y_{2i})\mathbb{P}\left[Y_{1i}^* > 0 \mid y_{2i}\right] \\
&= f(\epsilon_{2i})\mathbb{P}\left[\epsilon_{1i} > -\eta_{1i} \mid \epsilon_{2i}\right] \\
&= \frac{1}{\sigma_2}\phi\left(\frac{y_{2i}-\eta_{2i}}{\sigma_2}\right)\int_{-\eta_{1i}}^{+\infty} f(\epsilon_{1i}|\epsilon_{2i})\mathrm{d}\epsilon_{1i} \\
&= \frac{1}{\sigma_2}\phi\left(\frac{y_{2i}-\eta_{2i}}{\sigma_2}\right)\int_{-\eta_{1i}}^{+\infty} \frac{1}{\sqrt{1-\rho_{12}^2}}\phi\left(\frac{\epsilon_{1i}-\rho_{12}(y_{2i}-\eta_{2i})/\sigma_2}{\sqrt{1-\rho_{12}^2}}\right)\mathrm{d}\epsilon_{1i} \\
&= \frac{1}{\sigma_2}\phi\left(\frac{y_{2i}-\eta_{2i}}{\sigma_2}\right)\Phi\left(\frac{\eta_{1i}+\rho_{12}\left(y_{2i}-\eta_{2i}\right)/\sigma_2}{\sqrt{1-\rho_{12}^2}}\right).
\end{aligned}$$

Similarly, when $y_{1i} = 0$ we observe y_{3i}, and the contribution to the likelihood corresponds to

$$\begin{aligned}
\mathbb{P}\left[Y_{1i}^* \leq 0, Y_{3i}\right] &= f(y_{3i})\mathbb{P}\left[Y_{1i}^* \leq 0 \mid y_{3i}\right] \\
&= f(\epsilon_{3i})\mathbb{P}\left[\epsilon_{1i} \leq -\eta_{1i} \mid \epsilon_{3i}\right] \\
&= \frac{1}{\sigma_3}\phi\left(\frac{y_{3i}-\eta_{3i}}{\sigma_3}\right)\int_{-\infty}^{-\eta_{1i}} f(\epsilon_{1i}|\epsilon_{3i})\mathrm{d}\epsilon_{1i} \\
&= \frac{1}{\sigma_3}\phi\left(\frac{y_{3i}-\eta_{3i}}{\sigma_3}\right)\int_{-\infty}^{-\eta_{1i}} \frac{1}{\sqrt{1-\rho_{13}^2}}\phi\left(\frac{\epsilon_{1i}-\rho_{13}(y_{3i}-\eta_{3i})/\sigma_3}{\sqrt{1-\rho_{13}^2}}\right)\mathrm{d}\epsilon_{1i} \\
&= \frac{1}{\sigma_3}\phi\left(\frac{y_{3i}-\eta_{3i}}{\sigma_3}\right)\left[1-\Phi\left(\frac{\eta_{1i}+\rho_{13}\left(y_{3i}-\eta_{3i}\right)/\sigma_3}{\sqrt{1-\rho_{13}^2}}\right)\right].
\end{aligned}$$

Combining both results obtains the likelihood function given in equation (2.14).

A.2 Analytical gradient and Hessian of the semi-parametric ESR model

Recall that the semi-parametric ESR model presented in Chapter 2 relaxes the functional form specification of the deterministic model components using semi-parametric additive predictors (in which the effects of continuous covariates are represented via penalized regression splines) while retaining the distributional assumptions of the classical approach, that is, the error terms are assumed to follow a trivariate normal distribution.

The model log-likelihood function is given by

$$\begin{aligned}\ell(\boldsymbol{\beta}) = &\sum_{i=1}^{n} y_{1i} \left\{ -\log \sigma_2 + \log \phi(a_{2i}) + \log \Phi(A_{2i}) \right\} \\ &+ \sum_{i=1}^{n} (1 - y_{1i}) \left\{ -\log \sigma_3 + \log \phi(a_{3i}) + \log \left[\Phi(-A_{3i}) \right] \right\},\end{aligned}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T}, \boldsymbol{\beta}_3^\mathsf{T}, \sigma_2, \sigma_3, \rho_{12}, \rho_{13})^\mathsf{T}$ is the parameter vector, $a_{mi} = \frac{y_{mi} - \eta_{mi}}{\sigma_m}$, $A_{mi} = \frac{\eta_{1i} + \rho_{1m}(y_{mi} - \eta_{mi})/\sigma_m}{\sqrt{1 - \rho_{1m}^2}}$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively, for $m = 2, 3$, and $\eta_{mi} = \mathbf{x}_{mi}^\mathsf{T} \boldsymbol{\beta}_m$ denotes the semi-parametric additive predictor defined in Section 2.2.1, for $m = 1, 2, 3$. The analytical expressions of the gradient and Hessian are derived using the properties of the standard normal distribution and the IMR. In particular, recall that $\phi(\tilde{x}) = \phi(-\tilde{x})$ and $\phi'(\tilde{x}) = -\tilde{x}\phi(\tilde{x})$, and note that the first order derivative of the IMR is given by $\lambda'(\tilde{a}) = -\tilde{a}\lambda(\tilde{a}) - \lambda^2(\tilde{a})$.

The components of the gradient are

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} = \sum_{i=1}^{n} \left\{ y_{1i} \left[\frac{\lambda(A_{2i})}{\sqrt{1 - \rho_{12}^2}} \right] + (1 - y_{1i}) \left[\frac{-\lambda(-A_{3i})}{\sqrt{1 - \rho_{13}^2}} \right] \right\} \mathbf{x}_{1i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} = \sum_{i=1}^{n} y_{1i} \left\{ \frac{a_{2i}}{\sigma_2} - \frac{\rho_{12}\lambda(A_{2i})}{\sigma_2 \sqrt{1 - \rho_{12}^2}} \right\} \mathbf{x}_{2i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3} = \sum_{i=1}^{n} (1 - y_{1i}) \left\{ \frac{a_{3i}}{\sigma_3} + \frac{\rho_{13}\lambda(-A_{3i})}{\sigma_3 \sqrt{1 - \rho_{13}^2}} \right\} \mathbf{x}_{3i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \sigma_2^*} = \sum_{i=1}^{n} y_{1i} \left\{ -1 + a_{2i}^2 - \frac{a_{2i}\rho_{12}\lambda(A_{2i})}{\sqrt{1 - \rho_{12}^2}} \right\},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \sigma_3^*} = \sum_{i=1}^{n} (1 - y_{1i}) \left\{ -1 + a_{3i}^2 + \frac{a_{3i}\rho_{13}\lambda(-A_{3i})}{\sqrt{1-\rho_{13}^2}} \right\},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \rho_{12}^*} = \sum_{i=1}^{n} y_{1i} \frac{\lambda(A_{2i})\left[a_{2i} + \rho_{12}\eta_{1i}\right]}{\sqrt{1-\rho_{12}^2}},$$

and

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \rho_{13}^*} = \sum_{i=1}^{n} (1 - y_{1i}) \frac{-\lambda(-A_{3i})\left[a_{3i} + \rho_{13}\eta_{1i}\right]}{\sqrt{1-\rho_{13}^2}}.$$

