Penalized Likelihood Estimation of Trivariate Additive Binary Models

Doctoral Thesis

Author:

Panagiota Filippou

Supervisors:

Prof. Giampiero Marra

Prof. Rosalba Radice

A thesis submitted for the Degree of Doctor of Philosophy in the

Faculty of Mathematical and Physical Sciences

Department of Statistical Science

University College London

October 2017

Declaration

I, Panagiota Filippou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

To mum, dad

& Filippos

Ἡ Ἰθάκη σ᾽ ἔδωσε τ᾽ ὡραῖο ταξίδι.
Χωρὶς αὐτὴν δὲν θἄβγαινες στὸν δρόμο.
Ἄλλα δὲν ἔχει νὰ σὲ δώσει πιά.

Κι ἂν πτωχικὴ τὴν βρῇς, ἡ Ἰθάκη δὲν σὲ γέλασε.
Ἔτσι σοφὸς ποὺ ἔγινες, μὲ τόση πείρα,
ἤδη θὰ τὸ κατάλαβες ἡ Ἰθάκες τί σημαίνουν.

Κ. Π. Καβάφης, Ἰθάκη

Ithaka gave you the marvelous journey.
Without her you wouldn't have set out.
She has nothing left to give you now.

And if you find her poor, Ithaka won't have fooled you.
Wise as you will have become, so full of experience,
you'll have understood by then what these Ithakas mean.

C. P. Cavafy, 'Ithaka'
Translated by Edmund Keeley

Acknowledgements

As this journey is coming to an end I would like to say a few words of thanks to those who contributed to this piece of work.

Foremost, I would like to thank my supervisors Prof. Giampiero Marra and Prof. Rosalba Radice for the continuous support from day one. They are great academics but more importantly they are great people. It has been a pleasure working with you. Many thanks to Giampiero who believed in me and gave me the opportunity to work on this project. His guidance, care and help enormously contributed to the success of this thesis.

Studying at the Department of Statistical Science at UCL has been a brilliant experience, where I enjoyed good times with Francesco, Raphael, Rui, Robson and Verena. I would like to thank Marianna, Eleftheria, Vassilis, Menelaos and Theodoros for the great lunch and coffee breaks, during which we had discussions about almost everything. A big thanks goes to Marianna and Eleftheria who have helped me in numerous ways during various stages in my PhD. Valentina and Jose, thank you for the nice moments we have shared during your stay in London.

Outside of work, I would like to thank my friends Michalis, Myria, Yiannis, Efi, Eleftheria, Marianna and Vassilis for all those wonderful times that we all shared together the past four years.

I would also like to thank Prof. Thomas Kneib for suggesting the method developed in Chapter 6, and guiding me through its implementation.

Many thanks to my flatmates Elina and Sotia for helping me let go off work for a while with trips, drinks, parties and lots of laughs. You were simply great company! Special thanks to Sotia who is my family in London and a person who strongly supported me through this entire process.

Thanks to my beloved friends Andrea, Lia and Stella for sharing the good times with me when I was going back home. A very special mention goes to my best friend Elena who always found a way to cheer me up during those stressful days and has always been there for me.

I am eternally grateful to my parents and my brother Filippos for their continual love and support throughout my life. Thank you for always believing in me, supporting me and pushing me to follow my dreams. This journey would not have been possible without you.

I am also grateful to UCL and to the EPSRC (Engineering and Physical Sciences Research Council) for the financial support they provided during my PhD studies.

Abstract

In many empirical situations, modelling simultaneously three or more outcomes as well as their dependence structure can be of considerable relevance. Trivariate modelling is continually gaining in popularity (e.g., Genest et al., 2013; Król et al., 2016; Zhong et al., 2012) because of the appealing property to account for the endogeneity issue and non-random sample selection bias, two issues that commonly arise in empirical studies (e.g., Zhang et al., 2015; Radice et al., 2013; Marra et al., 2017; Bärnighausen et al., 2011). The applied and methodological interest in trivariate modelling motivates the current thesis and the aim is to develop and estimate a generalized trivariate binary regression model, which accounts for several types of covariate effects (such as linear, nonlinear, random and spatial effects), as well as error correlations.

In particular, the thesis focuses on the following targets. First, we address the issue in estimating accurately the correlation coefficients, which characterize the dependence of the binary responses conditional on regressors. We found that this is not an unusual occurrence for trivariate binary models and as far as we know such a limitation is neither discussed nor dealt with. Based on this framework, we develop models for dealing with data suffering from endogeneity and/or non-random sample selection. Moreover, we propose trivariate Gaussian copula models where the link functions can in principle be derived from any parametric distribution and the parameters describing the association between the responses can be made dependent on several types of covariate effects. All the coefficients of the model are estimated simultaneously within a penalized likelihood framework based on a carefully structured trust region algorithm with integrated automatic multiple smoothing parameter selection. The developments have been incorporated in the function `SemiParTRIV()`/`gjrm()` in the `R` package `GJRM` (Marra & Radice, 2017). The extensive use of simulated data as well as real datasets illustrates each development in detail and completes the analysis.

Key Words: Trivariate system of equations; Binary responses; Correlation-based

penalty; Penalized regression splines; Unobservables.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Objectives of thesis

In statistical analysis, researchers are often interested in modelling binary responses. When the dependent variable of a regression model is dichotomous instead of continuous, standard estimation techniques like Ordinary Least Squares (OLS) are inefficient and may yield predicted probabilities for y, the response, being equal to 1 lying outside the $[0, 1]$ interval (Aldrich & Nelson, 1984). A popular solution is to redefine the problem by using, for instance, the cumulative distribution function (cdf) of a standard Gaussian $\Phi : \mathbb{R} \to [0, 1]$, in which case the model takes the form $\mathbb{P}(y = 1|\mathbf{x}) = \Phi(\mathbf{x}^\top \boldsymbol{\beta})$, where $\mathbf{x}$ denotes a vector of covariates and $\boldsymbol{\beta}$ is a vector of regression parameters. This is a probit model which belongs to the class of Generalized Linear Models (GLMs, McCullagh & Nelder, 1989). This model can also be written as $y^* = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon$, where y^* is a latent continuous variable and $\varepsilon \sim \mathcal{N}(0, 1)$. In this case, y can be viewed as an indicator variable which is equal to 1 when $y^* > 0$ (or $-\varepsilon < \mathbf{x}^\top \boldsymbol{\beta}$) and 0 otherwise.

While most regression models focus on explaining dependencies between covariates and one single response variable alone, interest in modern statistical applications has recently shifted towards simultaneously studying multiple response variables. The joint modelling is an active area of statistics research that has received a lot of attention in the recent years. The reason for increased interest is that joint

models can be used when we wish to either investigate the role of unobserved variables that may have on an outcome of interest or correct for non-random sample section bias or when we wish to account for the effect of an endogenous covariate. Although these models are used in a wide range of applications in many statistical fields, computational tools for fitting such models are limited due to the fact that are computationally intensive to fit.

This thesis focuses on the development and estimation of flexible trivariate binary models. This is achieved by specifying a trivariate system of non-linear equations where the residual dependence between the responses is characterized through unobserved variables. This representation extends the class of Generalized Additive Models (GAMs, Hastie & Tibshirani, 1990; Wood, 2006) to the multivariate dimension where the functional form of the covariate effects is represented through penalized regression splines.

In particular, the methodology introduced in this thesis is motivated by the modelling of several ways in which unobserved explanatory variables may affect the outcome of interest. Specifically this thesis deals with the omission of common variables in the joint analysis of three responses, unmeasured confounding and non-random sample selection of individuals into (or out) a sample. Several models which build around an extension of this approach are discussed and are demonstrated through simulation studies and/or relevant empirical applications. The proceeding work includes a computationally stable and efficient estimation approach that accurately estimates all the coefficients of the model. All the necessary computational routines are available through the function `SemiParTRIV()`/`gjrm()` in the `R` package `GJRM`.

1.2 Outline

The current thesis is organized in the following way. In Chapter 2 we set up some of the notation that is used throughout the thesis and introduce the trivariate probit model with additive predictors which is our starting point for the development of the proposed methodology. We review some common examples for representing

different types of covariate effects, while a detailed description of the algorithm that is used for fitting trivariate probit models is described. Further, based on some simulation results we acknowledge the difficulty in estimating accurately the correlation coefficients, a problem that commonly arises in trivariate binary models at small or moderate samples. To the best of our knowledge, no research exists discussing or addressing such a limitation. In Chapter 3 we propose a new approach via penalized likelihood for addressing this difficulty, where we provide inferential tools for this framework and illustrate the approach using simulated data. The proposal is illustrated by jointly analysing multiple births, premature birth and low birth weight in North Carolina.

In Chapter 4 we develop several models for dealing with data suffering from endogeneity and/or non-random sample selection. In particular, we deal with these issues by developing three models: (i) the endogenous trivariate model that accounts for two sources of endogeneity; (ii) the double sample selection model which accounts for two layers of sample selection; and (iii) the endogenous-sample selection model that controls for both endogeneity and sample selection simultaneously. An application concerning the effect of two chronic diseases on labour force participation in United States (U.S.) is also discussed.

Chapter 5 extends the models of Chapters 2, 3 and 4 by allowing the link functions to be virtually derived from any parametric distribution. That is, we allow for the use of link functions other than probit. The additional links implemented for this work are the logit and complementary log-log. The representation and estimation of the model is discussed, while a simulation study assessing the performance of the model is also provided.

In Chapter 6, we extend and therefore enhance the trivariate additive binary regression model by allowing the model's association parameters to depend on several types of covariate effects. This extension is of some relevance since it can help to gain insights into the way the residual association between the responses is modified by the presence of covariates. The performance of the method is evaluated in simulations. Furthermore, the flexibility of the model is illustrated in an application

that uses birth data in North Carolina.

Chapter 7 reviews several copula approaches for modelling non-Gaussian error dependence in trivariate additive binary models and outlines the advantages and disadvantages of each method.

A summary of the main results is given in Chapter 8, where we also present some related open topics for further work in the area.

The developments contained in the thesis have been collected in the following papers:

- Filippou P, Kneib T, Marra G, Radice R, A trivarite additive regression model with arbitrary link functions and varying correlation matrix. (*submitted*).
- Filippou P, Marra G, Radice R (2017), Penalized likelihood estimation of a trivariate additive probit model. *Biostatistics*, 18(3), 569–585.

Chapter 2

Penalized likelihood estimation of a trivariate additive probit model

This chapter proposes a penalized likelihood method to estimate a trivariate probit model, which accounts for several types of covariate effects (such as linear, nonlinear, random and spatial effects), as well as error correlations. The parameters of the model are estimated within a penalized likelihood framework based on a carefully structured trust region algorithm with integrated automatic multiple smoothing parameter selection.

2.1 Introduction

Regression models usually involve one response variable and a set of covariates. However, modelling simultaneously more responses in a regression setting can be of considerable empirical relevance. The particular case of trivariate models has been addressed in the literature in various applied and methodological contexts. For example, counties with high rates of pre-term births are more likely to exhibit high rates of low birth weight and this dependence may not be attributed entirely to observed covariates; joint modelling of these responses will yield better calibrated outcome probabilities (Neelon et al., 2014). Loureiro et al. (2010) assessed the effect of parental smoking habits on their children's smoking habits

by estimating a three-equation probit regression model, whereas Kasteridis et al. (2010) employed a trivariate binary-ordered probit model to analyse the demand for cigarettes that identifies non-smokers, potential smokers, quitters and actual smokers. Using a trivariate probit-like approach, Zhong et al. (2012) evaluated the safety of a treatment and identified an optimal dose by jointly modelling the probabilities of toxicity, efficacy, and surrogate efficacy given a specific dose. Król et al. (2016) examined the response to a treatment on patients with metastatic colorectal cancer by analysing simultaneously three types of data: a longitudinal marker, recurrent events, and a terminal event. Rous et al. (2004) discussed a full-information MLE technique, the discrete factor method, to estimate the birth-weight-prenatal care relationship and at the same time to control for the potential biases arising from the selection of the pregnancy-resolution decision and the endogeneity of prenatal care. Zimmer & Trivedi (2006) employed a mixture of powers copula-based approach to model jointly three binary and discrete outcomes. Zhang et al. (2015) developed a Bayesian algorithm to estimate trivariate probit-ordered models affected by double sample selection. Nikoloulopoulos (2015) employed a trivariate copula model for allowing bivariate meta-analysis of diagnostic test accuracy studies to account for disease prevalence.

This chapter is about trivariate probit models which can be traced back to the seminal article by Ashford & Sowden (1970) on multivariate probit models. Chib & Greenberg (1998) later proposed a Bayesian approach for estimating such models. In these works, non-parametric covariates effects are not allowed for. We address this issue by considering trivariate probit models with additive or semi-parametric predictors, hence allowing for several types of covariate effects (such as linear, non-linear, random and spatial effects). This may help uncover interesting structures in the data and reduce the risk and consequences of mis-specifying covariate-response relationships (e.g., Donat & Marra, 2017, and references therein). To implement this advance a reliable estimation algorithm needs to be developed. To this end, we extend to this context the penalized likelihood framework based on a trust region method with automatic smoothing parameter selection developed by Marra

et al. (2017). Such extension relies on the availability of the analytical score and Hessian components of the model's log-likelihood, which are derived in this chapter and represent a contribution in itself. While the analytical score vectors and Hessian matrices are readily available for bivariate binary models, they are not in the multivariate binary context.

This chapter also illustrates the use of `SemiParTRIV()`/`gjrm()` in the package `GJRM` (Marra & Radice, 2017) for the `R` environment (Team, 2017), which implements the advances discussed in this chapter. Current functions for fitting trivariate probit models are `triprobit()` (Terracol, 2002) or `mvprobit()` (Cappellari & Jenkins, 2003) in `STATA` (LP, 2017), and `mvProbit()` in the `R` `mvProbit` package (Henningsen, 2015). These implementations do not deal with the problems that this chapter addresses. Moreover, `mvProbit()` may be unusably slow (as the author points out) and it requires all equations to have the same set of covariates. Note that we have focused on trivariate binary models, however the formulation in Section 2.2 can in principle be extended to the multivariate case as is the proposed estimation framework (see, for instance, the lemma and propositions in Section 2.3).

The remainder of the chapter is organised as follows. Section 2.2 discusses the trivariate probit model with additive or semi-parametric predictors. Section 2.3 provides details on the penalized likelihood estimation algorithm and presents some simulation results whereas the last section concludes the chapter with some discussion.

2.2 Trivariate probit model with flexible covariate effects

Suppose the data consists of n observations on $(\boldsymbol{y}_i, \mathbf{x}_i)_{i=1,\ldots,n}$ with $\boldsymbol{y}_i = (y_{1i}, y_{2i}, y_{3i})^\top$ denoting three correlated binary responses on a single subject i and $\mathbf{x}_i = \text{diag}(\mathbf{x}_{1i}^\top, \mathbf{x}_{2i}^\top, \mathbf{x}_{3i}^\top)$ denoting the $3 \times P$ design matrix, where $\mathbf{x}_{mi}^\top = (1, x_{m2,i}, \ldots, x_{mP_m,i})$, $P = \sum_{m=1}^{3} P_m$ and P_m denotes the number of covariates in each $\mathbf{x}_{mi}$, $\forall m = 1, 2, 3$. Given the set of explanatory variables $\mathbf{x}_i$, the model assumes that the three responses

are observed indicators determined by Gaussian latent continuous variables (as for the univariate case mentioned in the previous chapter). Then, we can define the marginal probability that $y_{mi} = 1$ as $\mathbb{P}(y_{mi} = 1|\mathbf{x}_{mi}) = \Phi(\eta_{mi})$, where in the classic case, $\eta_{mi} = \mathbf{x}_{mi}^{\top}\boldsymbol{\beta}_m$, $\boldsymbol{\beta}_m = (\beta_{m1}, \beta_{m2}, \ldots, \beta_{mP_m})^{\top}$ is a $P_m \times 1$ vector of parameters and $\Phi(\eta_{mi}) = \mathbb{E}(y_{mi}) = \mu_{mi}$ is the mean response for each y_{mi}.

Predictor η_{mi} can be extended to allow for several types of covariate effects. This can be achieved by introducing in η_{mi} some unspecified smooth functions $s_{m\nu_m} : \mathbb{R} \to \mathbb{R}$, $\nu_m = 1, \ldots, \tilde{N}_m$, where $\tilde{N}_m$ is the number of smooth components in the m^{th} equation. As in GAMs, we can therefore write

$$\eta_{mi} = \Phi^{-1}(\mu_{mi}) = \mathbf{v}_{mi}^{\top}\boldsymbol{\gamma}_m + s_{m1}(z_{m1i}) + s_{m2}(z_{m2i}) + \ldots + s_{m\tilde{N}_m}(z_{m\tilde{N}_m i}),$$

where Φ^{-1} is the quantile function of the standard Gaussian and $z_{m\nu_m i}$ is a continuous covariate, $\forall \nu_m, i$. This is essentially a GLM-like linear predictor involving some smooth functions covariates where $\mathbf{x}_{mi}$ is partitioned into two parts: the parametric component which is specified via $\mathbf{v}_{mi}$, with coefficient vector $\boldsymbol{\gamma}_m$, and the non-parametric part which is made up of smooth functions. The combination of these two parts gives raise to an additive or semi-parametric predictor.

Based on the above and using the latent variable formulation of the binary model, we specify the regressions for the responses as

$$y_{mi}^{*} = \mathbf{v}_{mi}^{\top}\boldsymbol{\gamma}_m + \sum_{\nu_m=1}^{\tilde{N}_m} s_{m\nu_m}(z_{m\nu_m i}) + \varepsilon_{mi}, \ \forall m = 1, 2, 3,$$

where $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})^{\top} = \boldsymbol{\varepsilon}_i \overset{iid}{\sim} \mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma})$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \vartheta_{12} & \vartheta_{13} \\ \vartheta_{21} & 1 & \vartheta_{23} \\ \vartheta_{31} & \vartheta_{32} & 1 \end{pmatrix}.$$

The error variances in $\boldsymbol{\Sigma}$ are normalized to unity (e.g., Greene, 2003, pp. 728), while the off-diagonal elements represent the correlations between the error terms

and $\vartheta_{kz} = \vartheta_{zk}$, $\forall z = 1, 2,\ k = 2, 3, z \neq k$.

2.2.1 Smooth function representation

Smooth functions can be specified in several ways; see Ruppert et al. (2003) for details. We opted for the regression spline approach popularized by Eilers & Marx (1996) because of its computational efficiency, theoretical properties and flexibility in representing several types of covariate effects (e.g., Wood, 2006; Yoshida & Naito, 2014). Using this approach, $s_{m\nu_m}(z_{m\nu_m i})$ is approximated by a linear combination of known basis functions $b_{m\nu_m j}(z_{m\nu_m i})$ and regression parameters $\alpha_{m\nu_m j}$. That is,

$$s_{m\nu_m}(z_{m\nu_m i}) \approx \sum_{j=1}^{J_m} \alpha_{m\nu_m,j} b_{m\nu_m,j}(z_{m\nu_m i}) = \mathbf{L}_{m\nu_m}(z_{m\nu_m i})\boldsymbol{\alpha}_{m\nu_m}, \tag{2.1}$$

where $\mathbf{L}_{m\nu_m}(z_{m\nu_m i})$ is a vector containing the J_m basis functions evaluated at $z_{m\nu_m i}$, i.e. $\mathbf{L}_{m\nu_m}(z_{m\nu_m i}) = \{b_{m\nu_m,1}(z_{m\nu_m i}), b_{m\nu_m,2}(z_{m\nu_m i}), \ldots, b_{m\nu_m,J_m}(z_{m\nu_m i})\}$, and $\boldsymbol{\alpha}_{m\nu_m}$ is the corresponding parameter vector defined as $\boldsymbol{\alpha}_{m\nu_m} = (\alpha_{m\nu_m,1}, \alpha_{m\nu_m,2}, \ldots, \alpha_{m\nu_m,J_m})^\top$, $\forall m, \nu_m$. Moreover, each $\boldsymbol{\alpha}_{m\nu_m}$ has an associated quadratic penalty $\lambda_{m\nu_m}\boldsymbol{\alpha}_{m\nu_m}^\top \mathbf{S}_{m\nu_m}\boldsymbol{\alpha}_{m\nu_m}$ which enforces specific properties on the $m\nu_m^{th}$ function, such as smoothness. Smoothing parameter $\lambda_{m\nu_m} \in [0, \infty)$ controls the trade-off between fit and smoothness. The overall penalty can be written as $\boldsymbol{\alpha}^\top \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3)^\top$, $\boldsymbol{\alpha}_m^\top = \left(\boldsymbol{\alpha}_{m1}^\top, \ldots, \boldsymbol{\alpha}_{m\tilde{N}_m}^\top\right)$ $\forall m$, $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{m=1}^{3}\sum_{\nu_m=1}^{\tilde{N}_m} \lambda_{m\nu_m}\mathbf{S}_{m\nu_m}$ and $\mathbf{S}_{m\nu_m}$ are positive definite or semi-definite symmetric known square matrices. Centering constraint $\sum_i s_{m\nu_m}(z_{m\nu_m i}) = 0$ is imposed on all smooth terms in the model for identification purposes; such an approach is applied automatically in estimation via the method discussed in Wood (2006, pp. 165–166). The above formulation allows us to represent many types of covariate effects. This will depend on the nature of the covariate(s) considered and some common examples are described in the next paragraphs. In what follows subscript m is omitted to avoid cluttering the notation.

Non-linear effects For continuous variables the smooth functions are represented using the regression spline approach popularized by Eilers & Marx (1996). B-splines

can be used for this purpose. In general, a spline is a function that is piecewise-defined by polynomial functions which are joint together. The points at which the functions join are known as the knots of the spline. Assume that J denotes the number of spline bases, and thus regression coefficients used to represent $s_\nu(z_\nu)$. To define a J parameter B-spline basis, we first introduce a sequence of $J + D + 1$ knots $z^*_{\nu,1} < z^*_{\nu,2} < \ldots < z^*_{\nu,J+D+1}$, where the spline function is evaluated within the interval $[z^*_{\nu,D+2}, z^*_{\nu,J}]$. The B-basis is strictly local as each basis function is non-zero over the intervals between $D+1$ adjacent knots, where $D+1$ denotes the order of the basis (e.g., $D = 2$ corresponds to the cubic spline). The $(D + 1)^{th}$ order spline can be represented as $\sum_{j=1}^{J} \alpha_{\nu,j} b^D_{\nu,j}(z_\nu)$, where the B-spline basis functions are defined recursively as

$$b^D_{\nu,j}(z_\nu) \;=\; \frac{z_\nu - z^*_{\nu,j}}{z^*_{\nu,j+D+1} - z^*_{\nu,j}} b^{D-1}_{\nu,j}(z_\nu) + \frac{z^*_{\nu,j+D+2} - z^*_\nu}{z^*_{\nu,j+D+2} - z^*_\nu z^*_{\nu,j+1}} b^{D-1}_{\nu,j+1}(z_{\nu i}),$$

and $b^{-1}_{\nu,j}(z_{\nu i}) = 1$ if $z^*_{\nu,j} \leq z_\nu < z^*_{\nu,j+1}$ and 0 otherwise.

Eilers & Marx (1996) developed the penalized B-splines (P-splines) which combine B-spline bases (usually defined on evenly spaced knots) with a difference penalty that is applied to the parameters $\alpha_{\nu,j}$ to control for function's roughness. For example, if one decides to penalize the squared difference between adjacent $\alpha_{\nu,j}$ then the penalty would look like

$$\lambda_\nu \sum_{j=1}^{J-1} (\alpha_{\nu,j+1} - \alpha_{\nu,j})^2 = \lambda_\nu \left\{ \alpha^2_{\nu,1} - 2\alpha_{\nu,1}\alpha_{\nu,2} + 2\alpha^2_{\nu,2} - 2\alpha_{\nu,2}\alpha_{\nu,3} + \ldots + \alpha^2_{\nu,J} \right\},$$

where λ_ν is defined as above. The penalty can equivalently be written as $\lambda_\nu \boldsymbol{\alpha}^\top_\nu \mathbf{S}_\nu \boldsymbol{\alpha}_\nu$, where $\mathbf{S}_\nu$ is defined as

$$\mathbf{S}_\nu \;=\; \begin{pmatrix} 1 & -1 & 0 & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

Thin-plate splines (Duchon, 1977) are an alternative to P-splines. Suppose we wish to estimate $s_\nu(\boldsymbol{z}_{\nu i})$ from n observations $(y_{\nu i}, \boldsymbol{z}_{\nu i})$ such that

$$y_{\nu i} \;\; = \;\; s_\nu(\boldsymbol{z}_{\nu i}) + \varepsilon_{\nu i}$$

where $\boldsymbol{z}_{\nu i}$ is a d-vector with $d \leq n$. Thin plate smoothing estimates $\hat{s}_\nu(\boldsymbol{z}_{\nu i})$ can be obtained by finding a function $f_\nu(\boldsymbol{z}_{\nu i})$ that minimizes

$$\|\boldsymbol{y}_\nu - \boldsymbol{f}_\nu\|^2 + \lambda_\nu \int \ldots \int_{\mathbb{R}^d} \sum_{\upsilon_1 + \ldots + \upsilon_d = D} \frac{D!}{\upsilon_1! \ldots \upsilon_d!} \left(\frac{\partial^D f_\nu}{\partial z_1^{\upsilon_1} \ldots \partial z_d^{\upsilon_d}} \right)^2 dz_1 \ldots dz_d, \quad (2.2)$$

where $\boldsymbol{y}_\nu = (\boldsymbol{y}_{\nu 1}, \ldots, \boldsymbol{y}_{\nu n})^\top$, $\boldsymbol{f}_\nu = (f_\nu(\boldsymbol{z}_{\nu 1}), \ldots, f_\nu(\boldsymbol{z}_{\nu n}))^\top$ and the d-variate integrals correspond to a penalty function that measures the wiggliness of f_ν. For the case of a smooth function of one predictor ($d = 1$) with wiggliness measured using second order derivatives ($D = 2$), expression (2.2) becomes

$$\|\boldsymbol{y}_\nu - \boldsymbol{f}_\nu\|^2 + \lambda_\nu \int_{\mathbb{R}} \left(\frac{\partial^2 f_\nu}{\partial z_1^2} \right)^2 dz_1.$$

By assuming that $2D > d$, then the function that minimizes (2.2) is

$$\hat{f}_\nu(\boldsymbol{z}_\nu) \;\; = \;\; \sum_{i=1}^{n} \alpha_{\nu,i} \bar{\kappa} \left(\|\boldsymbol{z}_\nu - \boldsymbol{z}_{\nu i}\| \right) + \sum_{\bar{q}=1}^{\bar{Q}} \zeta_{\nu,\bar{q}} \tilde{\varphi}_{\nu,\bar{q}}(\boldsymbol{z}_\nu), \quad (2.3)$$

where $\boldsymbol{\alpha}_\nu = (\alpha_{\nu,1}, \ldots, \alpha_{\nu,n})^\top$ and $\boldsymbol{\zeta}_\nu = (\zeta_{\nu,1}, \ldots, \zeta_{\nu,n})^\top$ are coefficient vectors to be estimated. The former is subject to the constraint $\tilde{\mathbf{T}}_\nu^\top \boldsymbol{\alpha}_\nu = \mathbf{0}$ for $\tilde{\mathbf{T}}_{\nu,\bar{q},i} = \tilde{\varphi}_{\nu,\bar{q}}(\boldsymbol{z}_{\nu i})$. Function $\tilde{\varphi}_{\nu,\bar{q}}(\boldsymbol{z}_{\nu i})$ spans the space of functions for which the penalty function is zero, i.e. the null space of the penalty, $\forall \bar{q}$. The exact expression for the basis functions $\bar{\kappa}(\cdot)$ can be found in Wood (2006, pp. 154). Introducing matrix $\mathbf{E}_\nu$ defined by $\mathrm{E}_{\nu,\bar{q},i} \equiv \bar{\kappa}\left(\|\boldsymbol{z}_\nu - \boldsymbol{z}_{\nu i}\|\right)$, the fitting problem becomes

$$\text{minimize}_{\boldsymbol{\alpha}_\nu, \boldsymbol{\zeta}_\nu} \|\boldsymbol{y}_\nu - \mathbf{E}_\nu \boldsymbol{\alpha}_\nu - \tilde{\mathbf{T}}_\nu \boldsymbol{\zeta}_\nu\|^2 + \lambda_\nu \boldsymbol{\alpha}_\nu^\top \mathbf{E}_\nu \boldsymbol{\alpha}_\nu, \;\; \text{subject to} \;\; \tilde{\mathbf{T}}_\nu^\top \boldsymbol{\alpha}_\nu = \mathbf{0}. \quad (2.4)$$

The knots' positions as well as the basis functions do not have to be selected, as

both are defined via the mathematical statement of the smoothing problem. Moreover, thin plate splines can smooth with respect to any number of covariates. A disadvantage of thin-plate splines, however, is computational cost as they require $\mathcal{O}(n^3)$ operations. Wood (2003) addressed this issue by proposing low rank approximations to thin-plate splines. The main idea of Wood's method is to truncate the space of the wiggly components $\boldsymbol{\alpha}_\nu$, while leaving the zero-wiggliness components $\boldsymbol{\zeta}_\nu$ unchanged. Suppose that $\mathbf{E}_\nu = \tilde{\mathbf{C}}_\nu \mathbf{P}_\nu \tilde{\mathbf{C}}_\nu^\top$ is the eigen-decomposition of $\mathbf{E}_\nu$, $\mathbf{P}_\nu$ is a diagonal matrix of eigenvalues of $\mathbf{E}_\nu$ arranged such that $|\mathrm{P}_{\nu,i,i}| \geq |\mathrm{P}_{\nu,i-1,i-1}|$ and the columns of $\tilde{\mathbf{C}}_\nu$ correspond to the eigen-vectors of $\mathbf{E}_\nu$. Let $\tilde{\mathbf{C}}_{\nu,\tilde{s}}$ be a matrix consisting of the first $\tilde{s}$ columns of $\tilde{\mathbf{C}}_\nu$ and $\mathbf{P}_{\nu,\tilde{s}}$ denotes the left $\tilde{s} \times \tilde{s}$ sub-matrix of $\mathbf{P}_\nu$. By expressing $\boldsymbol{\alpha}_\nu = \tilde{\mathbf{C}}_{\nu,\tilde{s}} \boldsymbol{\alpha}_{\nu,\tilde{s}}$, i.e. restricting $\boldsymbol{\alpha}_\nu$ to the column space of $\tilde{\mathbf{C}}_{\nu,\tilde{s}}$, then (2.4) becomes

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\alpha}_{\nu,\tilde{s}},\boldsymbol{\zeta}_\nu} \|\boldsymbol{y}_\nu - \tilde{\mathbf{C}}_{\nu,\tilde{s}} \mathbf{P}_{\nu,\tilde{s}} \boldsymbol{\alpha}_{\nu,\tilde{s}} - \tilde{\mathbf{T}}_\nu \boldsymbol{\zeta}_\nu\|^2 + \lambda_\nu \boldsymbol{\alpha}_{\nu,\tilde{s}}^\top \mathbf{P}_{\nu,\tilde{s}} \boldsymbol{\alpha}_{\nu,\tilde{s}}, \\ &\text{subject to } \ \tilde{\mathbf{T}}_\nu^\top \tilde{\mathbf{C}}_{\nu,\tilde{s}} \boldsymbol{\alpha}_{\nu,\tilde{s}} = \mathbf{0}. \end{aligned}$$

The above constrained problem can be transformed into an unconstrained problem by finding an orthogonal column basis $\tilde{\mathbf{Z}}_{\nu,\tilde{s}}$ such that $\tilde{\mathbf{T}}_\nu^\top \tilde{\mathbf{C}}_{\nu,\tilde{s}} \tilde{\mathbf{Z}}_{\nu,\tilde{s}} = \mathbf{0}$ (which can be done by taking the QR decomposition of $\tilde{\mathbf{C}}_{\nu,\tilde{s}}^\top \tilde{\mathbf{T}}_\nu$) and then restricting $\boldsymbol{\alpha}_{\nu,\tilde{s}}$ to this space, i.e. $\boldsymbol{\alpha}_{\nu,\tilde{s}} = \tilde{\mathbf{Z}}_{\nu,\tilde{s}} \tilde{\boldsymbol{\alpha}}_\nu$. That is,

$$\text{minimize}_{\tilde{\boldsymbol{\alpha}}_{\nu,\tilde{s}},\boldsymbol{\zeta}_\nu} \|\boldsymbol{y}_\nu - \tilde{\mathbf{C}}_{\nu,\tilde{s}} \mathbf{P}_{\nu,\tilde{s}} \tilde{\mathbf{Z}}_{\nu,\tilde{s}} \tilde{\boldsymbol{\alpha}}_\nu - \tilde{\mathbf{T}}_\nu \boldsymbol{\zeta}_\nu\|^2 + \lambda_\nu \tilde{\boldsymbol{\alpha}}_\nu^\top \mathbf{S}_\nu \tilde{\boldsymbol{\alpha}}_\nu,$$

which has a computational cost of $\mathcal{O}(\tilde{s}^3)$, for $\mathbf{S}_\nu = \tilde{\mathbf{Z}}_{\nu,\tilde{s}}^\top \mathbf{P}_{\nu,\tilde{s}}^\top \tilde{\mathbf{Z}}_{\nu,\tilde{s}}$. After fitting the model, the spline can be evaluated via (2.3) using $\boldsymbol{\alpha}_\nu = \tilde{\mathbf{C}}_{\nu,\tilde{s}} \tilde{\mathbf{Z}}_{\nu,\tilde{s}} \tilde{\boldsymbol{\alpha}}_\nu$. The choice of $\tilde{\mathbf{C}}_{\nu,\tilde{s}}$ plays a key role in the approximation method as it makes the minimum possible perturbation to the fitted values of the spline and at the same time makes the minimum possible change to the shape of the fitted spline (Wood, 2003). For more details about splines we refer the reader to Wood (2006, Ch. 4) and references therein.

Linear and Random Effects In general, no penalty is assigned to the parametric part of the model. That is, when $\mathbf{v}_{mi}$ is composed of binary and categorical variables, the entries in the penalty matrix that correspond to these variables are equal to zero. However, if the coefficients of, for instance, some factor variables in the model are weakly or not identified by the data then some penalization on the effects of these variables may be required. This can be achieved by employing, for instance, a Ridge-type penalty (which is made up of a smoothing parameter and an identity penalty matrix). This is equivalent to the assumption that the coefficients of the factor variable are i.i.d. normal random effects with unknown variance (e.g., Ruppert et al., 2003; Wood, 2006).

Spatial Effects To allow the probabilities of the responses to co-vary smoothly across, say, the regions of a country we can include in the model a variable that can exploit the spatial dependence of observations in neighbouring areas. For instance, pre-term births and low birth weights may co-vary smoothly over a country because of environmental influences such as poor air quality and neighbourhood poverty (e.g., Neelon et al., 2014, and references therein). Spatially adjacent regions are also more likely to share similar effects. When a geographic area is divided into discrete contiguous geographic units, the spatial information can be modelled via a Markov random field smoother. In this case, the spatial regional effects can be represented as $s_\nu(z_{\nu i}) = \mathbf{L}_\nu(z_{\nu i})\boldsymbol{\alpha}_\nu$, where $\boldsymbol{\alpha}_\nu = (\alpha_{\nu,1}, \ldots, \alpha_{\nu,\mathfrak{R}})^\top$ denotes the vector of spatial effects, $\mathfrak{R}$ is the total number of regions and $\mathbf{L}_\nu(z_{\nu i})$ is a set of area labels. The $[i, \mathfrak{r}]^{th}$ entry of the corresponding design matrix, that links observation i with the corresponding spatial effect, is equal to 1 if the observation belongs to region $\mathfrak{r}$ and 0 otherwise, $\forall \mathfrak{r} = 1, \ldots, \mathfrak{R}$. Following the assumption that spatially adjacent regions share similar effects, we form the smoothing penalty based on the neighbourhood

structure of the geographic units as

$$\mathbf{S}_\nu[\mathfrak{r},\mathfrak{q}] \;=\; \begin{cases} -1 & \text{if} \quad \mathfrak{r}\neq\mathfrak{q} \wedge \mathfrak{r} \text{ and } \mathfrak{q} \text{ are adjacent neighbors} \\ 0 & \text{if} \quad \mathfrak{r}\neq\mathfrak{q} \wedge \mathfrak{r} \text{ and } \mathfrak{q} \text{ are not adjacent neighbors} \\ K_\mathfrak{r} & \text{if} \quad \mathfrak{r}=\mathfrak{q} \wedge \mathfrak{r}\sim\mathfrak{q} \end{cases},$$

where $K_\mathfrak{r}$ is the total number of neighbours for region $\mathfrak{r}$. In a stochastic interpretation, this penalty is equivalent to the assumption that $\boldsymbol{\alpha}_\nu$ follows a Gaussian Markov random field (e.g., Rue & Held, 2005).

2.2.2 Compact formulation of the model

Using regression spline representation (2.1), we can re-express the model in a more compact way as

$$y^*_{mi} = \mathbf{v}^\top_{mi}\boldsymbol{\gamma}_m + \mathbf{L}^\top_{mi}\boldsymbol{\alpha}_m + \varepsilon_{mi} = \eta_{mi} + \varepsilon_{mi},$$

where $\eta_{mi} = \mathbf{v}^\top_{mi}\boldsymbol{\gamma}_m + \mathbf{L}^\top_{mi}\boldsymbol{\alpha}_m = \left(\mathbf{v}^\top_{mi}, \mathbf{L}^\top_{mi}\right)\left(\boldsymbol{\gamma}_m, \boldsymbol{\alpha}_m\right)^\top = \mathbf{x}^\top_{mi}\boldsymbol{\beta}_m$ and $\mathbf{L}^\top_{mi} = \{\mathbf{L}_{m1}(z_{m1i})^\top, \ldots, \mathbf{L}_{m\tilde{N}_m}(z_{m\tilde{N}_m i})^\top\}$, $\forall m, \nu_m$. After gathering all observations, we define $\mathbf{Y}_{3n\times 1}=(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^\top$, $\mathbf{Y}^*_{3n\times 1} = (\boldsymbol{y}^*_1, \boldsymbol{y}^*_2, \ldots, \boldsymbol{y}^*_n)^\top$, $\mathbf{V}_{3n\times\tilde{P}}=(\mathbf{V}_1; \mathbf{V}_2; \ldots \mathbf{V}_n)^\top$, $\mathbf{L}_{3n\times\tilde{N}} = (\mathbf{L}_1; \mathbf{L}_2; \ldots; \mathbf{L}_n)^\top$ and $\mathbf{X}_{3n\times P}=(\mathbf{X}_1; \mathbf{X}_2; \ldots; \mathbf{X}_n)^\top$. Thus, the trivariate system of equations can be expressed in matrix notation as follows

$$\mathbf{Y}^* = \mathbf{V}\boldsymbol{\gamma} + \mathbf{L}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} = \begin{bmatrix}\mathbf{V} & \mathbf{L}\end{bmatrix}\begin{bmatrix}\boldsymbol{\gamma} & \boldsymbol{\alpha}\end{bmatrix}^\top + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.5}$$

where $\mathbf{X} = \begin{bmatrix}\mathbf{V} & \mathbf{L}\end{bmatrix}$ is a block matrix, with corresponding parameter vector $\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\gamma} & \boldsymbol{\alpha}\end{bmatrix}^\top$, the error term is defined as $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^\top$, and $\tilde{P}$ and $\tilde{N}$ denote the total number of parametric and non-parametric components respectively. For each observation, variable $\boldsymbol{y}^*_i$ is distributed as $\mathcal{N}_3(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$ or equivalently as $\mathbf{Y}^* \sim \mathcal{N}_{3n}(\mathbf{X}\boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}})$ where $\mathbf{X}\boldsymbol{\beta}$ is a $3n \times 1$ vector and $\tilde{\boldsymbol{\Sigma}}$ denotes the $3n \times 3n$ covariance

block diagonal matrix

$$\tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \boldsymbol{\Sigma}_{nn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \boldsymbol{\Sigma} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{11} = \ldots = \boldsymbol{\Sigma}_{nn} = \boldsymbol{\Sigma}$ as all observations are assumed to follow the same covariance structure. This means that the n components of $\boldsymbol{\varepsilon}$ are mutually independent, which in turn means that the off-diagonals $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{13} = \ldots = \boldsymbol{\Sigma}_{n-1,n} = \mathbf{0}$. It could be the case that $\boldsymbol{\Sigma}_{11} \neq \ldots \neq \boldsymbol{\Sigma}_{nn}$ and $\boldsymbol{\Sigma}_{12} \neq \boldsymbol{\Sigma}_{13} \neq \ldots \neq \boldsymbol{\Sigma}_{n-1,n} \neq \mathbf{0}$, for instance. This is beyond the scope of this work and the feasibility of such an extension will be addressed in future research.

2.3 Parameter estimation

Because of the presence of flexible additive predictors in model (2.5), classical MLE is not appropriate for parameter estimation as over-fitting is likely to occur in practical situations. This issue is overcome by adopting a penalized approach where a penalty term, controlling for the model's smoothness, is added to the original objective function. Simultaneous estimation of all parameters of the trivariate additive probit model is therefore achieved by penalized MLE (PMLE) through problem

$$\hat{\boldsymbol{\delta}} \;:=\; \arg\min_{\boldsymbol{\delta}} -\ell_p(\boldsymbol{\delta}) = \arg\min_{\boldsymbol{\delta}} -\{\log \mathcal{L}(\mathbf{Y};\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\alpha}\}, \tag{2.6}$$

where $\boldsymbol{\delta} = (\boldsymbol{\beta}^\top, \boldsymbol{\vartheta}^\top)^\top$, $\boldsymbol{\vartheta} = (\vartheta_{12}, \vartheta_{13}, \vartheta_{23})^\top$, $\mathbf{S}_{\boldsymbol{\lambda}}$ is defined in Section 2.2.1, $\boldsymbol{\alpha}^\top \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\alpha} = \boldsymbol{\delta}^\top \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\delta}$ with $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}} = \text{diag}\left(\mathbf{0}^\top_{\tilde{P}_1}, \lambda_{1\nu_1}\mathbf{S}_{1\nu_1}, \ldots, \lambda_{1\tilde{N}_1}\mathbf{S}_{1\tilde{N}_1}, \mathbf{0}^\top_{\tilde{P}_2}, \lambda_{2\nu_2}\mathbf{S}_{2\nu_2}, \ldots, \lambda_{2\tilde{N}_2}\mathbf{S}_{2\tilde{N}_2}, \mathbf{0}^\top_{\tilde{P}_3}, \lambda_{3\nu_3}\mathbf{S}_{3\nu_3}, \ldots, \lambda_{3\tilde{N}_3}\mathbf{S}_{3\tilde{N}_3}, 0, 0, 0\right)$, $\mathbf{0}^\top_{\tilde{P}_m} = \left(0_{m1}, \ldots, 0_{m\tilde{P}_m}\right)$ and $\tilde{P}_m$ denotes the number of parametric components in the m^{th} equation, $\forall m$. For a 3-D binary response vector we have 2^3 trivariate probabilities expressed via the cdf of the trivariate normal

distribution. The likelihood is given by the joint density of observed outcomes

$$\mathcal{L}(\mathbf{Y};\boldsymbol{\delta}) = \prod_{i=1}^{n}\prod_{\tilde{k}=1}^{2^3}\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i;\boldsymbol{\delta}) = \prod_{i=1}^{n}\prod_{\tilde{k}=1}^{2^3}\boldsymbol{\Psi}_{i\tilde{k}}^{\mathcal{Y}_{i\tilde{k}}},$$

where $\mathcal{L}_{i\tilde{k}}$, is derived from Lemma 2.3.1 for $M = 3$. Term $\mathcal{Y}_{i\tilde{k}}$ denotes an indicator variable for the $\tilde{k}^{th}$ combination of the three possible events $y_{1i} = \bar{e}_1, y_{2i} = \bar{e}_2, y_{3i} = \bar{e}_3$ with $\bar{e}_m \in \{0, 1\}$ $\forall m$ and $\boldsymbol{\Psi}_{i\tilde{k}}$ is the corresponding trivariate normal cdf. For instance, if $\tilde{k} = 3$ corresponds to events $y_{1i} = y_{3i} = 1$ and $y_{2i} = 0$ then $\mathcal{Y}_{i3} = y_{1i}(1 - y_{2i})y_{3i}$ and $\boldsymbol{\Psi}_{i3} = \mathbb{P}(y_{1i} = 1, y_{2i} = 0, y_{3i} = 1)$.

Lemma 2.3.1. *Quantity $\mathcal{L}_{i\tilde{k}}$ evaluated at the vector $(\boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i)_{\tilde{k}}$ is equal to the cdf of a multivariate standardized normal vector with correlation matrix $(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)_{\tilde{k}}$, that is*

$$\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i;\boldsymbol{\delta}) = \boldsymbol{\Psi}_{i\tilde{k}}^{\mathcal{Y}_{i\tilde{k}}} = \{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}} = \{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{w}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\Upsilon}_i)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}},$$

where $\boldsymbol{w}_i = \boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i = (w_{1,i}, w_{2,i}, \ldots, w_{M,i})^\top$, $\boldsymbol{\Upsilon}_i = \boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i$, $w_{m,i} = \tilde{y}_{mi}\eta_{mi}$, for $\tilde{y}_{mi} = (2y_{mi} - 1)$, $\eta_{mi} = \boldsymbol{x}_{mi}^\top\boldsymbol{\beta}_m$, $\boldsymbol{\eta}_i = (\eta_{1i}, \eta_{2i}, \ldots, \eta_{Mi})^\top$ and $\boldsymbol{\mathcal{B}}_i$ denotes a diagonal $M \times M$ matrix with main diagonal elements $\tilde{y}_{mi} = (2y_{mi} - 1)$, that is $\boldsymbol{\mathcal{B}}_i = \text{diag}(2y_{1i} - 1, 2y_{2i} - 1, \ldots, 2y_{Mi} - 1)$.

Proof. See Appendix A.1. □

We can therefore express the log-likelihood function for model (2.5) as

$$\log\mathcal{L}(\mathbf{Y};\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) = \sum_{i=1}^{n}\sum_{\tilde{k}=1}^{2^3}\ell_{i\tilde{k}}(\boldsymbol{\delta}) = \sum_{i=1}^{n}\sum_{\tilde{k}=1}^{4}\left\{\mathcal{Y}_{i\tilde{k}}\log\boldsymbol{\Psi}_{i\tilde{k}} + \mathcal{Y}_{i(4+\tilde{k})}\log\boldsymbol{\Psi}_{i(4+\tilde{k})}\right\},$$

where $\boldsymbol{\Psi}_{i\tilde{k}} = \Phi_{3,\boldsymbol{\varepsilon}_i}((\boldsymbol{w}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\Upsilon}_i)_{\tilde{k}})$, $\boldsymbol{\Psi}_{i(4+\tilde{k})} = \Phi_{3,\boldsymbol{\varepsilon}_i}(-(\boldsymbol{w}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\Upsilon}_i)_{\tilde{k}})$, $\Phi_{3,\boldsymbol{\varepsilon}_i}$ corresponds to trivariate normal integrals, and $\boldsymbol{w}_i$ and $\boldsymbol{\Upsilon}_i$ are defined in Lemma 2.3.1. Note that for each $\tilde{k}$ the form of $\boldsymbol{w}_i$ and $\boldsymbol{\Upsilon}_i$ is different as their structure depends on the $\tilde{k}^{th}$ combination of the three possible events. In general, there are no exact methods for calculating the multivariate normal (MVN) probabilities $\Phi_{M,\boldsymbol{\varepsilon}_i}$, for $M \geq 2$. Accurate approximations, however, can be obtained via `ghkvec()`

in `bayesm` (Rossi, 2015), `pCopula()` in `copula` (Marius Hofert & Yan, 2017) and `pmnorm()` in `mnormt` (Azzalini, 2016), all implemented in the `R` environment. We adopted the latter approach as it was found to be more efficient than the former ones. Function `pmnorm()` evaluates the multivariate integrals by making a suitable call to function `sadmvn()`, a subroutine in `Fortran-77`. The problem is first defined in its general form and then a multivariate integration technique (based on a sequence of three transformations) is applied which simplifies the problem and places it into a form that allows for efficient calculation using standard numerical multiple integration algorithms (Genz, 1992). Appendix A.2.1 provides a detailed description of the algorithm for the reader's convenience. Although accurate results can be obtained via `pmnorm()`, computing time can become burdensome as n increases. As pointed out by Connors et al. (2014), who compared several approximation techniques, there is interest in lower-cost approximation approaches for computing MVN integrals. A possibility would be to employ the method by Trinh & Genz (2015) which consists of writing the MVN probabilities as the product of bivariate conditional probabilities. As compared to `pmnorm()`, this approach gains computational speed but becomes less accurate for highly correlated responses. The full description of the algorithm can be found in Appendix A.2.2; this has been implemented in `SemiParTRIV()`/`gjrm()`. Once $\mathbb{P}(y_{1i}=1, y_{2i}=1, y_{3i}=1)$ has been obtained for all observations, the remaining probabilities can be efficiently calculated using relationship $\sum_{i=1}^{n}\{p_{111i}+p_{110i}+p_{101i}+p_{011i}+p_{000i}+p_{001i}+p_{010i}+p_{100i}\} = \sum_{i=1}^{n}\{p_{11i}+p_{10i}+p_{01i}+p_{00i}\} = \sum_{i=1}^{n}\{p_{1i}+p_{0i}\} = 1$, where $p_{\bar{e}_1\bar{e}_2\bar{e}_3 i} = \mathbb{P}(y_{1i}=\bar{e}_1, y_{2i}=\bar{e}_2, y_{3i}=\bar{e}_3)$, $p_{\bar{e}_1\bar{e}_2 i} = \mathbb{P}(y_{1i}=\bar{e}_1, y_{2i}=\bar{e}_2)$ and $p_{\bar{e}_1 i} = \mathbb{P}(y_{1i}=\bar{e}_1)$. For example, $\mathbb{P}(y_{1i}=1, y_{2i}=1, y_{3i}=0)$ can be computed as $p_{110i} = p_{11i} - p_{111i}$ and $\mathbb{P}(y_{1i}=1, y_{2i}=0, y_{3i}=0)$ as $p_{100i} = p_{1i} - p_{11i} - p_{101i}$.

Restrictions on the correlation parameters The model requires the inclusion of two types of restrictions on the correlation parameters. First, because $\vartheta_{zk} \in [-1, 1]$ we use Fisher transformation $\vartheta^*_{zk} = \tanh^{-1}(\vartheta_{zk})$ and redefine parameter vector $\boldsymbol{\delta}$ as $(\boldsymbol{\beta}^\top, \boldsymbol{\vartheta}^{*\top})^\top$. This is convenient as it ensures that in optimization $\boldsymbol{\delta} \in \mathbb{R}^Q$,

where $\boldsymbol{\vartheta}^* = (\vartheta^*_{12}, \vartheta^*_{13}, \vartheta^*_{23})^\top$ and Q is the total number of parameters in $\boldsymbol{\delta}$. Second, when dealing with correlation matrices, the inclusion of range restrictions on their parameters is needed in order to ensure positive-definiteness. Such constraints have been discussed in the previous literature: a proof on this was first given by Stanley & Wang (1969), while novel geometric proofs were provided by Glass & Collins (1970) and Leung & Lam (1975). Based on the property of positive-definiteness of correlation matrices, Hubert (1972) also provided a proof for the bounds. For a trivariate distribution if two correlations are fixed then the remaining one should be restricted. That is, if ϑ_{13} and ϑ_{23} are known then ϑ_{12} is restricted as follows

$$\vartheta_{13}\vartheta_{23} - \sqrt{(1-\vartheta^2_{13})(1-\vartheta^2_{23})} < \vartheta_{12} < \vartheta_{13}\vartheta_{23} + \sqrt{(1-\vartheta^2_{13})(1-\vartheta^2_{23})}. \tag{2.7}$$

By doing so, the correlation matrix space is a subset of the hyper-cube $[-1,1]^3$. The geometric proof of (2.7) is provided in Appendix A.3 for the reader's convenience. We impose the above restriction using the eigenvalue method. Specifically, assume that a positive-definite correlation matrix $(\boldsymbol{\Upsilon}_i)_{\tilde{k}}$ is expressed as $(\boldsymbol{\Upsilon}_i)_{\tilde{k}} = \bar{\mathbf{P}}\bar{\mathbf{D}}\bar{\mathbf{P}}^\top, \forall \tilde{k}$, where $\bar{\mathbf{D}}$ is a diagonal matrix containing the eigenvalues of $(\boldsymbol{\Upsilon}_i)_{\tilde{k}}$ and $\bar{\mathbf{P}}$ is an orthogonal matrix of corresponding eigenvectors. When $(\boldsymbol{\Upsilon}_i)_{\tilde{k}}$ is not positive-definite, some eigenvalues are negative and typically not large in absolute sense. According to Rousseeuw & Molenberghs (1993), a common approach for transforming a non-positive-definite matrix into a positive-definite one is to replace the negative eigenvalues by their absolute values; that is re-express $(\boldsymbol{\Upsilon}_i)_{\tilde{k}}$ as $(\boldsymbol{\Upsilon}_i)'_{\tilde{k}} = \bar{\mathbf{P}}\mathbf{D}'\bar{\mathbf{P}}^\top$ where $\mathbf{D}'$ contains positive eigenvalues. The diagonal elements of $(\boldsymbol{\Upsilon}_i)'_{\tilde{k}}$ will not necessarily be equal to 1. To this end, we transform $(\boldsymbol{\Upsilon}_i)'_{\tilde{k}}$ to $(\tilde{\boldsymbol{\Upsilon}}_i)_{\tilde{k}} = \tilde{\mathbf{D}}'(\boldsymbol{\Upsilon}_i)'_{\tilde{k}}\tilde{\mathbf{D}}'^\top$ where $\tilde{\mathbf{D}}'$ is the diagonal matrix with diagonal elements equal to $1/\sqrt{r'_{mm,i}}$ and $r'_{mm,i}$ denotes the diagonal element of $(\boldsymbol{\Upsilon}_i)'_{\tilde{k}}, \forall \tilde{k}, m$. For more details see Rousseeuw & Molenberghs (1993).

Joint estimation of $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ via (2.6) would clearly lead to severe over-fitting as the optimal value of $\ell_p(\boldsymbol{\delta})$ would be reached when $\hat{\boldsymbol{\lambda}} = \mathbf{0}$ (e.g., Ruppert et al., 2003). Following Gu (2002), Marra et al. (2017) and Wood (2004), we estimate the model

and smoothing parameters using a two stage approach; one step concerns estimation of $\boldsymbol{\delta}$ conditional on $\boldsymbol{\lambda}$ and the other estimation of $\boldsymbol{\lambda}$ conditional on $\boldsymbol{\delta}$. Note that such an approach is philosophically very similar to the Bayesian estimation method discussed, for instance, by Klein & Kneib (2016a) where Bayesian sampling is used to estimate $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ conditional on each other.

2.3.1 Step 1: Estimating $\boldsymbol{\delta}$ given smoothing parameters

Holding $\boldsymbol{\lambda}$ fixed at a vector of values, we seek to minimize $-\ell_p(\boldsymbol{\delta})$. This is achieved via a trust-region algorithm which has generally proved to be more stable and faster than standard numerical optimization procedures when fitting simultaneous systems of equations (e.g., Donat & Marra, 2017; Radice et al., 2016). Each iteration $\varkappa$ of the trust-region algorithm solves the sub-problem

$$\begin{aligned} \min_{\boldsymbol{s}} \mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]}) \quad &:= \quad -\left\{ \ell_p\left(\boldsymbol{\delta}^{[\varkappa]}\right) + \boldsymbol{s}^\top \boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa]}) + \frac{1}{2}\boldsymbol{s}^\top \boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}^{[\varkappa]})\boldsymbol{s} \right\} \\ &\quad \text{subject to } \|\boldsymbol{s}\| \leq \boldsymbol{\Delta}^{[\varkappa]}, \\ \boldsymbol{\delta}^{[\varkappa+1]} \quad &= \quad \arg\min_{\boldsymbol{s}} \mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]}) + \boldsymbol{\delta}^{[\varkappa]}, \end{aligned} \tag{2.8}$$

where $\mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]})$ is a quadratic approximation of ℓ_p at $\boldsymbol{\delta}^{[\varkappa]}$, $\boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa]})$ denotes the penalized score function defined as $\boldsymbol{g}(\boldsymbol{\delta}^{[\varkappa]}) - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}^{[\varkappa]}$, $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}^{[\varkappa]})$, the penalized Hessian matrix, is given by $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}^{[\varkappa]}) - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}$, $\|\cdot\|$ denotes the Euclidean norm and $\boldsymbol{\Delta}^{[\varkappa]}$ is the radius of the trust region. The analytical score function, $\boldsymbol{g}_i(\boldsymbol{\delta}^{[\varkappa]}) = \nabla_{\boldsymbol{\delta}}\ell_i\left(\boldsymbol{\delta}^{[\varkappa]}\right)$, and Hessian matrix, $\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}^{[\varkappa]}) = \nabla_{\boldsymbol{\delta}}\nabla_{\boldsymbol{\delta}}^\top\ell_i\left(\boldsymbol{\delta}^{[\varkappa]}\right)$, required to implement the trust region

approach are computed using

$$\nabla_{\boldsymbol{\delta}}\ell_i(\boldsymbol{\delta}) = \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \frac{\partial\ell_i(\boldsymbol{\delta})}{\partial\bar{\boldsymbol{\eta}}_i} = \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i}\right\}, \tag{2.9}$$

$$\begin{aligned}\nabla_{\boldsymbol{\delta}}\nabla_{\boldsymbol{\delta}}^\top\ell_i(\boldsymbol{\delta}) &= \frac{\partial\ell_i(\boldsymbol{\delta})}{\partial\bar{\boldsymbol{\eta}}_i}\frac{\partial^2\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top} + \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \frac{\partial^2\ell_i(\boldsymbol{\delta})}{\partial\bar{\boldsymbol{\eta}}_i\partial\bar{\boldsymbol{\eta}}_i^\top}\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}} \\ &= \left\{\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i}\right\}\frac{\partial^2\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top} + \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{-\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}\boldsymbol{\Psi}_{i\tilde{k}}^\top}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i}\left(\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i^\top}\right)^\top + \right. \\ &\quad \left.\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial^2\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i\partial\bar{\boldsymbol{\eta}}_i}^\top\right\}\left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right),\end{aligned} \tag{2.10}$$

where, for each i, $\partial\bar{\boldsymbol{\eta}}_i/\partial\boldsymbol{\delta} = \text{diag}\,(\partial\eta_{1i}/\partial\boldsymbol{\beta}_1, \partial\eta_{2i}/\partial\boldsymbol{\beta}_2, \partial\eta_{3i}/\partial\boldsymbol{\beta}_3, \partial\vartheta^*_{12}/\partial\vartheta^*_{12}, \partial\vartheta^*_{13}/\partial\vartheta^*_{13}, \partial\vartheta^*_{23}/\partial\vartheta^*_{23}) = \text{diag}\,(\partial\eta_{1i}/\partial\boldsymbol{\beta}_1, \partial\eta_{2i}/\partial\boldsymbol{\beta}_2, \partial\eta_{3i}/\partial\boldsymbol{\beta}_3\,,\, 1, 1, 1)$ and $\partial\ell(\boldsymbol{\delta})/\partial\bar{\boldsymbol{\eta}}_i = (\partial\ell(\boldsymbol{\delta})/\partial\eta_{1i}\,,$ $\partial\ell(\boldsymbol{\delta})/\partial\eta_{2i}, \partial\ell(\boldsymbol{\delta})/\partial\eta_{3i}\,,\ \partial\ell(\boldsymbol{\delta})/\partial\vartheta^*_{12}, \partial\ell(\boldsymbol{\delta})/\partial\vartheta^*_{13}, \partial\ell(\boldsymbol{\delta})/\partial\vartheta^*_{23})^\top$. Predictor $\bar{\boldsymbol{\eta}}_i$ is functionally dependent on the Q-vector $\boldsymbol{\delta}$, that is $\bar{\boldsymbol{\eta}}_i = \bar{\boldsymbol{\eta}}_i(\boldsymbol{\delta})$, and is defined as $\bar{\boldsymbol{\eta}}_i = (\eta_{1i}, \eta_{2i}, \eta_{3i}, \eta_{4i}, \eta_{5i}, \eta_{6i})^\top$, where $(\eta_{4i}, \eta_{5i}, \eta_{6i}) = (\vartheta^*_{12}, \vartheta^*_{13}, \vartheta^*_{23})$. The difficulty with deriving analytical expressions for the derivative components in (2.9) and (2.10) is that they require working with trivariate integrals, which is not straightforward. This is addressed using the decomposition approach which consists of breaking the trivariate integrals into lower-order integrals which are then solved separately, and a method by Plackett (1954) which is based on the reduction formula which progressively simplifies the integrals until they can be evaluated. Using these techniques, we derive propositions 2.3.2 and 2.3.3 which show that the derivatives of a multivariate normal cdf, Φ_M, with respect to the model parameters require the evaluation of $M-1$ integrals, $\forall M \geq 3$. Specifically, we provide the key derivatives for the log-likelihood function of a generic multivariate probit model with correlation matrix structured as

$$\boldsymbol{\Upsilon}^*_i = \begin{pmatrix} 1 & r^*_{12,i} & r^*_{13,i} & \dots & r^*_{1M,i} \\ r^*_{12,i} & 1 & r^*_{23,i} & \dots & r^*_{2M,i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r^*_{1M,i} & r^*_{2M,i} & r^*_{3M,i} & \dots & 1 \end{pmatrix},$$

where $r^*_{zk,i} = \tanh(\vartheta^*_{zk})(2y_{zi}-1)(2y_{ki}-1)$, $\forall z,k,i$. The propositions below have been used to implement expressions (2.9) and (2.10) after setting $M=3$.

Proposition 2.3.2. *Assume that $\boldsymbol{w}_i$ is a multivariate standardized normal vector with correlation matrix equal to $\boldsymbol{\Upsilon}^*_i$. Then the first-order derivative of the M-variate normal cdf $\Phi_M(\boldsymbol{w_i}; \boldsymbol{0}, \boldsymbol{\Upsilon}^*_i)$ with respect to $\boldsymbol{\beta}_m$, $\forall m = 1,\dots,M$, can be expressed as*

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \boldsymbol{0}, \boldsymbol{\Upsilon}^*_i)}{\partial \boldsymbol{\beta}_m} \;=\; \phi(w_{m,i};0,1)\Phi_{M-1}(\boldsymbol{w}_{-m,i}|w_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i)(2y_{mi}-1)\boldsymbol{x}^\top_{mi},$$

where M denotes the total number of equations under a multivariate probit framework, $w_{m,i}$ denotes the linear predictor of the m^{th} equation and is equal to $(2y_{mi}-1)\boldsymbol{x}^\top_{mi}\boldsymbol{\beta}_m$, $\boldsymbol{\beta}_m$ denotes the parameter vector of covariate vector $\boldsymbol{x}_{mi}$ and the vector of linear predictors $\boldsymbol{w}_{-m,i}$ is defined as $(w_{1,i}, w_{2,i}, \dots, w_{m-1,i},\ w_{m+1,i}, \dots, w_{M,i})^\top$. The mean $\boldsymbol{M}^{*m}_i$ and variance-covariance matrix $\boldsymbol{\Theta}^{*m}_i$ is equal to $\boldsymbol{\Theta}^{*m}_{21,i} w_{m,i}$ and $\boldsymbol{\Theta}^{*m}_{22,i} - \boldsymbol{\Theta}^{*m}_{21,i}\boldsymbol{\Theta}^{*m}_{12,i}$, respectively, with $\boldsymbol{\Theta}^{*m}_{12,i}$, $\boldsymbol{\Theta}^{*m}_{21,i}$ and $\boldsymbol{\Theta}^{*m}_{22,i}$ defined by re-ordering $\boldsymbol{\Upsilon}^*_i$ as follows

$$\boldsymbol{\Upsilon}^{*m}_i = \left(\begin{array}{c|c} \overbrace{\Theta^{*m}_{11,i}}^{1\times 1} & \overbrace{\boldsymbol{\Theta}^{*m}_{12,i}}^{1\times(M-1)} \\ \hline \underbrace{\boldsymbol{\Theta}^{*m}_{21,i}}_{(M-1)\times 1} & \underbrace{\boldsymbol{\Theta}^{*m}_{22,i}}_{(M-1)\times(M-1)} \end{array} \right).$$

The element $\Theta^{*m}_{11,i}$ is equal to 1, the off-diagonal blocks $\boldsymbol{\Theta}^{*m}_{12,i}$ and $\boldsymbol{\Theta}^{*m}_{21,i}$ consist of the correlations $r^*_{m\varpi,i} = \tanh(\vartheta^*_{m\varpi})(2y_m-1)(2y_\varpi-1)$, $\forall\ \varpi \in \{1:M\}\setminus m$, for $m \neq \varpi$ and the symmetric sub-matrix $\boldsymbol{\Theta}^{*m}_{22,i}$ has main diagonal elements equal to 1 and off-diagonals equal to $r^*_{\bar\varphi\varpi,i} = \tanh(\vartheta^*_{\bar\varphi\varpi})(2y_{\bar\varphi}-1)(2y_\varpi-1)$, $\forall \bar\varphi, \varpi \in \{1:M\}\setminus m$, for $\bar\varphi \neq \varpi$.

Proof. See Appendix A.4.1. □

Proposition 2.3.3. *Assume that $\boldsymbol{w}_i$ is a multivariate standardized normal vector with correlation matrix equal to $\boldsymbol{\Upsilon}^*_i$. Then the first-order derivative of the M-variate normal cdf $\Phi_M(\boldsymbol{w_i}; \boldsymbol{0}, \boldsymbol{\Upsilon}^*_i)$ with respect to ϑ^*_{zk}, $\forall z = 1,\dots,M-1, k = z+1,\dots M$,*

can be expressed as

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk}^*} &= \phi_2(\boldsymbol{w}_{zk,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*zk}) \Phi_{M-2}(\boldsymbol{w}_{-zk,i} | \boldsymbol{w}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk}) \times \\ &\quad (2y_{zi}-1)(2y_{ki}-1)\frac{4e^{2\vartheta_{zk}^*}}{(e^{2\vartheta_{zk}^*}+1)^2},\end{aligned}$$

where M denotes the total number of equations under a multivariate probit framework, $\boldsymbol{w}_{zk,i} = (w_{z,i}, w_{k,i})^\top$, $w_{z,i}$ and $w_{k,i}$ refer to the linear predictors of the z^{th} and k^{th} equations respectively and are equal to $(2y_{mi}-1)\boldsymbol{x}_{mi}^\top \boldsymbol{\beta}_m$, $\forall m = z, k$, and $\boldsymbol{\beta}_m$ denotes the parameter vector of covariate vector $\boldsymbol{x}_{mi}$. The vector of linear predictors $\boldsymbol{w}_{-zk,i}$ is defined as $(w_{1,i}, w_{2,i}, \ldots, w_{z-1,i}, \; w_{z+1,i}, \ldots, w_{k-1,i}, w_{k+1,i}, \ldots w_{M,i})^\top$, while parameter $\vartheta_{zk}^* = \tanh^{-1}(\vartheta_{zk})$ where ϑ_{zk} denotes the correlation coefficient between the z^{th} and k^{th} responses. The variance-covariance matrix $\boldsymbol{\Theta}_i^{*zk}$ is equal to $\boldsymbol{\Theta}_{11,i}^{*zk}$, while the mean $\boldsymbol{M}_i^{*-zk}$ and variance-covariance matrix $\boldsymbol{\Theta}_i^{*-zk}$ is equal to $\boldsymbol{\Theta}_{21,i}^{*zk}\left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1}\boldsymbol{w}_{zk}$ and $\boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk}\left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1}\boldsymbol{\Theta}_{12,i}^{*zk}$, respectively. The sub-matrices $\boldsymbol{\Theta}_{11,i}^{*zk}$, $\boldsymbol{\Theta}_{12,i}^{*zk}$, $\boldsymbol{\Theta}_{21,i}^{*zk}$ and $\boldsymbol{\Theta}_{22,i}^{*zk}$ are defined by re-ordering $\boldsymbol{\Upsilon}_i^*$ as follows

$$\boldsymbol{\Upsilon}_i^{*zk} = \left(\begin{array}{c:c} \overbrace{\boldsymbol{\Theta}_{11,i}^{*zk}}^{2\times 2} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*zk}}^{2\times(M-2)} \\ \hdashline \underbrace{\boldsymbol{\Theta}_{21,i}^{*zk}}_{(M-2)\times 2} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*zk}}_{(M-2)\times(M-2)} \end{array}\right).$$

The sub-matrix $\boldsymbol{\Theta}_{11,i}^{*zk}$ has unit diagonals and off-diagonals defined as $r_{zk,i}^* = \tanh(\vartheta_{zk}^*)$ $(2y_z-1)(2y_k-1)$. The first row (column) of $\Theta_{12,i}^{*zk}$ ($\Theta_{21,i}^{*zk}$) contains the correlations $r_{z\bar{\varrho},i}^*$, for $\bar{\varrho} \in \{1:M\} \setminus z$, while the second row (column) of $\Theta_{12,i}^{*zk}$ ($\Theta_{21,i}^{*zk}$) contains the correlations $r_{\bar{\upsilon}k,i}^*$, for $\bar{\upsilon} \in \{1:M\} \setminus k$. The diagonal block $\boldsymbol{\Theta}_{22,i}^{*zk}$ is a symmetric matrix with unit diagonals and off-diagonal elements equal to $r_{\bar{\chi}\bar{\psi},i}^*$, $\forall\ \bar{\chi}, \bar{\psi} \in \{1:M\} \setminus \{z, k\}$ for $\bar{\chi} \neq \bar{\psi}$.

Proof. See Appendix A.4.2. □

The analytical derivatives have been verified via numerical differentiation using the `R` package `numDeriv` (Gilbert & Varadhan, 2016). Full matrices $\boldsymbol{\Upsilon}_i^{*m}$ and $\boldsymbol{\Upsilon}_i^{*zk}$ can be found in Appendix A.5.

Line-search methods compute $\boldsymbol{s}^{[\varkappa]}$ by minimising the unconstrained problem (2.8). The current solution $\boldsymbol{\delta}^{[\varkappa+1]}$ is then updated by scaling the step $\boldsymbol{s}^{[\varkappa]}$ by a factor $\boldsymbol{\tau}^{[\varkappa]}$ that approximately minimizes $-\ell_p(\boldsymbol{\delta})$ along the line that passes through $\boldsymbol{\delta}^{[\varkappa]}$ in the direction of $\boldsymbol{s}^{[\varkappa]}$, $\boldsymbol{\delta}^{[\varkappa+1]} = \boldsymbol{\delta}^{[\varkappa]} + \boldsymbol{\tau}^{[\varkappa]}\boldsymbol{s}^{[\varkappa]}$. If the function is non-convex then the optimizer may search far away from $\boldsymbol{\delta}^{[\varkappa]}$ but still chooses $\boldsymbol{\delta}^{[\varkappa+1]}$ to be close to $\boldsymbol{\delta}^{[\varkappa]}$. In some cases the function will be evaluated so far away from $\boldsymbol{\delta}^{[\varkappa]}$ that it will not be finite and the algorithm will fail. On the contrary, trust-region methods use a maximum distance for the move from $\boldsymbol{\delta}^{[\varkappa]}$ to $\boldsymbol{\delta}^{[\varkappa+1]}$ based on a region $\mathcal{R}^{[\varkappa]}$ around the current iterate $\boldsymbol{\delta}^{[\varkappa]}$ in which the algorithm 'trusts' that model function $\mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]})$ behaves like objective function $\ell_p(\boldsymbol{\delta})$. Current iteration $\boldsymbol{\delta}^{[\varkappa]}$ is updated with $\boldsymbol{s}^{[\varkappa]}$ if this step produces an improvement over the objective function $\ell_p(\boldsymbol{\delta})$, $\boldsymbol{\delta}^{[\varkappa+1]} = \boldsymbol{\delta}^{[\varkappa]} + \boldsymbol{s}^{[\varkappa]}$. Since points outside $\mathcal{R}^{[\varkappa]}$ are not considered, the algorithm never runs too far from the current iteration. The trust-region is shrunken if the proposed point in the region is not better than the current point, in which case the new problem is solved with smaller region. If the quadratic model is a good representation of the original objective function, then trial point $\boldsymbol{\delta}^{[\varkappa+1]}$ becomes the new iterate and the trust-region is enlarged, i.e. the iteration is successful. A detailed description of trust-region and line search techniques can be found in Nocedal & Wright (2006, Chap. 3, 4). The trust-region algorithm is summarised in Algorithm 1.

2.3.2 Step 2: Estimating $\boldsymbol{\lambda}$

There are several ways for estimating automatically multiple smoothing parameters (e.g., Wood, 2004, 2008, 2011; Radice et al., 2016; Marra et al., 2017). One way is to minimise a mean squared error criterion which can be shown to be equivalent to an approximate Akaike Information Criterion (AIC). In this work, we adopt this idea as well as a parametrization of the smoothing criterion discussed by Marra et al. (2017) which makes estimation more stable and efficient.

Suppose that $\boldsymbol{\delta}^{[\varkappa+1]}$ is the 'true' parameter value, and thus $\boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa+1]}) = \mathbf{0}$. By using a Taylor expansion for $\boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa+1]})$ at $\boldsymbol{\delta}^{[\varkappa]}$ it follows that $\mathbf{0} = \boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa+1]}) \approx$

Algorithm 1 (Trust Region Algorithm)

Require:
$\boldsymbol{\Delta}_{\max} > 0$, $\boldsymbol{\delta}^{[0]}$, $\boldsymbol{s}^{[0]}$, $\boldsymbol{\Delta}^{[0]} \in (0, \boldsymbol{\Delta}_{\max})$
Ensure:
$\|\boldsymbol{s}^{[\varkappa+1]}\| \geq 1.490116 \times 10^{-8}$ or $\varkappa \leq 100$

for $\varkappa = 0, 1, 2, \ldots$ **do**
  $\boldsymbol{s}^{[\varkappa+1]} := \arg\min_{\boldsymbol{s}} \mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]})$, subject to $\|\boldsymbol{s}\| \leq \boldsymbol{\Delta}^{[\varkappa]}$
  $\tilde{r}^{[\varkappa+1]} = \left\{\ell_p(\boldsymbol{\delta}^{[\varkappa]}) - \ell_p\left(\boldsymbol{\delta}^{[\varkappa]} + \boldsymbol{s}^{[\varkappa]}\right)\right\} / \left\{\ell_p\left(\boldsymbol{\delta}^{[\varkappa]}\right) - \mathcal{Q}_p\left(\boldsymbol{s}^{[k\varkappa+1]}\right)\right\}$
  if $\tilde{r}^{[\varkappa]} < 1/4$ **then**
    $\boldsymbol{\Delta}^{[\varkappa+1]} = \boldsymbol{\Delta}^{[\varkappa]}/4$
  else if $\tilde{r}^{[\varkappa]} > 3/4$ and $\|\boldsymbol{s}^{[\varkappa]}\| = \boldsymbol{\Delta}^{[\varkappa]}$ **then**
    $\boldsymbol{\Delta}^{[\varkappa+1]} = \min\left(2\boldsymbol{\Delta}^{[\varkappa]}, \boldsymbol{\Delta}_{\max}\right)$
  else
    $\boldsymbol{\Delta}^{[\varkappa+1]} = \boldsymbol{\Delta}^{[\varkappa]}$
  end if
  if $\tilde{r}^{[\varkappa]} \geq 1/4$ **then**
    $\boldsymbol{\delta}^{[\varkappa+1]} = \boldsymbol{s}^{[\varkappa+1]} + \boldsymbol{\delta}^{[\varkappa]}$
  else
    $\boldsymbol{\delta}^{[\varkappa+1]} = \boldsymbol{\delta}^{[\varkappa]}$
  end if
end for

$\boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa]}) + \boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}^{[\varkappa]})(\boldsymbol{\delta}^{[\varkappa+1]} - \boldsymbol{\delta}^{[\varkappa]})$. Solving for $\boldsymbol{\delta}^{[\varkappa+1]}$ yields, after some manipulation,

$$\boldsymbol{\delta}^{[\varkappa+1]} \;=\; \left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right)^{-1} \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\bar{\mathbf{z}}^{[\varkappa]}, \tag{2.11}$$

where $\boldsymbol{\mathcal{I}}^{[\varkappa]} = -\boldsymbol{\mathcal{H}}^{[\varkappa]}$ and $\bar{\mathbf{z}}^{[\varkappa]} = \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\boldsymbol{\delta}^{[\varkappa]} + \bar{\boldsymbol{\epsilon}}^{[\varkappa]}$ with $\bar{\boldsymbol{\epsilon}}^{[\varkappa]} = \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}^{-1}\boldsymbol{g}^{[\varkappa]}$. From standard likelihood theory $\bar{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\mathbf{z}} \sim \mathcal{N}(\boldsymbol{\mu}_{\bar{\mathbf{z}}}, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix, $\boldsymbol{\mu}_{\bar{\mathbf{z}}} = \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_0$ is the true parameter vector. The above representation allows us to estimate the smoothing parameters based on a parametrization of $\bar{\mathbf{z}}$ that uses $\boldsymbol{g}$ and $\boldsymbol{\mathcal{H}}$ as a whole instead of the n components that make them up. As argued by Marra et al. (2017), this is advantageous in estimation problems involving simultaneous systems of equations.

Now let $\hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}$ be the predicted value vector for $\bar{\mathbf{z}}$ defined as $\hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}} = \mathbf{C}_{\hat{\boldsymbol{\lambda}}}\bar{\mathbf{z}}$ where $\mathbf{C}_{\hat{\boldsymbol{\lambda}}} = \sqrt{\boldsymbol{\mathcal{I}}}\left(\boldsymbol{\mathcal{I}} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}}$, the influence matrix or hat matrix of the fitting problem which depends on the smoothing parameter vector. An appealing way of estimating

$\boldsymbol{\lambda}$ is to minimise the distance between $\hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}$ and the truth $\boldsymbol{\mu}_{\bar{\mathbf{z}}}$. This can be achieved using

$$\mathbb{E}(\|\boldsymbol{\mu}_{\bar{\mathbf{z}}} - \hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}\|^2) \;=\; \mathbb{E}(\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\|^2) - \tilde{n} + 2\mathrm{tr}(\mathbf{C}_{\boldsymbol{\lambda}}), \tag{2.12}$$

where $\tilde{n} = 6n$ and $\mathrm{tr}(\mathbf{C}_{\boldsymbol{\lambda}})$ is the number of estimated degrees of freedom (*edf*) of the penalized model which measures the flexibility of the fitted model. The *edf* of the model is defined as the sum of the *edf* of the smooth functions. Note that the RHS of (2.12) depends on the smoothing parameter through $\mathbf{C}_{\boldsymbol{\lambda}}$, while $\bar{\mathbf{z}}$ is associated with the un-penalized part of the model. In practice, smoothing parameters are selected by minimizing an estimate of (2.12), that is

$$\mathcal{V}(\boldsymbol{\lambda}) = \widehat{\|\boldsymbol{\mu}_{\bar{\mathbf{z}}} - \hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}\|^2} = \|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\|^2 - \tilde{n} + 2\mathrm{tr}(\mathbf{C}_{\boldsymbol{\lambda}}),$$

which is approximately equivalent to the AIC, defined as $2\mathrm{tr}(\mathbf{C}_{\boldsymbol{\lambda}}) - 2\ell(\hat{\boldsymbol{\delta}})$, where $-2\ell(\hat{\boldsymbol{\delta}})$ can be approximated as $\approx -2\ell(\boldsymbol{\delta}) - \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2$. Given $\boldsymbol{\delta}^{[\varkappa+1]}$, the estimation problem can be expressed as

$$\boldsymbol{\lambda}^{[\varkappa+1]} \;=\; \arg\min_{\boldsymbol{\lambda}} \mathcal{V}(\boldsymbol{\lambda}) := \|\bar{\mathbf{z}}^{[\varkappa+1]} - \mathbf{C}_{\boldsymbol{\lambda}}^{[\varkappa+1]}\bar{\mathbf{z}}^{[\varkappa+1]}\|^2 - \tilde{n} + 2\mathrm{tr}(\mathbf{C}_{\boldsymbol{\lambda}}^{[\varkappa+1]}),$$

which is solved by adapting the approach by Wood (2004) to the current context. This method implements a stable and efficient Newton method for estimating $\log(\boldsymbol{\lambda})$. Working with the logarithm of $\boldsymbol{\lambda}$ ensures that the smoothing parameter estimates are positive. The derivation of the above results can be found in Appendices A.6.1, A.6.2 and A.6.3.

The two steps are iterated until the algorithm satisfies the criterion $\left\{|\ell(\boldsymbol{\delta}^{[\varkappa+1]}) - \ell(\boldsymbol{\delta}^{[\varkappa]})|\right\} / \left\{0.1 + |\ell(\boldsymbol{\delta}^{[\varkappa+1]})|\right\} < 10^{-7}$. At convergence, well founded point-wise confidence intervals (CIs) for linear and non-linear functions of the model coefficients can be obtained using result $\boldsymbol{\delta} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}_p^{-1}\right)$. The rationale for using this result is provided in Marra & Wood (2012) for GAMs, whereas some examples of interval construction are given in Radice et al. (2016). For general smooth mod-

els, such as the one considered in this chapter, this result can be justified using the distribution of $\bar{\mathbf{z}}$ discussed in Marra et al. (2017), making the large sample assumption that $\boldsymbol{\mathcal{I}}$ can be treated as fixed, and making the usual Bayesian assumption on the prior of $\boldsymbol{\delta}$ for smooth models (e.g., Wood, 2006). Note that this result neglects smoothing parameter uncertainty. However, as argued by Marra & Wood (2012) this is not problematic provided that heavy oversmoothing is avoided (so that the bias is not too large a proportion of the sampling variability) and in our experience we found that this result works well in practice. The problem of testing smooth components for equality to zero is approached using the results discussed in Wood (2013a) and Wood (2013b).

2.3.3 Simulation study I

A simulation study was conducted to investigate the practical performance of the proposed approach as compared to the alternative routine `mvprobit()` available in `STATA`.

DGP1

In order to compare the results obtained from `SemiParTRIV()`/`gjrm()` and `mvprobit()`, we employed a Data Generating Process (DGP) based on the fully parametric model $\mathbf{Y}^* = \mathbf{V}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, with $\mathbf{V}$ containing binary and continuous variables with parametric effects. Exact simulation settings and the code used to generate the data can be found in Appendix A.7.1. The syntax used to fit trivariate probit models is

```
out <- SemiParTRIV(formula = f.l, data = dat)
```

where `f.l` consists of a list of three equations

```
eqn1 <- y1 ~ v1 + z1; eqn2 <- y2 ~ v1 + z1; eqn3 <- y3 ~ v1 + z1
f.l <- list(eqn1, eqn2, eqn3)
```

and `v1` and `z1` denote the binary and continuous covariates, respectively. Argument `data` refers to the data frame containing the variables in the model.

Figures 2.1 and 2.2 summarise the results. The regression coefficient estimates of both methods are satisfactory and converge to their true values as n increases. As expected, the variability of the estimates decreases as the sample size grows large. As for the correlation parameters, `SemiParTRIV()`/`gjrm()` considerably outperforms `mvprobit()` whose estimates do not improve as n increases. This may have important inferential implications; for instance, obtaining unbiased joint outcome probabilities requires accurate estimation of the correlation coefficient (e.g., Neelon et al., 2014). The `STATA` and `R` codes used to run the models for the above study are given in Appendix A.7.1.

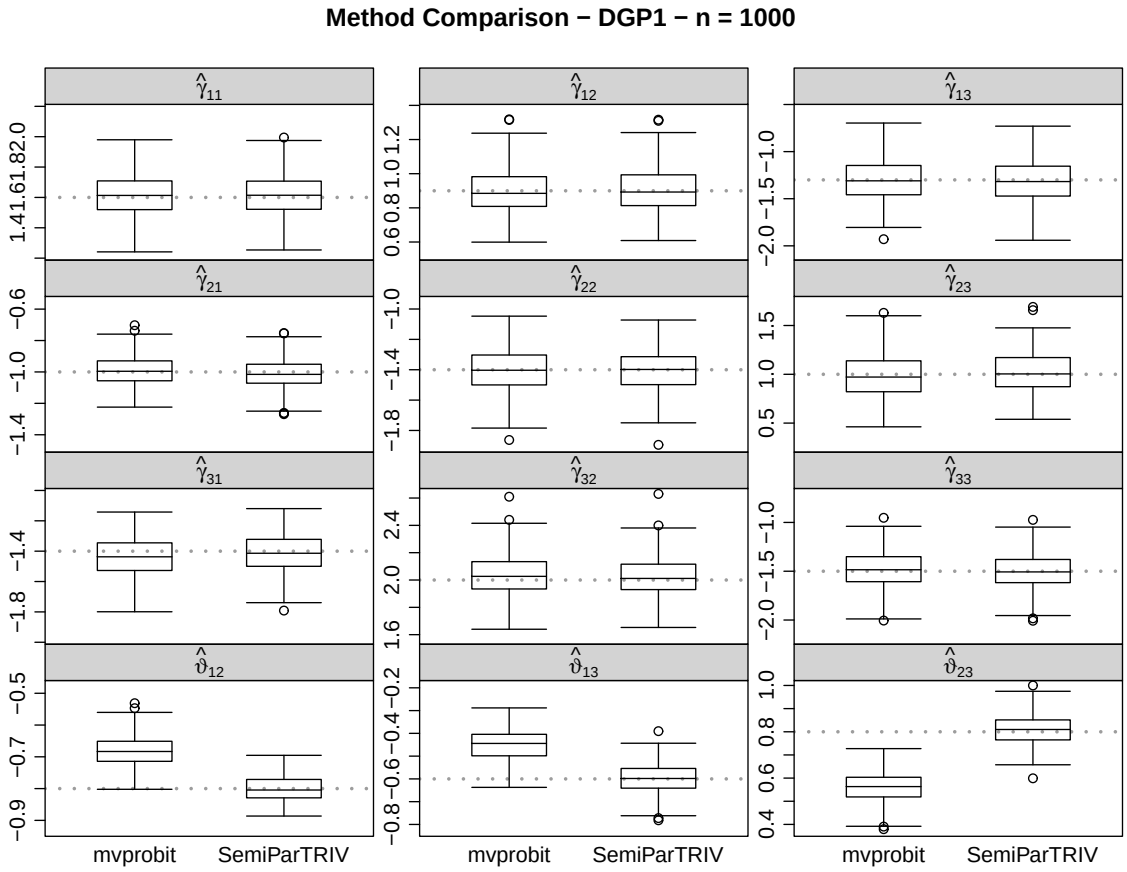

Figure 2.1: Boxplots of parameter estimates obtained applying `mvprobit()` and `SemiParTRIV()`/`gjrm()` to 250 datasets simulated using the settings described in Appendix A.7.1. The sample size was equal to 1000 and the true parameter values are represented by horizontal gray dotted lines.