The expressions for the non-zero Hessian components are

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^\intercal} = \sum_{i=1}^{n} \left\{ y_{1i} \frac{\lambda'(A_{2i})}{1-\rho_{12}^2} + (1 - y_{1i}) \frac{\lambda'(-A_{3i})}{1-\rho_{13}^2} \right\} \mathbf{x}_{1i}^\intercal \mathbf{x}_{1i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2^\intercal} = \sum_{i=1}^{n} y_{1i} \frac{-\rho_{12}\lambda'(A_{2i})}{\sigma_2(1-\rho_{12}^2)} \mathbf{x}_{1i}^\intercal \mathbf{x}_{2i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_3^\intercal} = \sum_{i=1}^{n} (1 - y_{1i}) \frac{-\rho_{13}\lambda'(-A_{3i})}{\sigma_3(1-\rho_{13}^2)} \mathbf{x}_{1i}^\intercal \mathbf{x}_{3i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \sigma_2^*} = \sum_{i=1}^{n} y_{1i} \frac{-\rho_{12} a_{2i}\lambda'(A_{2i})}{1-\rho_{12}^2} \mathbf{x}_{1i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \sigma_3^*} = \sum_{i=1}^{n} (1 - y_{1i}) \frac{-\rho_{13} a_{3i}\lambda'(-A_{3i})}{1-\rho_{13}^2} \mathbf{x}_{1i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \rho_{12}^*} = \sum_{i=1}^{n} y_{1i} \frac{\lambda'(A_{2i})\left[a_{2i} + \rho_{12}\eta_{1i}\right] + \rho_{12}\sqrt{1-\rho_{12}^2}\lambda(A_{2i})}{1-\rho_{12}} \mathbf{x}_{1i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \rho_{13}^*} = \sum_{i=1}^{n} (1 - y_{1i}) \frac{\lambda'(-A_{3i})\left[a_{3i} + \rho_{13}\eta_{1i}\right] - \rho_{13}\sqrt{1-\rho_{13}^2}\lambda(-A_{3i})}{1-\rho_{12}} \mathbf{x}_{1i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^\intercal} = \sum_{i=1}^{n} y_{1i} \left\{ -\frac{1}{\sigma_2^2} + \frac{\rho_{12}^2 \lambda'(A_{2i})}{\sigma_2^2(1-\rho_{12}^2)} \right\} \mathbf{x}_{2i}^\intercal \mathbf{x}_{2i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2 \partial \sigma_2^*} = \sum_{i=1}^{n} y_{1i} \left\{ -\frac{2a_{2i}}{\sigma_2} + \frac{\rho_{12}^2 a_{2i} \lambda'(A_{2i})}{\sigma_2(1-\rho_{12}^2)} + \frac{\rho_{12}\lambda(A_{2i})}{\sigma_2\sqrt{1-\rho_{12}^2}} \right\} \mathbf{x}_{2i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2 \partial \rho_{12}^*} = \sum_{i=1}^{n} y_{1i} \left\{ \frac{-\lambda(A_{2i}) - \lambda'(A_{2i})\rho_{12}\sqrt{1-\rho_{12}^2}\,[a_{2i} + \rho_{12}\eta_{1i}]}{\sigma_2\sqrt{1-\rho_{12}^2}} \right\} \mathbf{x}_{2i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3 \partial \boldsymbol{\beta}_3^\intercal} = \sum_{i=1}^{n} (1-y_{1i}) \left\{ -\frac{1}{\sigma_3^2} + \frac{\rho_{13}^2 \lambda'(-A_{3i})}{\sigma_3^2(1-\rho_{13})} \right\} \mathbf{x}_{3i}^\intercal \mathbf{x}_{3i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3 \partial \sigma_3^*} = \sum_{i=1}^{n} (1-y_{1i}) \left\{ -\frac{2a_{3i}}{\sigma_3} - \frac{\rho_{13}^2 a_{3i} \lambda'(-A_{3i})}{\sigma_3(1-\rho_{13}^2)} - \frac{\rho_{13}\lambda(-A_{3i})}{\sigma_3\sqrt{1-\rho_{13}^2}} \right\} \mathbf{x}_{3i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3 \partial \rho_{13}^*} = \sum_{i=1}^{n} (1-y_{1i}) \left\{ \frac{\lambda(-A_{3i}) - \lambda'(-A_{3i})\rho_{13}\sqrt{1-\rho_{13}^2}\,[a_{3i} + \rho_{13}\eta_{1i}]}{\sigma_3\sqrt{1-\rho_{13}^2}} \right\} \mathbf{x}_{3i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \sigma_2^* \partial \sigma_2^*} = \sum_{i=1}^{n} y_{1i} \left\{ -2a_{2i}^2 + \frac{\rho_{12} a_{2i} \lambda(A_{2i})}{\sqrt{1-\rho_{12}^2}} + \frac{a_{2i}^2 \rho_{12}^2 \lambda'(A_{2i})}{1-\rho_{12}^2} \right\},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \sigma_2^* \partial \rho_{12}^*} = \sum_{i=1}^{n} -y_{1i} \left\{ \lambda(A_{2i}) + \frac{\lambda'(A_{2i})\rho_{12}[a_{2i} + \rho_{12}\eta_{1i}]}{\sqrt{1-\rho_{12}^2}} \right\} \frac{a_{2i}}{\sigma_2(\sqrt{1-\rho_{12}^2})^3},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \sigma_3^* \partial \sigma_3^*} = \sum_{i=1}^{n} (1-y_{1i}) \left\{ -2a_{3i}^2 - \frac{\rho_{13} a_{3i} \lambda(-A_{3i})}{\sqrt{1-\rho_{13}^2}} - \frac{a_{3i}^2 \rho_{13}^2 \lambda'(-A_{3i})}{1-\rho_{13}^2} \right\},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \sigma_3^* \partial \rho_{13}^*} = \sum_{i=1}^{n} (1-y_{1i}) \left\{ \lambda(-A_{3i}) - \frac{\lambda'(-A_{3i})\rho_{13}[a_{3i} + \rho_{13}\eta_{1i}]}{\sqrt{1-\rho_{13}^2}} \right\} \frac{a_{3i}}{\sigma_3(\sqrt{1-\rho_{13}^2})^3},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \rho_{12}^* \partial \rho_{12}^*} = \sum_{i=1}^{n} y_{1i} \left\{ \lambda'(A_{2i}) \frac{(a_{2i} + \rho_{12}\eta_{1i})^2}{(1-\rho_{12}^2)^3} + \lambda(A_{2i}) \frac{\eta_{1i}(1+\rho_{12}^2) + 3\rho_{12}a_{2i}}{(\sqrt{1-\rho_{12}^2})^5} \right\},$$

and

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \rho_{13}^* \partial \rho_{13}^*} = \sum_{i=1}^{n} (1 - y_{1i}) \left\{ \lambda'(-A_{3i}) \frac{(a_{3i} + \rho_{13}\eta_{1i})^2}{(1-\rho_{13}^2)^3} - \lambda(-A_{3i}) \frac{\eta_{1i}(1+\rho_{13}^2) + 3\rho_{13}a_{3i}}{(\sqrt{1-\rho_{13}^2})^5} \right\}.$$

Appendix B

Complements to Chapter 3

B.1 An overview of copula-based modelling

In this section we outline several results from the copula literature that serve as a general overview of copula-based modelling. For completion, we repeat the definition of a bivariate copula function and the main results in connection with statistical modelling. We then describe two measures of association that summarise the dependence between two random variables and their representation in terms of an underlying copula function. We conclude by reviewing the main characteristics of several copula families. For an in-depth treatment, we refer the interested reader to standard references from which this overview draws from such as Joe (1997) and Nelsen (2006) for a theoretical treatment, and Trivedi & Zimmer (2007) for applications.

A 2-dimensional copula is a bivariate cdf with standard uniform margins, i.e., a function $\mathcal{C}\colon [0,1]^2 \to [0,1]$ defined as

$$\mathcal{C}(u_1, u_2) = \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2), \text{ where } U_1, U_2 \sim \mathcal{U}(0,1),$$

satisfying the following conditions

(C.1) $\mathcal{C}(0,u) = \mathcal{C}(u,0) = 0$, for every $u \in [0,1]$,

(C.2) $\mathcal{C}(1,u) = \mathcal{C}(u,1) = u$, for every $u \in [0,1]$,

(C.3) $\mathcal{C}$ is $2-$increasing.

Any bivariate copula is point-wise bounded by the Fréchet-Hoeffding bounds, defined by

$$\max\Big\{u_1 + u_2 - 1, 0\Big\} \leq \mathcal{C}\left(u_1, u_2\right) \leq \min\{u_1, u_2\}, \text{ for every } (u_1, u_2) \in [0,1]^2.$$

Given the univariate margins, both upper and lower bounds correspond to a bivariate cdf (and therefore a copula) that describes perfect positive dependence and perfect negative dependence, respectively.

The work of Sklar (1959) and Patton (2006) provide the most important results for statistical modelling with copulas. The former obtains a general representation of a multivariate distribution function as a composition of a copula and its univariate margins, while the latter extends Sklar's theorem to the situation where the univariate margins are conditional distribution functions. In a bivariate context, their results state that, given a random vector (Y_1, Y_2) and a set of covariates $\mathbf{z} = (\mathbf{z}_1^\intercal, \mathbf{z}_2^\intercal)^\intercal$ with joint cdf $F(y_1, y_2 \mid \mathbf{z})$ and marginal cdfs $F_1(y_1 \mid \mathbf{z}_1)$ and $F_2(y_2 \mid \mathbf{z}_2)$, there exists a 2-dimensional copula $\mathcal{C}$ such that

$$F(y_1, y_2 \mid \mathbf{z}) = \mathcal{C}\big(F_1(y_1 \mid \mathbf{z}_1), F_2(y_2 \mid \mathbf{z}_2) \mid \delta\big), \quad (y_1, y_2) \in \mathbb{R}^2, \tag{B.1}$$

where δ is a parameter that quantifies the dependence between the margins.

As a consequence of these results, statistical modelling using copula functions can be approached in two steps: first, specify a model for each of the marginal cdf of Y_1 and Y_2 (that do not need to belong to the same family); second, choose an appropriate copula function that links its univariate components together and captures their dependence structure.

Parameter estimation via maximum likelihood requires the derivation of the joint probability density function from the joint cdf given in Equation (B.1), which can be written as follows

$$f(y_1, y_2 \mid \mathbf{z}) = \frac{\partial\, \mathcal{C}\left(F_1(y_1 \mid \mathbf{z}_1), F_2(y_2 \mid \mathbf{z}_2) \mid \delta\right)}{\partial F_1(y_1 \mid \mathbf{z}_1) \partial F_2(y_2 \mid \mathbf{z}_2)} f_1(y_1 \mid \mathbf{z}_1) f_2(y_2 \mid \mathbf{z}_2). \tag{B.2}$$

The literature provides a large number of bivariate copula families. The main properties that make certain copula families appealing for modelling are their interpretability, the range of their dependence parameter, and whether they can be written in closed form. It is also desirable that a copula family can describe independence, and at least one of the Fréchet-Hoeffding bounds, usually through limiting cases of their dependence parameter (see, for example, Joe, 2014, for a review of copula functions in terms of their properties).

For some copula families the interpretation of their dependence parameter is not straightforward and it is useful (and common in practice) to transform it to other measures of dependence such as the Kendall's tau (τ) and/or the Spearman's rho (ρ_s). Both measures depend only on the underlying copula and not on the marginal cdfs, they are symmetric, invariant to increasing trans-

formations of the variables, and their range varies from -1 to 1. A value of 1 indicates perfect positive dependence (the copula coincides with the upper Fréchet bound), whereas a value of -1 implies perfect negative dependence (the copula coincides with the lower Fréchet bound). When the margins are independent, both measures take a value of 0.

The Kendall's τ is defined as the difference between the probabilities of concordance and discordance of the random pairs (Y_1, Y_2) and $(\tilde{Y}_1, \tilde{Y}_2)$ that is, $\tau = \mathbb{P}[(Y_1 - \tilde{Y}_1)(Y_2 - \tilde{Y}_2) > 0] - \mathbb{P}[(Y_1 - \tilde{Y}_1)(Y_2 - \tilde{Y}_2) < 0]$. This measure of dependence can be expressed in terms of a bivariate copula as follows

$$\tau = 4 \int_0^1 \int_0^1 \mathcal{C}(u_1, u_2) \mathrm{d}\mathcal{C}(u_1, u_2) - 1. \tag{B.3}$$

Given a sample of observations $(y_{1i}, y_{2i})_{i=1}^n$, an estimate of the Kendall's τ is computed using

$$\hat{\tau} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i \leq j \leq n} \text{sign}\big((y_{1i} - y_{1j})(y_{2i} - y_{2j})\big),$$

where $\text{sign}(\tilde{d}) = -1$ if $\tilde{d} > 0$, $\text{sign}(\tilde{d}) = 0$ if $\tilde{d} = 0$, and $\text{sign}(\tilde{d}) = 1$ if $\tilde{d} > 0$.