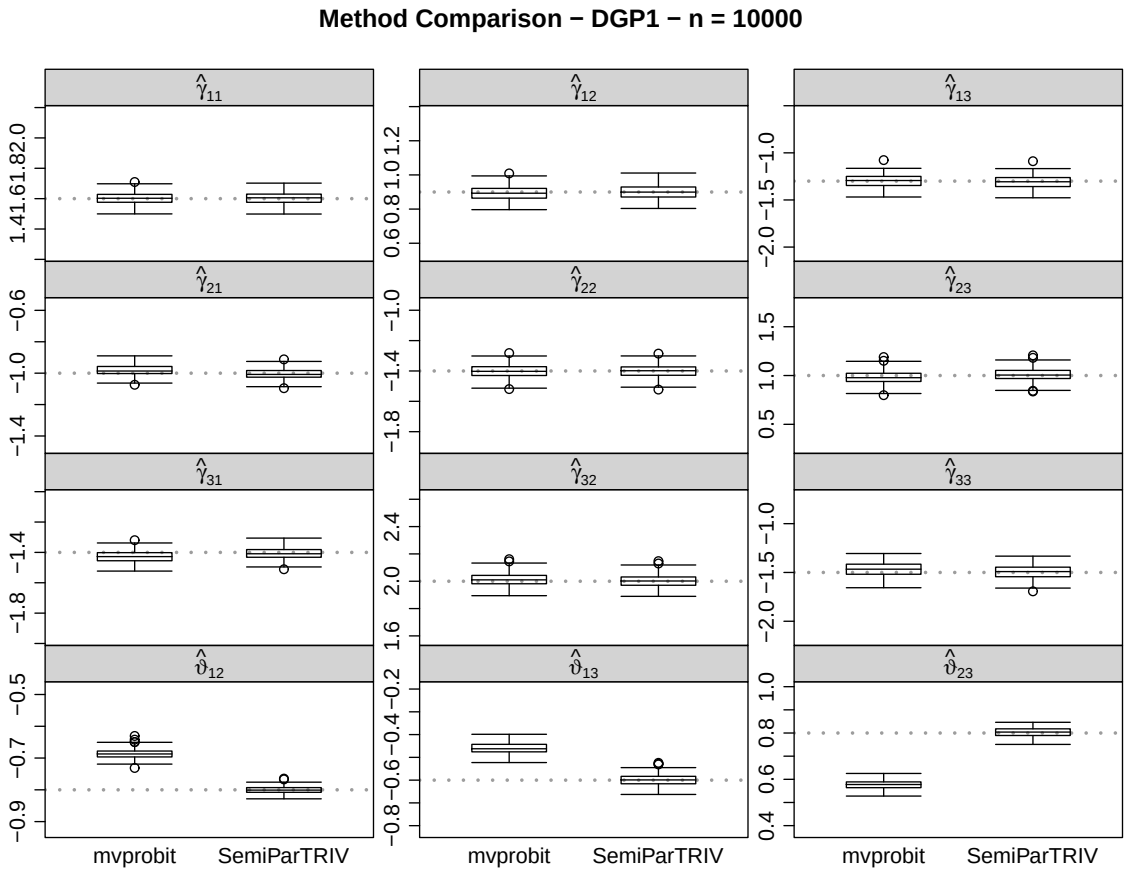

Figure 2.2: Boxplots of parameter estimates obtained applying `mvprobit()` and `SemiParTRIV()`/`gjrm()` on 250 datasets simulated using the settings described in Appendix A.7.1. The sample size was equal to 10000 and the true parameter values are represented by horizontal gray dotted lines.

Remark The unsatisfactory performance of `mvprobit()` in estimating the correlation parameters may be attributed to the method used for evaluating normal trivariate integrals, namely the Geweke-Hadjivassiliou-Keane (GHK) smooth recursive simulator (Geweke, 1991; Hajivassiliou & McFadden, 1991; Keane, 1990). Broadly speaking, the GHK approach first applies a Cholesky decomposition on the model's correlation matrix and then expresses the trivariate integrals as a product of three univariate probabilities defined in terms of truncated standard normal variables; Trinh & Genz (2015) introduced similar approximations which were found not to yield satisfactory results for highly correlated responses. Furthermore, Cappellari & Jenkins (2003) pointed out that if the correlation matrix obtained at a given

iteration of the optimization is not positive-definite then the GHK method uses the most recent positive-definite estimate of the correlation matrix; this runs the risk of delivering estimates that are far from the optimal values. When we tried different scenarios with higher and lower values for the correlation coefficients, we found that the stronger the magnitude of the correlations the worse the estimation results.

DGP2

The proposed approach does have some limitations, however. On occasion, the algorithm does not satisfy the first and second order necessary conditions for convergence (that is zero gradient and positive definite Hessian matrix). When this occurs, we observed that the non-zero gradient components and/or negative eigenvalues of the Hessian matrix are typically associated with the correlation parameters. To shed light on this issue, we conducted more simulation studies based on different configurations of the correlation matrix. We refer to the simulation settings of one such study as DGP2 whose description is given in Appendix A.7.1. Table 2.1 displays the percentage biases and root mean squared errors (RMSEs) for the estimates of the ϑ_{zk} (calculated as $\text{RMSE}(\hat{\vartheta}_{zk}) = \sqrt{\frac{1}{250}\sum_{\iota=1}^{250}\{\hat{\vartheta}_{zk,\iota} - \vartheta_{zk0}\}^2}$ where $\hat{\vartheta}_{zk,\iota}$ denotes the ι-th estimated value and ϑ_{zk0} is the true one). The results show that the estimation performance improves as n grows large, however at $n = 1000$ the method is not deemed to perform satisfactorily. Although not shown here, the estimated regression coefficients were similar to those of the previous study at both sample sizes. The `R` code used for this study is given in Appendix A.7.1.

To gain more insights into the above mentioned issue, we looked at the log-likelihood behaviour over the correlation parameters. For instance, we produced univariate transects through ℓ by evaluating $\ell(\boldsymbol{\delta})$ at the optimal MLE values for $\boldsymbol{\beta}$, ϑ_{12}^* and ϑ_{13}^*, for a grid of ϑ_{23}^* values. Figure 2.3 shows the corresponding $\ell(\boldsymbol{\delta})$ versus ϑ_{23}^*, based on 10 replicates, from which we observe a minimum that tends to be very shallow. This suggests that at small sample sizes the log-likelihood (and thus the model) may provide little information with which one can make inferences. Greater uncertainty is also expected. When this happens the parameter is weakly or not

	DGP2			
	$n = 1000$		$n = 10000$	
Estimator	Bias (%)	RMSE	Bias (%)	RMSE
$\hat{\vartheta}_{12}$	11.36	0.0935	-0.79	0.0262
$\hat{\vartheta}_{13}$	13.53	0.1204	1.86	0.0320
$\hat{\vartheta}_{23}$	-2.02	0.0567	0.16	0.0129

Table 2.1: Percentage biases and root mean squared errors (RMSEs) of the correlation estimates obtained applying `SemiParTRIV()`/`gjrm()` to 250 datasets simulated according to DGP2.

identified. The methodology described in the next chapter addresses this issue.

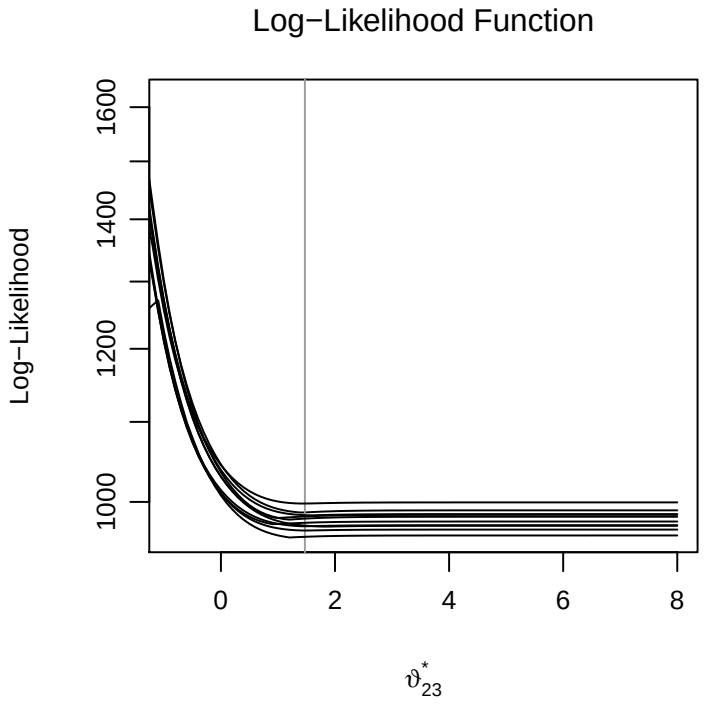

Figure 2.3: Profile log-likelihood function of the trivariate probit model for correlation parameter ϑ^*_{23}, for 10 data sets of sample size 1000 generated using DGP2 settings. The true value is represented by the vertical grey line.

2.4 Discussion

We have introduced a penalized likelihood method to estimate a trivariate system of probit regressions that incorporate additive or semi-parametric effects. Previous implementations of trivariate probit models are limited in many respects and we proposed a more general approach where several types of covariate effects are allowed for via the use of the regression spline methodology. The proposed development is backed by a reliable estimation method which requires analytical information on the score vector and Hessian matrix of the model's log-likelihood. Such information is not readily available in the literature and has been provided in this chapter. We have also developed the necessary computational tools which have been incorporated in the `R` package `GJRM` through function `SemiParTRIV()`/`gjrm()`.

Our simulations showed that the MLE results in some situations are unsatisfactory, a problem that commonly arises when the sample size is small or moderate. Next chapter proposes penalized MLE to deal with this difficulty.

Chapter 3

Correlation-based penalty approach to trivariate probit models

A penalized likelihood estimation approach is developed to address the difficulty in estimating accurately the correlation coefficients, which characterize the dependence of binary responses conditional on covariates. In this way, the issue with problematic flat likelihood functions is dealt and more efficient estimates are obtained. Issues related to practical implementation of the proposed approach are also discussed. The relevant numerical computation can be easily carried out using the `SemiParTRIV()`/`gjrm()` function in the `R` package `GJRM`. The proposed method is illustrated through a case study whose aim is to model jointly adverse birth binary outcomes in North Carolina.

3.1 Introduction

As discussed in Chapter 2, an issue with trivariate binary models is that the MLEs for the parameters in the correlation matrix may have large variance because the likelihood function near the optimum is flat. We propose a penalized likelihood approach for estimating accurately trivariate binary models when classical ML estima-

tion results are unsatisfactory. Penalization of log-likelihood functions is employed in various contexts for correcting the undesirable behaviour of regular MLE. This has been and still is an intensive research area in the statistical literature and has a large number of applications. An example is given in Section 2.3.2 where penalties are required to avoid over-fitting in curve estimation. Other examples include the development of penalized algorithms for high-dimensional problems (e.g, Kim et al., 2006; Park & Hastie, 2007), and the introduction of regularised regression approaches such as Ridge regression (Hoerl & Kennard, 1970), Bridge regression (Frank & Friedman, 1993), the Lasso approach (Tibshirani, 1996), the Smoothly Clipped Absolute Deviation (SCAD), Elastic-Net and Adaptive Lasso methods (Fan & Li, 2001; Zou & Hastie, 2005; Zou, 2006).

This chapter extends the semi-parametric trivariate probit model presented in Chapter 2 by addressing the difficulty in estimating the correlation coefficients that characterize the dependence of the binary responses conditional on regressors. We found that this is not an unusual occurrence for trivariate binary models and as far as we know such a limitation is neither discussed nor dealt with. Estimating such parameters accurately is crucial to obtain unbiased joint outcome probabilities, for instance. Moreover, to solve the issue with not continuously differentiable optimization problems we employ a local quadratic approximation (LQA) approach that is based on algorithms of Fan & Li (2001) and Ulbricht (2010). Asymptotic arguments of the proposed estimator are also provided. Note that in the bivariate binary case (see, for instance, Radice et al., 2016, and references therein) it is not necessary to penalize the correlation coefficient since the behavior of the respective log-likelihood function suggests that there is enough information that can be exploited in estimation. Parameter estimation is achieved within the penalized likelihood framework discussed in Chapter 2 using the trust region algorithm with integrated automatic multiple smoothing parameter selection. All the necessary computational routines are incorporated in the `R` function `SemiParTRIV()`/`gjrm()` that accompanies this chapter.

The chapter is organised as follows. Section 3.2 addresses the difficulty in esti-

mating the correlation coefficients of the trivariate model. Section 3.3 provides some asymptotic arguments and Section 3.4 applies the proposed approach to a case study that uses data from North Carolina whose aim is to model jointly plural births, low birth weight and premature birth. Conclusions are drawn in Section 3.5.

3.2 Correlation-based penalty

The aim of this section is to further augment the penalized log-likelihood function by introducing a penalty which addresses the difficulty in estimating the correlation parameters. The PMLE problem (2.6) then becomes

$$\hat{\boldsymbol{\delta}} \;:=\; \arg\min_{\boldsymbol{\delta}} -\{\ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\top \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\delta} - \mathcal{P}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})\}, \tag{3.1}$$

where $\mathcal{P}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ is a penalty acting on the correlations that depends on λ_{ϑ^*} which determines the amount of shrinkage required for ϑ^*_{zk}, $\forall z, k$. In this work, we employ the Ridge, Lasso and Adaptive Lasso approaches.

Suppose that $\mathbf{R}_q = \text{diag}\,(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ where the value of 1 on the $(q,q)^{th}$ entry of the matrix corresponds to the q^{th} parameter in $\boldsymbol{\delta}$, $\forall q = 1, \ldots, Q$, where Q denotes the total number of model parameters. Then, the penalties can be expressed as follows

$$\textbf{Lasso: } \mathcal{P}^{\text{L}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) = \mathcal{P}^{\text{L}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_1) = \lambda_{\vartheta^*}\left(|\vartheta^*_{12}| + |\vartheta^*_{13}| + |\vartheta^*_{23}|\right), \tag{3.2}$$

$$\textbf{Ridge: } \mathcal{P}^{\text{R}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) = \mathcal{P}^{\text{R}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_2^2) = \frac{1}{2}\lambda_{\vartheta^*}\left(\vartheta^{*2}_{12} + \vartheta^{*2}_{13} + \vartheta^{*2}_{23}\right), \tag{3.3}$$

$$\textbf{Ad. Lasso: } \mathcal{P}^{\text{AL}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) = \mathcal{P}^{\text{AL}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_1) = \lambda_{\vartheta^*}\left(\frac{|\vartheta^*_{12}|}{|\hat{\vartheta}^{*\text{MLE}}_{12}|^{\bar{\gamma}}} + \frac{|\vartheta^*_{13}|}{|\hat{\vartheta}^{*\text{MLE}}_{13}|^{\bar{\gamma}}} + \frac{|\vartheta^*_{23}|}{|\hat{\vartheta}^{*\text{MLE}}_{23}|^{\bar{\gamma}}}\right), \tag{3.4}$$

$\forall q = Q-2, Q-1, Q$, where superscripts L, R, and AL refer to the Lasso, Ridge and Adaptive Lasso penalties, respectively. The expression for the Adaptive Lasso is obtained as follows. Suppose that $\hat{\boldsymbol{\delta}}$ is a root-n-consistent estimator for $\boldsymbol{\delta}$, in which case we can use $\hat{\boldsymbol{\delta}}^{\text{MLE}}$. Then by picking a $\bar{\gamma} > 0$ it is possible to define adaptive

weights as $w_q = 1/|\mathbf{R}_q\hat{\boldsymbol{\delta}}^{\text{MLE}}|^{\bar{\gamma}}$ (Zou, 2006). Thus, we have that $w_{Q-2} = 1/|\hat{\vartheta}_{12}^{*\text{MLE}}|^{\bar{\gamma}}$, $w_{Q-1} = 1/|\hat{\vartheta}_{13}^{*\text{MLE}}|^{\bar{\gamma}}$ and $w_Q = 1/|\hat{\vartheta}_{23}^{*\text{MLE}}|^{\bar{\gamma}}$. Based on simulation studies, we found that $\bar{\gamma} = 1$ works well in most situations, however a sensitivity analysis trying different values for this parameter could be carried out. Note that when using Adaptive Lasso different amounts of shrinkage for each correlation are used and thus each coefficient is weighted differently. The derivation of expressions (3.2)-(3.4) can be found in Appendix B.1.1.

The main idea behind all penalties is similar: they shrink the correlation parameters towards zero as λ_{ϑ^*} increases. Simplified examples for the shapes of the three penalty functions are shown in Figure 3.1. As it can be seen from Figure 3.1, Lasso penalizes more than Ridge for instance. Using all penalty definitions and assessing the sensitivity of results to the different approaches is generally advisable.

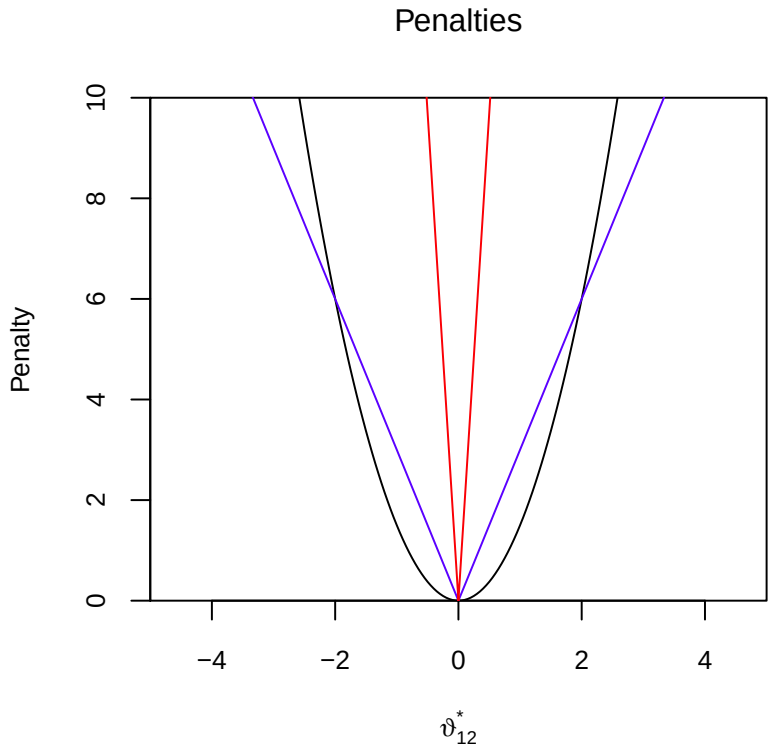

Figure 3.1: Shape of penalty functions for Ridge (——), Lasso (——) and Adaptive Lasso (——) for $\lambda_{\vartheta^*} = 3$.

3.2.1 Computational aspects

As pointed out by Ulbricht (2010), a penalty function should satisfy the following properties:

(P.1) $\mathcal{P}_{\lambda_{\vartheta^*}} : \mathbb{R}^+ \to \mathbb{R}^+$, $\mathcal{P}_{\lambda_{\vartheta^*}}(\mathbf{0}) = \mathbf{0}$,

(P.2) $\mathcal{P}_{\lambda_{\vartheta^*}}$ is continuous and strictly monotone in $\mathbf{R}_q^\top \boldsymbol{\delta}$,

(P.3) $\mathcal{P}_{\lambda_{\vartheta^*}}$ is continuously differentiable, $\forall \mathbf{R}_q\boldsymbol{\delta} \neq \mathbf{0}$, such that $\partial \mathcal{P}_{\lambda_{\vartheta^*}} / \partial \mathbf{R}_q\boldsymbol{\delta} > \mathbf{0}$.

The Ridge penalty is a quadratic function and satisfies (P.1)-(P.3). By contrast, the Lasso and Adaptive Lasso penalties are singular at $\boldsymbol{\delta} = \mathbf{0}$ (and thus not differentiable at this point) and non-concave with respect to $\boldsymbol{\delta}$. This can also be seen in Figure 3.1, where the curves of Lasso and Adaptive Lasso create a sharp point at the origin. In these cases, it would be unfeasible to maximize the penalized likelihood function using the approach described in Section 2.3.2. We therefore elect to approximate these two non-differentiable penalties by differentiable ones. Such approximations are available in the literature. For instance, Fan & Li (2001) approximated quadratically the non-convex SCAD penalty, while Ulbricht (2010) applied this idea to Lasso penalties. Rippe et al. (2012) approximated quadratically the L_0-type penalty by employing a weighted Ridge penalty. In this work, we employ the LQA approach.

Approximations of non-differentiable norms

The non-differentiability of L_1-type penalties such as Lasso and Adaptive Lasso can be avoided by approximating a norm at the critical point $\|\mathbf{R}_q\boldsymbol{\delta}\|_1 = \mathbf{0}$. Let $\|\mathbf{R}_q\boldsymbol{\delta}\|_1 = \|\boldsymbol{\xi}_q\|_1$. As in Koch (1996), norm $\|\boldsymbol{\xi}_q\|_1$ in a penalty function can be approximated by $\left(\boldsymbol{\xi}_q^\top \boldsymbol{\xi}_q + \bar{c}\right)^{1/2}$, where $\bar{c}$ is a small positive real number which controls how close the approximation and the exact function are; Oelker & Tutz (2013) argue that $\bar{c} \approx 10^{-8}$ works well in most cases. Similarly as in Oelker & Tutz (2013), we combine this approximation with a trick by Fan & Li (2001) as well as an idea introduced by Ulbricht (2010).

We assume that an approximation to each norm $\|\boldsymbol{\xi}_q\|_l$ exists such that

$$\|\boldsymbol{\xi}_q\|_l = \mathcal{K}_l(\boldsymbol{\xi}_q, \mathcal{C}) = \lim_{\bar{\mathcal{C}} \to \mathcal{C}} \mathcal{K}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}}),$$

where $\bar{\mathcal{C}}$ represents a set of possible tuning parameters, $\mathcal{C}$ is the set of boundary values for $\|\boldsymbol{\xi}_q\|_l$ and $\mathcal{K}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}})$ should be at least twice differentiable $\forall l \geq 1$. Additionally, for all $\boldsymbol{\xi}_q$, for which the derivative $\partial \|\boldsymbol{\xi}_q\|_l / \partial \boldsymbol{\xi}_q$ is defined, we assume that

$$\frac{\partial \|\boldsymbol{\xi}_q\|_l}{\partial \boldsymbol{\xi}_q} = \lim_{\bar{\mathcal{C}} \to \mathcal{C}} \mathcal{D}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}}),$$

where $\mathcal{D}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}}) = \partial \mathcal{K}_l(\boldsymbol{\xi}_q, \bar{\mathcal{C}}) / \partial \boldsymbol{\xi}_q \ \forall l$. We further assume that $\mathcal{D}_l(\mathbf{0}, \bar{\mathcal{C}}) = \mathbf{0}$. As mentioned above, the L_1 norm is approximated by $\mathcal{K}_1(\boldsymbol{\xi}_q, \bar{\mathcal{C}}) = (\boldsymbol{\xi}_q^\top \boldsymbol{\xi}_q + \bar{c})^{1/2}$. The first derivative $\mathcal{D}_1(\boldsymbol{\xi}_q, \bar{\mathcal{C}}) = \left(\boldsymbol{\xi}_q^\top \boldsymbol{\xi}_q + \bar{c}\right)^{-1/2} \boldsymbol{\xi}_q$ is a continuous approximation for the first-order derivative of the L_1 norm. In general, $\mathcal{K}_1(\boldsymbol{\xi}_q, \bar{\mathcal{C}})$ deviates only slightly from $\mathcal{K}_1(\boldsymbol{\xi}_q, \mathcal{C})$. That is, for $\boldsymbol{\xi}_q = \mathbf{0}$ the deviation is $\sqrt{\bar{c}}$, while for any other value of $\boldsymbol{\xi}_q$ the deviation is $< \sqrt{\bar{c}}$. Figure 3.2 shows approximation $\mathcal{K}_1(\boldsymbol{\xi}_q, \bar{\mathcal{C}})$ and its derivative $\mathcal{D}_1(\boldsymbol{\xi}_q, \bar{\mathcal{C}})$. Since the pictorial representation of vectorial norms requires plotting in more than two dimensions, we keep things simple and use a scalar argument in the L_1 norm which is approximated in the same way as a vector, that is $\|\xi\|_1 = |\xi| = (\xi^2)^{1/2} \approx (\xi^2 + \bar{c})^{1/2}$, where ξ can be any correlation parameter in $\boldsymbol{\delta}$. For illustrative purposes we used $\bar{c} = 0.1$.

Penalty $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$, for $\mathcal{G} = \{\mathrm{L}, \mathrm{AL}\}$, can be locally approximated by a quadratic function as follows. Suppose that $\tilde{\boldsymbol{\delta}}$ is an initial value close to $\hat{\boldsymbol{\delta}}$. Then we approximate $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ by a Taylor expansion of order 1 at $\tilde{\boldsymbol{\delta}}$, i.e.,

$$\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) \approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \nabla_{\tilde{\boldsymbol{\delta}}} \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})^\top (\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}). \tag{3.5}$$

As proved in Appendix B.1.2, $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ can be approximated as

$$\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) \approx \frac{1}{2} \boldsymbol{\delta}^\top \left\{ \nabla_{\|\mathbf{R}_q \tilde{\boldsymbol{\delta}}\|_1} \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q \tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q \tilde{\boldsymbol{\delta}}\right)^\top} \mathbf{R}_q \mathbf{R}_q^\top \right\} \boldsymbol{\delta} \approx \frac{1}{2} \boldsymbol{\delta}^\top \boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}} \boldsymbol{\delta},$$

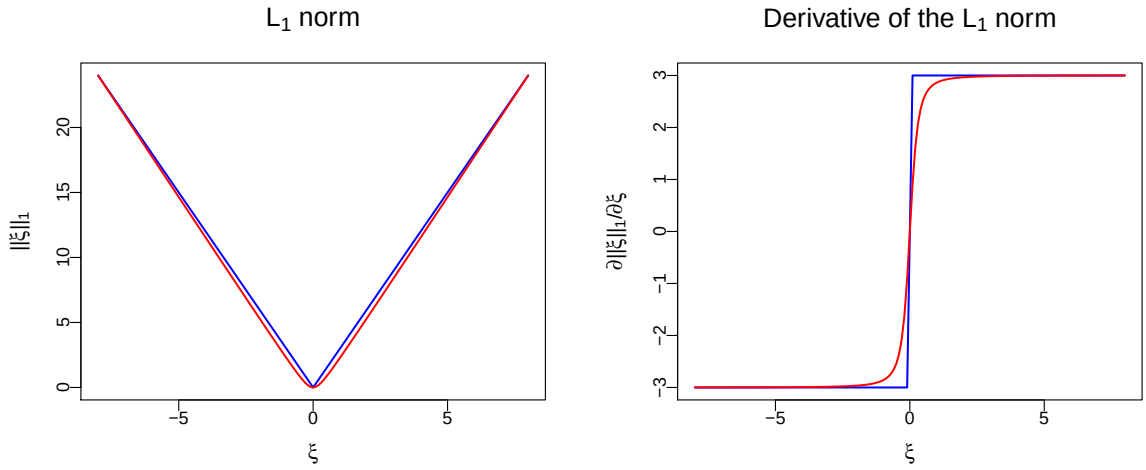

Figure 3.2: Graphical representation for the approximation of the L_1 norm (left panel) and its derivatives (right panel) with respect to ξ_q. The blue lines refer to the exact norms and derivatives based on sub-derivatives at $\xi_q = 0$, while the red lines correspond to the related approximations.

where $\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) = \partial\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})/\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1$, $\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}}) = \partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1/\partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}$, $\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ has the following form

$$\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}} = \begin{pmatrix} \mathbf{0}_{Q\times Q} & \mathbf{0}_{Q\times 3} \\ \mathbf{0}_{3\times Q} & \mathbf{A}^{\mathcal{G}}_{\lambda_{\vartheta^*}} \end{pmatrix},$$

and $\mathbf{A}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ is a 3×3 diagonal matrix that corresponds to the correlation parameters that have to be penalized, $\forall\mathcal{G}$. The expressions for the penalty matrices of Lasso and Adaptive Lasso are

$$\boldsymbol{\Lambda}^{\mathrm{L}}_{\lambda_{\vartheta^*}} = \lambda_{\vartheta^*}\,\mathrm{diag}\left(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, \frac{1}{\sqrt{\vartheta^{*2}_{12}+\bar{c}}}, \frac{1}{\sqrt{\vartheta^{*2}_{13}+\bar{c}}}, \frac{1}{\sqrt{\vartheta^{*2}_{23}+\bar{c}}}\right), \tag{3.6}$$

$$\boldsymbol{\Lambda}^{\mathrm{AL}}_{\lambda_{\vartheta^*}} = \lambda_{\vartheta^*}\,\mathrm{diag}\left(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{12}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{12}+\bar{c}}}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{13}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{13}+\bar{c}}}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{23}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{23}+\bar{c}}}\right). \tag{3.7}$$

Note that $\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ needs to be updated at each iteration of the algorithm as it depends on the estimated coefficients. In the Ridge penalty case, we simply have $\boldsymbol{\Lambda}^{\mathrm{R}}_{\lambda_{\vartheta^*}} = \lambda_{\vartheta^*}\,\mathrm{diag}\,(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 1, 1, 1)$. The derivations of (3.6) and (3.7) are given in Appendix B.1.3.

It follows that the penalized log-likelihood, score and Hessian matrix can be expressed as

$$\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}, \quad \boldsymbol{g}_p(\boldsymbol{\delta}) = \boldsymbol{g}(\boldsymbol{\delta}) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}, \quad \boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}) = \boldsymbol{\mathcal{H}}(\boldsymbol{\delta}) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}},$$

where $\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}} = \tilde{\mathbf{S}}_{\boldsymbol{\lambda}} + \boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ or $\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}} = \tilde{\mathbf{S}}_{\boldsymbol{\lambda}} + \boldsymbol{\Lambda}^{\mathrm{R}}_{\lambda_{\vartheta^*}}$ and $\bar{\boldsymbol{\lambda}}$ includes both $\boldsymbol{\lambda}$ and λ_{ϑ^*}. Problem (3.1) can now be solved using the approach described in Section 2.3 where matrix $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$ is replaced by $\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}$. If $\mathcal{P}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) = \mathbf{0}$ then $\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}$ clearly reduces to $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$.

3.2.2 Simulation study II

The aim of this simulation study is to assess the performance of the correlation-based penalty approach described above. We will use DGP2 from Section 2.3.3. Finally, the effectiveness of the method in estimating smooth function components will be explored.

DGP2

Recall from Simulation Study I in Section 2.3.3 that the correlation parameter estimates were not deemed satisfactory at $n = 1000$. Here, we re-examine this case by employing trivariate probit models with penalized correlations, using

```
outR  <- SemiParTRIV(f.l, data = dat, penCor = "ridge" )
outL  <- SemiParTRIV(f.l, data = dat, penCor = "lasso" )
outAL <- SemiParTRIV(f.l, data = dat, w.alasso = w.alasso,
                     penCor = "alasso")
```

where `f.l` and `data` are defined in Section 2.3.3. Argument `penCor` specifies the type of penalty used for the correlation parameters (`ridge`, `lasso` or `alasso`) and `w.alasso` denotes a 3×1 vector including the adaptive weights chosen as

```
w.alasso = c(theta12.ML, theta13.ML, theta23.ML)
```

with `theta12.ML, theta13.ML` and `theta23.ML` corresponding to $\hat{\vartheta}_{12}^{\mathrm{MLE}}$, $\hat{\vartheta}_{13}^{\mathrm{MLE}}$ and $\hat{\vartheta}_{23}^{\mathrm{MLE}}$. Table 3.1 shows substantial gains in accuracy and precision when penalizing

the correlation parameters. Compared to the unpenalized approach, the bias is negligible and the RMSE small. In this case, using `lasso` produced better overall performances as compared to `alasso` and `ridge`, although such differences may be judged as negligible.

		DGP2, n=1000	
Estimator	Correlation-based penalty	Bias (%)	RMSE
$\hat{\vartheta}_{12}$	Unpenalized	11.36	0.0935
	Ridge	0.10	0.0903
	Lasso	0.02	0.0835
	Adaptive Lasso	-0.31	0.0862
$\hat{\vartheta}_{13}$	Unpenalized	13.53	0.1204
	Ridge	0.13	0.1158
	Lasso	0.07	0.1092
	Adaptive Lasso	0.03	0.1142
$\hat{\vartheta}_{23}$	Unpenalized	-2.02	0.0567
	Ridge	-0.03	0.0551
	Lasso	-0.02	0.0475
	Adaptive Lasso	0.01	0.0428

Table 3.1: Percentage biases and root mean squared errors (RMSEs) of the correlation estimates obtained applying `SemiParTRIV()`/`gjrm()` to 250 datasets simulated according to DGP2 when the unpenalized approach and Ridge, Lasso and Adaptive Lasso correlation-based penalties are employed.

The good performance of the proposed approach can be justified visually by Figure 3.3 which shows that, in contrast to the unpenalized approach, penalizing the correlation parameters leads to a more pronounced optimum, hence less parameter variability and a reduced tendency to multiple minima.

DGP3

To assess the ability of `SemiParTRIV()`/`gjrm()` in estimating smooth function components, we modified slightly DGP2 by introducing non-linear effects for the continuous variable in the model. Estimation was achieved using the same syntax as that shown in the previous section but with equations specified as

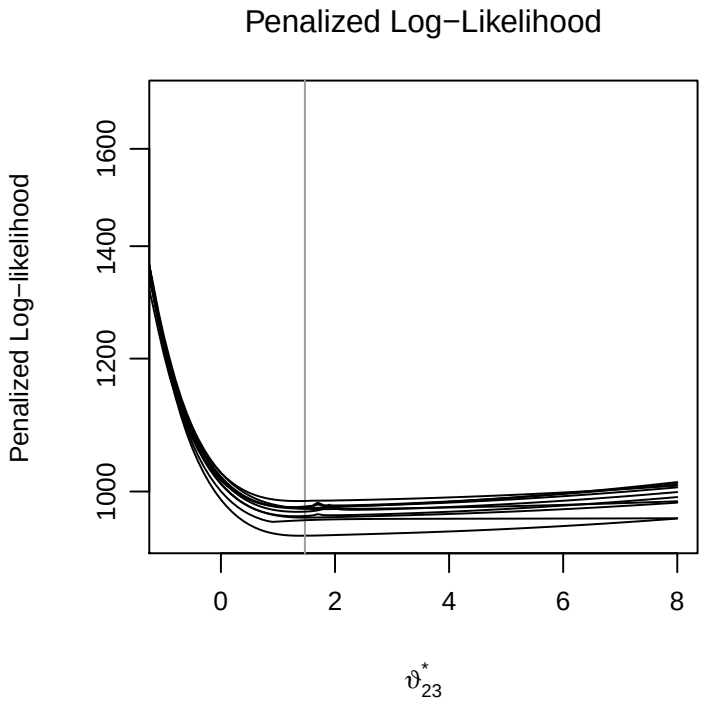

Figure 3.3: Profile penalized log-likelihood function of the trivariate probit model for correlation parameter ϑ^*_{23}, for 10 data sets of sample size 1000 generated using DGP2 settings. The true value is represented by the vertical grey line and the penalty used is Ridge.

```
eqn1 ~ v1 + s(z1); eqn2 ~ v1 + s(z1); eqn3 ~ v1 + s(z1)
```

where `s(z1)` defines a smooth function of the continuous covariate `z1`. A detailed description of DGP3 as well as the corresponding `R` code can be found in Appendix B.2.1. In this case, the coefficients of the spline bases and the correlations were penalized. The Lasso-type correlation-based penalty was employed (using Ridge and Adaptive Lasso produced virtually identical results). The estimates for the correlations and parametric part of the model were very similar to those of the previous study.

The estimated curves recover the true functions reasonably well (results are reported in Figure 3.4). For $n = 1000$, the estimates are rather variable and there are cases where the estimated functions are either wigglier or smoother than they should

be. This does not come as a surprise recalling that we are dealing with simultaneous binary models and as the sample size grows large the results improve considerably.

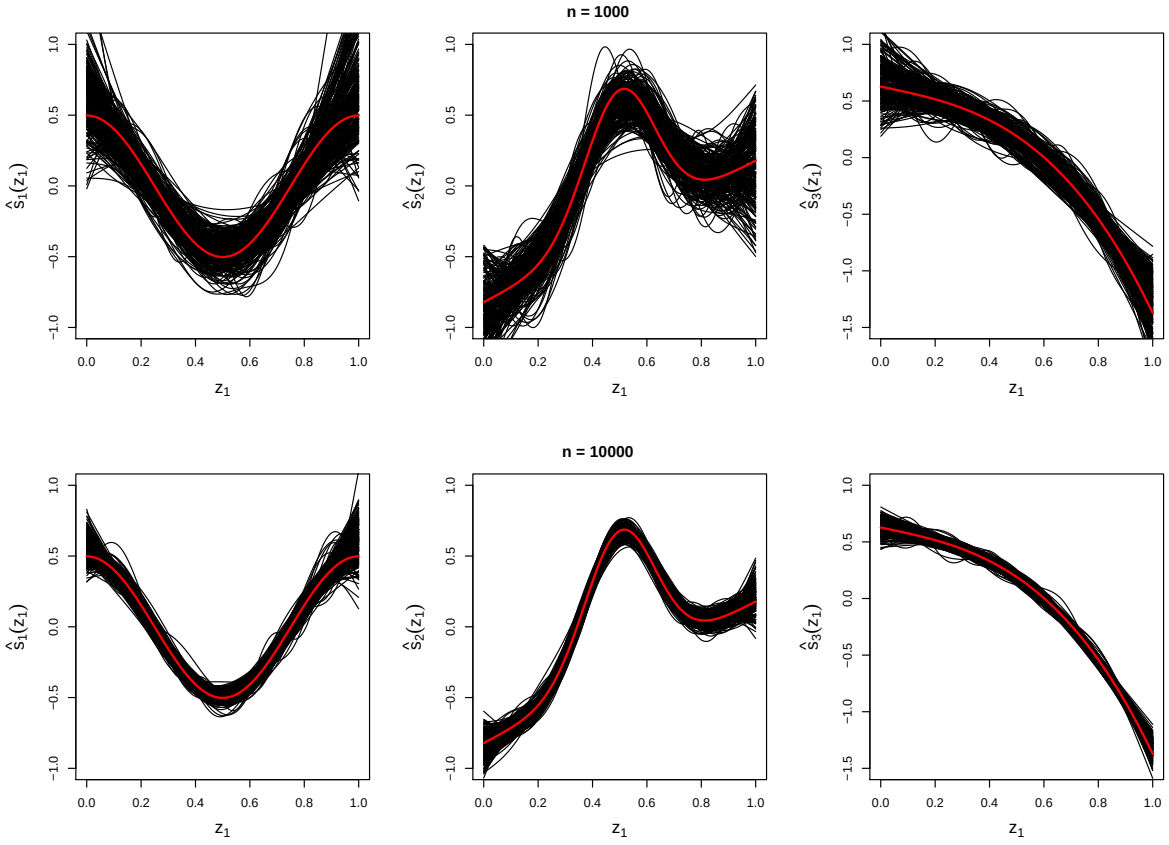

Figure 3.4: Estimated smooth functions for $s_1(z_1)$, $s_2(z_1)$ and $s_3(z_1)$ obtained applying `SemiParTRIV()`/`gjrm()` on 250 simulated datasets. The first row shows the estimated curves obtained from samples of 1000 observations, whereas those in the second row correspond to samples of 10000 observations. The black lines represent the estimated smooth functions over all replicates and the red solid lines show the true functions.

By using the result given in Section 2.3.2, that is $\boldsymbol{\delta} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}_p^{-1}\right)$, we quantify the uncertainty in the estimated smooth curves $\hat{s}_1(z_1)$, $\hat{s}_2(z_1)$ and $\hat{s}_3(z_1)$ by constructing CIs in order to obtain coverage probabilities for the non-linear terms in the model. Table 3.2 shows coverage probability results for the estimated curves at sample size equal to 1000 and 10000, when employing the Lasso-type penalty (similar results were obtained when using the Ridge and the Adaptive Lasso penalty). The coverage probabilities appear to be fairly close to their nominal values for all smooth functions, even at small sample sizes.

	Coverage Probabilities (%)	
	$n = 1000$	$n = 10000$
$\hat{s}_1(z_1)$	97.01	95.67
$\hat{s}_2(z_1)$	94.88	96.22
$\hat{s}_3(z_1)$	96.65	97.97

Table 3.2: Coverage probability results for $\hat{s}_1(z_1)$, $\hat{s}_2(z_1)$ and $\hat{s}_3(z_1)$ at two sample sizes, for the nominal level 95% when the Lasso-type penalty is employed.

The proposed approach generally proved effective. However, one should bear in mind that if the observed proportions of some trivariate binary events are very low then estimation may become challenging if not infeasible in some cases.

3.3 Theoretical aspects of the PMLE

In the following we assume that $s_{m\nu_m}(z_{m\nu_m i})$ is approximated by a spline basis with fixed high dimension, $\forall m, \nu_m, i$. Although this may be regarded as a strong assumption, in practice estimation is achieved with finite bases which, if rich enough, will allow one to assume that, compared to estimation variability, the modelling bias resulting from this approximation may be ignored (Kauermann, 2005). We also assume that both $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$ and $\boldsymbol{\Lambda}_{\lambda_{\vartheta^*}}$ (superscripts $\mathcal{G}$ and R have been suppressed to avoid clutter) are employed and denote the MLE as $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ and the PMLE as $\hat{\boldsymbol{\delta}}$.

Theorem 3.3.1. *Under certain regularity conditions, it can be proved that*

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1}\right),$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) = -\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ *and* $\boldsymbol{\delta}_0$ *denotes the true value vector of* $\boldsymbol{\delta}$.

Proof. See Appendix B.3.1. □

Note that although $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ is unbiased, when $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)$ is near singular then $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ has a large covariance matrix.

In what follows we consider the following assumptions (Cox & Barndorff-Nielsen, 1994, Ch. 3, pp. 82-83): (i) $\boldsymbol{g}(\boldsymbol{\delta}_0) \equiv \sqrt{n}\bar{\boldsymbol{g}}(\boldsymbol{\delta}_0) = \mathcal{O}_P(n^{1/2})$; (ii) $\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) = -\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) \equiv$

$-n\boldsymbol{\mathcal{I}}_i(\boldsymbol{\delta}_0) = \mathcal{O}(n)$; (iii) $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) - \mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) = \mathcal{O}_P(n^{1/2})$; (iv) $\bar{\boldsymbol{\lambda}} = o(n^{1/2})$, where $\bar{\boldsymbol{g}}(\boldsymbol{\delta}_0) = \mathcal{O}_P(1)$, $\boldsymbol{\mathcal{I}}_i(\boldsymbol{\delta}_0) = \mathcal{O}(1)$, $\bar{\boldsymbol{g}}(\boldsymbol{\delta}_0)$ is a normalized score function defined as $\bar{\boldsymbol{g}}(\boldsymbol{\delta}_0) = 1/n\boldsymbol{g}(\boldsymbol{\delta}_0) - \mathbb{E}\boldsymbol{g}(\boldsymbol{\delta}_0) = 1/n\boldsymbol{g}(\boldsymbol{\delta}_0)$ for $\mathbb{E}\boldsymbol{g}(\boldsymbol{\delta}_0) \approx \mathbf{0}$, and $\boldsymbol{\mathcal{I}}_i(\boldsymbol{\delta}_0)$ and $\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0)$ denote the expected and the observed Fisher information for a single observation, respectively, for $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) \equiv n\boldsymbol{\mathcal{I}}_i(\boldsymbol{\delta}_0)$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) \equiv n\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0)$. Assumption (iii) results by decomposing $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ in its mean and stochastic part, that is $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) = \mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\epsilon}$ where we assume that $\boldsymbol{\epsilon} = \mathcal{O}_P(n^{1/2})$ (Kauermann, 2005). Assumptions (i) - (iii) are the classical conditions for the consistency of the MLE, while assumption (iv) ensures that the smoothing parameter increases with the sample size; this is equivalent to $\boldsymbol{\Gamma}_{\bar{\lambda}} = o(n^{1/2})$.