The Spearman's ρ_s is defined by the linear correlation coefficient of the random vector $(F_1(Y_1), F_2(Y_2))$, that is, $\rho_s = \mathsf{Cor}(F_1(Y_1), F_2(Y_2))$. In terms of a bivariate copula, the Spearman's ρ_s can be written as

$$\rho_s = 12 \int_0^1 \int_0^1 \left[\mathcal{C}\left(u_1, u_2\right) - u_1 u_2\right] \mathrm{d}u_1 \mathrm{d}u_2 - 3.$$

Given a sample $(y_{1i}, y_{2i})_{i=1}^n$, an estimate of the Spearman's ρ_s is computed using the rank of the observations, i.e,

$$\hat{\rho}_s = \frac{\sum_{i=1}^n (r_{1i} - \bar{r}_1)(r_{2i} - \bar{r}_2)}{\sqrt{\sum_{i=1}^n (r_{1i} - \bar{r}_1)^2} \sqrt{\sum_{i=1}^n (r_{2i} - \bar{r}_2)^2}},$$

where r_{1i} and r_{2i} represent the ranks, and $\bar{r}_1 = \bar{r}_2 = (n+1)/2$.

From a practical point of view, the Kendall's τ is preferred over the Spearman's ρ_s since the former has explicit analytical expressions based on the copula parameter δ for several copula families.

Lastly, we summarise the main characteristics of several copula families that are frequent in the literature, namely, Ali-Mikhail-Haq (AMH; Ali et al., 1978), Frank (Frank, 1979), Gaussian, Clayton (Clayton, 1978), Gumbel (Gumbel, 1960), and Joe (Joe, 1993). For each of these families, Table B.1 shows the analytical expression, the range of the dependence parameter δ, and the corresponding expression of the Kendall's τ together with its range. In order to illustrate graph-

ically the dependence structure implied by these copula families, Figure B.1 shows their contour density plots based on standard normal univariate margins. The dependence parameters have been set to corresponding values of Kendall's τ similar to those that have been obtained in the empirical applications presented in this thesis.

Copula	$\mathcal{C}(u_1, u_2; \delta)$	Range of δ	Kendall's τ	Range of τ
AMH	$\frac{u_1 u_2}{1-\delta(1-u_1)(1-u_2)}$	$\delta \in [-1, 1]$	$1 - \frac{2}{3\delta^2}\{\delta + (1-\delta)^2 \log(1-\delta)\}$	$-0.181 \leq \tau \leq 1/3$
Clayton	$\left(u_1^{-\delta} + u_2^{-\delta} - 1\right)^{-1/\delta}$	$\delta \in (0, \infty)$	$\frac{\delta}{\delta+2}$	$0 < \tau < 1$
Frank	$-\delta^{-1} \log\left[1 + (e^{-\delta u_1} - 1)(e^{-\delta u_2} - 1)/(e^{-\delta} - 1)\right]$	$\delta \in \mathbb{R}\backslash\{0\}$	$1 - \frac{4}{\delta}\left[1 - D_1(\delta)\right]$	$-1 < \tau < 1$
Gaussian	$\Phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2) \mid \delta\right)$	$\delta \in [-1, 1]$	$\frac{2}{\pi}\sin^{-1}(\delta)$	$-1 \leq \tau \leq 1$
Gumbel	$\exp\left\{-\left[(-\log u_1)^\delta + (-\log u_2)^\delta\right]^{1/\delta}\right\}$	$\delta \in [1, \infty)$	$1 - \frac{1}{\delta}$	$0 \leq \tau < 1$
Joe	$1 - \left[(1-u_1)^\delta + (1-u_2)^\delta - (1-u_1)^\delta(1-u_2)^\delta\right]^{1/\delta}$	$\delta \in [1, \infty)$	$1 + \frac{2}{2-\delta}\left(\psi(2) - \psi(\frac{2}{\delta} + 1)\right)$	$0 < \tau < 1$

Table B.1: Analytical expressions, range of the association parameter δ, and expression and range of the Kendall's τ for several copula families. $D_1(x) = x^{-1}\int_0^x t\left(e^t - 1\right)^{-1} \mathrm{d}t$ is known as the Debye function of order one, and $\psi(\cdot)$ corresponds to the digamma function. The implementation of these copula families and their derivatives is available in the `GJRM` package.

The AMH, Frank, and Gaussian families are symmetric and allow to model positive and negative dependence between their margins. Both Gaussian and Frank copulas achieve the Fréchet-Hoeffding lower and upper bounds, and can describe independence however, the Gaussian copula exhibits a stronger dependence in the tails when compared to the Frank copula. Due to the limited range of the Kendall's tau given by the AMH family, using this copula in applications is usually restricted to situations where the dependence between the margins is modest (Trivedi & Zimmer, 2007). On the other hand, the Clayton, Joe, and Gumbel families are asymmetric and can only model positive dependence due to the restrictions in the range of their association parameters. They all obtain the upper Fréchet bound as $\delta \to \infty$, whereas independence is achieved as their parameter reaches the lower end of its range. The Gumbel and Joe families are similar but the former has a 'thinner' tail. A particular property of these three copula families is that they can be rotated by 90, 180 or 270 degrees to allow modelling data that exhibit negative dependence (Brechmann & Schepsmeier, 2013).

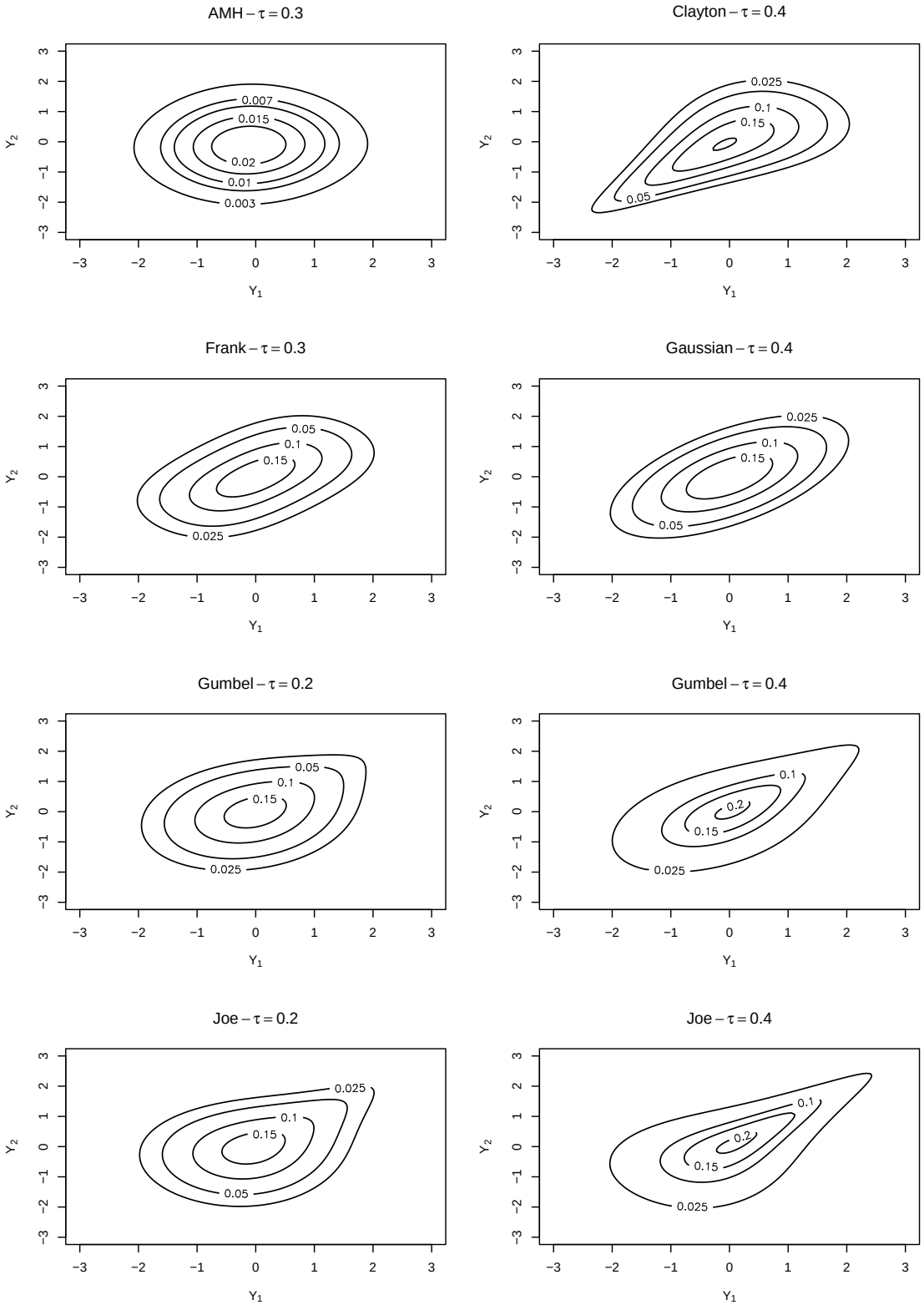

Figure B.1: Contour density plots of the AMH, Clayton, Frank, Gaussian, Gumbel, and Joe copula families based on standard normal margins. The dependence parameters have been set to corresponding values of Kendall's τ similar to those that have been obtained in the empirical applications.

B.2 Analytical gradient and Hessian of the copula-based ESR model

Let us re-write the log-likelihood function of the copula-based ESR model given in (3.10) as follows

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_{1i} \left\{ \log f_2(y_{2i}) + \log \left(1 - h_i^{(12)}\right) \right\} + (1 - y_{1i}) \left\{ \log f_3(y_{3i}) + \log h_i^{(13)} \right\},$$

where

$$h_i^{(1m)} = \frac{\partial \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi})}, \qquad m = 2, 3.$$

Let us also denote $\frac{\partial \eta_{\varpi i}}{\partial \boldsymbol{\beta}_\varpi} = \bar{\mathbf{x}}_{\varpi i}$, that is, the i^{th} row of overall design matrix associated with the additive predictor of each of the distribution parameters where $\varpi \in \{\mu_1, \mu_2, \mu_3, \sigma_2, \sigma_3, \nu_2, \nu_3\}$.

The analytical components of the gradient are given by

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1}} = \sum_{i=1}^{n} \left[y_{1i} \frac{-1}{1 - h_i^{(12)}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_1}} + (1 - y_{1i}) \frac{1}{h_i^{(13)}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_1}} \right] \bar{\mathbf{x}}_{\mu_1 i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2}} = \sum_{i=1}^{n} y_{1i} \left[\frac{1}{f_2(y_{2i})} \frac{\partial f_2(y_{2i})}{\partial \eta_i^{\mu_2}} - \frac{1}{1 - h_i^{(12)}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_2}} \right] \bar{\mathbf{x}}_{\mu_2 i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3}} = \sum_{i=1}^{n} (1 - y_{1i}) \left[\frac{1}{f_3(y_{3i})} \frac{\partial f_3(y_{3i})}{\partial \eta_i^{\mu_3}} + \frac{1}{h_i^{(13)}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_3}} \right] \bar{\mathbf{x}}_{\mu_3 i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_2}} = \sum_{i=1}^{n} y_{1i} \left[\frac{1}{f_2(y_{2i})} \frac{\partial f_2(y_{2i})}{\partial \eta_i^{\sigma_2}} - \frac{1}{1 - h_i^{(12)}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\sigma_2}} \right] \bar{\mathbf{x}}_{\sigma_2 i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_3}} = \sum_{i=1}^{n} (1 - y_{1i}) \left[\frac{1}{f_3(y_{3i})} \frac{\partial f_3(y_{3i})}{\partial \eta_i^{\sigma_3}} + \frac{1}{h_i^{(13)}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\sigma_3}} \right] \bar{\mathbf{x}}_{\sigma_3 i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_2}} = \sum_{i=1}^{n} y_{1i} \left[\frac{1}{f_2(y_{2i})} \frac{\partial f_2(y_{2i})}{\partial \eta_i^{\nu_2}} - \frac{1}{1-h_i^{(12)}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\nu_2}} \right] \bar{\mathbf{x}}_{\nu_2 i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_3}} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{1}{f_3(y_{3i})} \frac{\partial f_3(y_{3i})}{\partial \eta_i^{\nu_3}} + \frac{1}{h_i^{(13)}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\nu_3}} \right] \bar{\mathbf{x}}_{\nu_3 i},$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \delta_{12}} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{1-h_i^{(12)}} \frac{\partial h_i^{(12)}}{\partial \delta_{12}} \right],$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \delta_{13}} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{1}{h_i^{(13)}} \frac{\partial h_i^{(13)}}{\partial \delta_{13}} \right],$$

where, for $\varpi \in \{\mu, \sigma, \nu\}$ and $m = 2, 3$,

$$\frac{\partial h_i^{(1m)}}{\partial \eta_i^{\mu_1}} = \frac{\partial^2 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_1(0) \partial F_m(y_{mi})} \frac{\partial F_1(0)}{\partial \eta_i^{\mu_1}},$$

$$\frac{\partial h_i^{(1m)}}{\partial \eta_i^{\varpi_m}} = \frac{\partial^2 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial F_m(y_{mi})} \frac{\partial F_m(y_{mi})}{\partial \eta_i^{\varpi_m}},$$

and

$$\frac{\partial h_i^{(1m)}}{\partial \delta_{1m}} = \frac{\partial^2 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial \delta_{1m}}.$$