Theorem 3.3.2. *Under certain regularity conditions, the PMLE has the following asymptotic distribution*

$$\sqrt{n}\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}\left[\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 + \left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\lambda}}\boldsymbol{\delta}_0\right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, n\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right),$$

and thus the asymptotic covariance of $\hat{\boldsymbol{\delta}}$ *is equal to* $\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}^{-1}$ *while its asymptotic bias is* $-\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\lambda}}\boldsymbol{\delta}_0$.

Proof. See Appendix B.3.2. □

Under assumptions (i)-(iv) we have that $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 = \mathcal{O}_P(n^{-1/2})$, while assumptions (ii) and (iv) imply that $\mathbf{Cov}(\hat{\boldsymbol{\delta}}) = \mathcal{O}(n^{-1})$ and $\mathbf{Bias}(\hat{\boldsymbol{\delta}}) = o(n^{-1/2})$. The derivation of these results can be found in Appendix B.3.3. Note that when $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)$ is near singular then $\mathbf{Cov}(\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}) \to \infty$ and $\mathbf{Cov}(\hat{\boldsymbol{\delta}}) \to \mathbf{0}$. This verifies that asymptotically the PMLE has smaller variance than the MLE and thus may perform better.

Theorem 3.3.3. *If* $\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0| = o(n^{1/2})$ *and* $\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}| = o(n^{1/2})$, *then*

$$\sqrt{n}\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\right) \sim \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1}\right).$$

Proof. See Appendix B.3.4. □

Theorem 3.3.4. *Suppose that* $\bar{\boldsymbol{\lambda}} \in [0, \infty)$ *is fixed. Then the PMLE* $\hat{\boldsymbol{\delta}}$ *that minimizes* $-\ell_p(\boldsymbol{\delta})$ *is consistent, that is* $\lim_{n\to\infty} \mathbb{P}(\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\|^2 > \bar{\varepsilon}) = 0$, $\forall \bar{\varepsilon} > 0$.

Proof. See Appendix B.3.5. □

The above theorems have mainly been adapted from Fan & Li (2001), Li & Sudjianto (2005) and Oelker et al. (2014).

Theorem 3.3.3 shows that as the sample size grows large, under certain conditions, the asymptotic distribution of the PMLE coincides with that of MLE. This is a desirable property as it is well-known that the MLE is the most efficient estimator. The above theorem also suggests that PMLE is essentially needed when the sample size is small. This is in line with the results obtained in the simulation studies in Sections 2.3.3 and 3.2.2.

3.4 Analysis of North Carolina data

Birth weight and gestational age are strongly related with infant morbidity and mortality (Paneth, 1995; Butler et al., 2007). Infant's low birth weight (LBW) is commonly defined as weight less than 2500 grams, whereas preterm birth (PTB) is typically defined as number of gestation weeks less than 37. Kiely (1998) and Martin et al. (1999) argued that multiple births (MB) such as twins and triplets are strongly related with PTB and LBW. These variables are typically influenced by geographic, demographic and behavioural characteristics (Blondel et al., 2002; Neelon et al., 2014; Miranda et al., 2009; South et al., 2012; Meng, 2010, e.g.,). This section illustrates the proposed modelling framework using 2007-2008 birth data from the North Carolina Center for Health Statistics (`http://www.schs.state.nc.us/`). In particular, the goal is to analyse jointly LBW, PTB and MB conditional on flexible functions of covariates and to account for residual dependence between the responses.

3.4.1 Model specifications and results

The data set consists of $61,426$ female newborns (similar results were obtained for male infants) which provides details on infant and maternal health, and parental

characteristics. The choice of variables included in the model was mainly driven by previous work on the subject (e.g., Miranda et al., 2009; South et al., 2012; Neelon et al., 2014). The responses are plurality (`mb`), a binary variable that takes value 1 for singleton birth and 0 for twins, triplets, quadruplets and quintuplets, infant's birth weight (`lbw`), an indicator variable with value 1 if infant's birth weight is ≤ 2500 and 0 otherwise, and preemie (`ptb`), a dummy variable that takes value 1 if the infant was born before completing the 37^{th} week of gestation and 0 otherwise. The covariates are maternal race categorised as non-white and white (`nwhite`), smoking status with 1 indicating a mother reported smoking during pregnancy (`smoker`), weight gained by mother during pregnancy in pounds (`gained`), age of mother in years (`mage`) and the county in which the birth occurred (`county`).

We employed `STATA`'s function `mvprobit()` and the proposed `SemiParTRIV()`/`gjrm()`. The model equations are

$$\begin{aligned}
\texttt{mb}_i^* &= \beta_{11} + \beta_{12}\texttt{nwhite}_i + \beta_{13}\texttt{smoker}_i + \texttt{gained}_i + \texttt{mage}_i + \texttt{county}_i + \varepsilon_{1i},\\
\texttt{lbw}_i^* &= \beta_{21} + \beta_{22}\texttt{nwhite}_i + \beta_{23}\texttt{smoker}_i + \texttt{gained}_i + \texttt{mage}_i + \texttt{county}_i + \varepsilon_{2i},\\
\texttt{ptb}_i^* &= \beta_{31} + \beta_{32}\texttt{nwhite}_i + \beta_{33}\texttt{smoker}_i + \texttt{gained}_i + \texttt{mage}_i + \texttt{county}_i + \varepsilon_{3i}.
\end{aligned}$$

In this case, parameter estimation of the proposed approach was carried out without the need of imposing a penalty on the correlation coefficients since no convergence issue signaling a possible issue with the identifiability of the correlations was encountered. In fact, using correlation-based penalties did not lead to different results. The regression coefficient estimates for the two competing methods were very similar. However, as shown in Table 3.3, the estimated correlations are different. Moreover, the proposed approach was faster and produced narrower intervals as compared to those of `STATA`'s routine. Figure 3.5 depicts the joint probabilities (averaged by county) that birth is multiple, infant's birth weight is normal and the baby is born full term when using the two approaches. The probabilities obtained using `mvprobit()` are overall higher than those obtained using `SemiParTRIV()`/`gjrm()`. This can be attributed to the different correlation estimates of the two methods. Our

simulations showed that `STATA`'s routine produces biased correlation coefficients, hence we would be reluctant to trust such results.

	`SemiParTRIV()`	`mvprobit()`
$\hat{\vartheta}_{12}$ (95% CI)	-0.7617 $(-0.7612, -0.7622)$	-0.5191 $(-0.5027, -0.5351)$
$\hat{\vartheta}_{13}$ (95% CI)	-0.6397 $(-0.6390, -0.6402)$	-0.4277 $(-0.4107, -0.4443)$
$\hat{\vartheta}_{23}$ (95% CI)	0.7853 (0.7850, 0.7856)	0.6796 (0.6692, 0.6897)
Execution Time	296.26	349.41

Table 3.3: Correlation parameter estimates obtained using `SemiParTRIV()`/`gjrm()` and `mvprobit()`. Corresponding 95% intervals (CIs) are reported in parentheses. The execution time (in seconds) for each method is reported at the bottom of the table.

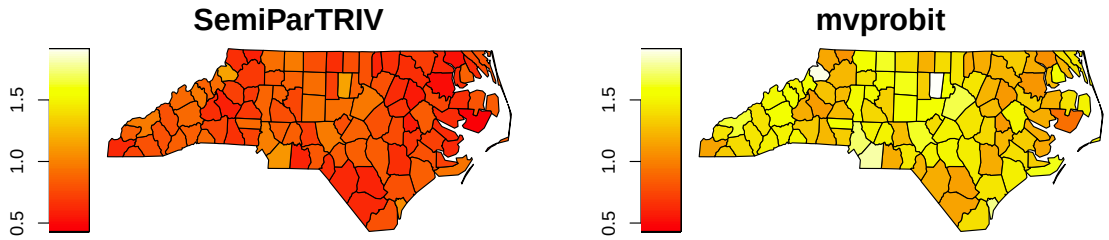

Figure 3.5: Joint probabilities (in %) that `mb` is multiple, `lbw` is $>$ 2500 grams and `ptb` is $>$ 37 weeks by county in North Carolina, obtained using `SemiParTRIV()`/`gjrm()` and `mvprobit()`.

Our approach allows for flexible functional dependence of the responses on the covariates. We therefore re-specify the model using the following equations

$$\begin{aligned}
\texttt{mb}_i^* &= \beta_{11} + \beta_{12}\texttt{nwhite}_i + \beta_{13}\texttt{smoker}_i + s_{11}(\texttt{gained}_i) + s_{12}(\texttt{mage}_i) + \\
&\quad s_{1\texttt{spatial}}(\texttt{county}_i) + \varepsilon_{1i}, \\
\texttt{lbw}_i^* &= \beta_{21} + \beta_{22}\texttt{nwhite}_i + \beta_{23}\texttt{smoker}_i + s_{21}(\texttt{gained}_i) + s_{22}(\texttt{mage}_i) + \\
&\quad s_{2\texttt{spatial}}(\texttt{county}_i) + \varepsilon_{2i}, \\
\texttt{ptb}_i^* &= \beta_{31} + \beta_{32}\texttt{nwhite}_i + \beta_{33}\texttt{smoker}_i + s_{31}(\texttt{gained}_i) + s_{32}(\texttt{mage}_i) + \\
&\quad s_{3\texttt{spatial}}(\texttt{county}_i) + \varepsilon_{3i},
\end{aligned}$$

where s_{m1} and s_{m2}, $\forall m = 1, 2, 3$, are smooth functions of $\texttt{gained}_i$ and $\texttt{mage}_i$ rep-

resented using penalized thin plate regression splines with twenty bases and second order penalties, and $s_{m\texttt{spatial}}$, for all m, models spatial regional effects using a Markov random field approach. Below we report and discuss some of the model results.

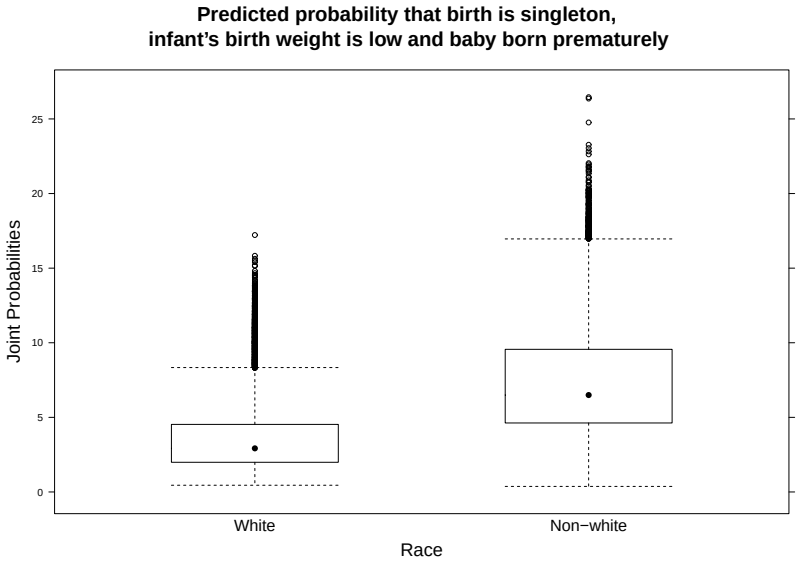

Figure 3.6: Joint prediction for singleton birth, infant's birth weight ≤ 2500 grams and baby born before completing the 37 gestational week, stratified by race, using the semi-parametric trivariate probit model.

Figure 3.6 presents a box plot for the predicted probability of singleton birth, birth weight ≤ 2500 grams and baby born before 37 weeks, stratified by race. It shows that the predicted probability of joint occurrence of babies born to non-white mothers is roughly twice than that of babies of white mothers. An example of estimated regression effects is shown in Figure 3.7 for the `lbw` equation. This suggests that the probability of low birth weight decreases with weight gained by the mother during pregnancy (with a pick at around 40 pounds) and then increases (although with quite some uncertainty). The effect of mother's age on the probability of lower infant's birth weight appears to be almost steady up to 30 years with a dramatic increase for women older than 40 years. Note that the estimated smooths are centered around zero because of centering identifiability constraints (see Section 2.2.2), however this does not affect interpretation. The point-wise CIs do not contain the zero line in most of the ranges of the `gained` and `mage` values. This suggests

that these two variables are important factors in determining `lbw`. The spatial map shows the effects of the county variable on the outcome, where darker colours correspond to decreased probability of low birth weight. P-values for testing smooth components for equality to zero were obtained by adapting the results discussed in Wood (2013a) and Wood (2013b) to the current context. These showed that the covariate effects are significant at least at the 5% level.

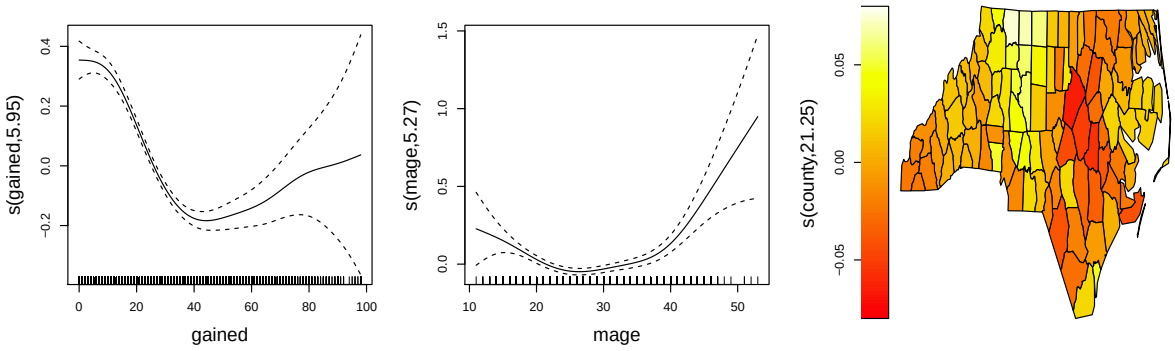

Figure 3.7: Smooth effects of `gained` and `mage` on `lbw` and associated 95% pointwise confidence intervals. The jittered rug plot, at the bottom of each graph, shows the covariate values. The numbers in brackets in the y-axis captions specify the *edf* of the smooth curve with $edf = 1$ corresponding to a straight line estimate; the higher the value the more complex the estimated curve. The map on the right hand side shows the magnitude of the estimates for the regional variable in each of the 100 counties in North Carolina.

3.5 Concluding remarks

In this chapter we have extended the penalized likelihood method introduced in Chapter 2 by penalizing the model's correlation coefficients via differentiable and approximations of non-differentiable penalties. This addresses the difficulty in estimating accurately the correlation parameters at small or modest sample sizes, an issue that has been neglected in the literature and that is likely to have a detrimental impact on the empirical performance of simultaneous binary models with more than two responses. Some asymptotic properties of the proposed estimator have also been discussed. The proposed model can be easily fitted using the

`SemiParTRIV()`/`gjrm()` function in the `R` package `GJRM`. The proposed method has been illustrated through simulations as well as a case study whose aim was to estimate a simultaneous model for three binary outcomes of newborn infants in North Carolina. Our results showed that joint outcome probabilities are affected by the way the model's parameters are estimated, especially the correlation coefficients.

The next chapter will look into the feasibility of modelling unobserved confounding through the trivariate probit model, where the aim is to correct for the presence of the endogeneity issue and non-random sample selection.

Chapter 4

Modelling unobserved confounding through additive trivariate probit models

This chapter discusses several models which can be obtained as byproduct of the framework developed in the previous chapters. These models deal with a problem which arises in observational studies when confounders (i.e., explanatory variables that are associated with treatment, or selection, and response) are unobserved (Heckman, 1978, 1979; Maddala, 1983; Van de Ven & Van Praag, 1981; Greene, 2003). This issue is known in the econometric literature as *endogeneity* and we will look at two cases: (I) endogeneity of a treatment variable and (II) endogenus or non-random sample selection of individuals into (or out) a sample. Several alternative approaches (not discussed here) are available in the literature to deal with (I) and (II) and the reader is referred to Clarke & Windmeijer (2012), Marra et al. (2017) and references therein for more details.

4.1 Introduction

In what follows we define problems (I) and (II), present some practical examples and review some existing literature on these issues.

(I) Endogeneity of Treatment In many statistical studies it is often of interest to examine the effect of a predictor, often referred to as treatment, on a particular response variable within a regression framework. To obtain an unbiased or consistent estimate of the treatment-response relationship, all confounders should be included in the model. Confounders can be observed and unobserved. Observed confounders can be measured and hence accounted for in the analysis. Such variables may be gender, race and educational background, for instance. However, there might be confounders such as motivation, ability and intelligence that cannot be observed and/or are difficulty to quantify. If we could include all confounders (observed and unobserved) in the model then standard estimation techniques, such as OLS regression, could be employed. If all relevant regressors can not be included in the model then confounding bias is expected (Cameron & Trivedi, 2005, Chapter 1.2.5, pp. 8). This problem is typically referred to as *endogeneity of the treatment*. This type of bias usually arises in observational studies, but even randomised controlled trials can be affected by this. For instance, in an observational study the treatment that each individual receives can not be randomly allocated (instead individuals typically assign themselves into a particular group) while randomised controlled trials may be affected by partial non-compliance. In both cases, observed and unobserved confounders need to be accounted for. In fact, conventional statistical methods controlling only for an observed source of confounding are likely to be of little use (Clarke & Windmeijer, 2012).

A wide range of applications discussing and addressing this issue are available in the literature. For example, Radice et al. (2013) studied the effect of obesity on the probability of employment in Italy accounting for the potential presence of observed and unobserved covariates (e.g., ability and motivation). Buscha & Conte (2014) examined the relationship between educational attainment in compulsory schooling and truancy. Here, truancy was considered to be endogenous because of the presence of unobserved covariates such as motivation and satisfaction that are likely to affect both truancy and educational outcomes. Colchero & Sosa-Rubí (2012) estimated the relationship between household income and lifestyle choices

with women's body mass index (BMI) controlling for the potential endogeneity of income arising from unobserved variables (e.g., productivity and self-control) that were deemed to be related with both income and BMI. Since depression may be both an antecedent and a consequence of smoking, Lie & Gardner (2016) analysed the reciprocal relationship between smoking and depression in Indonesia using a simultaneous equation estimation approach where both smoking and depression were treated as endogenous.

(II) Non-random Sample Selection When analysing data, it is typically assumed that a random sample from some underlying population is available and that if an outcome of interest is missing for some individuals it is common practice to assume that data are missing at random. That is, the probability that an outcome is missing depends only on observed variables and not unobserved ones (e.g., Heitjan & Basu, 1996). However, this is not always the case as there may be some individuals which are systematically less (or more) likely to be part of the sample. In this case a proportion of the whole population is not represented in the survey; the sample will include only the responders, hence the dependent variable of interest will be observed only for a restricted sample. If non-responders refuse to participate, for instance, in a survey because of some unobserved confounders (i.e., variables that are associated with both decision to participate and outcome), then non-random selection arises. If the responding and non-responding sub-samples share similar features then sample selection is not an issue. Thus, standard estimation methods can be employed. On the other hand, if the two sub-samples differ in some unobserved characteristics then selection bias will arise. Failure to account for sample selection may lead to inconsistent estimation results.

Sample selection bias can be viewed as a special case of endogeneity bias which occurs when the selection process generates endogeneity in the selected sub-sample. Practically, sample selection bias may manifest in two ways. First, the individuals or data units being investigated may have selected themselves out of the sample; for example individuals may feel that they do not want to participate in a survey.

Second, data analysts may have made sample selection decisions; that is analysts may have imposed some requirements for the entry of an individual in a study. In both cases there are two groups of individuals: those who participate and those who do not participate in the study. Non-random sample selection arises if the sample of individuals that participate differ in some characteristics from the sample consisting of non-participants.

A classic sample selection example is a model for the wages and employment of women, studied in the seminal works by Gronau (1973) and Heckman (1976), where hours worked are observed only for women who decide to participate in the labor force. Since then, many researchers have been focusing on modelling non-random samples. Sharma et al. (2013) examine the waiting time-socioeconomic status relationship within publicly-funded systems accounting for selection bias as richer patients are more likely to opt for private care when they expect high waiting times in public hospitals, thus leaving poor patients in public hospitals waiting longer. Marra et al. (2017) use sample selection models to correct for HIV prevalence estimates in Sub-Saharan African countries, where the data are affected by non-participation since some individuals choose not to participate in HIV testing.

The most common method to model data that are affected by unobserved confounders is the two-stage approach, which removes the bias by including an additional explanatory variable in the model representing an omitted variable (e.g. Wooldridge, 2002; Beck et al., 2003; Leigh & Schembr, 2004; Lindenl & Adams, 2006; Heckman, 1979). Many researchers, however, have argued that simultaneous likelihood estimation methods may be superior to conventional two-stage procedures in some cases (e.g., Wooldridge, 2002; Bhattacharya et al., 2006; Freedman & Sekhon, 2010). MLE methods address the issue of endogeneity of the treatment by setting up a bivariate recursive system of equations, for example similar to the model developed in Marra & Radice (2011). Recent approaches for tackling selection bias include the works by Chib et al. (2009) and Wiesenfarth & Kneib (2010), who introduce Bayesian frameworks allowing for flexible estimation of the covariate

effects, with a frequentist counterpart proposed by Marra et al. (2013). Computational routines for estimation of recursive and selection models are available in `STATA`'s routines `biprobit()` (using MLE methods) and `heckprobit()` (based on two-stage methods) respectively, whereas `SemiParBIVProbit()` (Marra & Radice, 2017) employs penalized MLE to fit models that allow for flexible estimation of the covariate effects and several shapes for the dependence structure of the model's equations.

If dealing separately with endogeneity and non-random sample selection then the above methods are adequate for accounting for these problems. In practice, however, there may be situations in which the two issues arise simultaneously. For example, the employment of a worker may depend on both the worker's decision to work and employment's decision to hire (e.g., Mohanty, 2001). Moreover, inferior endowments may cause problematic pregnancies (e.g., lower birth-weight) and therefore more prenatal care visits may be required; women's decision may positively affect birth-weight and prenatal care use if women who practice healthier behaviour during pregnancy are also more likely to give birth (e.g., Rous et al., 2004).

In this chapter, we develop three models: the endogenous trivariate probit model controlling for two sources of endogeneity, the double sample selection model where there are two layers of selection, and the endogenous-sample selection model controlling for both endogeneity of the treatment and non-random sample selection. Estimation of the above models has been discussed in the literature. Keay (2016) introduced a partial copula approach for models with multiple discrete endogenous variables, and models dealing with both endogeneity of the treatment and sample selection bias. Rous et al. (2004) employ a full-information MLE technique, the discrete factor method, controlling for potential biases arising from non-random sample selection and endogeneity of the treatment. Li (2011) extends the estimation technique of Chib et al. (2009) (which involves one selection mechanism) and proposes a Markov chain Monte Carlo estimation algorithm accounting for two layers of selection, whereas Zhang et al. (2015) develop a Bayesian sampling algorithm for estimating trivariate probit-ordered models with double rules of sample selection. The

proposed framework allows for flexible predictors' specifications through the inclusion of non-parametric and spatial covariate effects, making the models more flexible than the aforementioned approaches. The techniques implemented in this chapter are based on the framework described in the previous chapter. All the necessary computational routines are incorporated in the `R` function `SemiParTRIV()`/`gjrm()`.

The rest of the chapter is organised as follows. Sections 4.2, 4.3 and 4.4 discuss the endogenous trivariate probit model, the double sample selection model and the endogenous-sample selection model, respectively, and provide details on estimation and inference. Section 4.5 studies the performance of the double sample selection model using simulated data, whereas the endogenous trivariate probit model is applied to a case study whose aim is to jointly estimate the effect of two chronic diseases on labour force participation accounting for the potential endogeneity of the two diseases. The final section provides a discussion.

4.2 The endogenous trivariate probit model

4.2.1 Model specification

In economics, the endogeneity issue is commonly structured in terms of a regression model from which important regressors have been omitted and hence become a part of the model's error terms. Here we are interested in studying the effect of two endogenous treatments on the outcome variable accounting for unobserved confounding and flexible covariate effects. This extends the model proposed by Marra & Radice (2011) which can only deal with one endogenous variable at a time. The model structure builds on a first reduced form or treatment equation for the potentially endogenous dummy variable, a second treatment equation that describes the second potential endogenous dummy variable, and the outcome equation which determines the response variable. The model can be expressed in terms of latent

responses as

$$
\begin{aligned}
y_{1i}^{*} &= \mathbf{v}_{1i}^{\top}\boldsymbol{\gamma}_1 + \mathbf{L}_{1i}^{\top}\boldsymbol{\alpha}_1 + \varepsilon_{1i}, && (4.1)\\
y_{2i}^{*} &= \psi_1 y_{1i} + \mathbf{v}_{2i}^{\top}\boldsymbol{\gamma}_2 + \mathbf{L}_{2i}^{\top}\boldsymbol{\alpha}_2 + \varepsilon_{2i}, && (4.2)\\
y_{3i}^{*} &= \psi_2 y_{1i} + \psi_3 y_{2i} + \mathbf{v}_{3i}^{\top}\boldsymbol{\gamma}_3 + \mathbf{L}_{3i}^{\top}\boldsymbol{\alpha}_3 + \varepsilon_{3i},
\end{aligned}
$$

where latent variables y_{1i}^{*} and y_{2i}^{*} denote the two endogenous treatments, y_{3i}^{*} characterizes the outcome variable, ψ_1 is the effect of the first treatment on the second treatment on the scale of the linear predictor, and ψ_2 and ψ_3 denote the effect of the first and second treatments, respectively, on the outcome. The components in $\mathbf{v}_{mi}^{\top}\boldsymbol{\gamma}_m + \mathbf{L}_{mi}^{\top}\boldsymbol{\alpha}_m$ are the same for all m, and all exogenous and the three error terms are assumed to follow a standard trivariate normal distribution with zero mean and variance-covariance matrix equal to $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is defined in Section 2.2; $\vartheta_{zk} \neq 0$, $\forall z, k$, suggests that unobserved confounding is present and thus joint estimation of the three equations is required. Since the model includes only unidirectional effects (the treatment variables affect the outcome but the outcome does not affect the treatments), we refer to this system as 'recursive model'. The model is indeed a special case of the simultaneous equation system described in the previous chapter, while the recursive structure follows from the condition of logical consistency which states that only two observed endogenous variables are allowed on the right-hand side of the model. This is because the probabilities for the different combinations of the three binary variables have to sum to one (e.g., Maddala, 1983, pp. 118).

4.2.2 Identification of treatment effects

Although the recursive trivariate probit model is in principle capable at delivering consistent estimates of the treatment effects, their identification relies on functional form assumptions. This has been discussed extensively in the literature. Heckman (1978) states that in simultaneous equation models with endogenous dummy variables, only the full rank condition of the regressor matrix is needed for identification of the model parameters. On the other hand Maddala (1983, pp. 122) reports that

the parameters of the outcome equation in an endogenous binary probit model are not identified in the absence of an exclusion restriction (ER, an extra covariate in the model that is associated with the treatment, is not directly related to the outcome, and is independent of the unobserved confounders), while Wilde (2000) argues that Maddala's argument is only valid when the linear predictors of the two equations are both constants and demonstrates that as long as there exists at least one varying exogenous regressor in each equation the identification problem does not arise. As Wilde (2000) clearly states, however, if the model assumptions are met then identification is theoretically achieved (even if ERs are not included in the model) and the treatment effects will be consistently estimated. In practice, basing identification only on the assumed model's functional form may be problematic as the model is likely to be misspesified to some degree. Therefore, empirical identification is better achieved in the presence of ERs (e.g., Little, 1985; Sajaia, 2008; Buscha & Conte, 2014). This has also been confirmed by the recent works of Li et al. (2016) and Marra et al. (2017), where it has been shown that in the absence of valid ERs parameter estimates may be biased when the model is misspecified.

4.2.3 Parameter estimation

Since the error terms of the three equations are assumed to be correlated, simultaneous estimation is desirable. Let the linear predictor in equation (4.1) be defined as $\eta_{1i} = \mathbf{x}_{1i}^{\top}\boldsymbol{\beta}_1$, where $\mathbf{x}_{1i}$ includes $\mathbf{v}_{1i}^{\top}$ and $\mathbf{L}_{1i}$, while $\boldsymbol{\beta}_1$ contains $\boldsymbol{\gamma}_1$ and $\boldsymbol{\alpha}_1$. The quantities for the predictor in (4.2) are the same as those in (4.1) with the exception that $\mathbf{x}_{2i}^{\top}$ and $\boldsymbol{\beta}_2$ also include y_{1i} and parameter ψ_1, respectively. Similarly, $\mathbf{x}_{3i}^{\top}$ and $\boldsymbol{\beta}_3$ also include y_{1i} and y_{2i}, and ψ_2 and ψ_3 respectively. The joint distribution of the three responses conditional on $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$ and $\mathbf{x}_{3i}$, $p_{\bar{e}_1\bar{e}_2\bar{e}_3 i} = \mathbb{P}(y_{1i} = \bar{e}_1, y_{2i} = \bar{e}_2, y_{3i} = \bar{e}_3|\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i})$, has therefore eight elements: p_{111i}, p_{110i}, p_{101i}, p_{011i}, p_{000i}, p_{001i}, p_{010i}

and p_{100i} and thus the log-likelihood function can be expressed as

$$\begin{aligned}\ell \;=\; \sum_{i=1}^{n} \{ & y_{1i}y_{2i}y_{3i}p_{111i} + y_{1i}y_{2i}(1-y_{3i})p_{110i} + y_{1i}(1-y_{2i})y_{3i}p_{101i} + \\ & (1-y_{1i})y_{2i}y_{3i}p_{011i} + (1-y_{1i})(1-y_{2i})(1-y_{3i})p_{000i} + \\ & (1-y_{1i})(1-y_{2i})y_{3i}p_{001i} + (1-y_{1i})y_{2i}(1-y_{3i})p_{010i} + \\ & y_{1i}(1-y_{2i})(1-y_{3i})p_{100i}\}\,,\end{aligned}$$

which is essentially the log-likelihood function of the classic trivariate probit model described in Section 2.3. Consequently, its respective gradient and Hessian components have the same expressions which means that parameter estimation of the endogenous trivariate probit model is achieved using the PMLE method discussed in Section 3.2.

In this context, function `SemiParTRIV()`/`gjrm()` can be used as follows

```
eqn1 <- y1 ~ z1 + v1 + v2 + v3
eqn2 <- y2 ~ y1 + z1 + v1 + v2
eqn3 <- y3 ~ y1 + y2 + z1 + v1
f.l <- list(eqn1, eqn2, eqn3)
out <- SemiParTRIV(formula = f.l, data = dat)
```

where `v2` and `v3` denote the ERs.

4.2.4 Average treatment effect

In empirical applications the causal effect of a treatment variable, say y_{1i}, on the response probability $\mathbb{P}(y_{3i} = 1|y_{1i}, y_{2i}, \mathbf{v}_{3i}^{\top}, \mathbf{L}_{3i}^{\top})$ is of primary interest. For given values of y_{2i}, $\mathbf{v}_{3i}^{\top}$ and $\mathbf{L}_{3i}^{\top}$, this can be calculated using the following expression

$$\mathbb{P}(y_{3i}|y_{1i} = 1, y_{2i}, \mathbf{v}_{3i}^{\top}, \mathbf{L}_{3i}^{\top}) - \mathbb{P}(y_{3i}|y_{1i} = 0, y_{2i}, \mathbf{v}_{3i}^{\top}, \mathbf{L}_{3i}^{\top}),$$

where $\mathbb{P}(y_{3i} = 1|y_{1i} = 1, y_{2i}, \mathbf{v}_{3i}^{\top}, \mathbf{L}_{3i}^{\top}) = \Phi(\eta_{3i}^{(y_{1i}=1)})$, $\mathbb{P}(y_{3i} = 1|y_{1i} = 0, y_{2i}, \mathbf{v}_{3i}^{\top}, \mathbf{L}_{3i}^{\top}) = \Phi(\eta_{3i}^{(y_{1i}=0)})$, and $\eta_{3i}^{(y_{1i}=\bar{e}_1)}$ denotes the linear predictor in the outcome equation evaluated at $y_{1i} = \bar{e}_1$, $\forall \bar{e}_1 = \{0, 1\}$. Similarly, the impact of y_{1i} on y_{2i} is equal to

$\mathbb{P}(y_{2i}|y_{1i}=1,\mathbf{v}_{2i}^\top,\mathbf{L}_{2i}^\top) - \mathbb{P}(y_{2i}|y_{1i}=0,\mathbf{v}_{2i}^\top,\mathbf{L}_{2i}^\top)$ and the effect of y_{2i} on y_{3i} equals to $\mathbb{P}(y_{3i}|y_{2i}=1,y_{1i},\mathbf{v}_{3i}^\top,\mathbf{L}_{3i}^\top) - \mathbb{P}(y_{3i}|y_{2i}=0,y_{1i},\mathbf{v}_{3i}^\top,\mathbf{L}_{3i}^\top)$. This is known as the causal treatment effect (TE; e.g., Angrist et al., 1996) in the literature. It measures the causal difference in outcomes between individuals that receive the treatment ($y_{1i}=1$ or $y_{2i}=1$) and individuals who do not receive it ($y_{1i}=0$ or $y_{2i}=0$). For each individual only one of the two potential outcomes can be observed; the other outcome is the counterfactual. To measure the average TE (ATE) in a specific sample, we use $1/n\sum_{i=1}^{n}\text{TE}_i$ where TE_i denotes the TE of individual i (e.g., Abadie et al., 2004).

CIs for ATE can be obtained by simulation from the posterior distribution $\boldsymbol{\delta}\dot{\sim}\mathcal{N}\left(\hat{\boldsymbol{\delta}},-\hat{\boldsymbol{\mathcal{H}}}_p^{-1}\right)$ described in Section 3.2.2.

The ATE with corresponding CI can be computed using function `AT()` in `GJRM`. For example, the effect of y_{2i} on y_{3i} (in %) with corresponding 95% CI can be obtained as

```
AT(out, nm.end = "y2", eq = 3)
```

where `nm.end` denotes the endogenous variable and `eq` indicates the equation that contains the endogenous variable.

4.3 The double sample selection model

4.3.1 Model specification

The target here is to fit a regression model when some observations for the outcome variable are missing not at random. We consider three responses $(y_{1i},y_{2i},y_{3i}) \in \{0,1\}$, where y_{1i} and y_{2i} characterize whether or not an observation of the outcome variable y_{3i} is observed; unobserved values for the outcome are coded as 0. The situation considered is shown in Figure 4.1. The second selection mechanism y_{2i} is observed only if the individual passes the first selection mechanism (i.e., $y_{1i}=1$) and the outcome y_{3i} is observed only if the individual passes both stages (i.e., $y_{1i}=1$ and $y_{2i}=1$). To address the double sample selection bias problem we first write the

model in terms of three latent variables as

$$
\begin{aligned}
y_{1i}^{*} &= \mathbf{v}_{1i}^{\top}\boldsymbol{\gamma}_1 + \mathbf{L}_{1i}^{\top}\boldsymbol{\alpha}_1 + \varepsilon_{1i} && (4.3)\\
y_{2i}^{*} &= \left\{\mathbf{v}_{2i}^{\top}\boldsymbol{\gamma}_2 + \mathbf{L}_{2i}^{\top}\boldsymbol{\alpha}_2 + \varepsilon_{2i}\right\} \times y_{1i} && (4.4)\\
y_{3i}^{*} &= \left\{\mathbf{v}_{3i}^{\top}\boldsymbol{\gamma}_3 + \mathbf{L}_{3i}^{\top}\boldsymbol{\alpha}_3 + \varepsilon_{3i}\right\} \times y_{1i} \times y_{2i}, && (4.5)
\end{aligned}
$$

where y_{1i} is a binary variable taking value 0 or 1, and y_{2i} and y_{3i} are determined as

$$
y_{2i} = \begin{cases} 1 & \text{if} \quad (y_{2i}^{*} > 0 \;\; \& \;\; y_{1i} = 1) \\ 0 & \text{if} \quad (y_{2i}^{*} < 0 \;\; \& \;\; y_{1i} = 1) \\ - & \text{if} \quad y_{1i} = 0 \end{cases},
$$

and

$$
y_{3i} = \begin{cases} 1 & \text{if} \quad (y_{3i}^{*} > 0 \;\; \& \;\; y_{1i} = 1 \;\; \& \;\; y_{2i} = 1) \\ 0 & \text{if} \quad (y_{3i}^{*} < 0 \;\; \& \;\; y_{1i} = 1 \;\; \& \;\; y_{2i} = 1) \\ - & \text{if} \quad y_{1i} = 0 \;\; \text{or} \;\; (y_{1i} = 1 \;\; \& \;\; y_{2i} = 0) \end{cases}.
$$

We assume that the selection equations (4.3) and (4.4) are linked with the outcome equation (4.5) through unobservables and this link is formalized through a trivariate normal distribution with zero mean and variance-covariance matrix equal to $\boldsymbol{\Sigma}$. The same identification arguments discussed in Section 4.2.2 apply here as well.

4.3.2 Parameter estimation

Since the availability of the responses is determined according to y_{1i} and y_{2i}, it follows that the data identify the following possible events: (i) individuals who do not pass the first selection mechanism and thus y_{2i} and y_{3i} are not observed; (ii) individuals who pass the first selection mechanism but do not pass the second one and thus y_{3i} is not observed; (iii) individuals who pass both selection mechanisms and $y_{3i} = 0$; and (iv) individuals who pass both selection mechanisms and $y_{3i} = 1$.

The log-likelihood can therefore be expressed as

$$\ell = \sum_{i=1}^{n} \{(1-y_{1i})\log(p_{0i}) + y_{1i}(1-y_{2i})\log(p_{10i}) + y_{1i}y_{2i}(1-y_{3i})\log(p_{110i}) + y_{1i}y_{2i}y_{3i}\log(p_{111i})\}, \quad (4.6)$$

where $p_{\bar{e}_1}$, $p_{\bar{e}_1\bar{e}_2}$ and $p_{\bar{e}_1\bar{e}_2\bar{e}_3}$, for $\bar{e}_m \in \{0,1\}$, $\forall m$, are defined in Section 2.3. Since (4.6) is structured differently from the function of the trivariate model discussed in the previous chapter, it follows that its respective score and Hessian components need to be modified accordingly. Analytical derivative information can be obtained via expressions (2.9) and (2.10) and Propositions 2.3.2 and 2.3.3 in Section 2.3.1. Nevertheless, the estimation framework proposed in the previous chapter will be unaffected by such changes.

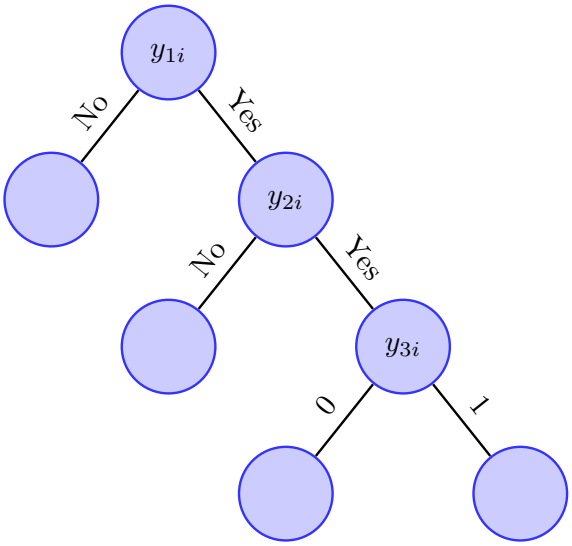

Figure 4.1: Diagram describing data affected by double sample selection rules. y_{1i} and y_{2i} correspond to the first and second selection mechanisms, while y_{3i} refers to the outcome of interest.

The model can be fitted using `SemiParTRIV()`/`gjrm()`, that is

```
out <- SemiParTRIV(formula = f.l, data = dat, Model = "TSS")
```

where `TSS` stands for the trivariate probit model with double sample selection, and `f.l` and `dat` have been previously defined.

4.3.3 Estimating the overall mean

An important quantity to estimate in this context is the prevalence or overall mean of the outcome. In surveys, prevalence estimates can be computed as a weighted average of individual predicted values with survey weights $\tilde{\mathrm{w}}_i$:

$$\hat{\mathbb{P}}(y_3 = 1) \;=\; \frac{\sum_{i=1}^{n} \left\{ \tilde{\mathrm{w}}_i \hat{\mathbb{P}}(y_{3i} = 1 | \mathbf{v}_{3i}^\top, \mathbf{L}_{3i}^\top, y_{2i}) \right\}}{\sum_{i=1}^{n} \tilde{\mathrm{w}}_i}, \tag{4.7}$$

where $\hat{\mathbb{P}}(y_{3i} = 1 | \mathbf{v}_{3i}^\top, \mathbf{L}_{3i}^\top, y_{2i}) = \Phi(\hat{\eta}_{3i})$. A corresponding CI can be derived using posterior simulation using the distributional result given in Section 4.2.4. Expression (4.7) with corresponding 95% CI can be computed using

```
prev(out)
```

4.3.4 Reducing the computational burden

In a double selection context, gains in speed can be obtained by reducing the computational time needed to evaluate ℓ, $\boldsymbol{g}$ and $\boldsymbol{\mathcal{H}}$ during the optimization process. This can be achieved by using three main indexes in the algorithm: (a) the first index indicates whether an individual passes the first selection mechanism; (b) the second index represents whether an individual passes the second selection mechanism; and (c) the third index relates only to the participants. By doing so, the log-likelihood in (4.6) can be re-expressed as a sum over three disjoint subsets of a sample: one for the observations who do not pass the first selection mechanism, one for the observations who pass the first selection mechanism but do not pass the second one, and the other for the remaining observations. That is,

$$\begin{aligned} \ell \;=\; & \sum_{i=1}^{n_1} \{(1 - y_{1i}) \log(p_{0i})\} + \sum_{i=n_1+1}^{n_2} \{y_{1i}(1 - y_{2i}) \log(p_{10i})\} + \\ & \sum_{i=n_2+1}^{n} \{y_{1i} y_{2i}(1 - y_{3i}) \log(p_{110i}) + y_{1i} y_{2i} y_{3i} \log(p_{111i})\}, \end{aligned}$$

where n_1 denotes the number of observations that do not pass the first stage, $n_2 - n_1$ indicates the number of observations who pass the first stage but not the second, $n - (n_1 + n_2)$ is the number of observed outcomes and n the total number of observations. Therefore instead of computing each component in ℓ for each i, we evaluate each component according to the individual's indexes. This is practically more efficient and hence the computation of the log-likelihood and related quantities is less expensive.