The non-zero components of the Hessian matrix are

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\mu_1}^{\intercal}} = \sum_{i=1}^{n} \left\{ y_{1i} \left[\frac{-1}{(1-h_i^{(12)})^2} \left(\frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_1}} \right)^2 - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\mu_1}} \right] \right.$$
$$\left. + (1-y_{1i}) \left[\frac{-1}{(h_i^{(13)})^2} \left(\frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_1}} \right)^2 + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\mu_1}} \right] \right\} \bar{\mathbf{x}}_{\mu_1 i}^{\intercal} \bar{\mathbf{x}}_{\mu_1 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\mu_2}^\intercal} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(1-h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_2}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_1}} - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\mu_2}} \right] \bar{\mathbf{x}}_{\mu_1 i}^\intercal \bar{\mathbf{x}}_{\mu_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\mu_3}^\intercal} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_3}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_1}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\mu_3}} \right] \bar{\mathbf{x}}_{\mu_1 i}^\intercal \bar{\mathbf{x}}_{\mu_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\sigma_2}^\intercal} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(1-h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\sigma_2}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_1}} - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\sigma_2}} \right] \bar{\mathbf{x}}_{\mu_1 i}^\intercal \bar{\mathbf{x}}_{\sigma_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\sigma_3}^\intercal} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\sigma_3}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_1}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\sigma_3}} \right] \bar{\mathbf{x}}_{\mu_1 i}^\intercal \bar{\mathbf{x}}_{\sigma_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\nu_2}^\intercal} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(1-h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\nu_2}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_1}} - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\nu_2}} \right] \bar{\mathbf{x}}_{\mu_1 i}^\intercal \bar{\mathbf{x}}_{\nu_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \boldsymbol{\beta}_{\nu_3}^\intercal} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\nu_3}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_1}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\nu_3}} \right] \bar{\mathbf{x}}_{\mu_1 i}^\intercal \bar{\mathbf{x}}_{\nu_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \delta_{12}} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(1-h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \delta_{12}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_1}} - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\mu_1} \partial \delta_{12}} \right] \bar{\mathbf{x}}_{\mu_1 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_1} \partial \delta_{13}} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \delta_{13}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_1}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\mu_1} \partial \delta_{13}} \right] \bar{\mathbf{x}}_{\mu_1 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2} \partial \boldsymbol{\beta}_{\mu_2}^\intercal} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(f_2(y_{2i}))^2} \left(\frac{\partial f_2(y_{2i})}{\partial \eta_i^{\mu_2}} \right)^2 + \frac{1}{f_2(y_{2i})} \frac{\partial^2 f_2(y_{2i})}{\partial \eta_i^{\mu_2} \partial \eta_i^{\mu_2}} + \frac{-1}{(1-h_i^{(12)})^2} \left(\frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_2}} \right)^2 - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\mu_2} \partial \eta_i^{\mu_2}} \right] \bar{\mathbf{x}}_{\mu_2 i}^\intercal \bar{\mathbf{x}}_{\mu_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2} \partial \boldsymbol{\beta}_{\sigma_2}^\intercal} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(f_2(y_{2i}))^2} \frac{\partial f_2(y_{2i})}{\partial \eta_i^{\sigma_2}} \frac{\partial f_2(y_{2i})}{\partial \eta_i^{\mu_2}} + \frac{1}{f_2(y_{2i})} \frac{\partial^2 f_2(y_{2i})}{\partial \eta_i^{\mu_2} \partial \eta_i^{\sigma_2}} + \frac{-1}{(1-h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\sigma_2}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_2}} - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\mu_2} \partial \eta_i^{\sigma_2}} \right] \bar{\mathbf{x}}_{\mu_2 i}^\intercal \bar{\mathbf{x}}_{\sigma_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2} \partial \boldsymbol{\beta}_{\nu_2}^\intercal} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(f_2(y_{2i}))^2} \frac{\partial f_2(y_{2i})}{\partial \eta_i^{\nu_2}} \frac{\partial f_2(y_{2i})}{\partial \eta_i^{\mu_2}} + \frac{1}{f_2(y_{2i})} \frac{\partial^2 f_2(y_{2i})}{\partial \eta_i^{\mu_2} \partial \eta_i^{\nu_2}} + \frac{-1}{(1-h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\nu_2}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_2}} - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\mu_2} \partial \eta_i^{\nu_2}} \right] \bar{\mathbf{x}}_{\mu_2 i}^\intercal \bar{\mathbf{x}}_{\nu_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_2} \partial \delta_{12}} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(1-h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \delta_{12}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\mu_2}} - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\mu_2} \partial \delta_{12}} \right] \bar{\mathbf{x}}_{\mu_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3} \partial \boldsymbol{\beta}_{\mu_3}^\intercal} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(f_3(y_{3i}))^2} \left(\frac{\partial f_3(y_{3i})}{\partial \eta_i^{\mu_3}} \right)^2 + \frac{1}{f_3(y_{3i})} \frac{\partial^2 f_3(y_{3i})}{\partial \eta_i^{\mu_3} \partial \eta_i^{\mu_3}} - \frac{1}{(h_i^{(13)})^2} \left(\frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_3}} \right)^2 + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\mu_3} \partial \eta_i^{\mu_3}} \right] \bar{\mathbf{x}}_{\mu_3 i}^\intercal \bar{\mathbf{x}}_{\mu_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3} \partial \boldsymbol{\beta}_{\sigma_3}^\intercal} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(f_3(y_{3i}))^2} \frac{\partial f_3(y_{3i})}{\partial \eta_i^{\sigma_3}} \frac{\partial f_3(y_{3i})}{\partial \eta_i^{\mu_3}} + \frac{1}{f_3(y_{3i})} \frac{\partial^2 f_3(y_{3i})}{\partial \eta_i^{\mu_3} \partial \eta_i^{\sigma_3}} - \frac{1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\sigma_3}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_3}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\mu_3} \partial \eta_i^{\sigma_3}} \right] \bar{\mathbf{x}}_{\mu_3 i}^\intercal \bar{\mathbf{x}}_{\sigma_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3} \partial \boldsymbol{\beta}_{\nu_3}^\intercal} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(f_3(y_{3i}))^2} \frac{\partial f_3(y_{3i})}{\partial \eta_i^{\nu_3}} \frac{\partial f_3(y_{3i})}{\partial \eta_i^{\mu_3}} + \frac{1}{f_3(y_{3i})} \frac{\partial^2 f_3(y_{3i})}{\partial \eta_i^{\mu_3} \partial \eta_i^{\nu_3}} - \frac{1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\nu_3}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_3}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\mu_3} \partial \eta_i^{\nu_3}} \right] \bar{\mathbf{x}}_{\mu_3 i}^\intercal \bar{\mathbf{x}}_{\nu_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mu_3} \partial \delta_{13}} = \sum_{i=1}^{n} (1 - y_{1i}) \left[\frac{-1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \delta_{13}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\mu_3}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\mu_3} \partial \delta_{13}} \right] \bar{\mathbf{x}}_{\mu_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_2} \partial \boldsymbol{\beta}_{\sigma_2}^\intercal} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(f_2(y_{2i}))^2} \left(\frac{\partial f_2(y_{2i})}{\partial \eta_i^{\sigma_2}} \right)^2 + \frac{1}{f_2(y_{2i})} \frac{\partial^2 f_2(y_{2i})}{\partial \eta_i^{\sigma_2} \partial \eta_i^{\sigma_2}} \right.$$
$$\left. + \frac{-1}{(1 - h_i^{(12)})^2} \left(\frac{\partial h_i^{(12)}}{\partial \eta_i^{\sigma_2}} \right)^2 - \frac{1}{1 - h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\sigma_2} \partial \eta_i^{\sigma_2}} \right] \bar{\mathbf{x}}_{\sigma_2 i}^\intercal \bar{\mathbf{x}}_{\sigma_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_2} \partial \boldsymbol{\beta}_{\nu_2}^\intercal} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(f_2(y_{2i}))^2} \frac{\partial f_2(y_{2i})}{\partial \eta_i^{\nu_2}} \frac{\partial f_2(y_{2i})}{\partial \eta_i^{\sigma_2}} + \frac{1}{f_2(y_{2i})} \frac{\partial^2 f_2(y_{2i})}{\partial \eta_i^{\sigma_2} \partial \eta_i^{\nu_2}} \right.$$
$$\left. + \frac{-1}{(1 - h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\nu_2}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\sigma_2}} - \frac{1}{1 - h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\sigma_2} \partial \eta_i^{\nu_2}} \right] \bar{\mathbf{x}}_{\sigma_2 i}^\intercal \bar{\mathbf{x}}_{\nu_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_2} \partial \delta_{12}} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(1 - h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \delta_{12}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\sigma_2}} - \frac{1}{1 - h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\sigma_2} \partial \delta_{12}} \right] \bar{\mathbf{x}}_{\sigma_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_3} \partial \boldsymbol{\beta}_{\sigma_3}^\intercal} = \sum_{i=1}^{n} (1 - y_{1i}) \left[\frac{-1}{(f_3(y_{3i}))^2} \left(\frac{\partial f_3(y_{3i})}{\partial \eta_i^{\sigma_3}} \right)^2 + \frac{1}{f_3(y_{3i})} \frac{\partial^2 f_3(y_{3i})}{\partial \eta_i^{\sigma_3} \partial \eta_i^{\sigma_3}} \right.$$
$$\left. - \frac{1}{(h_i^{(13)})^2} \left(\frac{\partial h_i^{(13)}}{\partial \eta_i^{\sigma_3}} \right)^2 + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\sigma_3} \partial \eta_i^{\sigma_3}} \right] \bar{\mathbf{x}}_{\sigma_3 i}^\intercal \bar{\mathbf{x}}_{\sigma_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_3} \partial \boldsymbol{\beta}_{\nu_3}^\intercal} = \sum_{i=1}^{n} (1 - y_{1i}) \left[\frac{-1}{(f_3(y_{3i}))^2} \frac{\partial f_3(y_{3i})}{\partial \eta_i^{\nu_3}} \frac{\partial f_3(y_{3i})}{\partial \eta_i^{\sigma_3}} + \frac{1}{f_3(y_{3i})} \frac{\partial^2 f_3(y_{3i})}{\partial \eta_i^{\sigma_3} \partial \eta_i^{\nu_3}} \right.$$
$$\left. - \frac{1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\nu_3}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\sigma_3}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\sigma_3} \partial \eta_i^{\nu_3}} \right] \bar{\mathbf{x}}_{\sigma_3 i}^\intercal \bar{\mathbf{x}}_{\nu_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\sigma_3} \partial \delta_{13}} = \sum_{i=1}^{n} (1 - y_{1i}) \left[\frac{-1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \delta_{13}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\sigma_3}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\sigma_3} \partial \delta_{13}} \right] \bar{\mathbf{x}}_{\sigma_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_2} \partial \boldsymbol{\beta}_{\nu_2}^\intercal} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(f_2(y_{2i}))^2} \left(\frac{\partial f_2(y_{2i})}{\partial \eta_i^{\nu_2}} \right)^2 + \frac{1}{f_2(y_{2i})} \frac{\partial^2 f_2(y_{2i})}{\partial \eta_i^{\nu_2} \partial \eta_i^{\nu_2}} + \frac{-1}{(1-h_i^{(12)})^2} \left(\frac{\partial h_i^{(12)}}{\partial \eta_i^{\nu_2}} \right)^2 - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\nu_2} \partial \eta_i^{\nu_2}} \right] \bar{\mathbf{x}}_{\nu_2 i}^\intercal \bar{\mathbf{x}}_{\nu_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_2} \partial \delta_{12}} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(1-h_i^{(12)})^2} \frac{\partial h_i^{(12)}}{\partial \delta_{12}} \frac{\partial h_i^{(12)}}{\partial \eta_i^{\nu_2}} - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \eta_i^{\nu_2} \partial \delta_{12}} \right] \bar{\mathbf{x}}_{\nu_2 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_3} \partial \boldsymbol{\beta}_{\nu_3}^\intercal} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(f_3(y_{3i}))^2} \left(\frac{\partial f_3(y_{3i})}{\partial \eta_i^{\nu_3}} \right)^2 + \frac{1}{f_3(y_{3i})} \frac{\partial^2 f_3(y_{3i})}{\partial \eta_i^{\nu_3} \partial \eta_i^{\nu_3}} - \frac{1}{(h_i^{(13)})^2} \left(\frac{\partial h_i^{(13)}}{\partial \eta_i^{\nu_3}} \right)^2 + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\nu_3} \partial \eta_i^{\nu_3}} \right] \bar{\mathbf{x}}_{\nu_3 i}^\intercal \bar{\mathbf{x}}_{\nu_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\nu_3} \partial \delta_{13}} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(h_i^{(13)})^2} \frac{\partial h_i^{(13)}}{\partial \delta_{13}} \frac{\partial h_i^{(13)}}{\partial \eta_i^{\nu_3}} + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \eta_i^{\nu_3} \partial \delta_{13}} \right] \bar{\mathbf{x}}_{\nu_3 i},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \delta_{12} \partial \delta_{12}} = \sum_{i=1}^{n} y_{1i} \left[\frac{-1}{(1-h_i^{(12)})^2} \left(\frac{\partial h_i^{(12)}}{\partial \delta_{12}} \right)^2 - \frac{1}{1-h_i^{(12)}} \frac{\partial^2 h_i^{(12)}}{\partial \delta_{12} \partial \delta_{12}} \right],$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \delta_{13} \partial \delta_{13}} = \sum_{i=1}^{n} (1-y_{1i}) \left[\frac{-1}{(h_i^{(13)})^2} \left(\frac{\partial h_i^{(13)}}{\partial \delta_{13}} \right)^2 + \frac{1}{h_i^{(13)}} \frac{\partial^2 h_i^{(13)}}{\partial \delta_{13} \partial \delta_{13}} \right],$$