4.4 The endogenous-sample selection model

4.4.1 Model specification

Let $(y_{1i}, y_{2i}, y_{3i}) \in \{0, 1\}$, where y_{1i} characterizes whether or not an observation for the outcome variable y_{3i} is observed and y_{2i} is endogenous to y_{3i}. The target is to estimate a model controlling for the potential biases surrounding both non-random sample selection and endogeneity of the treatment. The situations considered are depicted in Figures 4.2 and 4.3. In the former, outcome y_{3i} and treatment y_{2i} are observed only if the individual passes the selection stage (i.e., $y_{1i} = 1$), otherwise both y_{2i} and y_{3i} are labeled as missing. In the latter, outcome y_{3i} is observed only if the individual passes the selection stage while information on y_{2i} is available even if the individual does not pass the first stage. Note that the diagrams depict situations in which an endogenous-sample selection model can be employed; importantly, the availability of y_{2i} for non-participants depends on the study at hand. In both cases, endogeneity of the treatment and non-random sample selection can be addressed using a trivariate model with partial observability. The model consists of the selection equation that indicates whether the individual takes part in the study, an equation controlling for the endogenous nature of the binary treatment and the outcome equation for the binary outcome. Using the latent variable representation,

the model for which y_{2i} is not always observed can be expressed as

$$\begin{aligned} y_{1i}^* &= \mathbf{v}_{1i}^\top \boldsymbol{\gamma}_1 + \mathbf{L}_{1i}^\top \boldsymbol{\alpha}_1 + \varepsilon_{1i} \\ y_{2i}^* &= \left\{ \mathbf{v}_{2i}^\top \boldsymbol{\gamma}_2 + \mathbf{L}_{2i}^\top \boldsymbol{\alpha}_2 + \varepsilon_{2i} \right\} \times y_{1i} \\ y_{3i}^* &= \left\{ \psi y_{2i} + \mathbf{v}_{3i}^\top \boldsymbol{\gamma}_3 + \mathbf{L}_{3i}^\top \boldsymbol{\alpha}_3 + \varepsilon_{3i} \right\} \times y_{1i}, \end{aligned}$$

where y_{2i} and y_{3i} are determined as

$$y_{mi} = \begin{cases} 1 & \text{if} \quad (y_{mi}^* > 0 \ \ \& \ \ y_{1i} = 1) \\ 0 & \text{if} \quad (y_{mi}^* < 0 \ \ \& \ \ y_{1i} = 1) \\ - & \text{if} \quad y_{1i} = 0 \end{cases},$$

$\forall m = 1, 2$, if data follow the process shown in Figure 4.2. The model for which y_{2i} is always observed can be expressed as

$$\begin{aligned} y_{1i}^* &= \mathbf{v}_{1i}^\top \boldsymbol{\gamma}_1 + \mathbf{L}_{1i}^\top \boldsymbol{\alpha}_1 + \varepsilon_{1i} \\ y_{2i}^* &= \mathbf{v}_{2i}^\top \boldsymbol{\gamma}_2 + \mathbf{L}_{2i}^\top \boldsymbol{\alpha}_2 + \varepsilon_{2i} \\ y_{3i}^* &= \left\{ \psi y_{2i} + \mathbf{v}_{3i}^\top \boldsymbol{\gamma}_3 + \mathbf{L}_{3i}^\top \boldsymbol{\alpha}_3 + \varepsilon_{3i} \right\} \times y_{1i}, \end{aligned}$$

where

$$y_{2i} = \begin{cases} 1 & \text{if} \quad (y_{2i}^* > 0 \ \ \& \ \ (y_{1i} = 1 \text{ or } y_{1i} = 0)) \\ 0 & \text{if} \quad (y_{2i}^* < 0 \ \ \& \ \ (y_{1i} = 1 \text{ or } y_{1i} = 0)) \end{cases},$$

and

$$y_{3i} = \begin{cases} 1 & \text{if} \quad (y_{3i}^* > 0 \ \ \& \ \ y_{1i} = 1) \\ 0 & \text{if} \quad (y_{3i}^* < 0 \ \ \& \ \ y_{1i} = 1) \\ - & \text{if} \quad y_{1i} = 0 \end{cases},$$

when data follow the process shown in Figure 4.3. Parameter ψ indicates the effect of the treatment on the outcome. The errors are assumed to follow a trivariate normal distribution with zero mean and variance-covariance equal to $\boldsymbol{\Sigma}$. The same identification arguments discussed in Section 4.2.2 apply here as well.

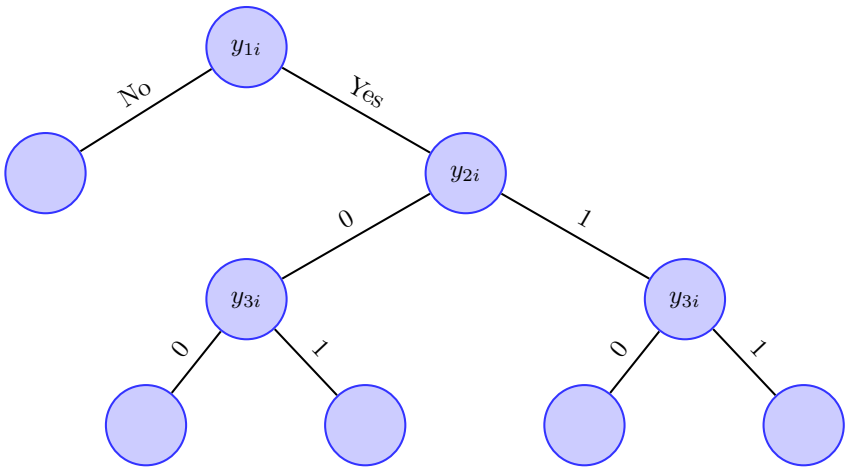

Figure 4.2: Diagram describing data affected by non-random sample selection and endogenity of a treatment. y_{1i} corresponds to the selection mechanism, y_{2i} denotes the binary endogenous variable and y_{3i} is the binary outcome. Variable y_{2i} is not available for non-participants.

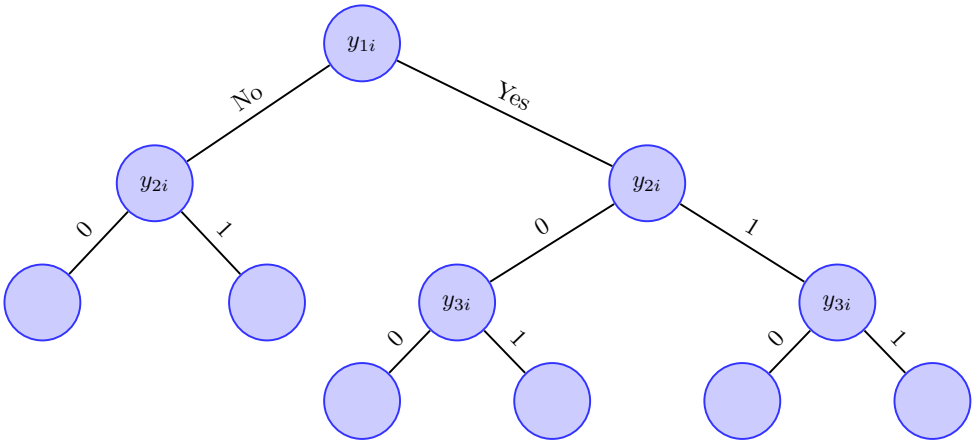

Figure 4.3: Diagram describing data affected by non-random sample selection and endogenity of a treatment. y_{1i} corresponds to the selection mechanism, y_{2i} denotes the binary endogenous variable and y_{3i} is the binary outcome. Variable y_{2i} is available for non-participants.

4.4.2 Parameter estimation

As a consequence of the missing outcomes, the construction of the log-likelihood function is analogous to the one presented in Section 4.3. In this case we deal only with single selectivity and since the availability of y_{2i} depends on the study at hand,

the data identify either five or six possible events. In the first case, we have: (i) individual does not pass the selection mechanism and thus y_{2i} and y_{3i} are not observed; (ii) individual participates in the survey and $(y_{2i}, y_{3i}) = (1, 1)$; (iii) individual participates and $(y_{2i}, y_{3i}) = (1, 0)$; (iv) individual participates and $(y_{2i}, y_{3i}) = (0, 0)$; (v) individual participates and $(y_{2i}, y_{3i}) = (0, 1)$. In the second case, (i) individual does not pass the selection step, $y_{2i} = 1$ and y_{3i} is not observed; (ii) individual does not pass the first stage, $y_{2i} = 0$ and y_{3i} is not observed; (iii) individual participates and $(y_{2i}, y_{3i}) = (1, 1)$; (iv) individual participates and $(y_{2i}, y_{3i}) = (1, 0)$; (v) individual participates and $(y_{2i}, y_{3i}) = (0, 0)$; (vi) individual participates and $(y_{2i}, y_{3i}) = (0, 1)$. The log-likelihood function can be expressed for the former case as

$$\begin{aligned} \ell \;=\; & \sum_{i=1}^{n} \{(1 - y_{1i}) \log(p_{0i}) + y_{1i} y_{2i} y_{3i} \log(p_{111i}) + y_{1i} y_{2i} (1 - y_{3i}) \log(p_{110i}) + \\ & y_{1i} (1 - y_{2i}) (1 - y_{3i}) \log(p_{100i}) + y_{1i} (1 - y_{2i}) y_{3i} \log(p_{101i})\} \,, \end{aligned} \tag{4.8}$$

and for the latter case as

$$\begin{aligned} \ell \;=\; & \sum_{i=1}^{n} \{(1 - y_{1i}) y_{2i} \log(p_{01i}) + (1 - y_{1i}) (1 - y_{2i}) \log(p_{00i}) + y_{1i} y_{2i} y_{3i} \log(p_{111i}) + \\ & y_{1i} y_{2i} (1 - y_{3i}) \log(p_{110i}) + y_{1i} (1 - y_{2i}) (1 - y_{3i}) \log(p_{100i}) + \\ & y_{1i} (1 - y_{2i}) y_{3i} \log(p_{101i})\} \,. \end{aligned} \tag{4.9}$$

In particular, ℓ is equal to (4.8) if the endogenous variable y_{2i} is not available after individual's non-participation and ℓ is equal to (4.9) if y_{2i} is observed under non-participation. Similar to the double sample selection model, ℓ (and thus $\boldsymbol{g}$ and $\boldsymbol{\mathcal{H}}$) has a different structure from the log-likelihood function of trivariate model with fully observed responses. Analytical derivative information for the models can be obtained using the expressions (2.9) and (2.10) and Propositions 2.3.2 and 2.3.3 given in Section 2.3.1, while the model can be fitted using the PMLE approach discussed in the previous chapters.

Using `SemiParTRIV()`/`gjrm()`, the model can be used as follows

```
out <- SemiParTRIV(formula = f.l, data = dat, Model = ESS)
```

where `ESS` stands for the endogenous-sample selection model, and `f.l` and `dat` have been previously defined. The function uses by default log-likelihood function (4.8). The ATEs and prevalence estimates with corresponding CIs can be computed as already discussed in Sections 4.2.4 and 4.3.3.

4.4.3 Reducing the computational burden

The computational time required for estimating the model can be reduced by employing the technique discussed in Section 4.3.4. Since we deal with single selectivity here, we re-express ℓ, $\boldsymbol{g}$ and $\boldsymbol{\mathcal{H}}$ based only on an index which in this case indicates whether an individual participates in the study. The log-likelihood function of the model presented in Figures 4.2 and 4.3 can therefore be written as

$$\begin{aligned}\ell &= \sum_{i=1}^{n_1}\{(1-y_{1i})\log(p_{0i})\} + \sum_{i=n_1+1}^{n}\{y_{1i}y_{2i}y_{3i}\log(p_{111i}) + y_{1i}y_{2i}(1-y_{3i})\log(p_{110i}) + \\ &\quad y_{1i}(1-y_{2i})(1-y_{3i})\log(p_{100i}) + y_{1i}(1-y_{2i})y_{3i}\log(p_{101i})\},\end{aligned}$$

and

$$\begin{aligned}\ell &= \sum_{i=1}^{n_1}\{(1-y_{1i})y_{2i}\log(p_{01i}) + (1-y_{1i})(1-y_{2i})\log(p_{00i})\} + \\ &\quad \sum_{i=n_1+1}^{n}\{y_{1i}y_{2i}y_{3i}\log(p_{111i}) + y_{1i}y_{2i}(1-y_{3i})\log(p_{110i}) + \\ &\quad y_{1i}(1-y_{2i})(1-y_{3i})\log(p_{100i}) + y_{1i}(1-y_{2i})y_{3i}\log(p_{101i})\}.\end{aligned}$$

The first subset in both expressions corresponds to non-participants and the second one to participants; n_1 denotes the number of individuals in the former subset, while the number of participants is $n - n_1$. In a similar way, this also applies to the components in $\boldsymbol{g}$ and $\boldsymbol{\mathcal{H}}$.

4.5 Simulations and real data illustration

This secton has two aims: assessing the empirical effectiveness of the double sample selection model via simulation, and applying the recursive trivariate model to a case study.

4.5.1 Simulation study

The simulation study employs the DGP3 settings described in Appendix B.2.1, where the model specification used to generate the data includes two ERs, v_{2i} and v_{3i}. The equations, in `R` notation, are specified as

```
eqn1 ~ v1 + s(z1) + v2 + v3; eqn2 ~ v1 + s(z1) + v2; eqn3 ~ v1 + s(z1)
```

where `v1` is a binary variable, `v2` and `v3` denote the binary ERs and `s(z1)` defines a smooth function of the continuous covariate `z1`. We employed the PMLE approach discussed in the previous chapter, where the coefficients of the spline bases are penalized and the correlations are also penalized using a Lasso approach.

Figure 4.4 shows the estimated smooth curves obtained from 250 replicates using sample sizes of 5000 and 15000. In general, the method appears to be effective in recovering the true functions. The variability that characterizes the curve estimates for $n = 5000$ does not come as a surprise given the considerable loss of information in a double selection context. Table 4.1 shows the percentage biases and RMSEs of the correlation coefficients and prevalence estimates. The experiment shows that the estimated correlation coefficients are affected by some bias, especially at $n = 5000$. This is not unexpected given the complexity of the model and substantial loss of information that a double selection process implies. Overall, biases and RMSEs reduce as n increases. As for the prevalence estimates, both bias and RMSE become negligible as n grows (see also Figure 4.5 which shows that, as the sample size increases, the prevalence estimates approach their true value). For model comparison purposes, we also present the percentage biases and RMSEs of the correlations and prevalence estimates obtained using the unpenalized approach (i.e., no penalization

on the correlation parameters was imposed). The results presented in Table 4.2 show overall that the bias and RMSE of the estimates are higher, compared to the corresponding quantities in Table 4.1. This suggests that the parameters can better be estimated when they are penalized.

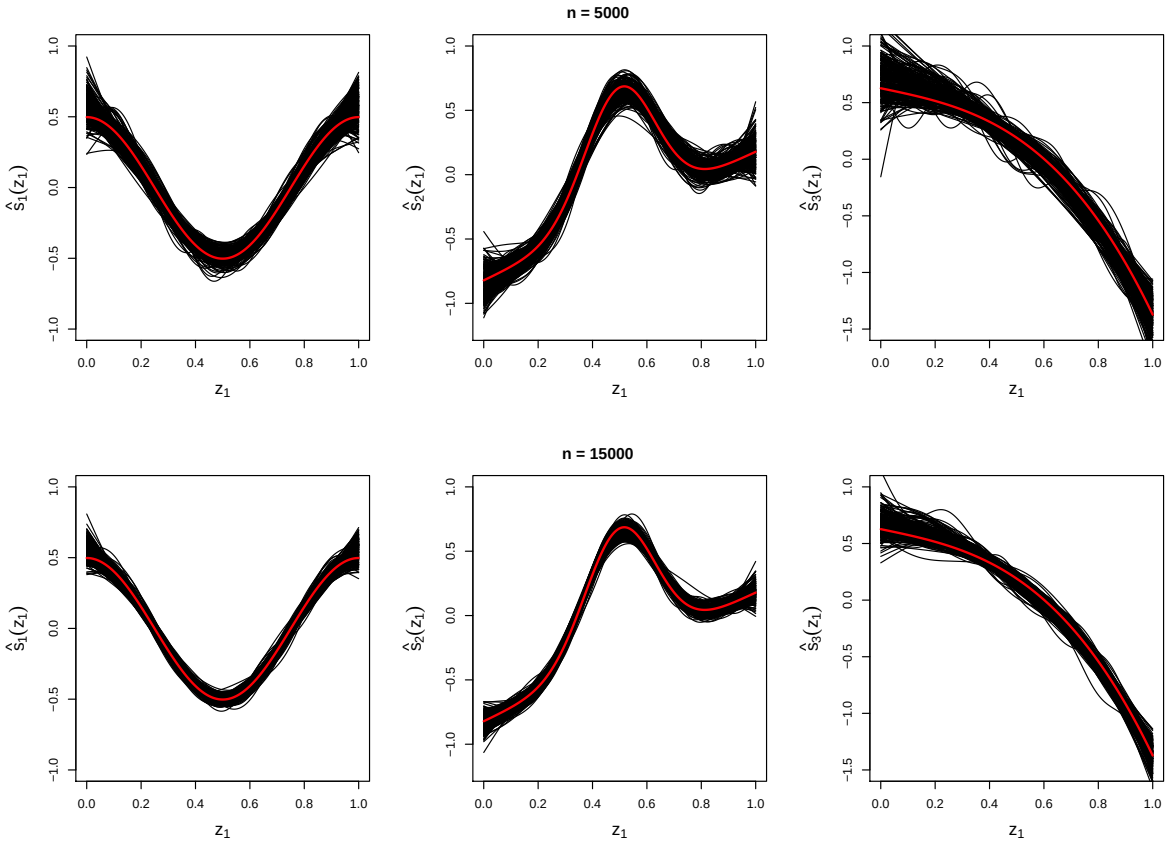

Figure 4.4: Estimated smooth functions for $s_1(z_1)$, $s_2(z_1)$ and $s_3(z_1)$ obtained applying `SemiParTRIV()`/`gjrm()` on 250 simulated datasets. The first row shows the estimated curves obtained from samples of 5000 observations, whereas those in the second row correspond to samples of 15000 observations. The black lines represent the estimated smooth functions over all replicates and the red solid lines show the true functions.

4.5.2 Labor force data analysis

Chronic diseases are considered to be important conditions in the developing and developed countries. In 1997, 124 million people worldwide were estimated to have diabetes (Amos et al., 1997), while one in five adults in the U.S. were found to have multiple chronic diseases (e.g., Ward & Schiller, 2013; Ward et al., 2014).

	$n = 5000$		$n = 15000$	
Estimator	Bias (%)	RMSE	Bias (%)	RMSE
$\hat{\vartheta}_{12}$	20.08	0.0875	9.16	0.0612
$\hat{\vartheta}_{13}$	15.39	0.1179	14.86	0.0892
$\hat{\vartheta}_{23}$	-11.28	0.1499	-6.74	0.0990
$\hat{\mathbb{P}}(y_3 = 1)$	6.29	0.0151	0.93	0.0066

Table 4.1: Percentage biases and root mean squared errors (RMSEs) of the correlation estimates and prevalence estimate obtained applying the double sample selection model to 250 datasets simulated according to DGP3, where the correlation parameters are penalized via the Lasso penalty.

	$n = 5000$		$n = 15000$	
Estimator	Bias (%)	RMSE	Bias (%)	RMSE
$\hat{\vartheta}_{12}$	25.23	0.0989	17.10	0.0705
$\hat{\vartheta}_{13}$	21.19	0.1284	16.39	0.0921
$\hat{\vartheta}_{23}$	-14.27	0.1783	-7.80	0.1115
$\hat{\mathbb{P}}(y_3 = 1)$	6.97	0.0166	0.32	0.0064

Table 4.2: Percentage biases and root mean squared errors (RMSEs) of the correlation estimates and prevalence estimate obtained applying the double sample selection model to 250 datasets simulated according to DGP3, where the correlation parameters are not penalized.

The prevalence of multiple chronic diseases has been increasing over the past decade (Ward & Schiller, 2013). Whiting et al. (2011) suggest that, worldwide, people living with diabetes will increase by 50.7% by the year 2030.

Chronic health problems do not only affect the health care system, but have also a negative impact on labour force participation. Among U.S. adults, having multiple (≥ 2) chronic conditions reduces the employment probability by $11 - 29\%$ (Ward, 2015). Individual chronic diseases, such as diabetes (Bastida & Pagán, 2002; Tunceli et al., 2005; Minor, 2011) and rheumatoid arthritis (Kessler et al., 2008), were found to be associated with work-related outcomes. Treating the incidence of chronic illness as exogenous may lead to imprecise estimates. For instance, diseases such as diabetes and heart disease may be correlated through unobserved covariates that are also related to labour force participation. That is, personal motivation is positively associated with labour force participation, motivation may influence lifestyle choices

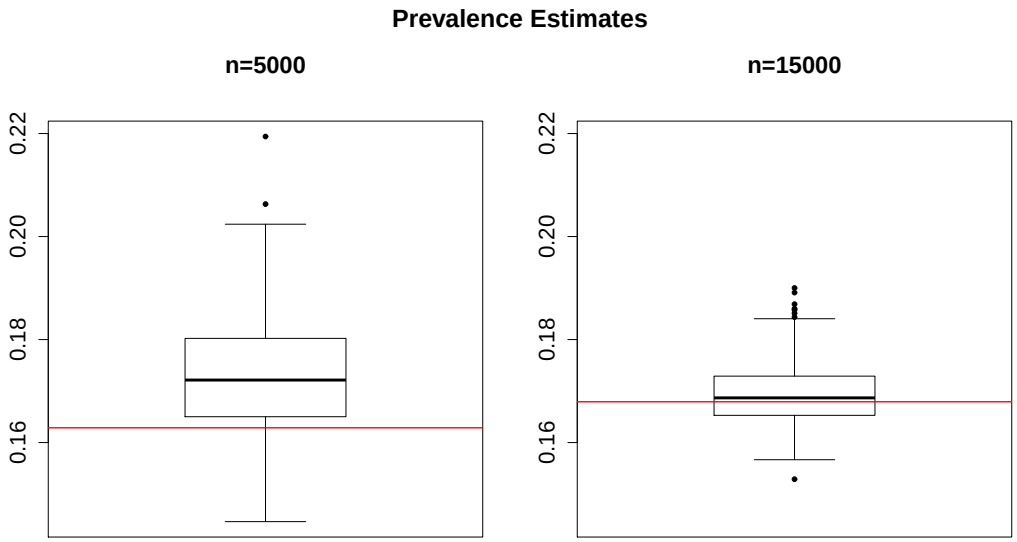

Figure 4.5: Boxplots corresponding to the prevalence estimates of the semi-parametric double sample selection model for sample sizes equal to 5000 and 15000. Results are obtained from 250 replications of DGP3 and the horizontal red lines represent the true prevalence.

and a healthy lifestyle may decrease the probability of having diabetes (Latif, 2009). Moreover if only one chronic disease is accounted for, say diabetes, then this means that we assume that other illnesses do not affect the decision to participate in the labour force and are uncorrelated with diabetes; studies have shown that people with diabetes are at an increased risk of having heart disease (e.g., Harris, 1998; Black et al., 1999; AIHW, 2006). Thus, joint estimation of multiple chronic diseases and work outcomes can lead to improved estimation results and inference.

In this section we jointly analyse diabetes, heart disease and labour force participation conditional on flexible functions on covariates and account for the potential endogeneity of diabetes and heart disease for the decision to work, and the potential endogeneity of diabetes. The empirical analysis was carried out using the endogenous trivariate probit model described in Section 4.2.

Data and Empirical Analysis

The study examines data from the 2012 Medical Expenditure Panel Survey (MEPS), which are collected by the Agency for Healthcare Research and Quality (AHRQ). The survey considered a sample of $38,974$ individuals containing a nationally representative sample of the U.S. civilian non-institutionalized population and providing information of individuals health status, demographic and socio-economic characteristics, employment and satisfaction with health care. Individuals were part of one of the two MEPS panels: Rounds 3, 4, and 5 of Panel 16 or Rounds 1, 2, and 3 of Panel 17. Here we focus on Rounds 4 and 2 (R4/2) and we aim at estimating the effect of two major chronic diseases on the probability of labour force participation. Individuals who did not have a complete set of the variables or aged < 17 were excluded from the original sample. After exclusions, the final sample includes $23,295$ observations.

Tables 4.3 and 4.4 report empirical bivariate densities of the dependent variables of interest. Employment status refers to whether the person was employed during the round. As shown in the tables, the majority of those who are employed have not been diagnosed with diabetes and/or heart disease, while only few of them have been diagnosed with the disease(s).

	Diabetes		
Employment Status	0	1	Total
Employed	13648 (58.59%)	916 (3.93%)	14564 (62.52%)
Non-employed	7290 (31.29%)	1441 (6.19%)	8731 (37.48%)
Total	20938 (89.88%)	2357 (10.12%)	23295 (100.00%)

Table 4.3: Empirical density for diabetes and employment status. The proportions in brackets show the corresponding proportions in the sample.

Our model specification follows the studies of Harris (2009) and Zhang et al. (2009) who estimate recursive simultaneous probit models accounting for the potential endogeneity of the incidence of chronic conditions. The exact definition of the

Employment Status	Heart Disease		Total
	0	1	
Employed	13377 (57.42%)	1187 (5.10%)	14564 (62.52%)
Non-employed	6669 (28.63%)	2062 (8.85%)	8731 (37.48%)
Total	20046 (86.05%)	3249 (13.95%)	23295 (100.00%)

Table 4.4: Observed density for heart disease and employment status. The proportions in brackets show the corresponding proportions in the sample.

variables used in the analysis is given in Table 4.5. Note that other chronic diseases are excluded from our analysis due to the fact that diabetes and heart disease are the most common physical chronic diseases and share common factors that are not clearly related to other chronic diseases such as cancer or asthma.

Variable	Explanation
`diab`	whether individual has been diagnosed with diabetes
`heartd`	whether individual has been diagnosed as having heart disease
`empl`	whether individual is currently employed
`age`	age in years
`educ`	classification of education (0: less than 1st Grade, ..., 16: Master, Doctorate or Other Professional Degree)
`usborn`	whether individual was born in US
`marital`	marital status (1: married, ..., 10: separated in round)
`engspk`	whether individual is comfortable conversing in English
`region`	region the respondent was living (1: north-east, ..., 4: west)
`health`	perceived health status (1: excellent, ..., 5: poor)
`hyper`	whether individual has been diagnosed as having high blood pressure
`cholest`	whether individual has been diagnosed as having high cholesterol
`smok`	whether individual currently smokes

Table 4.5: Description of the variables obtained in Round 4 of Panel 16 and Round 2 of Panel 17 in the MEPS dataset.

Our approach allows for the semi-parametric estimation of the covariate-response

relationships; thus we define the model as

$$
\begin{aligned}
\texttt{diab}_i^* &= \beta_{11} + \beta_{12}\texttt{educ}_i + \beta_{13}\texttt{usborn}_i + \beta_{14}\texttt{marital}_i + \beta_{15}\texttt{engspk}_i + \beta_{16}\texttt{region}_i + \\
&\quad \beta_{17}\texttt{health}_i + s_{11}(\texttt{age}_i) + \beta_{18}\texttt{hyper}_i + \beta_{19}\texttt{smok}_i + \beta_{1,10}\texttt{cholest}_i \\
\texttt{heartd}_i^* &= \beta_{21} + \psi_1\texttt{diab}_i + \beta_{22}\texttt{educ}_i + \beta_{23}\texttt{usborn}_i + \beta_{24}\texttt{marital}_i + \beta_{25}\texttt{engspk}_i + \\
&\quad \beta_{26}\texttt{region}_i + \beta_{27}\texttt{health}_i + s_{21}(\texttt{age}_i) + \beta_{28}\texttt{hyper}_i + \beta_{29}\texttt{smok}_i + \\
&\quad \beta_{2,10}\texttt{cholest}_i \\
\texttt{empl}_i^* &= \beta_{31} + \psi_1\texttt{diab}_i + \psi_2\texttt{heartd}_i + \beta_{32}\texttt{educ}_i + \beta_{33}\texttt{usborn}_i + \beta_{34}\texttt{marital}_i + \\
&\quad \beta_{35}\texttt{engspk}_i + \beta_{36}\texttt{region}_i + \beta_{37}\texttt{health42}_i + s_{31}(\texttt{age}_i) + \beta_{38}\texttt{hyper}_i + \\
&\quad +\beta_{39}\texttt{smok}_i + \beta_{3,10}\texttt{cholest}_i,
\end{aligned}
$$

where s_{m1} are smooth functions of $\texttt{age}_i$ represented using thin plate regression splines with twenty bases and second order penalties, $\forall m = 1, 2, 3$. Note that since the available data do not provide any valid ERs, the model specification does not include any of these variables. As described previously, however, identification and estimation of the model could be significantly improved using suitable ERs; thus one should be cautious when interpreting the results of this study. Next paragraph presents the most important results.

Results and Inference

Figure 4.6 shows the non-linear effects of `age` for the treatment and outcome equations. The incidence of chronic diseases is positively related to age indicating greater risk as individuals become older. However, the effect of age on the incidence of diabetes decreases after 70 years of age; a similar result was found in Zhang et al. (2009) where individuals in younger age bands ($50 - 64$ years of age) were more likely to have mental illnesses than those in the oldest band of $60 - 64$. As expected, labor force participation increases rabidly with age up to $28 - 30$ years after which the effect is almost steady up to around $50 - 55$ years and it decreases for people older than 60 years. The zero flat line is not contained within the CIs of the smooths,

indicating that `age` is a significant predictor for the responses. Table 4.6 presents

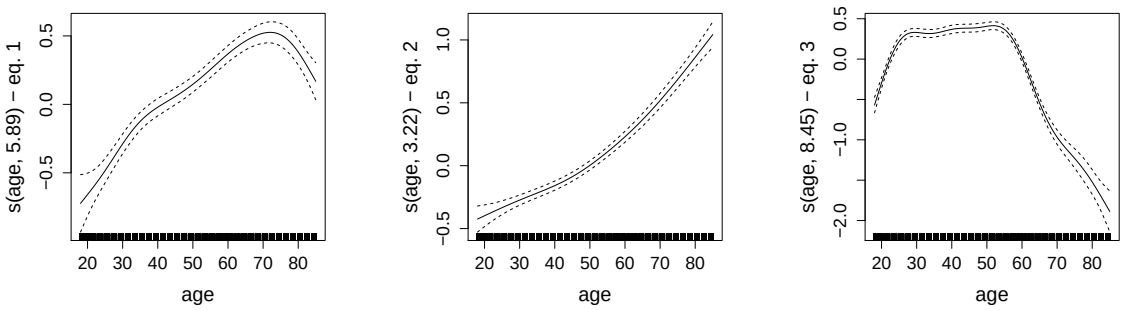

Figure 4.6: Function estimates obtained applying the endogenous trivariate model using the proposed fitting method. Dashed lines represent 95% Bayesian point-wise CIs. The first two curves correspond to the smooth term of `age` in the equations describing `diab` (eq. 1) and `heartd` (eq. 2), while the last one to the equation describing `empl` (eq. 3). The effective degrees of freedom are reported into brackets in the y-axis caption.

the correlation parameter estimates obtained when applying the semi-parametric recursive bivariate probit (SRBP) model and the semi-parametric recursive trivariate probit (SRTP) model with their corresponding 95% CIs. The estimated correlation between the two diseases ($\hat{\vartheta}_{12}$) and between labor force participation and heart disease ($\hat{\vartheta}_{23}$) is moderate, while the low value for $\hat{\vartheta}_{13}$ indicates that diabetes may not have a strong impact of labor force participation. The estimated ATEs for SRTP can be interpreted as follows. The probability of having been diagnosed as having heart disease increases by 6.40% for individuals who have been diagnosed with diabetes, while the probability that individual is currently employed decreases by 2.38% and 30% for people who have diabetes and heart disease, respectively. Although the estimates obtained from the two models are close to each other, there appears to be a slight overestimation of the parameters ϑ_{12} (SATE_{12}) and ϑ_{23} (SATE_{23}) while the inconsistency in ϑ_{13} (SATE_{13}) is more noticeable. This suggests that a trivariate system may better account for the dependencies between the treatment and outcome variables, hence providing more accurate results.

Estimator	SRBP	SRTP
$\hat{\vartheta}_{12}$	-0.144 $(-0.284, 0.032)$	-0.119 $(-0.258, 0.067)$
$\hat{\vartheta}_{13}$	0.147 $(0.032, 0.288)$	0.005 $(-0.133, 0.140)$
$\hat{\vartheta}_{23}$	0.409 $(0.280, 0.514)$	0.403 $(0.293, 0.514)$
$\widehat{\text{SATE}}_{12}$	7.39 $(2.21, 13.63)$	6.40 $(2.22, 13.34)$
$\widehat{\text{SATE}}_{13}$	-11.34 $(-21.44, -1.23)$	-2.38 $(-10.33, 6.41)$
$\widehat{\text{SATE}}_{23}$	-30.3 $(-37.1, -24.3)$	-30.0 $(-35.3, -23.4)$

Table 4.6: Estimates of the correlation coefficients and ATEs (in %) obtained applying the semi-parametric recursive bivariate probit (SRBP) model and the semi-parametric recursive trivariate probit (SRTP) model on the MEPS data. $\widehat{\text{SATE}}_{zk}$ corresponds to the estimated average treatment effect obtained using the z^{th} equation as the treatment equation and the k^{th} equation as the outcome equation, $\forall z = 1, 2, k = 2, 3, z \neq k$. 95% Bayesian CIs were obtained using 100 coefficient vectors simulated from the posterior distribution of the estimated model parameters.

4.6 Conclusions

In this chapter, we have discussed several models that can be derived from the general framework introduced in the previous chapter. These can deal with data suffering from endogeneity and/or non-random sample selection. The models include parametric and non-parametric components, allowing researchers to achieve a higher degree of flexibility in empirical modelling. We have provided inferential tools and discussed briefly model's identification. A technique for reducing computing time was also discussed. Parameter estimation of all models is achieved using the generic PMLE approach discussed in Chapter 3. We have also developed the necessary computational procedures which are incorporated in the `R` package `GJRM`.

A Monte Carlo experiment for the double sample selection model was conducted, showing the promising performance of the model. Using the endogenous trivariate model, we examined the effect of two chronic diseases on labor force participation using the 2012 MEPS dataset. The results have shown that both diseases affect negatively individual's employment status, while having diabetes increases the probability of having been diagnosed as having heart disease.

In the next chapter we aim at accommodating link functions other than probit.

That is, instead of specifying the marginal distributions of the three responses using the standard normal distribution we could use the logistic distribution for instance.

Chapter 5

Extending the additive trivariate binary model to non-probit margins

In this chapter, we consider the simultaneous estimation of three binary regressions using a three-equation system in which the trivariate distribution is defined by the Gaussian copula with arbitrary margins. In particular, we extend the trivariate additive probit model of Chapters 2, 3 and 4 to allow for arbitrary link functions. The estimation framework (and hence `SemiParTRIV()`/`gjrm()`) is extended accordingly to incorporate such a feature.

5.1 Introduction

All models considered so far use the probit transformation for the probabilities, but other choices are also possible. In fact, any transformation that maps probabilities into the real line could be used to produce a trivariate model, as long as the transformation is one-to-one, continuous and differentiable. This chapter extends the material presented in the previous chapters to allow for the flexible specification of the marginal links. Specifically, we employ the logistic and Gumbel distributions which give rise to the logit and complementary log-log links, respectively. Together

with the probit link, they are the most commonly used links in GLMs/GAMs for binary responses. These additional links are used extensively in numerous disciplines, including the medical and social sciences. In clinical research logit models are widely used as they provide direct information about which treatment has the best odds of benefiting a patient, for instance. Complementary log-log models have important applications in survival analysis where they can, for example, provide a clear insight into the relative reduction of risk for death or progression.

In general, the expected value of a response of interest, y_{mi}, conditioned to a set of explanatory variables (contained in linear predictor η_{mi}) can be represented through a generalized model using the so-called *link function* $g_m : [0,1] \to \mathbb{R}$ which links the random (y_{mi}) and systematic (η_{mi}) components of the model (McCullagh & Nelder, 1989, Ch. 2.2), $\forall m = 1, \ldots, M$. In the univariate case, this can be expressed as

$$\mathbb{E}(y_{mi}) = \mu_{mi} = g_m^{-1}(\eta_{mi}),$$

or $g_m(\mu_{mi}) = \eta_{mi}$, where η_{mi} is an additive predictor (made up of regression coefficients and covariates as described in Section 2.2). The link function specifies a non-linear transformation between the linear predictor and the mean of the distribution function. In Chapters 2, 3 and 4, we employed the probit link $g_m(\mu_{mi}) = \Phi^{-1}(\mu_{mi})$, $\forall m = 1, 2, 3$.

By using the fact that the inverse of any continuous univariate cdf can be used for the link g_m, we re-express the link function as

$$\eta_{mi} \;=\; \mathrm{F}_m^{-1}(\mu_{mi}), \tag{5.1}$$

or $\mu_{mi} = \mathrm{F}_m(\eta_{mi})$, where $\mathrm{F}_m : \mathbb{R} \to [0,1]$ is any univariate cdf. The logit and complementary log-log links can be specified as

$$\eta_{mi} = \log\left(\frac{\mu_{mi}}{1-\mu_{mi}}\right) \quad \text{and} \quad \eta_{mi} = \log\left(-\log(1-\mu_{mi})\right),$$

where μ_{mi} corresponds to the cdf of the logistic and Gumbel distribution defined $\forall m$ as

$$\mu_{mi} = \frac{\exp(\eta_{mi})}{1+\exp(\eta_{mi})} \quad \text{and} \quad \mu_{mi} = 1 - \exp(-\exp(\eta_{mi})),$$

respectively.

The probit, logit and complementary log-log functions share the feature of mapping the unit interval onto the real line. Looking at Figure 5.1, which depicts the three links, we observe that all functions are increasing, continuous and differentiable over $0 < \mathbb{P}(y_{mi} = 1) < 1$ and they are almost linearly related over $0.1 < \mathbb{P}(y_{mi} = 1) < 0.9$. However, when the probability of a successful outcome is extremely small or large, the linear relationship does not hold. In contrast to the complementary log-log function, the logit and probit links are both symmetric around 0.

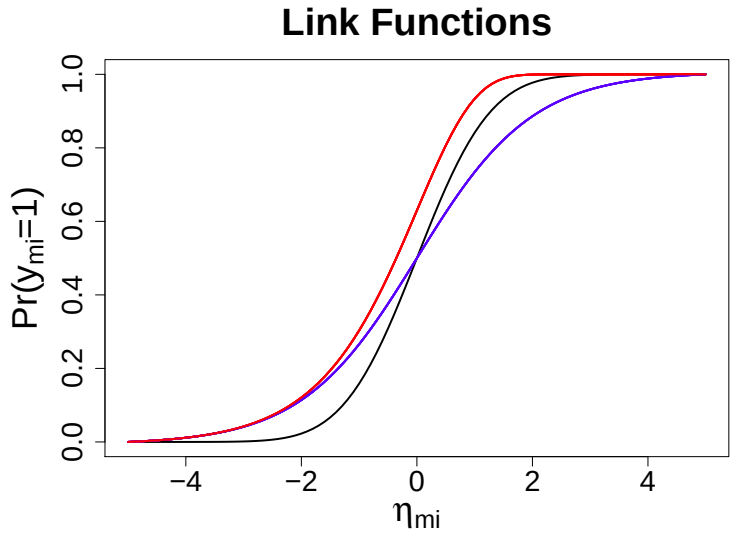

Figure 5.1: Probit (——), logit (——) and complementary log-log (——) functions. The y-axis corresponds to the probability of success $\mathbb{P}(y_{mi} = 1)$ and the x-axis denotes the generic m^{th} linear predictor η_{mi}.

In what follows, we allow for the logit and complementary log-log links into the trivariate models described in Chapters 2, 3 and 4, discuss some fitting details and present a simulation study examining the performance of the models. Conclusions

are drawn in Section 5.4.

5.2 Gaussian copula with arbitray margins

In order to employ arbitrary link functions in the model we need to re-express the univariate, bivariate and trivariate cdfs in such a way that each of the three margins belongs to a different distributional family. The univariate cdf has already been re-defined in (5.1), while the bivariate and trivariate cdfs can be re-expressed via a Gaussian copula. In general, a copula function can be described as a multivariate distribution function in which the marginal distributions may come from different families (Joe, 1997). This construction allows one to consider the marginal distributions and the dependence between them as two separate but related issues.

Suppose that $\tilde{\mathcal{C}}$ denotes a joint cdf whose support is contained in $[0,1]^3$ and whose one-dimensional margins are uniform (Dall'Aglio et al., 2012). Let $\mathrm{F}(\eta_{1i}, \eta_{2i}, \eta_{3i}) = \mathbb{P}(y^*_{1i} > 0, y^*_{2i} > 0, y^*_{3i} > 0)$ be a joint cdf and $\mathcal{U}_m^{-1} : (0,1) \to \mathbb{R}$ a quantile function. Then there exists a three-dimensional copula function $\tilde{\mathcal{C}} : [0,1]^3 \to [0,1]$ that represents the joint distribution function in terms of margins such that

$$\begin{aligned}
\tilde{\mathcal{C}}(\mu_{1i}, \mu_{2i}, \mu_{3i}) &= \tilde{\mathcal{C}}(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i})) \\
&= \mathrm{F}(\eta_{1i}, \eta_{2i}, \eta_{3i}) \\
&= \mathrm{F}\left(\mathcal{U}_1^{-1}(\mu_{1i}), \mathcal{U}_2^{-1}(\mu_{2i}), \mathcal{U}_3^{-1}(\mu_{3i})\right) \\
&= \mathrm{F}\left(\mathcal{U}_1^{-1}(\mathrm{F}_1(\eta_{1i})), \mathcal{U}_2^{-1}(\mathrm{F}_2(\eta_{2i})), \mathcal{U}_3^{-1}(\mathrm{F}_3(\eta_{3i}))\right), \qquad (5.2)
\end{aligned}$$

which satisfies the following conditions (Sklar, 1959)

(C.1) $\tilde{\mathcal{C}}\left(\mathrm{F}_1(\eta_{1i})), 1, 1\right) = \mathrm{F}_1(\eta_{1i})$, $\tilde{\mathcal{C}}\left(1, \mathrm{F}_2(\eta_{2i})), 1\right) = \mathrm{F}_2(\eta_{2i})$, $\tilde{\mathcal{C}}\left(1, 1, \mathrm{F}_3(\eta_{3i}))\right) = \mathrm{F}_3(\eta_{3i})$, $\forall \mathrm{F}_m(\eta_{mi}) \in [0,1]$ and $m \leq 3$;

(C.2) $\tilde{\mathcal{C}}\left(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i})\right) = 0$ if $\mathrm{F}_m(\eta_{mi}) = 0$ for any $m \leq 3$;

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

Condition (C.1) states that if the realizations of two variables are known each with

marginal probability one, then the joint probability of the three outcomes is the same as the probability of the remaining uncertain outcome. Condition (C.2) is sometimes referred to as the grounded property of a copula and states that the joint probability of all outcomes is zero if the marginal probability of any outcome is zero. Condition (C.3) means that the copula volume of any 3-dimensional interval is non-negative; in the bivariate setting, for instance, the volume between two points $[\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i})]$ and $\left[\tilde{\mathrm{F}}_1(\eta_{1i}), \tilde{\mathrm{F}}_2(\eta_{2i})\right]$, where $\mathrm{F}_m(\eta_{mi}) \geq \tilde{\mathrm{F}}_m(\eta_{mi})$, $\forall m = 1, 2$, takes the form $\tilde{\mathcal{C}}\left(\tilde{\mathrm{F}}_1(\eta_{1i}), \tilde{\mathrm{F}}_2(\eta_{2i})\right) - \tilde{\mathcal{C}}\left(\tilde{\mathrm{F}}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i})\right) - \tilde{\mathcal{C}}\left(\mathrm{F}_1(\eta_{1i}), \tilde{\mathrm{F}}_2(\eta_{2i})\right) + \tilde{\mathcal{C}}\left(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i})\right)$. A copula $\tilde{\mathcal{C}}$ is unique on the cartesian product of the ranges of the marginal cdfs $\mathbf{Ran}(\mathrm{F}_1(\eta_{1i})) \times \mathbf{Ran}(\mathrm{F}_2(\eta_{2i})) \times \mathbf{Ran}(\mathrm{F}_2(\eta_{3i}))$. The copula is unique if the margins $\mathrm{F}_m(\eta_{mi})$ are continuous, $\forall m$. Any copula lies always in the interval

$$\max\left\{\sum_{m=1}^{3} \mathrm{F}_m(\eta_{mi}) - 2, 0\right\} \leq \tilde{\mathcal{C}}\left(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i})\right) \leq \min\left\{\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i})\right\},$$

the so-called *Fréchet-Hoeffding* bounds. A desirable feature of a copula is that it should cover the sample space between the lower and upper bounds, and that as the correlation parameters approach the lower (upper) bound of its permissible ranges, the copula approaches the Fréchet-Hoeffding lower (upper) bound. Knowledge of the Fréchet-Hoeffding bounds is therefore important in selecting an appropriate copula. Depending on the copula one wishes to employ, the copula dependence parameters (which represent the dependence between the margins) can sometimes be difficult to interpret because they are not necessarily in the customary $[-1, 1]$ interval. Therefore, it is common to convert the dependence parameter to a familiar measure of association such as Kendall's tau or Spearman's rho. In this chapter we employ the trivariate Gaussian copula with dependence structure characterized by coefficients ϑ_{12}, ϑ_{13} and ϑ_{23} which form the model's correlation matrix $\boldsymbol{\Sigma}$. For full details on copulae see, for instance, Trivedi & Zimmer (2007, Ch. 2) and references therein. Note that the properties discussed above also apply to bivariate copulae, which can be formed as $\mathrm{F}\left(\mathcal{U}_1^{-1}(\mathrm{F}_1(\eta_{1i})), \mathcal{U}_2^{-1}(\mathrm{F}_2(\eta_{2i}))\right)$.

Although several copula functions can be used (e.g., Student-t, Frank), our interest in this chapter lies in making the marginals' specification flexible. Chapter 7 will look at alternative representations of trivariate dependence. Based on (5.2), we express the trivariate Gaussian copula as $\Phi_{3,\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)$, while the bivariate Gaussian copula is structured as $\Phi_{2,\boldsymbol{\varepsilon}_i}\left(\mathcal{W}_{z,i}, \mathcal{W}_{k,i}; (2y_{zi}-1)\tanh(\vartheta_{zk}^*)(2y_{ki}-1)\right)$, where $\boldsymbol{\mathcal{W}}_i = (\mathcal{W}_{1,i}, \mathcal{W}_{2,i}, \mathcal{W}_{i,3})$, $\mathcal{W}_{m,i} = (2y_{mi}-1)\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))$, $\mathrm{F}_m(\eta_{mi})$ can be either the normal, logistic or Gumbel univariate cdf and $\boldsymbol{\Upsilon}_i^*$ is defined in Section 2.3.1, $\forall z = 1,2, k = 2,3, z \neq k, m = 1,2,3$.

Parameter estimation

The estimation and inferential framework introduced in Chapters 2, 3 and 4 can be employed for trivariate models with non-probit link functions after some necessary amendments. These are described below.

Allowing for different marginal distributions requires the modification of quantity $\mathcal{L}_{i\tilde{k}}$ given in Lemma 2.3.1 in Section 2.3. That is, $\mathcal{L}_{i\tilde{k}}$ needs to be re-written in a more general way using the copula representation described above. Lemma 5.2.1 derives such expression, where the specification of the cdf $\mathrm{F}_m(\eta_{mi})$ depends on the marginal distribution one wishes to employ. In the case of probit margins, Lemma 5.2.1 reduces to Lemma 2.3.1.

Lemma 5.2.1. *Quantity $\mathcal{L}_{i\tilde{k}}$, evaluated at the vector $(\boldsymbol{\mathcal{B}}_i\mathbf{H}_i)_{\tilde{k}}$ is equal to the cdf of a multivariate standardized normal vector with correlation matrix $(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)_{\tilde{k}}$, that is*

$$\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i; \boldsymbol{\delta}) = \boldsymbol{\Psi}_{i\tilde{k}}^{\mathcal{Y}_{i\tilde{k}}} = \{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{\mathcal{B}}_i\mathbf{H}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}} = \{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{\mathcal{W}}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\Upsilon}_i^*)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}},$$

where $\boldsymbol{\mathcal{W}}_i = \boldsymbol{\mathcal{B}}_i\mathbf{H}_i = (\mathcal{W}_{1,i}, \ldots, \mathcal{W}_{M,i})^\top$, $\mathbf{H}_i = (\Phi^{-1}(\mathrm{F}_1(\eta_{1i})), \ldots, \Phi^{-1}(\mathrm{F}_M(\eta_{Mi})))^\top$, $\boldsymbol{\Upsilon}_i^* = \boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i$, $\mathcal{W}_{m,i} = \tilde{y}_{mi}\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))$, for $\tilde{y}_{mi} = (2y_{mi}-1)$, $\mathrm{F}_m(\eta_{mi})$ denotes the univariate cdf, $\eta_{mi} = \boldsymbol{x}_{mi}^\top\boldsymbol{\beta}_m$ and $\boldsymbol{\mathcal{B}}_i$ denotes a diagonal $M \times M$ matrix with main diagonal elements $\tilde{y}_{mi} = (2y_{mi}-1)$, that is $\boldsymbol{\mathcal{B}}_i = \mathrm{diag}(2y_{1i}-1, 2y_{2i}-1, \ldots, 2y_{Mi}-1)$.

Proof. See Appendix C.1. □

Estimation of the model parameters can be achieved by extending the efficient

and stable trust region algorithm with integrated automatic multiple smoothing parameter selection described in Chapter 3 to allow for the specification of virtually any parametric link function. This requires to amend the results presented in the previous chapters. Specifically, we compute the analytical score function $\nabla_{\boldsymbol{\delta}}\ell_i(\boldsymbol{\delta})$ and Hessian matrix $\nabla\nabla_{\boldsymbol{\delta}\boldsymbol{\delta}^\top}\ell_i(\boldsymbol{\delta})$ as

$$\begin{aligned}
\nabla_{\boldsymbol{\delta}}\ell_i(\boldsymbol{\delta}) &= \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \frac{\partial\ell_i(\boldsymbol{\delta})}{\partial\bar{\boldsymbol{\eta}}_i} \\
&= \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{\frac{\partial\ell_i(\boldsymbol{\delta})}{\partial\bar{\mathbf{F}}_i}\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right\} \\
&= \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i}\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right\}, \qquad (5.3) \\
\nabla\nabla_{\boldsymbol{\delta}\boldsymbol{\delta}^\top}\ell_i(\boldsymbol{\delta}) &= \left\{\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i}\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right\}\frac{\partial^2\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top} + \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{-\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}\boldsymbol{\Psi}_{i\tilde{k}}^\top}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i}\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right. \\
&\quad \left.\left(\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i}\right)^\top + \frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\left[\frac{\partial^2\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i\bar{\mathbf{F}}_i^\top}\left(\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right)^2 + \frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i}\frac{\partial^2\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}\partial\bar{\boldsymbol{\eta}}^\top}\right]\right\}\left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right), \qquad (5.4)
\end{aligned}$$

where $\bar{\mathbf{F}}_i = (\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i}), \mathrm{F}_4(\eta_{4i}), \mathrm{F}_5(\eta_{5i}), \mathrm{F}_6(\eta_{6i}))^\top$ with $(\mathrm{F}_4(\eta_{4i}), \mathrm{F}_5(\eta_{5i}), \mathrm{F}_6(\eta_{6i})) = (\vartheta^*_{12}, \vartheta^*_{13}, \vartheta^*_{23})$ and $\bar{\boldsymbol{\eta}}$ is defined in Section 2.3.1. The above expressions are similar to (2.9) and (2.10) in Section 2.3.1, except for the extra part corresponding to the derivatives of $\bar{\mathbf{F}}_i$. In the trivariate probit model, (5.3) and (5.4) reduce to (2.9) and (2.10), respectively, as $\partial\bar{\mathbf{F}}_i/\partial\bar{\boldsymbol{\eta}}_i = \mathbf{1}$ and $\partial^2\bar{\mathbf{F}}_i/\partial\bar{\boldsymbol{\eta}}\partial\bar{\boldsymbol{\eta}}^\top = \mathbf{0}$. Computation of (5.3) and (5.4) was achieved via Propositions 5.2.2 and 5.2.3 which generalize Propositions 2.3.2 and 2.3.3.

Proposition 5.2.2. *Assume that $\boldsymbol{\mathcal{W}}_i$ is a multivariate standardized normal vector with correlation matrix equal to $\boldsymbol{\Upsilon}^*_i$. Then the first-order derivative of the M-variate normal cdf $\Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)$ with respect to $\boldsymbol{\beta}_m$, $\forall m = 1, \ldots, M$, can be expressed as*

$$\begin{aligned}
\frac{\partial\Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)}{\partial\boldsymbol{\beta}_m} &= \phi(\mathcal{W}_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{\mathcal{W}}_{-m,i}|\mathcal{W}_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) \times \\
&\quad \frac{f_m(\eta_{mi})}{\phi\left(\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))\right)}(2y_{mi} - 1)\boldsymbol{x}^\top_{mi}
\end{aligned}$$

where M denotes the total number of equations under a multivariate binary frame-

work, $\mathcal{W}_{m,i}$ denotes the linear predictor of the m^{th} equation and is equal to $(2y_{mi}-1)\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))$, $\boldsymbol{\beta}_m$ denotes the parameter vector of covariate vector $\boldsymbol{x}_{mi}$, the vector of linear predictors $\boldsymbol{\mathcal{W}}_{-m,i}$ is defined as $(\mathcal{W}_{1,i},\ldots,\mathcal{W}_{m-1,i},\mathcal{W}_{m+1,i},\ldots,\mathcal{W}_{M,i})^\top$ and $f_m(\eta_{mi})$ and $\mathrm{F}_m(\eta_{mi})$ denote the univariate pdf and cdf respectively which can be specified via the normal, logistic and Gumbel distributions. The mean $\boldsymbol{M}_i^{*m}$ and variance-covariance matrix $\boldsymbol{\Theta}_i^{*m}$ is equal to $\boldsymbol{\Theta}_{21,i}^{*m}\mathcal{W}_{m,i}$ and $\boldsymbol{\Theta}_{22,i}^{*m}-\boldsymbol{\Theta}_{21,i}^{*m}\boldsymbol{\Theta}_{12,i}^{*m}$, respectively, with $\boldsymbol{\Theta}_{12,i}^{*m}$, $\boldsymbol{\Theta}_{21,i}^{*m}$ and $\boldsymbol{\Theta}_{22,i}^{*m}$ defined by re-ordering $\boldsymbol{\Upsilon}_i^*$ as follows

$$\boldsymbol{\Upsilon}_i^{*m}=\left(\begin{array}{c|c} \overbrace{\Theta_{11,i}^{*m}}^{1\times 1} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*m}}^{1\times(M-1)} \\ \hline \underbrace{\boldsymbol{\Theta}_{21,i}^{*m}}_{(M-1)\times 1} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*m}}_{(M-1)\times(M-1)} \end{array}\right).$$

The element $\Theta_{11,i}^{*m}$ is equal to 1, the off-diagonal blocks $\boldsymbol{\Theta}_{12,i}^{*m}$ and $\boldsymbol{\Theta}_{21,i}^{*m}$ consist of the correlations $r_{m\varpi,i}^*=\tanh(\vartheta_{m\varpi}^*)(2y_m-1)(2y_\varpi-1)$, $\forall\ \varpi\in\{1:M\}\setminus m$, for $m\neq\varpi$ and the symmetric sub-matrix $\boldsymbol{\Theta}_{22,i}^{*m}$ has main diagonal elements equal to 1 and off-diagonals equal to $r_{\bar\varphi\varpi,i}^*=\tanh(\vartheta_{\bar\varphi\varpi}^*)(2y_{\bar\varphi}-1)(2y_\varpi-1)$, $\forall\bar\varphi,\varpi\in\{1:M\}\setminus m$, for $\bar\varphi\neq\varpi$.

Proof. See Appendix C.2.1. □

Proposition 5.2.3. *Assume that $\boldsymbol{\mathcal{W}}_i$ is a multivariate standardized normal vector with correlation matrix equal to $\boldsymbol{\Upsilon}_i^*$. Then the first-order derivative of the M-variate normal cdf $\Phi_M(\boldsymbol{\mathcal{W}}_i;\mathbf{0},\boldsymbol{\Upsilon}_i^*)$ with respect to ϑ_{zk}^*, $\forall z=1,\ldots,M-1,k=z+1,\ldots M$, can be expressed as*

$$\begin{aligned}\frac{\partial\Phi_M(\boldsymbol{\mathcal{W}}_i;\mathbf{0},\boldsymbol{\Upsilon}_i^*)}{\partial\vartheta_{zk}^*} &= \phi_2\left(\boldsymbol{\mathcal{W}}_{zk,i};\mathbf{0},\boldsymbol{\Theta}_i^{*zk}\right)\Phi_{M-2}(\boldsymbol{\mathcal{W}}_{-zk,i}|\boldsymbol{\mathcal{W}}_{zk,i};\boldsymbol{M}^{*-zk},\boldsymbol{\Theta}_i^{*-zk})\times \\ &\quad (2y_{zi}-1)(2y_{ki}-1)\frac{4\mathrm{e}^{2\vartheta_{zk}^*}}{(\mathrm{e}^{2\vartheta_{zk}^*}+1)^2}\end{aligned}$$

where M denotes the total number of equations under a multivariate binary framework, $\boldsymbol{\mathcal{W}}_{zk,i}=(\mathcal{W}_{z,i},\mathcal{W}_{k,i})^\top$, $\boldsymbol{\mathcal{W}}_{-zk,i}=(\mathcal{W}_{1,i},\ldots,\mathcal{W}_{z-1,i},\mathcal{W}_{z+1,i},\ldots,\mathcal{W}_{k-1,i},\mathcal{W}_{k+1,i},\ldots,\mathcal{W}_{M,i})^\top$, $\mathcal{W}_{z,i}$ and $\mathcal{W}_{k,i}$ refer to the linear predictors of the z^{th} and k^{th} equations respectively and are equal to $(2y_{mi}-1)\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))$, $\forall m=z,k$, $\boldsymbol{\beta}_m$ denotes the

parameter vector of covariate vector $\boldsymbol{x}_{mi}$ *and* $f_m(\eta_{mi})$ *and* $\mathrm{F}_m(\eta_{mi})$ *denote the univariate pdf and cdf respectively which can be specified via the normal, logistic and Gumbel distributions. Parameter* $\vartheta^*_{zk} = \tanh^{-1}(\vartheta_{zk})$ *where* ϑ_{zk} *denotes the correlation coefficient between the* z^{th} *and* k^{th} *responses. The variance-covariance matrix* $\boldsymbol{\Theta}^{*zk}_i$ *is equal to* $\boldsymbol{\Theta}^{*zk}_{11,i}$*, while the mean* $\boldsymbol{M}^{*-zk}_i$ *and variance-covariance matrix* $\boldsymbol{\Theta}^{*-zk}_i$ *is equal to* $\boldsymbol{\Theta}^{*zk}_{21,i}\left(\boldsymbol{\Theta}^{*zk}_{11,i}\right)^{-1}\boldsymbol{\mathcal{W}}_{zk}$ *and* $\boldsymbol{\Theta}^{*zk}_{22,i} - \boldsymbol{\Theta}^{*zk}_{21,i}\left(\boldsymbol{\Theta}^{*zk}_{11,i}\right)^{-1}\boldsymbol{\Theta}^{*zk}_{12,i}$*, respectively. The sub-matrices* $\boldsymbol{\Theta}^{*zk}_{11,i}$*,* $\boldsymbol{\Theta}^{*zk}_{12,i}$*,* $\boldsymbol{\Theta}^{*zk}_{21,i}$ *and* $\boldsymbol{\Theta}^{*zk}_{22,i}$ *are defined by re-ordering* $\boldsymbol{\Upsilon}^*_i$ *as follows*

$$\boldsymbol{\Upsilon}^{*zk}_i = \left(\begin{array}{c:c} \overbrace{\boldsymbol{\Theta}^{*zk}_{11,i}}^{2\times 2} & \overbrace{\boldsymbol{\Theta}^{*zk}_{12,i}}^{2\times(M-2)} \\ \hdashline \underbrace{\boldsymbol{\Theta}^{*zk}_{21,i}}_{(M-2)\times 2} & \underbrace{\boldsymbol{\Theta}^{*zk}_{22,i}}_{(M-2)\times(M-2)} \end{array}\right).$$

The sub-matrix $\boldsymbol{\Theta}^{*zk}_{11,i}$ *has unit diagonals and off-diagonals defined as* $r^*_{zk,i} = \tanh(\vartheta^*_{zk})$ $(2y_z - 1)(2y_k - 1)$*. The first row (column) of* $\Theta^{*zk}_{12,i}$ *(*$\Theta^{*zk}_{21,i}$*) contains the correlations* $r^*_{z\bar{\varrho},i}$*, for* $\bar{\varrho} \in \{1:M\}\setminus z$*, while the second row (column) of* $\Theta^{*zk}_{12,i}$ *(*$\Theta^{*zk}_{21,i}$*) contains the correlations* $r^*_{\bar{\upsilon}k,i}$*, for* $\bar{\upsilon} \in \{1:M\}\backslash k$*. The diagonal block* $\boldsymbol{\Theta}^{*zk}_{22,i}$ *is a symmetric matrix with unit diagonals and off-diagonal elements equal to* $r^*_{\bar{\chi}\bar{\psi},i}$*,* $\forall\ \bar{\chi}, \bar{\psi} \in \{1:M\}\backslash\{z,k\}$ *for* $\bar{\chi} \neq \bar{\psi}$*.*

Proof. See Appendix C.2.2. □

All derivatives have been verified as in Chapter 2.

In this case, `SemiParTRIV()`/`gjrm()` can be used as follows

```
out <- SemiParTRIV(formula = f.l, data = dat, margins = margins,
                   Model = mod, penCor = PenFun)
```

where arguments `f.l`, `dat`, `PenFun` and `mod` have the same definitions as in Chapter 3, while `margins` specifies the link functions used for the three margins. Possible choices for `margins` are `"probit"`, `"logit"` and `"cloglog"`.

5.3 Simulation study

In this section, we conduct a simulation study to assess the practical performance of the trivariate Gaussian copula model when employing a mixture of three link functions using the DGP3 settings presented in Appendix B.2.1. The chosen link functions were complementary log-log, logit and probit for the first, second and third outcome, respectively. Parameter estimation was carried out using the PMLE approach discussed in Chapter 3, where the correlations were penalized via the Lasso penalty (the Ridge and the Adaptive Lasso provide similar results). The model specification and settings are the same as those employed in Sections 3.2.2 (DGP3) and 4.5.1.

The results are summarized in Figures 5.2 and 5.3, which depict the parametric and smooth component estimates obtained over 250 replicates for two different sample sizes. On average, the regression coefficient estimates approach their true values as n increases and their variability decreases as the sample size grows large. The study shows that the method is effective in recovering the true functions, although occasionally (especially when $n = 1000$) the estimated curves appear to be wigglier than they should be. This behaviour has been commented in Chapter 3.

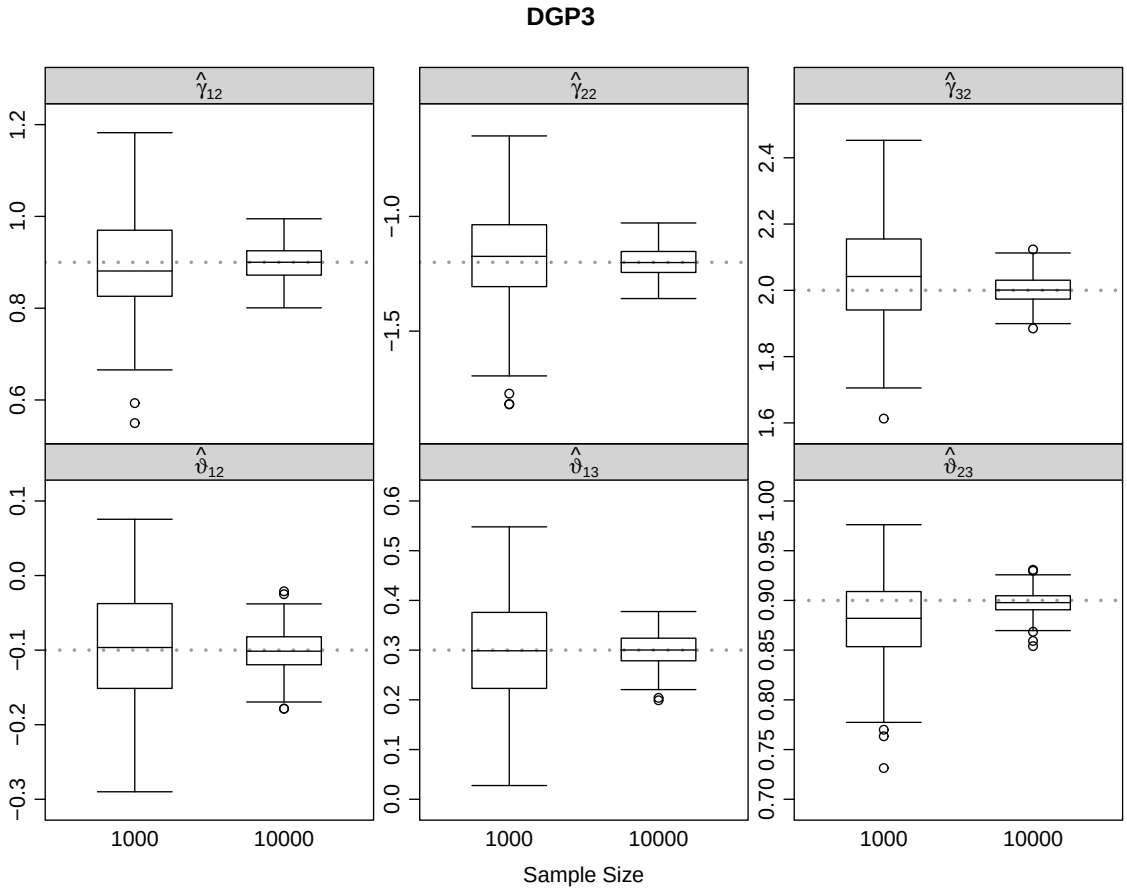

Figure 5.2: Boxplots of parameter estimates obtained applying the trivariate Gaussian copula model on 250 simulated datasets with complementary log-log, logit and probit links for sample sizes equal to 1000 and 10000. True parameter values are represented by horizontal gray dotted lines.

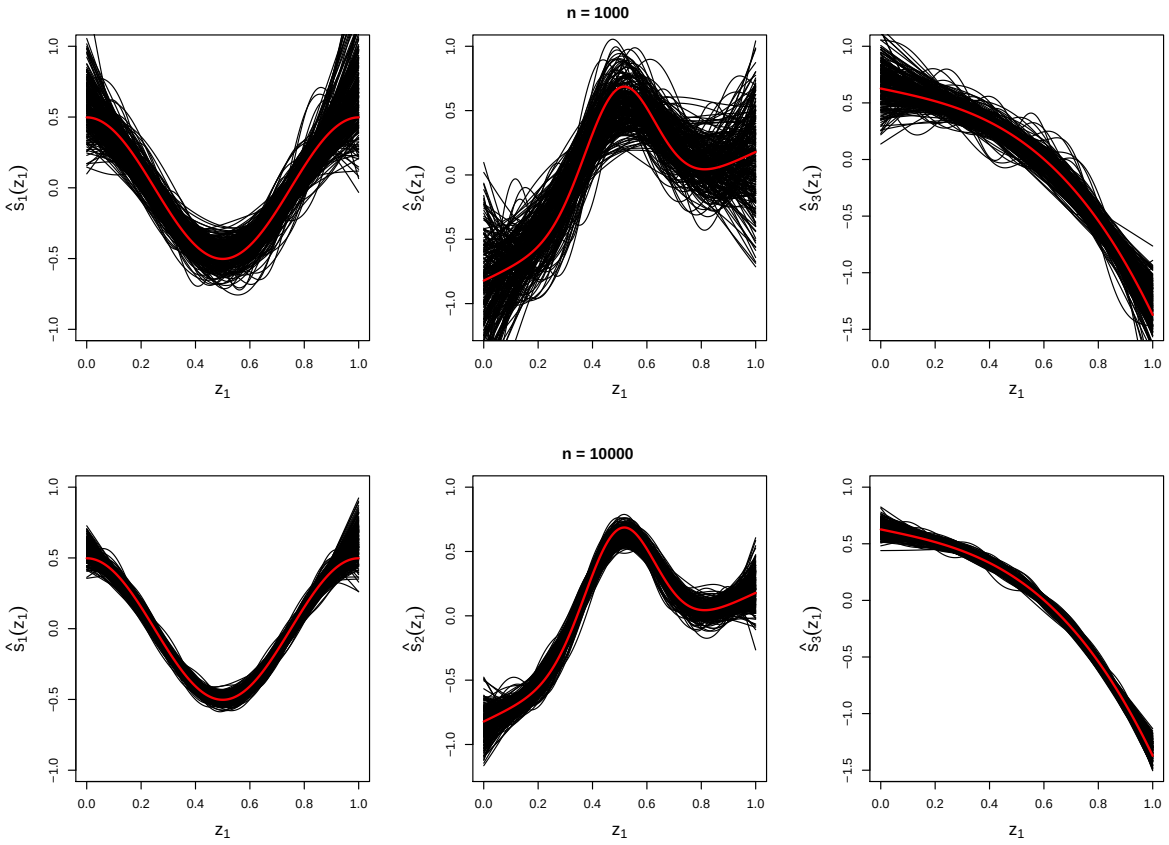

Figure 5.3: Estimated smooth functions for $s_1(z_1)$, $s_2(z_1)$ and $s_3(z_1)$ obtained applying the trivariate Gaussian copula model on 250 simulated datasets with complementary log-log, logit and probit links. The first row shows the estimated curves obtained from samples of 1000 observations, whereas those in the second row correspond to samples of 10000 observations. The black lines represent the estimated smooth functions over all replicates and the red solid lines show the true functions.

5.4 Conclusions

In this chapter, we have discussed the simultaneous estimation of three binary regressions where the trivariate distribution is specified by the Gaussian copula which allows for virtually any parametric link function. The functions considered were the probit, logit and complementary log-log links. Parameter estimation is carried out within a PMLE framework with integrated automatic multiple smoothing parameter selection, and the proposed models can be easily used via function `SemiParTRIV()`/`gjrm()`.

As mentioned in the introduction, the Gaussian trivariate copula binary model

with arbitrary margins can offer advantages in empirical modelling as compared to the fully Gaussian version. The next section will discuss another extension where the correlation coefficients of the trivariate Gaussian are modelled as functions of semi-parametric predictors.

Chapter 6

A trivariate additive regression model with varying correlation matrix

In this chapter, we propose a generalisation of a trivariate additive binary model where the parameters describing the association between the responses can be made dependent on several types of covariate effects (such as linear, nonlinear, random, and spatial effects). All necessary amendments made in estimation framework have been incorporated in `SemiParTRIV()`/`gjrm()`. The effectiveness of the model is assessed in simulation as well as empirically by modelling jointly three adverse birth binary outcomes in North Carolina.

6.1 Introduction

In the previous chapters, we assumed that the correlation structure that accounts for the dependence between the three response variables is fixed. However, it may be the case that the strength or direction of the dependence is modified by covariates. To reduce the risk of misspecification, therefore, we extend the material presented in the previous chapters to allow the model's association parameters to depend on several types of covariate effects. Within this framework, the systematic part of the

model is expanded to allow each correlation parameter to be modelled as a function of the available data. This can help to gain insights into the way the residual association between the responses is modified by the presence of covariates.

It is worth noting that our proposal can also be regarded as an extension of the bivariate regression approaches introduced by Marra et al. (2017), Klein & Kneib (2016b) and Radice et al. (2016) as well as of the popular GAMs and GAMs for location, scale and shape of Wood (2006) and Rigby & Stasinopoulos (2005). Function `SemiParTRIV()`/`gjrm()` in the `R` package `GJRM` (Marra & Radice, 2017) includes the developments in this chapter.

The remainder of the chapter is organized as follows. Section 6.2 introduces the proposed model and Section 6.3 provides the key estimation details. The proposal is empirically evaluated in a simulation study, presented in Section 6.4, and then applied to a case study in Section 6.5. Section 6.6 concludes the chapter.

6.2 Model specification

This section introduces an extension of the trivariate binary model that is based on a modified Cholesky decomposition of the model's correlation matrix.

To allow each association parameter to be expressed as function of an additive predictor, we re-express the correlation matrix as

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} 1 & \vartheta_{12,i} & \vartheta_{13,i} \\ \vartheta_{12,i} & 1 & \vartheta_{23,i} \\ \vartheta_{13,i} & \vartheta_{23,i} & 1 \end{pmatrix},$$

where $\vartheta_{zk,i}$ is the correlation coefficient between the z^{th} and k^{th} responses for subject i, $\forall z, k, i$. The challenge to address here is that the range of each correlation's additive predictor has to be unbounded to avoid constrained optimization and that the correlation matrix $\boldsymbol{\Sigma}_i$ must be positive definite with each of its coefficients taking values in $[-1, 1]$. This makes the parameter space of $\boldsymbol{\Sigma}_i$ somewhat complex with restrictions for each parameter depending on the values of the others. To this end,

we propose using a modified Cholesky decomposition approach which is described below.

6.2.1 Unconstrained parametrization for the correlation matrix

The standard Cholesky decomposition of a positive-definite correlation matrix $\mathbf{\Sigma}$ is of the form $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^\top$, where $\mathbf{C}$ is a unique lower-triangular matrix with positive diagonal entries. Modifications of the standard Cholesky decomposition can be found in the literature. For example, Pourahmadi (1999, 2000) shows that the modified Cholesky decomposition of $\mathbf{\Sigma}^{-1}$ offers a simple unconstrained reparametrization of the covariance matrix, while Chen & Dunson (2003) propose an alternative modified Cholesky decomposition to factorize the covariance matrix. As shown by Pourahmadi (2007), who provides an overview of the two methods, estimation of the new parameters in the latter decomposition may be more demanding computationally. In this chapter, we employ a modification of the work by Pourahmadi (1999, 2000), where we employ the modified Cholesky approach with unit variance constraints to deal with correlation matrices.

Let $\bar{\mathbf{\Sigma}}_i^*$ denote a symmetric positive-definite correlation matrix, $\forall i$, defined as

$$\bar{\mathbf{\Sigma}}_i^* = \bar{\mathbf{C}}_i^* \bar{\mathbf{C}}_i^{*\top} = \begin{pmatrix} 1 & \eta_{12,i} & \eta_{13,i} \\ \eta_{12,i} & 1+\eta_{12,i}^2 & \eta_{12,i}\eta_{13,i}+\eta_{23,i} \\ \eta_{13,i} & \eta_{12,i}\eta_{13,i}+\eta_{23,i} & 1+\eta_{13,i}^2\eta_{23,i}^2 \end{pmatrix},$$

where $\eta_{zk,i}$ is a function of parametric components and smooth functions defined as

$$\eta_{zk,i} = \mathbf{v}_{zk,i}^\top \boldsymbol{\gamma}_{zk} + \mathbf{L}_{zk,i}^\top \boldsymbol{\alpha}_{zk} = \mathbf{x}_{zk,i}^\top \boldsymbol{\beta}_{zk} \in \mathbb{R}, \tag{6.1}$$

$\forall z, k, i$, and $\bar{\mathbf{C}}_i^*$ is equal to

$$\bar{\mathbf{C}}_i^* = \begin{pmatrix} 1 & 0 & 0 \\ \eta_{12,i} & 1 & 0 \\ \eta_{13,i} & \eta_{23,i} & 1 \end{pmatrix}.$$

Terms $\mathbf{v}_{zk,i}$, $\boldsymbol{\gamma}_{zk}$, $\mathbf{L}_{zk,i}$, $\boldsymbol{\alpha}_{zk}$, $\mathbf{x}_{zk,i}$ and $\boldsymbol{\beta}_{zk}$ in (6.1) are defined similarly as $\mathbf{v}_{mi}$, $\boldsymbol{\gamma}_m$, $\mathbf{L}_{mi}$, $\boldsymbol{\alpha}_m$, $\mathbf{x}_{mi}$ and $\boldsymbol{\beta}_m$ in Section 2.2. Formulation (6.1) allows us to represent many types of covariate effects depending on the nature of the covariate(s) considered. These include linear, non-linear, random and spatial effects. By using the variance-correlation decomposition $\boldsymbol{\Sigma}_i = \mathbf{T}_i \bar{\boldsymbol{\Sigma}}_i^* \mathbf{T}_i$ with $\mathbf{T}_i = \text{diag}\left(1, \left(1+\eta_{12,i}^2\right)^{-1/2}, \left(1+\eta_{13,i}^2+\eta_{23,i}^2\right)^{-1/2}\right)$, we have that the correlation matrix $\boldsymbol{\Sigma}_i$ can be expressed as

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} 1 & \frac{\eta_{12,i}}{\sqrt{1+\eta_{12,i}^2}} & \frac{\eta_{13,i}}{\sqrt{1+\eta_{13,i}^2+\eta_{23,i}^2}} \\ \frac{\eta_{12,i}}{\sqrt{1+\eta_{12,i}^2}} & 1 & \frac{\eta_{12,i}\eta_{13,i}+\eta_{23,i}}{\sqrt{(1+\eta_{12,i}^2)(1+\eta_{12,i}^2+\eta_{23,i}^2)}} \\ \frac{\eta_{13,i}}{\sqrt{1+\eta_{13,i}^2+\eta_{23,i}^2}} & \frac{\eta_{12,i}\eta_{13,i}+\eta_{23,i}}{\sqrt{(1+\eta_{12,i}^2)(1+\eta_{12,i}^2+\eta_{23,i}^2)}} & 1 \end{pmatrix}.$$

The correlation parameters can therefore be defined as $\vartheta_{12,i} = \eta_{12,i}/\sqrt{1+\eta_{12,i}^2}$, $\vartheta_{13,i} = \eta_{13,i}/\sqrt{1+\eta_{13,i}^2+\eta_{23,i}^2}$ and $\vartheta_{23,i} = (\eta_{12,i}\eta_{13,i}+\eta_{23,i})/\sqrt{(1+\eta_{12,i}^2)(1+\eta_{12,i}^2+\eta_{23,i}^2)}$. It follows that

$$\eta_{12,i} = \sqrt{\frac{\vartheta_{12,i}^2}{1-\vartheta_{12,i}^2}}, \quad \eta_{13,i} = \sqrt{\frac{\vartheta_{13,i}^2\left(1+\frac{\mathsf{A}}{1-\mathsf{A}}\right)}{1-\vartheta_{13,i}^2}}, \quad \eta_{23,i} = \sqrt{\frac{\mathsf{A}}{1-\mathsf{A}}},$$

where $\mathsf{A} = \left(\frac{\vartheta_{23,i}\sqrt{1+\eta_{12,i}^2}-\eta_{12,i}\vartheta_{13,i}}{\sqrt{1-\vartheta_{13,i}^2}}\right)^2$. Therefore, by construction we have that $\vartheta_{zk,i} \in [-1,1]$, $\eta_{zk,i} \in \mathbb{R}$, $\forall z, k, i$ and the resulting correlation matrix is positive definite, as required.

6.3 Estimation details

Simultaneous estimation of all parameters of the trivariate additive binary model is achieved by solving

$$\hat{\boldsymbol{\delta}} := \arg\min_{\boldsymbol{\delta}} -\ell_p(\boldsymbol{\delta}) = \arg\min_{\boldsymbol{\delta}} -\{\log \mathcal{L}(\mathbf{Y};\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\top \tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\delta}\}, \tag{6.2}$$

where $\boldsymbol{\delta} = (\boldsymbol{\beta}^\top, \boldsymbol{\beta}_{\boldsymbol{\vartheta}}^\top)^\top$, $\boldsymbol{\beta}$ is defined as in Section 2.2, $\boldsymbol{\beta}_{\boldsymbol{\vartheta}} = (\boldsymbol{\beta}_{12}, \boldsymbol{\beta}_{13}, \boldsymbol{\beta}_{23})^\top$, $\boldsymbol{\beta}_{zk}$ denotes the coefficients in additive predictor $\eta_{zk,i}$, $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}} = \text{diag}\Big(\mathbf{0}_{\tilde{P}_1}^\top, \lambda_{1\nu_1}\mathbf{S}_{1\nu_1}, \ldots, \lambda_{1\tilde{N}_1}\mathbf{S}_{1\tilde{N}_1}, \mathbf{0}_{\tilde{P}_2}^\top, \lambda_{2\nu_2}\mathbf{S}_{2\nu_2}, \ldots, \lambda_{2\tilde{N}_2}\mathbf{S}_{2\tilde{N}_2}, \mathbf{0}_{\tilde{P}_3}^\top, \lambda_{3\nu_3}\mathbf{S}_{3\nu_3}, \ldots, \lambda_{3\tilde{N}_3}\mathbf{S}_{3\tilde{N}_3}, \mathbf{0}_{\tilde{P}_{12}}^\top, \lambda_{12\nu_{12}}\mathbf{S}_{12\nu_{12}}, \ldots, \lambda_{12\tilde{N}_{12}}\mathbf{S}_{12\tilde{N}_{12}}, \mathbf{0}_{\tilde{P}_{13}}^\top, \lambda_{13\nu_{13}}\mathbf{S}_{13\nu_{13}}, \ldots, \lambda_{13\tilde{N}_{13}}\mathbf{S}_{13\tilde{N}_{13}}, \mathbf{0}_{\tilde{P}_{23}}^\top, \lambda_{23\nu_{23}}\mathbf{S}_{23\nu_{23}}, \ldots, \lambda_{23\tilde{N}_{23}}\mathbf{S}_{23\tilde{N}_{23}}\Big)$, $\mathbf{S}_{zk\nu_{zk}}$ is defined following a similar construction as $\mathbf{S}_{m\nu_m}$, $\lambda_{zk\nu_{zk}}$ is defined similarly as $\lambda_{m\nu_m}$ and $\tilde{P}_{zk}$ denotes the number of parametric components in $\eta_{zk,i}$. Likelihood $\mathcal{L}_{i\tilde{k}}$ is derived from Lemma 6.3.1 for $M = 3$.

Lemma 6.3.1. *Quantity $\mathcal{L}_{i\tilde{k}}$, evaluated at the vector $(\boldsymbol{\mathcal{B}}_i\mathbf{H}_i)_{\tilde{k}}$ is equal to the cdf of a multivariate standardized normal vector with correlation matrix $(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i)_{\tilde{k}}$, that is*

$$\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i;\boldsymbol{\delta}) = \boldsymbol{\Psi}_{i\tilde{k}}^{\mathcal{Y}_{i\tilde{k}}} = \{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{\mathcal{B}}_i\mathbf{H}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}} = \{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{\mathcal{W}}_i)_{\tilde{k}}; \mathbf{0}, (\boldsymbol{\Upsilon}_i^*)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}},$$

where $\boldsymbol{\mathcal{W}}_i = \boldsymbol{\mathcal{B}}_i\mathbf{H}_i = (\mathcal{W}_{1,i}, \ldots, \mathcal{W}_{M,i})^\top$, $\mathbf{H}_i = (\Phi^{-1}(\text{F}_1(\eta_{1i})), \ldots, \Phi^{-1}(\text{F}_M(\eta_{Mi})))^\top$, $\boldsymbol{\Upsilon}_i^* = \boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i$, $\mathcal{W}_{m,i} = \tilde{y}_{mi}\Phi^{-1}(\text{F}_m(\eta_{mi}))$, for $\tilde{y}_{mi} = (2y_{mi} - 1)$, $\text{F}_m(\eta_{mi})$ denotes the univariate cdf, $\eta_{mi} = \boldsymbol{x}_{mi}^\top\boldsymbol{\beta}_m$ and $\boldsymbol{\mathcal{B}}_i$ denotes a diagonal $M \times M$ matrix with main diagonal elements $\tilde{y}_{mi} = (2y_{mi} - 1)$, that is $\boldsymbol{\mathcal{B}}_i = \text{diag}(2y_{1i} - 1, 2y_{2i} - 1, \ldots, 2y_{Mi} - 1)$.

Proof. See Appendix D.1. □

To minimize (6.2), we have extended the algorithm presented in Chapter 5 to allow for the correlation matrix to depend on covariate effects as described earlier. The practical success of this extension depends on the availability of the analytical score and Hessian matrix of the model which are fundamental for a reliable, stable and efficient implementation of the above mentioned algorithm. This requires to

amend and generalize the results presented in Chapter 5. In particular, we derive the analytical score function $\nabla_{\boldsymbol{\delta}}\ell_i(\boldsymbol{\delta})$ and Hessian matrix $\nabla\nabla_{\boldsymbol{\delta}\boldsymbol{\delta}^\top}\ell_i(\boldsymbol{\delta})$ as

$$\begin{aligned}
\nabla_{\boldsymbol{\delta}}\ell_i(\boldsymbol{\delta}) &= \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \frac{\partial\ell_i(\boldsymbol{\delta})}{\partial\bar{\boldsymbol{\eta}}_i} \\
&= \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{\frac{\partial\ell_i(\boldsymbol{\delta})}{\partial\bar{\mathbf{F}}_i}\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right\} \\
&= \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i}\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right\}, \qquad (6.3) \\
\nabla\nabla_{\boldsymbol{\delta}\boldsymbol{\delta}^\top}\ell_i(\boldsymbol{\delta}) &= \left\{\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i}\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right\}\frac{\partial^2\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top} + \left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right)^\top \left\{-\frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}\boldsymbol{\Psi}_{i\tilde{k}}^\top}\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i}\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right. \\
&\quad \left.\left(\frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\boldsymbol{\eta}}_i}\right)^\top + \frac{1}{\boldsymbol{\Psi}_{i\tilde{k}}}\left[\frac{\partial^2\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i\bar{\mathbf{F}}_i^\top}\left(\frac{\partial\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}_i}\right)^2 + \frac{\partial\boldsymbol{\Psi}_{i\tilde{k}}}{\partial\bar{\mathbf{F}}_i}\frac{\partial^2\bar{\mathbf{F}}_i}{\partial\bar{\boldsymbol{\eta}}\partial\bar{\boldsymbol{\eta}}^\top}\right]\right\}\left(\frac{\partial\bar{\boldsymbol{\eta}}_i}{\partial\boldsymbol{\delta}}\right), \qquad (6.4)
\end{aligned}$$

where $\bar{\boldsymbol{\eta}}_i = (\eta_{1i}, \eta_{2i}, \eta_{3i}, \eta_{12,i}, \eta_{!3,i}, \eta_{23,i})^\top$, $\bar{\mathbf{F}}_i = (\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i}), \mathrm{F}_4(\eta_{4i}), \mathrm{F}_5(\eta_{5i}), \mathrm{F}_6(\eta_{6i}))^\top$ with $(\mathrm{F}_4(\eta_{4i}), \mathrm{F}_5(\eta_{5i}), \mathrm{F}_6(\eta_{6i})) = (\vartheta_{12}, \vartheta_{13}, \vartheta_{23})$, $\partial\bar{\boldsymbol{\eta}}_i/\partial\boldsymbol{\delta} = \mathrm{diag}\,(\partial\eta_{1i}/\partial\boldsymbol{\beta}_1, \partial\eta_{2i}/\partial\boldsymbol{\beta}_2, \partial\eta_{3i}/\partial\boldsymbol{\beta}_3, \partial\eta_{12,i}/\partial\boldsymbol{\beta}_{12}, \partial\eta_{13,i}/\partial\boldsymbol{\beta}_{13}, \partial\eta_{23,i}/\partial\boldsymbol{\beta}_{23})$ and $\partial\ell(\boldsymbol{\delta})/\partial\bar{\boldsymbol{\eta}}_i = (\partial\ell(\boldsymbol{\delta})/\partial\eta_{1i}, \partial\ell(\boldsymbol{\delta})/\partial\eta_{2i}, \partial\ell(\boldsymbol{\delta})/\partial\eta_{3i}, \partial\ell(\boldsymbol{\delta})/\partial\eta_{12,i}, \partial\ell(\boldsymbol{\delta})/\partial\eta_{13,i}, \partial\ell(\boldsymbol{\delta})/\partial\eta_{23,i})^\top$. Implementation of (6.3) and (6.4) has been a tedious and non-trivial task. This extension required, for instance, the use of the multivariate chain rule which was employed as follows. As shown in Section 6.2.1, $\vartheta_{zk,i}$ may depend on $\eta_{zk,i}$ and $\boldsymbol{\eta}_{-zk,i}$, where $\boldsymbol{\eta}_{-zk,i} \in \tilde{\boldsymbol{\eta}}_i \setminus \eta_{zk,i}$, for $\tilde{\boldsymbol{\eta}}_i = (\eta_{12,i}, \eta_{13,i}, \eta_{23,i})^\top$. Hence, term $\partial\bar{\boldsymbol{F}}_i^*/\partial\tilde{\boldsymbol{\eta}}_i$, for $\bar{\boldsymbol{F}}_i^* = (\vartheta_{12,i}, \vartheta_{13,i}, \vartheta_{23,i})^\top$, is a 3×3 Jacobian matrix containing all the derivatives of $\bar{\boldsymbol{F}}_i^*$ with respect to $\bar{\boldsymbol{\eta}}_i$. That is,

$$\frac{\partial\bar{\boldsymbol{F}}_i^*}{\partial\tilde{\boldsymbol{\eta}}_i} = \begin{pmatrix} \frac{\partial\vartheta_{12,i}}{\eta_{12,i}} & \frac{\partial\vartheta_{12,i}}{\eta_{13,i}} & \frac{\partial\vartheta_{12,i}}{\eta_{23,i}} \\ \frac{\partial\vartheta_{13,i}}{\eta_{12,i}} & \frac{\partial\vartheta_{13,i}}{\eta_{13,i}} & \frac{\partial\vartheta_{13,i}}{\eta_{23,i}} \\ \frac{\partial\vartheta_{23,i}}{\eta_{12,i}} & \frac{\partial\vartheta_{23,i}}{\eta_{13,i}} & \frac{\partial\vartheta_{23,i}}{\eta_{23,i}} \end{pmatrix}.$$

The above accounts for the dependencies between $\vartheta_{zk,i}$ and $\eta_{zk,i}$ as well as $\boldsymbol{\eta}_{-zk,i}$. Second-order derivatives were derived in a similar way. More generically, implementation of the score function and Hessian matrix was achieved via Propositions 6.3.2 and 6.3.3 by setting $M = 3$.