where, for $\varpi \in \{\mu, \sigma, \nu\}$ and $m = 2, 3$,

$$\frac{\partial^2 h_i^{(1m)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\mu_1}} = \frac{\partial^3 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_1(0) \partial F_1(0) \partial F_m(y_{mi})} \left(\frac{\partial F_1(0)}{\partial \eta_i^{\mu_1}} \right)^2 + \frac{\partial^2 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_1(0) \partial F_m(y_{mi})} \frac{\partial^2 F_1(0)}{\partial \eta_i^{\mu_1} \partial \eta_i^{\mu_1}},$$

$$\frac{\partial^2 h_i^{(1m)}}{\partial \eta_i^{\mu_1} \partial \eta_i^{\varpi_m}} = \frac{\partial^3 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_1(0) \partial F_m(y_{mi}) \partial F_m(y_{mi})} \frac{\partial F_m(y_{mi})}{\partial \eta_i^{\varpi_m}} \frac{\partial F_1(0)}{\partial \eta_i^{\mu_1}},$$

$$\frac{\partial^2 h_i^{(1m)}}{\partial \eta_i^{\mu_1} \partial \delta_{1m}} = \frac{\partial^3 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_1(0) \partial F_m(y_{mi}) \partial \delta_{1m}} \frac{\partial F_1(0)}{\partial \eta_i^{\mu_1}},$$

$$\begin{aligned}\frac{\partial^2 h_i^{(1m)}}{\partial \eta_i^{\mu_m} \partial \eta_i^{\varpi_m}} =& \frac{\partial^3 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial F_m(y_{mi}) \partial F_m(y_{mi})} \frac{\partial F_m(y_{mi})}{\partial \eta_i^{\varpi_m}} \frac{\partial F_m(y_{mi})}{\partial \eta_i^{\mu_m}} \\ &+ \frac{\partial^2 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial F_m(y_{mi})} \frac{\partial^2 F_m(y_{mi})}{\partial \eta_i^{\mu_m} \partial \eta_i^{\varpi_m}},\end{aligned}$$

$$\frac{\partial^2 h_i^{(1m)}}{\partial \eta_i^{\varpi_m} \partial \delta_{1m}} = \frac{\partial^3 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial F_m(y_{mi}) \partial \delta_{1m}} \frac{\partial F_m(y_{mi})}{\partial \eta_i^{\varpi_m}}$$

for $\varpi \in \{\sigma, \nu\}$ and $m = 2, 3$,

$$\begin{aligned}\frac{\partial^2 h_i^{(1m)}}{\partial \eta_i^{\sigma_m} \partial \eta_i^{\varpi_m}} =& \frac{\partial^3 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial F_m(y_{mi}) \partial F_m(y_{mi})} \frac{\partial F_m(y_{mi})}{\partial \eta_i^{\varpi_m}} \frac{\partial F_m(y_{mi})}{\partial \eta_i^{\sigma_m}} \\ &+ \frac{\partial^2 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial F_m(y_{mi})} \frac{\partial^2 F_m(y_{mi})}{\partial \eta_i^{\sigma_m} \partial \eta_i^{\varpi_m}}\end{aligned}$$

and for $m = 2, 3$,

$$\begin{aligned}\frac{\partial^2 h_i^{(1m)}}{\partial \eta_i^{\nu_m} \partial \eta_i^{\nu_m}} =& \frac{\partial^3 \mathcal{C}_{12}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial F_m(y_{mi}) \partial F_m(y_{mi})} \frac{\partial F_m(y_{mi})}{\partial \eta_i^{\nu_m}} \frac{\partial F_m(y_{mi})}{\partial \eta_i^{\nu_m}} \\ &+ \frac{\partial^2 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial F_m(y_{mi})} \frac{\partial^2 F_m(y_{mi})}{\partial \eta_i^{\nu_m} \partial \eta_i^{\nu_m}}\end{aligned}$$

$$\frac{\partial^2 h_i^{(1m)}}{\partial \delta_{1m} \partial \delta_{1m}} = \frac{\partial^3 \mathcal{C}_{1m}\left(F_1(0), F_m(y_{mi})\right)}{\partial F_m(y_{mi}) \partial \delta_{1m} \partial \delta_{1m}}.$$

The parametric distributions, copula functions, and their derivatives are implemented in the `GJRM` package.

B.3 Summary of parametric distribution functions

Distribution (support of Y)	$F(y \mid \mu, \sigma, \nu)$	$f(y \mid \mu, \sigma, \nu)$	$\mathsf{E}(Y)$	$\mathsf{Var}(Y)$	Parameters' range
Gamma $(Y > 0)$	$\frac{\gamma(\sigma^{-2}, y\mu^{-1}\sigma^{-2})}{\Gamma(\sigma^{-2})}$	$\frac{y^{1/\sigma^2-1}e^{-y/(\mu\sigma^2)}}{(\sigma^2\mu)^{1/\sigma^2}\Gamma\left(1/\sigma^2\right)}$	μ	$\mu^2\sigma^2$	$\mu, \sigma > 0$
Gumbel $(-\infty < Y < \infty)$	$1 - \exp\left\{-\exp\left(\frac{y-\mu}{\sigma}\right)\right\}$	$\frac{1}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\}$	$\mu - 0.57722\sigma$	$\frac{\pi^2\sigma^2}{6}$	$-\infty < \mu < \infty, \sigma > 0$
Log-normal $(y > 0)$	$\Phi\left(\frac{\log y - \mu}{\sigma}\right)$	$\frac{1}{y\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\{\log y-\mu\}^2}{2\sigma^2}\right]$	$e^{\mu+\sigma^2/2}$	$e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$	$-\infty < \mu < \infty, \sigma > 0$
Normal $(-\infty < Y < \infty)$	$\Phi\left(\frac{y-\mu}{\sigma}\right)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$-\infty < \mu < \infty, \sigma > 0$
Weibull $(Y > 0)$	$1 - \exp\left\{-(y/\mu)^\sigma\right\}$	$\frac{\sigma y^{\sigma-1}}{\mu^\sigma}\exp\left(-(y/\mu)^\sigma\right)$	$\mu\Gamma\left(\frac{1}{\sigma}+1\right)$	$\mu^2\left[\Gamma\left(\frac{2}{\sigma}+1\right) - \left\{\Gamma\left(\frac{1}{\sigma}+1\right)\right\}^2\right]$	$\mu, \sigma > 0$
Singh-Maddala $(Y > 0)$	$1 - \left\{1 + \left(\frac{y}{\mu}\right)^\sigma\right\}^{-\nu}$	$\frac{\sigma\nu y^{\sigma-1}}{\mu^\sigma\left\{1+\left(\frac{y}{\mu}\right)^\sigma\right\}^{\nu+1}}$	$\mu\frac{\Gamma\left(1+\frac{1}{\sigma}\right)\Gamma\left(-\frac{1}{\sigma}+\nu\right)}{\Gamma(\nu)}$	$\mu^2\left\{\Gamma\left(1+\frac{2}{\sigma}\right)\Gamma(\nu)\Gamma\left(-\frac{2}{\sigma}+\nu\right) - \Gamma\left(1+\frac{1}{\sigma}\right)^2\Gamma\left(-\frac{1}{\sigma}+\nu\right)^2\right\}$	$\mu, \sigma, \nu > 0$

Table B.2: Summary of the parametric continuous distribution functions considered in this thesis. The distributions are defined and parametrized as in Rigby et al. (2019) and their implementation is in the `GJRM` package. The expectation and variance of the Singh-Maddala distribution are only defined when $\sigma\nu > 1$ and $\sigma\nu > 2$, respectively. $\Gamma(\cdot)$ denotes the gamma function and $\gamma(\cdot, \cdot)$ the lower incomplete gamma function.