Proposition 6.3.2. *Assume that $\boldsymbol{\mathcal{W}}_i$ is a multivariate standardized normal vector with correlation matrix equal to $\boldsymbol{\Upsilon}_i^*$. Then the first-order derivative of the M-variate normal cdf $\Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)$ with respect to $\boldsymbol{\beta}_m$, $\forall m = 1, \ldots, M$, can be expressed as*

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \boldsymbol{\beta}_m} &= \phi(\mathcal{W}_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{\mathcal{W}}_{-m,i}|\mathcal{W}_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m})\frac{f_m(\eta_{mi})}{\phi\left(\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))\right)} \\ &\quad (2y_{mi} - 1)\boldsymbol{x}_{mi}^\top\end{aligned}$$

where M denotes the total number of equations under a multivariate binary framework, $\mathcal{W}_{m,i}$ denotes the linear predictor of the m^{th} equation and is equal to $(2y_{mi} - 1)\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))$, $\boldsymbol{\beta}_m$ denotes the parameter vector of covariate vector $\boldsymbol{x}_{mi}$, the vector of linear predictors $\boldsymbol{\mathcal{W}}_{-m,i}$ is defined as $(\mathcal{W}_{1,i}, \ldots, \mathcal{W}_{m-1,i}, \mathcal{W}_{m+1,i}, \ldots, \mathcal{W}_{M,i})^\top$ and $f_m(\eta_{mi})$ and $\mathrm{F}_m(\eta_{mi})$ denote the univariate pdf and cdf respectively which can be specified via the normal, logistic and Gumbel distributions. The mean $\boldsymbol{M}_i^{*m}$ and variance-covariance matrix $\boldsymbol{\Theta}_i^{*m}$ is equal to $\boldsymbol{\Theta}_{21,i}^{*m}\mathcal{W}_{m,i}$ and $\boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\boldsymbol{\Theta}_{12,i}^{*m}$, respectively, with $\boldsymbol{\Theta}_{12,i}^{*m}$, $\boldsymbol{\Theta}_{21,i}^{*m}$ and $\boldsymbol{\Theta}_{22,i}^{*m}$ defined by re-ordering $\boldsymbol{\Upsilon}_i^*$ as follows

$$\boldsymbol{\Upsilon}_i^{*m} = \left(\begin{array}{c|c} \overbrace{\Theta_{11,i}^{*m}}^{1\times 1} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*m}}^{1\times(M-1)} \\ \hline \underbrace{\boldsymbol{\Theta}_{21,i}^{*m}}_{(M-1)\times 1} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*m}}_{(M-1)\times(M-1)} \end{array}\right).$$

The element $\Theta_{11,i}^{*m}$ is equal to 1, the off-diagonal blocks $\boldsymbol{\Theta}_{12,i}^{*m}$ and $\boldsymbol{\Theta}_{21,i}^{*m}$ consist of the correlations $r_{m\varpi,i}^* = \mathrm{t}_{mm,i}\mathrm{t}_{\varpi\varpi,i}\bar{\sigma}_{m\varpi,i}^*(2y_{mi} - 1)(2y_{\varpi i} - 1)$, where $\mathrm{t}_{mm,i}$ and $\mathrm{t}_{\varpi\varpi,i}$ denote the $(m, m)^{th}$ and $(\varpi, \varpi)^{th}$ element of matrix $\mathbf{T}_i$, respectively, $\forall\, \varpi \in \{1 : M\} \setminus m, m \neq \varpi$, and $\bar{\sigma}_{m\varpi,i}^*$ is the $(m, \varpi)^{th}$ element of matrix $\bar{\boldsymbol{\Sigma}}_i^*$ (matrices $\mathbf{T}_i$ and $\bar{\boldsymbol{\Sigma}}_i^*$ are defined Appendix D.2). The symmetric sub-matrix $\boldsymbol{\Theta}_{22,i}^{*m}$ has main diagonal elements equal to 1 and off-diagonals equal to $r_{\bar{\varphi}\varpi,i}^* = \mathrm{t}_{\bar{\varphi}\bar{\varphi},i}\mathrm{t}_{\varpi\varpi,i}\sigma_{\bar{\varphi}\varpi,i}^*(2y_{\bar{\varphi}i} - 1)(2y_{\varpi i} - 1)$, $\forall \bar{\varphi}, \varpi \in \{1 : M\} \setminus m$, for $\bar{\varphi} \neq \varpi$.

Proof. See Appendix C.2.1. □

Proposition 6.3.3. *Assume that $\boldsymbol{\mathcal{W}}_i$ is a multivariate standardized normal vector with correlation matrix equal to $\boldsymbol{\Upsilon}_i^*$. Then the first-order derivative of the M-variate*

normal cdf $\Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)$ *with respect to* $\boldsymbol{\beta}_{zk}$, $\forall z = 1, \ldots, M-1, k = z+1, \ldots M$, *can be expressed as*

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \boldsymbol{\beta}_{zk}} &= \big(\phi_2(\boldsymbol{\mathcal{W}}_{12,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*12}) \Phi_{M-2}(\boldsymbol{\mathcal{W}}_{-12,i} | \boldsymbol{\mathcal{W}}_{12,i}; \boldsymbol{M}_i^{*-12}, \boldsymbol{\Theta}_i^{*-12}), \ldots, \\ &\quad \phi_2(\boldsymbol{\mathcal{W}}_{M-1,M,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*M-1,M}) \Phi_{M-2}(\boldsymbol{\mathcal{W}}_{-M-1,M,i} | \boldsymbol{\mathcal{W}}_{M-1,M,i}; \\ &\quad \boldsymbol{M}_i^{*-M-1,M}, \boldsymbol{\Theta}_i^{*-M-1,M})\big) \times \left(\frac{\partial r_{12,i}^*}{\partial \eta_{zk,i}}, \ldots, \frac{\partial r_{M-1,M,i}^*}{\partial \eta_{zk,i}}\right)^\top \boldsymbol{x}_{zk,i}^\top,\end{aligned}$$

where M *denotes the total number of equations under a multivariate binary framework,* $\boldsymbol{\beta}_{zk}$ *denotes the parameter vector of covariate vector* $\boldsymbol{x}_{zk,i}$, $\boldsymbol{\mathcal{W}}_{zk,i} = (\mathcal{W}_{z,i}, \mathcal{W}_{k,i})^\top$, $\boldsymbol{\mathcal{W}}_{-zk,i} = (\mathcal{W}_{1,i}, \ldots, \mathcal{W}_{z-1,i}, \mathcal{W}_{z+1,i}, \ldots, \mathcal{W}_{k-1,i}, \mathcal{W}_{k+1,i}, \ldots, \mathcal{W}_{M,i})^\top$, $\forall z, k$, $\mathcal{W}_{z,i}$ *and* $\mathcal{W}_{k,i}$ *refer to the linear predictors of the* z^{th} *and* k^{th} *equations respectively and are equal to* $(2y_{mi} - 1)\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))$, $\forall m = z, k$, *and* $f_m(\eta_{mi})$ *and* $\mathrm{F}_m(\eta_{mi})$ *denote the univariate pdf and cdf respectively which can be specified via the normal, logistic and Gumbel distributions. The variance-covariance matrix* $\boldsymbol{\Theta}_i^{*zk}$ *is equal to* $\boldsymbol{\Theta}_{11,i}^{*zk}$, *while the mean* $\boldsymbol{M}_i^{*-zk}$ *and variance-covariance matrix* $\boldsymbol{\Theta}_i^{*-zk}$ *is equal to* $\boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{\mathcal{W}}_{zk}$ *and* $\boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{\Theta}_{12,i}^{*zk}$, *respectively,* $\forall z, k$. *The sub-matrices* $\boldsymbol{\Theta}_{11,i}^{*zk}$, $\boldsymbol{\Theta}_{12,i}^{*zk}$, $\boldsymbol{\Theta}_{21,i}^{*zk}$ *and* $\boldsymbol{\Theta}_{22,i}^{*zk}$ *are defined by re-ordering* $\boldsymbol{\Upsilon}_i^*$ *as follows*

$$\boldsymbol{\Upsilon}_i^{*zk} = \left(\begin{array}{c|c} \overbrace{\boldsymbol{\Theta}_{11,i}^{*zk}}^{2\times 2} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*zk}}^{2\times(M-2)} \\ \hline \underbrace{\boldsymbol{\Theta}_{21,i}^{*zk}}_{(M-2)\times 2} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*zk}}_{(M-2)\times(M-2)} \end{array}\right).$$

The sub-matrix $\boldsymbol{\Theta}_{11,i}^{*zk}$ *has unit diagonals and off-diagonals defined as* $r_{zk,i}^* = \mathrm{t}_{zz,i}$ $\mathrm{t}_{kk,i}\bar{\sigma}_{zk,i}^*(2y_{zi} - 1)(2y_{ki} - 1)$, *where* $\mathrm{t}_{mm,i}$ *denotes the* $(m,m)^{th}$ *element of matrix* $\mathbf{T}_i$, $\forall m = zk$, *and* $\bar{\sigma}_{zk,i}^*$ *is the* $(z,k)^{th}$ *element of matrix* $\bar{\boldsymbol{\Sigma}}_i^*$ *(matrices* $\mathbf{T}_i$ *and* $\bar{\boldsymbol{\Sigma}}_i^*$ *are defined in Appendix D.2). The first row (column) of* $\Theta_{12,i}^{*zk}$ *(*$\Theta_{21,i}^{*zk}$*) contains the correlations* $r_{z\bar{\varrho},i}^*$, *for* $\bar{\varrho} \in \{1 : M\} \setminus z$, *while the second row (column) of* $\Theta_{12,i}^{*zk}$ *(*$\Theta_{21,i}^{*zk}$*) contains the correlations* $r_{\bar{\upsilon}k,i}^*$, *for* $\bar{\upsilon} \in \{1 : M\} \setminus k$. *The diagonal block* $\boldsymbol{\Theta}_{22,i}^{*zk}$ *is a symmetric matrix with unit diagonals and off-diagonal elements equal to* $r_{\bar{\chi}\bar{\psi},i}^*$, $\forall$ $\bar{\chi}, \bar{\psi} \in \{1 : M\} \setminus \{z, k\}$ *for* $\bar{\chi} \neq \bar{\psi}$.

Proof. See Appendix D.3. □

All derivatives have been verified as in Chapter 2.

6.4 Simulation Study

To gain some insights into the practical performance of the proposed approach, we conducted a simulation study. We considered three binary outcomes, one binary covariate and one continuous regressor. The chosen link functions were logit, complementary log-log and probit. Exact simulation settings are given in the Appendix D.4. The syntax to fit the proposed trivariate binary model is

```
out  <- SemiParTRIV(formula = f.l, data = dat, Chol = TRUE,
                    margins = c("logit", "cloglog", "probit"))
```

where `f.l` consists of a list of six equations

```
eqn1 <- y1 ~ v1 + s(z1)
eqn2 <- y2 ~ v1 + s(z1)
eqn3 <- y3 ~ v1 + s(z1)
eqn12 <- ~ v1 + s(z1)
eqn13 <- ~ v1 + s(z1)
eqn23 <- ~ v1 + s(z1)
f.l <- list(eqn1, eqn2, eqn3, eqn12, eqn13, eqn23)
```

`v1` and `z1` denote the binary and continuous covariates, respectively, `s()` represents a smooth function that is set up using a penalized thin plate regression spline with 10 bases and penalty based on second order derivatives, the last three equations in `f.l` refer to the additive predictors for the correlation parameters ϑ_{12}, ϑ_{13} and ϑ_{23}, `dat` is a data frame containing the variables in the model, `Chol = TRUE` indicates that the modified Cholesky decomposition approach has to be employed and `margins` denotes the three link functions.

Figures 6.1 and 6.2 depict linear and non-linear estimates obtained when applying the proposed approach. Overall, the mean estimates are close to the true values and,

as expected, their variability decreases as the sample size grows large. The main exception is perhaps the parametric component of the additive predictor related to ϑ_{23}, where at $n = 1000$ the estimate exhibits some bias and a larger variability as compared to the other parameters. Also note that the uncertainty of the estimates for all the components in the correlations' additive predictors is higher than that of the estimates for the three marginal equations. This is not so surprising given the complexity of the proposed model and the fact that the correlation parameters are usually more difficult to estimate in a flexible regression setting when the outcomes are binary. Overall, the results improve considerably as n increases.

6.5 Empirical illustration

We illustrate the potential of the proposed model using 2007-2008 birth data from the North Carolina Center for Health Statistics. The data contain information on $64,690$ male newborns and build upon the analysis conducted in Chapter 3. As before, the choice of variables included in the model was mainly driven by previous work on the subject (e.g., South et al., 2012; Neelon et al., 2014). The responses are plurality (`mb`), infant's birth weight (`lbw`) and preterm birth (`ptb`), while the covariates are maternal race (`nwhite`), smoking status (`smoker`), weight gained by mother during pregnancy in pounds (`gained`), age of mother in years (`mage`) and county in which the birth occurred (`county`). For full description of the variables we refer the reader to Section 3.4.

In Section 3.4 we built a model for the joint analysis of `mb`, `lbw` and `ptb`, and showed the impacts that the model's covariates have on the responses as well as some joint probabilities of interest. Here, the focus is on alternative specifications for the link functions and on understanding how the association between the three outcomes is modified by the presence of covariates. We started off with the specification adopted in Section 3.4 where all model's additive predictors contained all the covariates available in the data. That is, all additive predictors included $\texttt{nwhite}_i$, $\texttt{smoker}_i$, $s(\texttt{gained}_i)$, $s(\texttt{mage}_i)$ and $s_{\texttt{spatial}}(\texttt{county}_i)$, where the smooth functions of

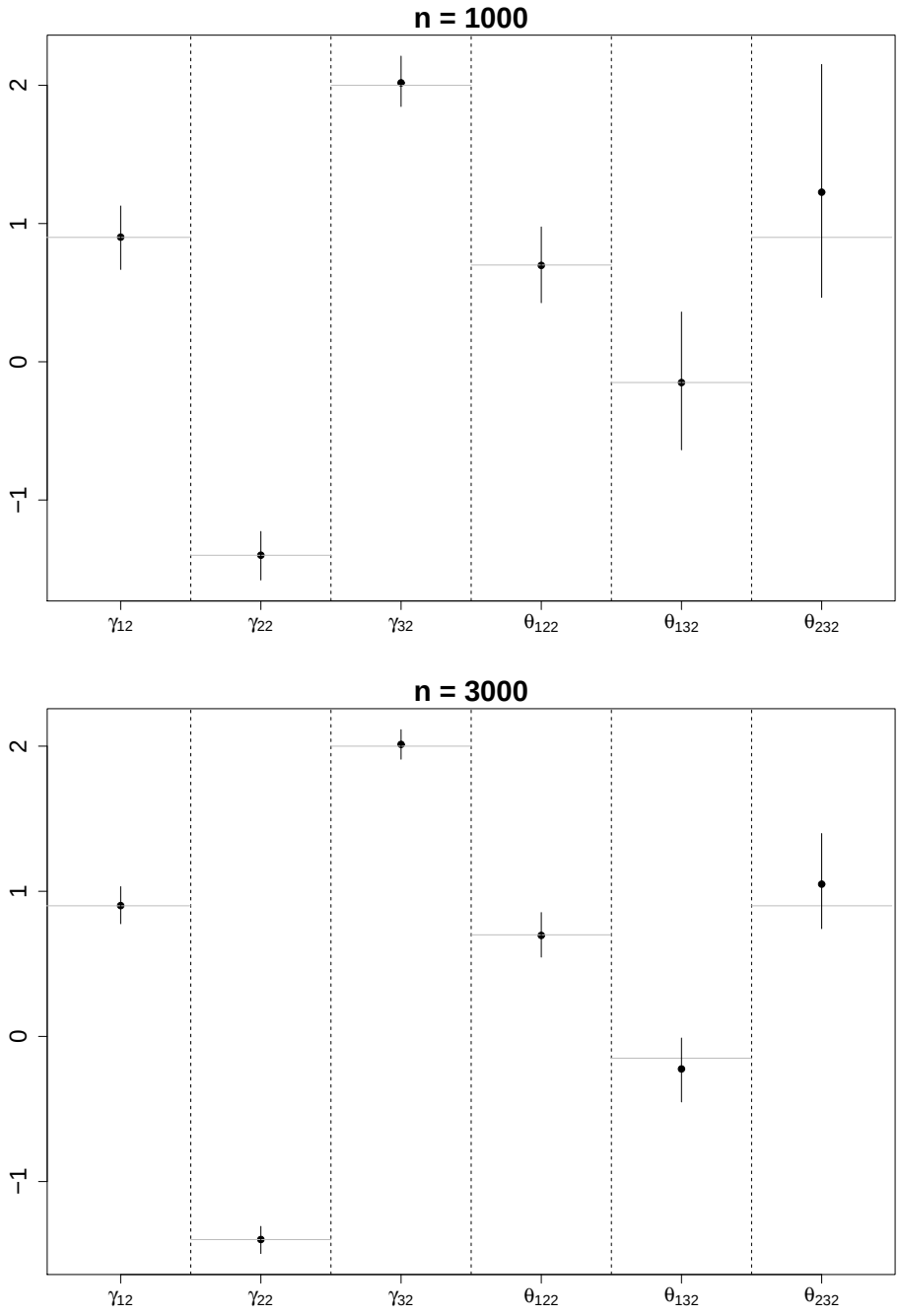

Figure 6.1: Linear coefficient estimates obtained by applying the proposed model to data simulated from a trivariate Gaussian copula model with logistic, Gumbel and normal margins. Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by gray horizontal lines.

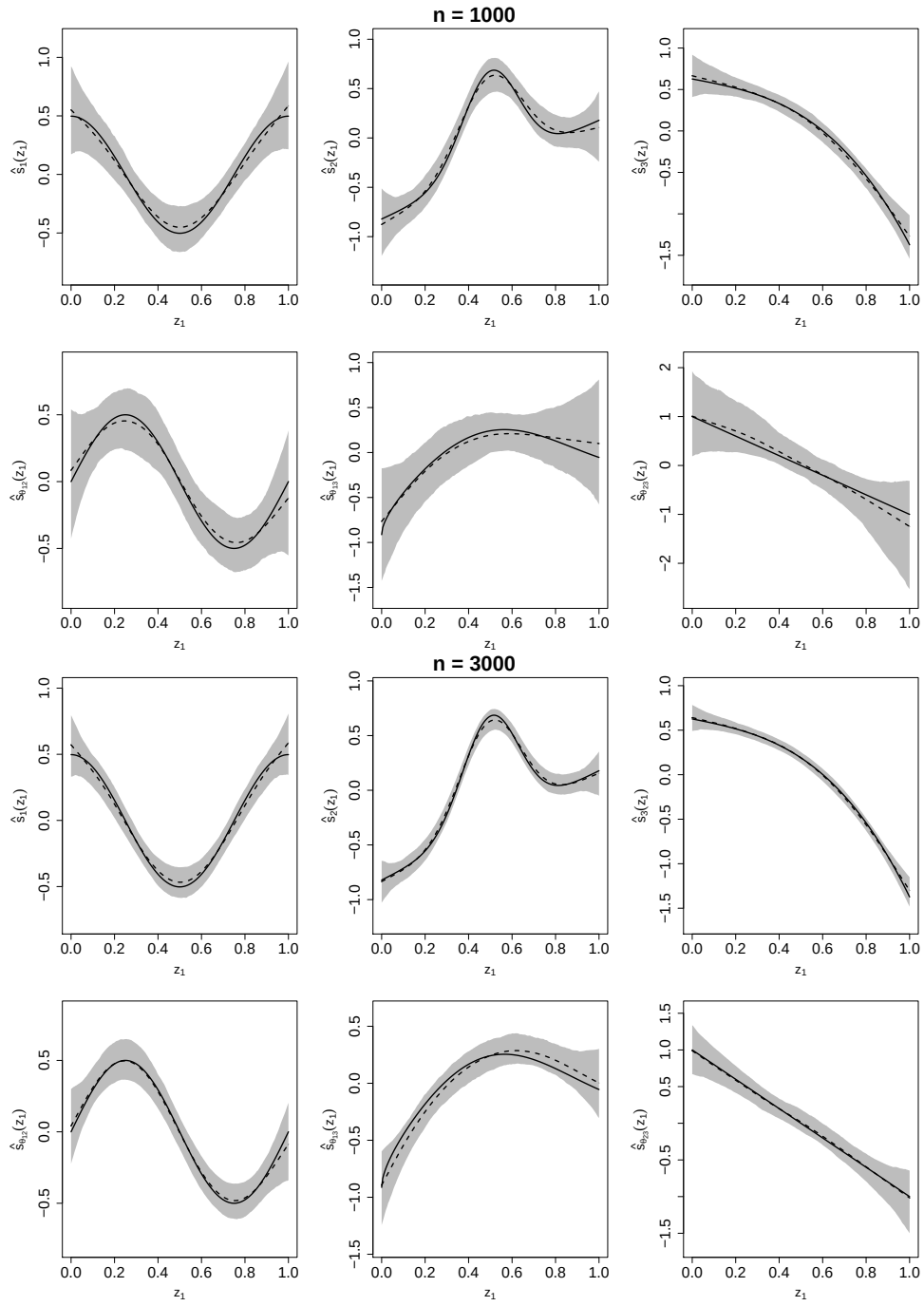

Figure 6.2: Smooth function estimates obtained by applying the proposed model to data simulated from a trivariate Gaussian copula model with logistic, Gumbel and normal margins. True functions are represented by black solid lines, mean estimates by dashed lines and point-wise ranges resulting from 5% and 95% quantiles by shaded areas.

$\texttt{gained}_i$ and $\texttt{mage}_i$ were represented using penalized thin plate regression splines, and the spatial smooth for the regional effects was set up using a Markov random field approach (Wood, 2006). To simplify the model building process we used the fact that the specification for the marginal models and their dependence can be addressed separately. For each margin we fitted three univariate GAMs based on the probit, logit and cloglog links. For each margin and link the covariate effects were always all significant. The links chosen were logit, logit and cloglog for `mb`, `lbw` and `ptb`, respectively. We then focused on the correlations' additive predictors and viewed all of their covariates effects as being part of a unique equation. We employed the classic backward selection procedure and also looked at the significance of the effects to favor more parsimonious specifications. The additive predictors for the six equations of the final model are:

$$
\begin{aligned}
\eta_{1i} &= \gamma_{11} + \gamma_{12}\texttt{nwhite}_i + \gamma_{13}\texttt{smoker}_i + s_{11}(\texttt{gained}_i) + s_{12}(\texttt{mage}_i) + s_{1\texttt{spatial}}(\texttt{county}_i),\\
\eta_{2i} &= \gamma_{21} + \gamma_{22}\texttt{nwhite}_i + \gamma_{23}\texttt{smoker}_i + s_{21}(\texttt{gained}_i) + s_{22}(\texttt{mage}_i) + s_{2\texttt{spatial}}(\texttt{county}_i),\\
\eta_{3i} &= \gamma_{31} + \gamma_{32}\texttt{nwhite}_i + \gamma_{33}\texttt{smoker}_i + s_{31}(\texttt{gained}_i) + s_{32}(\texttt{mage}_i) + s_{3\texttt{spatial}}(\texttt{county}_i),\\
\eta_{12i} &= \gamma_{12,1} + \gamma_{12,2}\texttt{nwhite}_i + s_{12}(\texttt{gained}_i) + s_{12\texttt{spatial}}(\texttt{county}_i),\\
\eta_{13i} &= \gamma_{13,1} + \gamma_{13,2}\texttt{nwhite}_i + \gamma_{13,3}\texttt{smoker}_i + s_{13,1}(\texttt{gained}_i) + s_{13,2}(\texttt{mage}_i) +\\
&\quad s_{13\texttt{spatial}}(\texttt{county}_i),\\
\eta_{23i} &= \gamma_{23,1} + s_{23,1}(\texttt{gained}_i) + s_{23,2}(\texttt{mage}_i),
\end{aligned}
$$

Some results are presented below.

Figure 6.3 shows the estimated model's correlations by county in North Carolina. Here, the effects for two binary predictors in the model were set to zero (since the majority of individuals are white and non smokers) while the continuous regressors were set at their average values. Figure 6.4 displays the estimated correlations by `gained` where the two binary predictors were set at 0, `mage` at its average value and `county` was randomly chosen (although results were very similar across counties).

Generally, the three binary outcomes are strongly correlated with each other even after accounting for covariates at marginal level. Interestingly, as shown in Figure

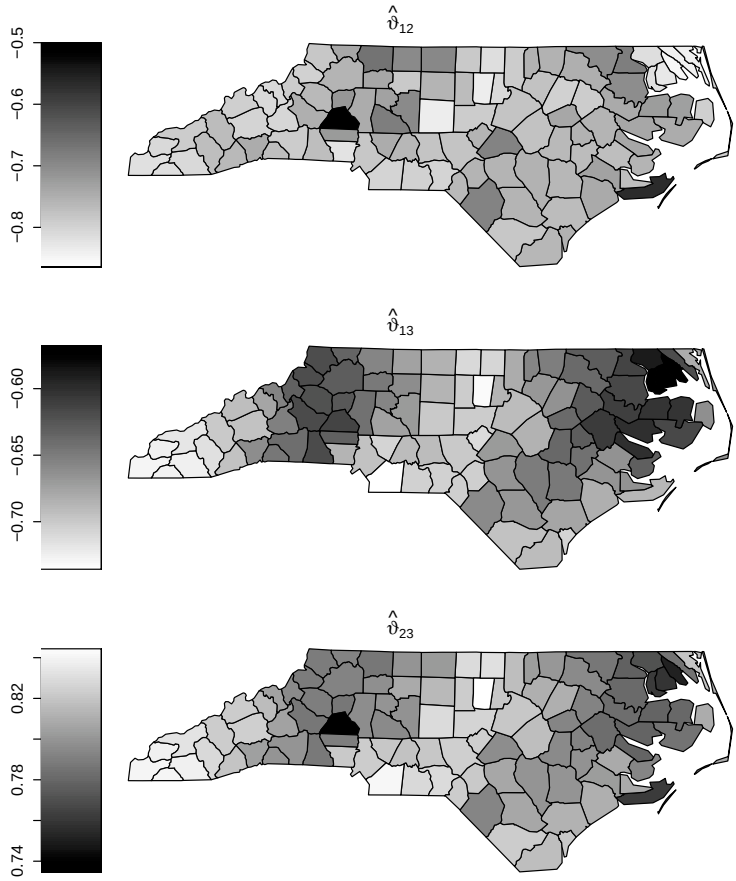

Figure 6.3: Spatially varying estimates of correlations ϑ_{12} ϑ_{13} and ϑ_{23} obtained by applying the proposed approach to North Carolina data.

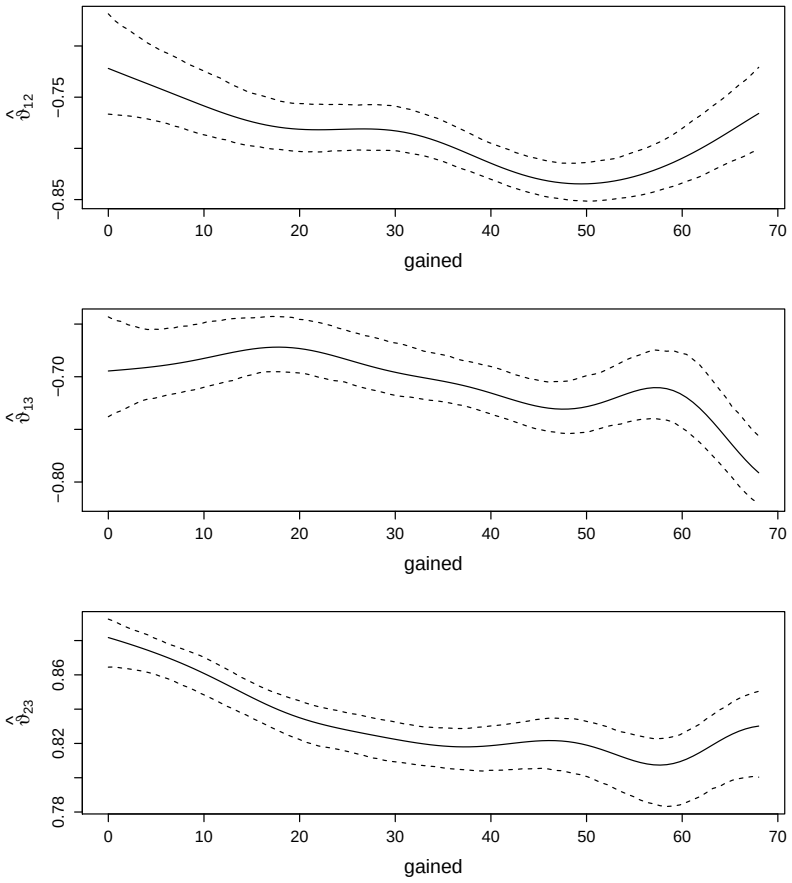

Figure 6.4: Estimates of correlations ϑ_{12} ϑ_{13} and ϑ_{23} by `gained` obtained by applying the proposed approach to North Carolina data. Point-wise 95% confidence intervals were obtained using the posterior simulation approach described in Section 2.3.2.

6.3, there is a good deal of spatial variation in the strength of the correlations. Specifically, the three responses seem to be more strongly related in the west and central areas of North Carolina than they are otherwise. Figure 6.4 suggests that the absolute association between `mb` and `lbw` increases for values of `gained` up to 50 and then decreases, the correlation between `mb` and `ptb` overall increases, and the dependence between `lbw` and `ptb` decreases for values of `gained` between 50 and 60 and then increases. These are new findings which open up questions for further research to elucidate the nature of such dependencies in North Carolina.

6.6 Discussion

This chapter proposed a generalisation of the trivariate additive binary model which allows for the model's correlation coefficients to depend on flexible additive predictors. The flexibility of the approach allows us to gain detailed insights into the unmeasured covariates accounting for a variety of regression effects. In this way, general forms of dependency (related with the correlation parameters) can be captured. The parameters of the model can be estimated simultaneously within a penalized likelihood framework based on a trust region algorithm with automatic smoothing parameter selection, and the model can be easily employed via the `SemiParTRIV()`/`gjrm()` in the `R` package `GJRM`. The potential of the approach has been demonstrated using simulated and real data.

To further enhance the flexibility of the model, an interesting extension is to allow for trivariate dependence structures other than Gaussian. This may be beneficial for obtaining a more accurate representation of the dependence between the responses which may lead to improved estimation. This will be addressed in the next chapter.

Chapter 7

Non-Gaussian Distributions

This chapter considers non-Gaussian dependencies between three binary responses. The model is described in terms of a copula-based extension where several methods are discussed.

7.1 Introduction

As shown in the previous chapters, modelling trivariate binary data based on the assumption of normality makes estimation feasible. The case of non-normal dependencies, however, is more cumbersome. This chapter discusses some copula-based possible extensions that allow for non-normal dependencies. Copula-based models allow one to form a joint multivariate distribution by specifying separately the marginal distributions and the dependence structure linking the marginals.

We consider several ways of modelling non-Gaussian error dependence by reviewing the growing literature on copula-based models for trivariate binary data. In general, this may be advantageous in empirical studies as such an extension would allow one to assume a dependence beyond that implied by the classical Gaussian distribution and at the same time to employ different marginals irrespective of the association linking them. This would consequently allow for a greater degree of flexibility in specifying and estimating the model.

In what follows, we discuss five different ways of modelling dependence for the

trivariate case: (i) Archimedean copulae; (ii) mixtures of powers; (iii) pair-copulae construction; (iv) trivariate Student t-distribution; and (v) composite likelihood approach. The advantages and disadvantages of each method are discussed, whereas some conclusions are drawn in the last section.

7.2 Copulae for trivariate binary models

For the sake of simplicity but without loss of generality, in what follows we assume that additive predictor $\eta_{zk,i}$ is a function of an intercept.

7.2.1 Trivariate Archimedean copulae

Any continuous multivariate cdf can be decomposed into univariate marginal cdfs that are connected by a copula function, which accounts for the dependence between the marginals and allows for a great deal of flexibility in specifying the joint distribution of the response variables. A class of multivariate copulae is the Archimedean copula (e.g., McNeil & Nešlehová, 2009; Noh et al., 2013; Nikoloulopoulos, 2016), which includes several popular families. Focusing on the trivariate case, we define the Archimedean copula $\tilde{\mathcal{C}}$ as

$$\tilde{\mathcal{C}}(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i}); \vartheta) = \mathscr{C}(\mathscr{C}^{-1}(\mathrm{F}_1(\eta_{1i})) + \mathscr{C}^{-1}(\mathrm{F}_2(\eta_{2i})) + \mathscr{C}^{-1}(\mathrm{F}_3(\eta_{3i})); \vartheta), \tag{7.1}$$

for some generator function $\mathscr{C} : [0 : 1] \to \mathbb{R}^+$ with $\mathscr{C}(0) = 1$ and $\mathscr{C}(\infty) = 0$. McNeil & Nešlehová (2009) provide necessary and sufficient conditions for $\mathscr{C}$ to generate a feasible Archimedean copula. The generator $\mathscr{C}$ is required to be 3-monotone, that is differentiable up to the first order with $(-1)^{\bar{d}}\mathscr{C}^{(\bar{d})}(\mathsf{L}) \geq 0$, $\forall \bar{d} = 0, 1$, for any $\mathsf{L} \in [0, \infty)$ and with $(-1)\mathscr{C}^{(1)}(\mathsf{L})$ being non-decreasing and convex on $[0, \infty)$. There are many families of Archimedean copulae; among the best known are the Clayton (Clayton, 1978), Frank (Frank, 1979) and Gumbel (Gumbel, 1960), whose form is presented in Table 7.1.

Copula	$\tilde{\mathcal{C}}(\bar{v}_1, \bar{v}_2, \bar{v}_3; \vartheta)$	Range of ϑ
Clayton	$\left(\bar{v}_1^{-\vartheta} + \bar{v}_2^{-\vartheta} + \bar{v}_3^{-\vartheta} - 2\right)^{-\frac{1}{\vartheta}}$	$\vartheta \in (0, \infty)$
Frank	$-\frac{1}{\vartheta} \log\left\{1 + \frac{(e^{-\vartheta\bar{v}_1}-1)(e^{-\vartheta\bar{v}_2}-1)(e^{-\vartheta\bar{v}_3}-1)}{\left(e^{-\vartheta}-1\right)^3}\right\}$	$\vartheta \in \mathbb{R} \setminus \{0\}$
Gumbel	$\exp\left\{-\left((-\log\bar{v}_1)^\vartheta + (-\log\bar{v}_2)^\vartheta + (-\log\bar{v}_3)^\vartheta\right)^{\frac{1}{\vartheta}}\right\}$	$\vartheta \in [1, \infty)$

Table 7.1: Definition of trivariate Archimedean copulae, with corresponding parameter range of association parameter ϑ.

In general, Archimedean copulae have the advantage of producing closed form expressions and also of yielding different kinds of asymmetries. The specifications in Table 7.1, however, can be rather restrictive in practical situations as they imply a symmetric dependence between the three pairs $(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}))$, $(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_3(\eta_{3i}))$ and $(\mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i}))$. That is, the association parameters that characterize the dependence between the three responses are assumed to be equal. This means that $\vartheta_{12} = \vartheta_{13} = \vartheta_{23} = \vartheta$ and thus a single dependence parameter can be estimated from the model. This assumption can rarely be satisfied in practice. The next section shows how trivariate copulae can be constructed in a less restrictive structure.

7.2.2 Mixtures of powers

Joe (1993) extended multivariate Archimedean copulae to a more flexible class using the mixtures of powers. This approach produces two dependence parameters for a trivariate copula. Based on bivariate Archimedean copulae and using Laplace transformations, the trivariate mixtures of powers representation is

$$\begin{aligned}\tilde{\mathcal{C}}(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i})) \;=\; & \int_0^\infty \int_0^\infty \mathrm{G}^{\bar{\alpha}_2}(\mathrm{F}_1(\eta_{1i}))\mathrm{G}^{\bar{\alpha}_2}(\mathrm{F}_2(\eta_{2i}))d\mathscr{M}_2(\bar{\alpha}_2; \bar{\alpha}_1) \\ & \mathrm{G}^{\bar{\alpha}_1}(\mathrm{F}_3(\eta_{3i}))d\mathscr{M}_1(\bar{\alpha}_1), \end{aligned} \tag{7.2}$$

where $\mathrm{G}(\mathrm{F}_1(\eta_{1i})) = \exp\left(-\mathscr{C}^{-1}\left(\mathrm{F}_1(\eta_{1i})\right)\right)$, $\mathrm{G}(\mathrm{F}_2(\eta_{2i})) = \exp\left(-\mathscr{C}^{-1}\left(\mathrm{F}_2(\eta_{2i})\right)\right)$, $\mathrm{G}(\mathrm{F}_3(\eta_{3i})) = \exp\left(-\mathscr{V}^{-1}\left(\mathrm{F}_3(\eta_{3i})\right)\right)$ and $\mathscr{V}$ is a Laplace transformation. Distribution $\mathscr{M}_1$ has Laplace transformation $\mathscr{C}(\cdot)$ and $\mathscr{M}_2$ has Laplace transformation $\left(\left(\mathscr{C}^{-1} \circ \mathscr{V}\right)^{-1}\left(-\bar{\alpha}_1^{-1}\log(\cdot)\right)\right)^{-1}$. In this formulation, $\bar{\alpha}_1$ can be thought of as the

unobserved variables that affect $\mathrm{F}_1(\eta_{1i})$, $\mathrm{F}_2(\eta_{2i})$ and $\mathrm{F}_3(\eta_{3i})$ while $\bar{\alpha}_2$ affects $\mathrm{F}_1(\eta_{1i})$ and $\mathrm{F}_2(\eta_{2i})$, $\forall \bar{\alpha}_1, \bar{\alpha}_2 > 0$. When $\mathscr{C} = \mathscr{V}$, expression (7.2) simplifies to (7.1). When $\mathscr{C} \neq \mathscr{V}$, the trivariate Archimedean copula corresponding to (7.2) can be formed as

$$\begin{aligned}\tilde{\mathcal{C}}(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i})) &= \mathscr{V}\left(\mathscr{V}^{-1} \circ \mathscr{C}\left(\mathscr{C}^{-1}(\mathrm{F}_1(\eta_{1i})) + \mathscr{C}^{-1}(\mathrm{F}_2(\eta_{2i}))\right) + \right. \\ &\quad \left. \mathscr{V}^{-1}(\mathrm{F}_3(\eta_{3i}))\right). \qquad (7.3)\end{aligned}$$

The derivation of the above expression can be found in Marshall & Olkin (1988). Table 7.2 reports the expressions for some trivariate copulae when applying the mixtures of powers approach. We refer the reader to Joe (1993) for details on deriving these expressions.

Copula	$\mathcal{C}(\bar{v}_1, \bar{v}_2, \bar{v}_3; \vartheta_1, \vartheta_2)$
Clayton	$\left[\left(\bar{v}_1^{-\vartheta_2} + \bar{v}_2^{-\vartheta_2} - 1\right)^{\frac{\vartheta_1}{\vartheta_2}} + \bar{v}_3^{-\vartheta_1} - 1\right]^{-\frac{1}{\vartheta_1}}$
Frank	$-\frac{1}{\vartheta_1}\log\left\{1 - \tilde{v}_1^{-1}\left(1 - \left[1 - \tilde{v}_2^{-1}(1 - e^{-\vartheta_2\bar{v}_1})(1 - e^{-\vartheta_2\bar{v}_2})\right]^{\frac{\vartheta_1}{\vartheta_2}}\right)(1 - e^{-\vartheta_1\bar{v}_3})\right\}$
Gumbel	$\exp\left\{-\left(\left[(-\log\bar{v}_1)^{\vartheta_2} + (-\log\bar{v}_2)^{\vartheta_2}\right]^{\frac{\vartheta_1}{\vartheta_2}} + (-\log\bar{v}_3)^{\vartheta_1}\right)^{\frac{1}{\vartheta_1}}\right\}$

Table 7.2: Definition of trivariate copulae obtained from the mixtures of powers approach. The association parameters ϑ_1 and ϑ_2 denote the association between $[\bar{v}_1, \bar{v}_2]$ and $\bar{v}_3$, and $\bar{v}_1$ and $\bar{v}_2$, respectively, while parameters $\tilde{v}_1$ and $\tilde{v}_2$ are equal to $1 - e^{-\vartheta_1}$ and $1 - e^{-\vartheta_2}$. The parameter ranges of ϑ_1 and ϑ_2 are the same as those in Table 7.1.

Although the specifications in Table 7.2 are not as restrictive as the specifications in Table 7.1, they are still not capable of modelling separately the dependence between all pairs. Instead, they are symmetric with respect to $(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}))$ which is often not the case in empirical applications. The partially symmetric formulation of (7.3) also requires the constraint $\vartheta_1 \leq \vartheta_2$, where $\vartheta_2 = \vartheta_{12}$ and $\vartheta_1 = \vartheta_{13} = \vartheta_{23}$. Moreover, the ordering of the marginals in (7.3) can change. For instance, instead of using the grouping $([\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}); \vartheta_2], \mathrm{F}_3(\eta_{3i}); \vartheta_1)$, one could employ $([\mathrm{F}_1(\eta_{1i}), \mathrm{F}_3(\eta_{3i}); \vartheta_2], \mathrm{F}_2(\eta_{2i}); \vartheta_1)$ which provides a different interpretation for ϑ_1 and ϑ_2. Presumably each grouping is justified by some set of assumptions

about dependence. This constitutes a potential weakness as it may be difficult to choose a priori the ordering of the marginals in empirical studies. For a more detailed description of the method we refer the reader to Joe (1997, Ch.5), Zimmer & Trivedi (2006) and Trivedi & Zimmer (2007, Ch.3).

7.2.3 Pair-copulae constructions in 3 dimensions

An alternative approach to model multivariate data is the *pair-copulae* construction (PCC), originally proposed by Joe (1996), which can be defined as a multivariate copula that is constructed from a cascade of bivariate copulae. That is, the joint distribution is obtained from using bivariate pair-copulae that may be conditional on a specific set of variables, allowing to model the dependence among the marginals. Due to their high flexibility and their simple structure, PCCs are becoming increasingly popular for constructing continuous multivariate distributions (e.g., Aas et al., 2009; Czado, 2010; Panagiotelis et al., 2012).

A PCC in 3 dimensions can be obtained by computing the trivariate cdf (Joe, 1996)

$$\tilde{\mathcal{C}}(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i})) = \int_{-\infty}^{\eta_{2i}} \tilde{\mathcal{C}}_{13|2} \left(\mathrm{F}_{1|2}(\eta_{1i}|l_{2i}; \vartheta_{12}), \mathrm{F}_{3|2}(\eta_{3i}|l_{2i}; \vartheta_{23}); \vartheta_{13|2}\right) \mathrm{F}_2(l_{2i}) dl_{2i}, \tag{7.4}$$

where $\tilde{\mathcal{C}}_{13|2}$ is a conditional bivariate copula, $\vartheta_{13|2}$ denotes the partial correlation coefficient defined as $(\vartheta_{13} - \vartheta_{12}\vartheta_{23}) \,/\, \left(\sqrt{1-\vartheta_{12}^2}\sqrt{1-\vartheta_{23}^2}\right)$ and $\mathrm{F}_{1|2}$ and $\mathrm{F}_{3|2}$ are conditional cdfs obtained from the bivariate cdfs $\tilde{\mathcal{C}}(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}); \vartheta_{12})$ and $\tilde{\mathcal{C}}(\mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i}); \vartheta_{23})$. Note that the above representation is based on the assumption that the conditional copula $\tilde{\mathcal{C}}_{13|2}$ depends on the conditioning variables only indirectly through the conditional distribution functions that constitute its arguments. This leads to the so-called *simplified* PCC. Here, the potentially complex dependence between variables that are conditioned on and the copula functions can be neglected, thus making PCCs tractable for inference. Further, the PCC is order dependent. That is, in (7.4) there are three possible ways of permuting $\mathrm{F}_1(\eta_{1i})$, $\mathrm{F}_2(\eta_{2i})$ and

$\mathrm{F}_3(\eta_{3i})$. A different choice of the variables' order leads to a different PCC and to a different factorisation of the joint trivariate distribution. This consequently implies a different interpretation of the correlation parameters. Therefore, selection of an appropriate PCC depends crucially on the study at hand; failure in selecting an appropriate construction may yield misleading results. This may make the approach inconvenient for practitioners as choosing an appropriate conditioning may not be straightforward in practical studies. Moreover the evaluation of (7.4) remains a challenging computational problem.

7.2.4 The trivariate Student-t distribution

The dependence of three responses can be characterized through a trivariate Student-t distribution, which shares similar features to the Gaussian. In this work we explored the benefits of using the trivariate Student-t copula $\mathcal{C}\left(\mathrm{F}_1(\eta_{1i}), \mathrm{F}_2(\eta_{2i}), \mathrm{F}_3(\eta_{3i})\right) = \mathrm{T}_{3,\tilde{\nu}}(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon})$, where $\tilde{\nu}$ denotes the degrees of freedom.

The trivariate Student-t copula has the appealing ability to allow for tail dependence. Similarly to the Gaussian case, the difficulty with this distribution is that the derivation of the analytical score vector and Hessian matrix requires working with trivariate integrals, which is not straightforward. As mentioned in Section 6.3, analytical derivative information is essential for the algorithm to work properly in a complex regression setting; preliminary work confirmed that the use of classic optimization techniques, implemented using the `R` functions `nlm()` and `optim()`, can be inefficient and unstable when compared to a trust-region algorithm using analytical first and second order derivatives.

Before attempting a full and proper implementation of this distribution, we experimented with it using a very simple simulation set up. Specifically, we employed a DGP based on the following system of equations

$$\begin{aligned} y^*_{1i} &= -0.04 + 0.5v_{1i} + \varepsilon_{1i}, \\ y^*_{2i} &= -0.20 + 0.4v_{1i} + \varepsilon_{2i}, \\ y^*_{3i} &= 0.05 - 0.2v_{1i} + \varepsilon_{3i}, \end{aligned}$$

where $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})^\top \sim \mathrm{T}_{3,\tilde{\nu}}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\tilde{\nu} = 3$, and v_{mi} denotes a binary regressor. The correlation parameters were set as $\vartheta_{12} = 0.2$, $\vartheta_{13} = 0.4$ and $\vartheta_{23} = 0.8$. We generated 250 datasets with sample size equal to 1000.

Figure 7.1 compares the parameter estimates obtained when using the trivariate Gaussian and Student-t copula models. The latter was implemented via the `R` routine `optim()` where numerical derivative information was used. The results are very similar across the two approaches, hence suggesting that there is no much to be gained by relaxing the Gaussian assumption in the current context.

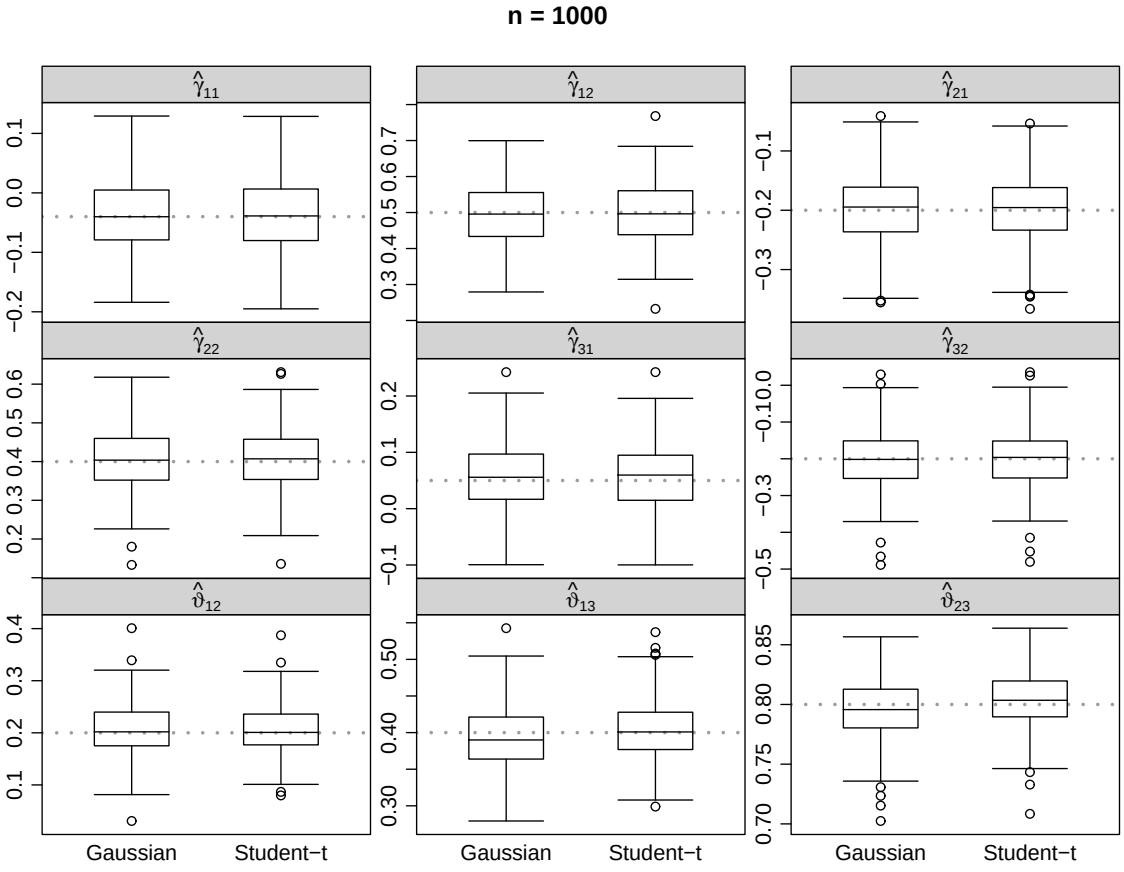

Figure 7.1: Boxplots of parameter estimates obtained by applying the trivariate Gaussian and Student-t copula models to 250 simulated datasets with sample size equal to 1000. The first two rows refer to the regression coefficient estimates and the last row to the estimated correlations. The true parameter values are represented by horizontal gray dotted lines.

7.2.5 Composite likelihood

The difficulty with evaluating high-dimensional integrals can be overcome by employing the *composite likelihood* (CL) technique by Zhao & Joe (2005), which is based on a two-stage method. In our case, estimates $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_3$ are first obtained by maximizing the CL function of univariate margins

$$\ell_1(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} \left\{ \ell_1(y_{1i}; \boldsymbol{\beta}_1) + \ell_1(y_{2i}; \boldsymbol{\beta}_2) + \ell_1(y_{3i}; \boldsymbol{\beta}_3) \right\},$$

where $\ell_1(y_{mi}; \boldsymbol{\beta}_m)$ corresponds to i^{th} contribution to the m^{th} univariate log-likelihood function. Next, with $\boldsymbol{\beta}_m$ fixed at the estimate of $\ell_1(\boldsymbol{\beta})$, we estimate ϑ_{zk}, $\forall z, k$, by maximizing the CL function of bivariate margins which is the summation of the log-likelihoods of pairs (y_{zi}, y_{ki}). That is,

$$\begin{aligned} \ell_2(\hat{\boldsymbol{\beta}}; \boldsymbol{\Sigma}) \;=\; & \sum_{i=1}^{n} \Big\{ \mathsf{w}_{12,i} \ell_2(y_{1i}, y_{2i}; \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \vartheta_{12}) + \mathsf{w}_{13,i} \ell_2(y_{1i}, y_{3i}; \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_3, \vartheta_{13}) + \\ & \mathsf{w}_{23,i} \ell_2(y_{2i}, y_{3i}; \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3, \vartheta_{23}) \Big\}, \end{aligned}$$

where $\mathsf{w}_{zk,i}$ are positive weights and $\ell_2(y_{zi}, y_{ki}; \hat{\boldsymbol{\beta}}_z, \hat{\boldsymbol{\beta}}_k, \vartheta_{zk})$ corresponds to the i^{th} contribution to the bivariate log-likelihood of (y_{zi}, y_{ki}). The choice of optimal weights, such that the loss of efficiency is as small as possible, is addressed in the works by Kuk & Nott (2000), Andersen (2004), Zhao & Joe (2005) and Joe & Lee (2009).

The CL method is a relatively simple approach that can deliver reasonable results when the log-likelihood function is computationally too difficult to implement. Hence the motivation for the use of this method is usually computational tractability and is commonly employed in the context of joint modelling of high-dimensional responses. In the trivariate context, however, tractability is not a big concern. A potential drawback of this approach is that the information in the data may not be fully exploited as parameter estimation is carried out in two steps, hence making the CL approach less efficient than the simultaneous parameter estimator.

7.3 Discussion

We have discussed the use of copula-based models for trivariate binary data and derived some results. The aim was to model dependence structure beyond the classical Gaussian distribution. Although the approaches discussed in this chapter allow for non-Gaussian structures, the majority of them make certain strong assumptions which may be regarded as acceptable only in specific applied contexts. In fact, such methods would limit the generality as well as applicability of the modelling approach presented here. The only suitable alternative would appear to be the trivariate Student-t distribution, however, as shown, there is not much to be gained by using such distribution in our context. In conclusion, the Gaussian copula seems to be a sensible and tractable modelling choice for the case of trivariate binary data.

Chapter 8

Final remarks

8.1 Summary of the thesis

The current thesis has been mainly motivated by the recent applied and methodological interest in modelling simultaneously more responses in a regression setting and we aimed to widen the applicability of the method by introducing a flexible modelling framework for trivariate Gaussian copula additive models that accounts for the presence of unobservables. Our target was two fold: (i) to develop the theory needed for fitting flexible trivariate equation models; and (ii) to make the developments available to the public use by implementing a reliable estimation algorithm in the `R` language.

In Chapter 2 we outlined a flexible joint modelling framework by considering trivariate probit models with additive predictors. We have shown that our extended framework provides improvement to model fit and also offers better prediction when compared to existing estimation approaches. We have also shown that under small or moderate sample size, the MLE results in some situations are unsatisfactory. Such problem has been tackled in Chapter 3 by introducing an approach for penalizing the correlation coefficients. The software for straightforward implementation of the proposed approach has been provided, while some asymptotic properties of the proposed estimator have also been discussed. The validity of the method has been confirmed via simulation studies.

As byproduct of the framework developed in Chapters 2 and 3, Chapter 4 introduced a flexible framework to model unmeasured confounding and non-random sample selection where several models are discussed. A Monte Carlo experiment showed the promising performance of the double sample selection model, while the endogenous trivariate model has been used for the analysis of two chronic diseases on labor force participation.

Chapter 5 proposed a framework which allows researchers to estimate trivariate binary models with arbitrary link functions. We explored the possibility of modelling the margins using probit, logit and complementary log-log links through the use of Gaussian copulae. The model was illustrated through simulated data.

In Chapter 6, we further enhanced the trivariate Gaussian binary model by allowing the model's association parameters to depend on several types of covariate effects. The practical performance of the approach was assessed via real data, where we jointly analysed multiple births, premature birth and low birth weight in North Carolina using a triariate Gaussian copula additive model that permits each correlation parameter to be specified as a function of an additive predictor.

In Chapter 7 we have discussed some copula based possible extensions to model the dependence structure beyond the classical Gaussian distribution. After a review of the available methods, we concluded that such an extension may not be particularly interesting for the class of models we consider. It looks like that maintaining the assumption of the normality is not too problematic for trivariate binary data.

8.2 Topics for future research

Although in this thesis we have restricted to the case of binary responses only, it is conceivable that other types of marginal outcomes might be of interest (e.g., continuous, discrete). Therefore, an interesting extension of the proposed methodology would be to account for outcome types other than binary. This will considerably extend the scope and applicability of the trivariate modelling approach introduced in this thesis. Such an extension will require deriving the model's log-likelihood and

its respective score and Hessian components.

A second extension of the proposed model would be to consider systems involving more than three responses. The parameter estimation of such a model can be employed via the developed methodology presented in Chapters 2 and 3, where the log-likelihood function as well as the analytical derivative components need to be recomputed. This can be implemented via the propositions presented throughout the thesis by replacing M with the total number of equations one wishes to use.

Appendix A

Complements to Chapter 2

A.1 Proof of Lemma 2.3.1

Proof. For convenience we ignore index $\tilde{k}$ and term $\mathcal{Y}_{i\tilde{k}}$. By definition,

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{y}_i;\boldsymbol{\delta}) &= \mathbb{P}(-\tilde{y}_{1i}y^*_{1i} \leq 0,\ldots,-\tilde{y}_{Mi}y^*_{Mi} \leq 0) \\
&= \mathbb{P}(-\tilde{y}_{1i}(\eta_{1i}+\varepsilon_{1i}) \leq 0,\ldots,-\tilde{y}_{Mi}(\eta_{Mi}+\varepsilon_{Mi}) \leq 0) \\
&= \mathbb{P}(-\tilde{y}_{1i}\eta_{1i}-\tilde{y}_{1i}\varepsilon_{1i} \leq 0,\ldots,-\tilde{y}_{Mi}\eta_{Mi}-\tilde{y}_{Mi}\varepsilon_{Mi} \leq 0) \\
&= \mathbb{P}(-\tilde{y}_{1i}\varepsilon_{1i} \leq \tilde{y}_{1i}\eta_{1i},\ldots,-\tilde{y}_{Mi}\varepsilon_{Mi} \leq \tilde{y}_{Mi}\eta_{Mi}) \\
&= \Phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i;\mathbf{0},\boldsymbol{\Sigma}) \\
&= \int_{-\infty}^{\tilde{y}_{Mi}\eta_{Mi}} \cdots \int_{-\infty}^{\tilde{y}_{1i}\eta_{1i}} \phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\Sigma}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}. \qquad \text{(A.1)}
\end{aligned}$$

Since $\tilde{y}_{mi}$ is either equal to -1 or 1, it follows that $\boldsymbol{\mathcal{B}}_i = \boldsymbol{\mathcal{B}}_i^{-1}$ and $|\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i| = |\boldsymbol{\Sigma}|$. In addition, the pdf of a multivariate normal vector $-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\Sigma}$ can be re-expressed as the pdf of a multivariate normal vector

$\boldsymbol{\varepsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i$, that is

$$\begin{aligned}\phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(-\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i)^\top(\boldsymbol{\Sigma})^{-1}(-\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i)\right\}\\ &= |2\pi(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{l}_i^\top(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)^{-1}\boldsymbol{l}_i\right\}\\ &= \phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i).\end{aligned}$$

Therefore, equation (A.1) can be written as

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{y}_i;\boldsymbol{\delta}) &= \int_{-\infty}^{\tilde{y}_{Mi}\eta_{Mi}}\cdots\int_{-\infty}^{\tilde{y}_{1i}\eta_{1i}}\phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)\prod_{\tilde{c}=1}^{M}dl_{\tilde{c},i}\\ &= \Phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)\\ &= \Phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{w}_i;\mathbf{0},\boldsymbol{\Upsilon}_i),\end{aligned}$$

where

$$\boldsymbol{\Upsilon}_i = \begin{pmatrix}1 & r_{12,i} & \dots & r_{1M,i}\\ r_{12,i} & 1 & \dots & r_{2M,i}\\ \vdots & \vdots & \ddots & \vdots\\ r_{1M,i} & r_{2M,i} & \dots & 1\end{pmatrix},$$

for $r_{zk,i} = \vartheta_{zk}(2y_{zi}-1)(2y_{ki}-1)$, $\forall z,k,i$. Note that the above derivation applies to all $\tilde{k}$s, thus the likelihood $\mathcal{L}_{i\tilde{k}}$ is equal to

$$\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i;\boldsymbol{\delta}) = \{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{w}_i)_{\tilde{k}};\mathbf{0},(\boldsymbol{\Upsilon}_i)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}},$$

as required. □

A.2 Computation of trivariate normal integrals

A.2.1 Numerical computation of multivariate normal integrals

In what follows we describe in detail the numerical method used in `pmnorm()` in `R` package `mnormt` (Azzalini, 2016) for the evaluation of multivariate normal integrals.

Introduction

Let $(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{J}}) = (\mathcal{E}_1, \mathcal{J}_1) \times (\mathcal{E}_2, \mathcal{J}_2) \times \ldots \times (\mathcal{E}_M, \mathcal{J}_M)$ be a M-dimensional rectangle. Then the problem is to find

$$\Phi_M(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{J}}) = \frac{1}{\sqrt{|\boldsymbol{\Upsilon}_i|(2\pi)^M}} \int_{\mathcal{E}_1}^{\mathcal{J}_1} \ldots \int_{\mathcal{E}_M}^{\mathcal{J}_M} \exp\left(-\frac{1}{2}\boldsymbol{l}_i^\top \boldsymbol{\Upsilon}_i^{-1} \boldsymbol{l}_i\right) d\boldsymbol{l}_i, \quad \text{(A.2)}$$

where $|\boldsymbol{\Upsilon}_i|$ denotes the determinant of $\boldsymbol{\Upsilon}_i$. Since we are interested in the value of the distribution function $\Phi_M(\boldsymbol{w}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}_i)$, we have that $\boldsymbol{\mathcal{E}} = (-\infty, \ldots, -\infty)$; this reduces the number of variables in the problem and makes the evaluation of Φ_M simpler (see next section for more details). In addition, the upper bound $\boldsymbol{\mathcal{J}}$ is equal to $(w_{1,i}, \ldots, w_{M,i})$. For $M = 1$ and $M = 2$, a reliable way to calculate the distribution function is via `pnorm()` in `stats` (Team & contributors worldwide, 2015) and `pbinorm()` in `VGAM` (Yee, 2015). Here, we assume $M > 2$ and we describe Genz's approach for computing Φ_M which uses numerical integration software based on subregion adaptive methods. A problem, however, that arises with these methods is that they assume finite integration limits. Because infinite limits are used in our case, we need to handle them: we apply a sequence of transformations to turn the problem into a form that allows for efficient computation of Φ_M. Note that even if $\boldsymbol{\mathcal{E}}$ is finite, the transformations are also applied in order to make the numerical computation of the integral easy. The set of transformations that are employed are described in the next section.

Genz's method

The basic idea of this method is to transform the original domain of integration $(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{J}})$ to $[0,1]^M = [0,1] \times [0,1] \times \ldots \times [0,1]$. First we will keep the domain of integration general and assume that both $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{J}}$ are finite. Then we move onto our case where $\mathcal{E}_m = -\infty$ and $\mathcal{J}_m = w_{m,i}$, $\forall m$. Genz's method can be employed using the following sequence of three transformations.

(T.1) We begin by employing the Cholesky decomposition transformation $\boldsymbol{l}_i = \mathbf{C}_i^* \boldsymbol{a}_i$, where $\mathbf{C}_i^*$ denotes the Cholesky factor of the covariance matrix $\boldsymbol{\Upsilon}_i$, such that $\mathbf{C}_i^*$ is a lower triangular matrix and $\boldsymbol{\Upsilon}_i = \mathbf{C}_i^* \mathbf{C}_i^{*\top}$. Vector $\boldsymbol{a}_i$ consists of univariate standard normal random variables that are independent of each other. Applying this transformation to equation (A.2) leads to

$$
\begin{aligned}
\Phi_M(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{J}}) &= 1/\sqrt{|\boldsymbol{\Upsilon}_i|(2\pi)^M} \int_{\mathcal{E}_1'}^{\mathcal{J}_1'} \int_{\mathcal{E}_2'(a_1)}^{\mathcal{J}_2'(a_1)} \ldots \int_{\mathcal{E}_M'(a_1,\ldots,a_{M-1})}^{\mathcal{J}_M'(a_1,\ldots,a_{M-1})} \exp\left(-\frac{1}{2} (\mathbf{C}_i^* \boldsymbol{a}_i)^\top \right. \\
&\quad \left. (\mathbf{C}_i^* \mathbf{C}_i^{*\top})^{*-1} (\mathbf{C}_i^* \boldsymbol{a}_i) \right) |\mathbf{C}_i^*| d\boldsymbol{a}_i \\
&= 1/\left(|\boldsymbol{\Upsilon}_i^{\frac{1}{2}}| (2\pi)^{\frac{M}{2}} \right) \int_{\mathcal{E}_1'}^{\mathcal{J}_1'} \int_{\mathcal{E}_2'(a_1)}^{\mathcal{J}_2'(a_1)} \ldots \int_{\mathcal{E}_M'(a_1,\ldots,a_{M-1})}^{\mathcal{J}_M'(a_1,\ldots,a_{M-1})} \exp\big(\\
&\quad \left. -\frac{1}{2} \boldsymbol{a}_i^\top \mathbf{C}_i^{*\top} \mathbf{C}_i^{*-\top} \mathbf{C}_i^{*-1} \mathbf{C}_i^* \boldsymbol{a}_i \right) |\boldsymbol{\Upsilon}_i^{\frac{1}{2}}| d\boldsymbol{a}_i \\
&= \frac{1}{(2\pi)^{\frac{M}{2}}} \int_{\mathcal{E}_1'}^{\mathcal{J}_1'} \int_{\mathcal{E}_2'(a_1)}^{\mathcal{J}_2'(a_1)} \ldots \int_{\mathcal{E}_M'(a_1,\ldots,a_{M-1})}^{\mathcal{J}_M'(a_1,\ldots,a_{M-1})} \exp\left(-\frac{1}{2} \boldsymbol{a}_i^\top \boldsymbol{a}_i \right) d\boldsymbol{a}_i \\
&= \int_{\mathcal{E}_1'}^{\mathcal{J}_1'} \frac{1}{\sqrt{2\pi}} e^{-\frac{a_1^2}{2}} \int_{\mathcal{E}_2'(a_1)}^{\mathcal{J}_2'(a_1)} \frac{1}{\sqrt{2\pi}} e^{-\frac{a_2^2}{2}} \ldots \int_{\mathcal{E}_M'(a_1,\ldots,a_{M-1})}^{\mathcal{J}_M'(a_1,\ldots,a_{M-1})} \frac{1}{\sqrt{2\pi}} e^{-\frac{a_M^2}{2}} \\
&\quad da_M \ldots da_1 \\
&= \int_{\mathcal{E}_1'}^{\mathcal{J}_1'} \phi(a_1) \int_{\mathcal{E}_2'(a_1)}^{\mathcal{J}_2'(a_1)} \phi(a_2) \ldots \int_{\mathcal{E}_M'(a_1,\ldots,a_{M-1})}^{\mathcal{J}_M'(a_1,\ldots,a_{M-1})} \phi(a_M) da_M \ldots da_1,
\end{aligned}
\tag{A.3}
$$

where the limits $\mathcal{E}_m'(a_1, \ldots, a_{M-1})$ and $\mathcal{J}_m'(a_1, \ldots, a_{M-1})$ come from inequality $\boldsymbol{\mathcal{E}} \leq \boldsymbol{l}_i = \mathbf{C}_i^* \boldsymbol{a}_i \leq \boldsymbol{\mathcal{J}}$. Specifically, for $m = 1$

$$
\mathcal{E}_1' = \mathcal{E}_1 \leq a_1 \leq \mathcal{J}_1 = \mathcal{J}_1',
$$

while for $m = 2, \ldots, M$

$$\mathcal{E}'_m = \frac{\left(\mathcal{E}_m - \sum_{h=1}^{m-1} c^*_{mh,i} a_h\right)}{c^*_{mm,i}} \leq a_m \leq \frac{\left(\mathcal{J}_m - \sum_{h=1}^{m-1} c^*_{mh,i} a_h\right)}{c^*_{mm,i}} = \mathcal{J}'_m,$$

where $\mathcal{E}'_m = \mathcal{E}'_m(a_1, \ldots, a_{M-1})$ and $\mathcal{J}'_m = \mathcal{J}'_m(a_1, \ldots, a_{M-1})$. The elements $c^*_{mh,i}$ and $c^*_{mm,i}$ refer to the components of the lower triangular matrix $\mathbf{C}^*_i$, that is

$$\mathbf{C}^*_i = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ c^*_{21,i} & c^*_{22,i} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ c^*_{M1,i} & c^*_{M2,i} & \ldots & c^*_{MM,i} \end{pmatrix}.$$

The element $c^*_{11,i}$ is equal to 1 because of the following relation: $c^*_{11,i} = \sqrt{r_{11,i}}$ where $r_{11,i}$ refers to the (1, 1) diagonal element of $\mathbf{\Upsilon}_i$. Since $r_{11,i} = 1$, it follows that $c^*_{11,i} = \sqrt{1} = 1$.

(T.2) Next we transform the a_m's by using $a_m = \Phi^{-1}(\mathcal{Z}_m)$, where $\Phi(a_m)$ is the standard univariate normal distribution. Therefore equation (A.3) becomes

$$\begin{aligned} \Phi_M(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{J}}) &= \int_{\mathcal{S}_1}^{\mathcal{T}_1} \int_{\mathcal{S}_2(\mathcal{Z}_1)}^{\mathcal{T}_2(\mathcal{Z}_1)} \ldots \int_{\mathcal{S}_M(\mathcal{Z}_1,\ldots,\mathcal{Z}_{M-1})}^{\mathcal{T}_M(\mathcal{Z}_1,\ldots,\mathcal{Z}_{M-1})} \phi(a_1)\phi(a_2)\ldots\phi(a_M) \\ &\quad \frac{d\mathcal{Z}_M \ldots d\mathcal{Z}_1}{\phi(a_1)\ldots\phi(a_M)} \\ &= \int_{\mathcal{S}_1}^{\mathcal{T}_1} \int_{\mathcal{S}_2(\mathcal{Z}_1)}^{\mathcal{T}_2(\mathcal{Z}_1)} \ldots \int_{\mathcal{S}_M(\mathcal{Z}_1,\ldots,\mathcal{Z}_{M-1})}^{\mathcal{T}_M(\mathcal{Z}_1,\ldots,\mathcal{Z}_{M-1})} d\mathcal{Z}_M \ldots d\mathcal{Z}_1, \end{aligned} \tag{A.4}$$

where the limits for $m = 1$ can be defined as

$$\mathcal{S}_1 = \Phi(\mathcal{E}_1) \leq \mathcal{Z}_1 \leq \Phi(\mathcal{J}_1) = \mathcal{T}_1,$$

while for $m = 2, \ldots, M$

$$\mathcal{S}_M = \Phi\left(\frac{\mathcal{E}_m - \sum_{h=1}^{m-1} c^*_{mh,i} \Phi^{-1}(\mathcal{Z}_h)}{c^*_{mm,i}}\right) \leq \mathcal{Z}_m \leq \Phi\left(\frac{\mathcal{J}_m - \sum_{h=1}^{m-1} c^*_{mh,i} \Phi^{-1}(\mathcal{Z}_h)}{c^*_{mm,i}}\right) \mathcal{T}_M,$$

where $\mathcal{S}_M$ and $\mathcal{T}_M$ refer to $\mathcal{S}_M(\mathcal{Z}_1,\ldots,\mathcal{Z}_{M-1})$ and $\mathcal{T}_M(\mathcal{Z}_1,\ldots,\mathcal{Z}_{M-1})$.

(T.3) Even though (A.4) is much simpler than (A.3), the integration region is more complicated. To overcome this, Genz (1992) suggested the transformation $\mathcal{Z}_m = \mathcal{S}_m + \omega_m(\mathcal{T}_m - \mathcal{S}_m)$, which standardizes this region, that is $0 \leq \omega_m \leq 1$, $\forall m$. In addition,

$$\frac{d\mathcal{Z}_m}{d\omega_m} = \mathcal{T}_m - \mathcal{S}_m \implies d\mathcal{Z}_m = (\mathcal{T}_m - \mathcal{S}_m)d\omega_m.$$

Therefore, (A.4) can be expressed as

$$\begin{aligned}\Phi_M(\boldsymbol{\mathcal{J}}) &= \int_0^1\int_0^1\ldots\int_0^1 (\mathcal{T}_1-\mathcal{S}_1)(\mathcal{T}_2-\mathcal{S}_2)\ldots(\mathcal{T}_M-\mathcal{S}_M)d\omega_M\ldots d\omega_1 \\ &= (\mathcal{T}_1-\mathcal{S}_1)\int_0^1(\mathcal{T}_2-\mathcal{S}_2)\ldots\int_0^1(\mathcal{T}_M-\mathcal{S}_M)d\boldsymbol{\omega},\end{aligned}$$

where $\mathcal{S}_m = \Phi\left(\left(\mathcal{E}_m - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{S}_h + \omega_h(\mathcal{T}_h - \mathcal{S}_h))\right)/c^*_{mm,i}\right)$ and $\mathcal{T}_m = \Phi\left(\left(\mathcal{J}_m - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{S}_h + \omega_h(\mathcal{T}_h - \mathcal{S}_h))\right)/c^*_{mm,i}\right)$. Since both $\mathcal{S}_m$ and $\mathcal{T}_m$ do not depend on ω_m, the innermost integral is equal to 1 and the number of integration variables can be reduced to $M-1$. Therefore, standard numerical integration methods can be applied for the computation of

$$\Phi_M(\boldsymbol{\mathcal{J}}) = \int_0^1\ldots\int_0^1 \tilde{f}(\omega_1,\ldots,\omega_{M-1})d\boldsymbol{\omega},$$

for $\tilde{f}(\omega_1,\ldots,\omega_{M-1}) = (\mathcal{T}_1-\mathcal{S}_1)(\mathcal{T}_2(\omega_1)-\mathcal{S}_2(\omega_1))\ldots(\mathcal{T}_M(\omega_1,\ldots,\omega_{M-1})-\mathcal{S}_M(\omega_1\ldots,\omega_{M-1}))$.

Computation of $\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i)$ using Genz's method

Since we are interested in the computation of the multivariate normal distribution function $\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i)$, $\mathcal{E}_m = -\infty$ and $\mathcal{J}_m = w_{m,i}$, $\forall m$. Therefore,

$$\mathcal{S}_m = \Phi\left(\frac{-\infty - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{S}_h + \omega_h(\mathcal{T}_h - \mathcal{S}_h))}{c^*_{mm,i}}\right) \to 0,$$

and

$$\begin{aligned}\mathcal{T}_m &= \Phi\left(\frac{w_{m,i} - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{S}_h + \omega_h(\mathcal{T}_h - \mathcal{S}_h))}{c^*_{mm,i}}\right) \\ &= \Phi\left(\frac{w_{m,i} - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\omega_h\mathcal{T}_h))}{c^*_{mm,i}}\right),\end{aligned}$$

since $\mathcal{S}_h = 0$, for all h. It follows that

$$\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i) = \int_0^1 \dots \int_0^1 \mathcal{T}_1\mathcal{T}_2(\omega_1)\dots\mathcal{T}_M(\omega_1, \dots, \omega_{M-1})d\boldsymbol{\omega}. \quad (A.5)$$

Once we get the transformed expression (A.5), the sub-region adaptive method is applied (see next section) and thus the cdf Φ_M is obtained.

The algorithm

The algorithm that is used in the subroutine `sadmvn()` in `Fortran-77` for the numerical computation of the multivariate normal distribution function $\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i)$ is based on subdivisions of $[0, 1]$, where each sub-region is used to provide a better approximation to $\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i)$. As previously described, we set $\mathcal{S}_m = 0$ to avoid wasteful evaluation of Φ.

The basic algorithm can be described as follows. Suppose that $\tilde{\epsilon}$ denotes the global absolute error and $\bar{\mathrm{N}}_{\max}$ is the maximum number of sub-regions. The algorithm starts with region $\tilde{\mathrm{R}}_{11} = [0, 1]^M$. At the $\bar{\mathrm{N}}^{th}$ step, $[0, 1]$ is partitioned into $\bar{\mathrm{N}}$ sub-regions $\tilde{\mathrm{R}}_{\bar{\mathrm{N}}1}, \dots \tilde{\mathrm{R}}_{\bar{\mathrm{N}}\bar{\mathrm{N}}}$ and in each sub-region we get estimates $\tilde{\mathrm{I}}_{\bar{\mathrm{N}}1}, \dots, \tilde{\mathrm{I}}_{\bar{\mathrm{N}}\bar{\mathrm{N}}}$ of the corresponding integrals by applying quadrature rules. Moreover, we obtain absolute error estimates $\tilde{E}_{\bar{\mathrm{N}}1}, \dots, \tilde{E}_{\bar{\mathrm{N}}\bar{\mathrm{N}}}$. If $\tilde{E}_{\bar{\mathrm{N}}1} + \dots + \tilde{E}_{\bar{\mathrm{N}}\bar{\mathrm{N}}} < \tilde{\epsilon} = 10^{-6}$ or $\bar{\mathrm{N}} \geq \bar{\mathrm{N}}_{\max} = 2000 \times M$ then the algorithm stops. Otherwise a new subdivision has to be determined and the above procedure is repeated. Further details about the algorithm can be found in Genz (1991), Genz (1992), Genz & Kass (1997) and Genz & Bretz (2002).

A.2.2 Bivariate conditioning approximation for trivariate normal integrals

This section describes the bivariate conditioning algorithm applied for the evaluation of trivariate normal integrals, which is based on the work by Trinh & Genz (2015). As described in Section 2.3, the computation of triple integrals is required only for the evaluation of $\mathbb{P}(y_{1i} = 1, y_{2i} = 1, y_{3i} = 1)$; the remaining probabilities can be evaluated via univariate and bivariate normal integrals. Thus, the aim is to provide accurate and low computational cost methods for approximating the triple integrals

$$\Phi_3\left(\eta_{1i}, \eta_{2i}, \eta_{3i}; \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|(2\pi)^3}} \int_{-\infty}^{\eta_{1i}} \int_{-\infty}^{\eta_{2i}} \int_{-\infty}^{\eta_{3i}} \exp\left(-\frac{1}{2}\boldsymbol{l}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{l}_i\right) d\boldsymbol{l}_i,$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Upsilon}_i)_1$.

The algorithm is based on the Cholesky decomposition of the correlation matrix $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\top$, where $\mathbf{C}$ is a lower triangular matrix. Note that the decomposition always exists as $\boldsymbol{\Sigma}$ is symmetric and positive-definite because of the restrictions imposed on the correlation parameters. Based on this, we have that $\boldsymbol{l}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{l}_i = \boldsymbol{l}_i^\top \mathbf{C}^{-\top} \mathbf{C}^{-1} \boldsymbol{l}_i$ and by using transformation $\boldsymbol{l}_i = \mathbf{C}\mathbf{a}_i$ we get $\boldsymbol{l}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{l}_i = \mathbf{a}_i^\top \mathbf{a}_i$ with $d\boldsymbol{l}_i = |\mathbf{C}| d\mathbf{a}_i = \sqrt{|\boldsymbol{\Sigma}|} d\mathbf{a}_i$. The integrals are transformed according to $-\infty \leq \mathbf{C}\mathbf{a}_i \leq \boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i = (\eta_{1i}, \eta_{2i}, \eta_{3i})^\top$. Specifically, the limits can be determined as follows

$$\begin{aligned}
-\infty &\leq \mathrm{a}_{1i} \leq \frac{\eta_{1i}}{c_{11}} = \frac{\eta_{1i}}{\sqrt{\sigma_{11}}} = \eta'_{1i} \\
-\infty &\leq \mathrm{a}_{2i} \leq \frac{\eta_{2i} - c_{21}\mathrm{a}_{1i}}{c_{22}} = \frac{\eta_{2i} - c_{21}\mathrm{a}_{1i}}{\sqrt{\sigma_{22}}} = \eta'_{2i} \\
-\infty &\leq \mathrm{a}_{2i} \leq \frac{\eta_{3i} - c_{31}\mathrm{a}_{1i} - c_{32}\mathrm{a}_{2i}}{c_{33}} = \frac{\eta_{3i} - c_{31}\mathrm{a}_{1i} - c_{32}\mathrm{a}_{2i}}{\sqrt{\sigma_{33}}} = \eta'_{3i}.
\end{aligned}$$

The a_{zi} values, $\forall z = 1, 2$, cannot be computed directly, so they are approximated using their truncated expected values:
$\tilde{\mu}_{\mathrm{a}_{zi}} = \mathbb{E}(-\infty, \eta'_{zi}) = \left(\phi(-\infty) - \phi(\eta'_{zi})\right) / \left(\Phi(\eta'_{zi}) - \Phi(-\infty)\right) = -\phi(\eta'_{zi})/\Phi(\eta'_{zi})$. The basic idea of this replacement is that these values are the average values that the a_{zi}s would have if we simulated a_{zi}s with values taken from truncated univariate

distributions. In order to improve the accuracy of the results the authors apply variable re-ordering. They specify that these orderings do not change the value of the probabilities as long as the integration limits and corresponding rows and columns of $\mathbf{\Sigma}$ are also permuted. Specifically, sorting the variables so that those with the shortest integration interval widths are the outer integration variables reduces the overall variation of the integrand and thus makes the numerical integration problem easier.

The algorithm is structured as follows.

Step 1 First, we need to select the outermost integration variable. This can be done by choosing the variable ς so that

$$\varsigma = \text{argmin}_{1\leq\varsigma\leq 3}\left\{\Phi\left(\frac{\eta_{\varsigma i}}{\sqrt{\sigma_{\varsigma\varsigma}}}\right) - \Phi(-\infty)\right\} = \text{argmin}_{1\leq\varsigma\leq 3}\left\{\Phi\left(\frac{\eta_{\varsigma i}}{\sqrt{\sigma_{\varsigma\varsigma}}}\right)\right\}.$$

The rows and columns of $\mathbf{\Sigma}$ as well as the integration limits for variables 1 and ς are interchanged. The elements in the first column of $\mathbf{C}$ are computed as follows: $c_{11} = \sqrt{\sigma_{11}}$, $c_{21} = \sigma_{21}/c_{11}$ and $c_{31} = \sigma_{31}/c_{11}$, where $\sigma_{..}$ denotes the $(\cdot,\cdot)^{th}$ element of $\mathbf{\Sigma}$. Then, we set $\hat{\eta}_{1i} = \eta'_{1i}$ and $\tilde{\mu}_{\mathrm{a}_{1i}} = -\phi(\hat{\eta}_{1i})/\Phi(\hat{\eta}_{1i})$.

Step 2 Next, ς is chosen such that

$$\begin{aligned}\varsigma &= \text{argmin}_{2\leq\varsigma\leq 3}\left\{\Phi\left(\frac{\eta_{\varsigma i} - c_{\varsigma 1}\tilde{\mu}_{\mathrm{a}_{1i}}}{\sqrt{\sigma_{\varsigma\varsigma} - c_{\varsigma 1}^2}}\right) - \Phi(-\infty)\right\} \\ &= \text{argmin}_{2\leq\varsigma\leq 3}\left\{\frac{\eta_{\varsigma i} - c_{\varsigma 1}\tilde{\mu}_{\mathrm{a}_{1i}}}{\sqrt{\sigma_{\varsigma\varsigma} - c_{\varsigma 1}^2}}\right\}.\end{aligned}$$

The rows and columns of $\mathbf{\Sigma}$, the integration limits, and c_{12} and $c_{\varsigma 2}$ are interchanged. The elements in the second column of $\mathbf{C}$ are computed as follows: $c_{22} = \sqrt{\sigma_{22} - c_{21}^2}$, $c_{32} = (\sigma_{32} - c_{21}c_{31})/c_{22}$. Then we let $\hat{\eta}_{2i} = (\eta_{2i} - c_{21}\tilde{\mu}_{\mathrm{a}_{1i}})/c_{22}$ and $\tilde{\mu}_{\mathrm{a}_{2i}} = -\phi(\hat{\eta}_{2i})/\Phi(\hat{\eta}_{2i})$.

Step 3 At this step, we calculate the $(3,3)^{th}$ element of $\mathbf{C}$ as $c_{33} = \sqrt{\sigma_{33} - c_{31}^2 - c_{32}^2}$ and we set $\hat{\eta}_{3i} = (\eta_{3i} - c_{31}\tilde{\mu}_{\mathrm{a}_{1i}} - c_{32}\tilde{\mu}_{\mathrm{a}_{2i}})/c_{33}$.

Step 4 Based on the resulting $\mathbf{C}$ matrix, we can determine $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{D}}$ using the relation $\tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{L}}^{\top} = \mathbf{C}\mathbf{C}^{\top}$, where $\tilde{\mathbf{D}} = \mathbf{C}_{\tilde{\mathbf{D}}}\mathbf{C}_{\tilde{\mathbf{D}}}^{\top}$, $\tilde{\mathbf{L}} = \mathbf{C}\mathbf{C}_{\tilde{\mathbf{D}}}^{-1}$ and $\mathbf{C}_{\tilde{\mathbf{D}}}$ denotes the block diagonal matrix of $\mathbf{C}$.

Step 5 Once we obtain $\mathbf{C}$, $\tilde{\mathbf{L}}$, $\tilde{\mathbf{D}}$ and the new upper integration limits, say $\tilde{\eta}_{1i}$, $\tilde{\eta}_{2i}$ and $\tilde{\eta}_{3i}$, the next step is the computation of the bivariate normal approximation. In particular, based on a similar transformation that has been discussed above, we obtain the updated upper integration limits as follows

$$\tilde{\eta}_{1i} = \frac{\hat{\eta}_{1i}}{\sqrt{\tilde{d}_{11}}}, \tilde{\eta}_{2i} = \frac{\hat{\eta}_{2i}}{\sqrt{\tilde{d}_{22}}}, \tilde{\eta}_{3i} = \frac{\hat{\eta}_{3i} - \tilde{g}_3}{\sqrt{\tilde{d}_{33}}},$$

where $\tilde{g}_3 = \tilde{l}_{31}\tilde{e}_1 + \tilde{l}_{32}\tilde{e}_2$, $\tilde{e}_1 = \bar{\mu}_1\sqrt{\tilde{d}_{11}}$, $\tilde{e}_2 = \bar{\mu}_2\sqrt{\tilde{d}_{22}}$, $\bar{\mu}_1 = 1/\mathcal{F}\left\{-\rho\phi(\tilde{\eta}_{2i})\Phi\left((\tilde{\eta}_{1i} - \rho\tilde{\eta}_{2i})/\tilde{q}\right) - \phi(\tilde{\eta}_{1i})\Phi\left((\tilde{\eta}_{2i} - \rho\tilde{\eta}_{1i})/\tilde{q}\right)\right\}$, $\bar{\mu}_2 = 1/\mathcal{F}\left\{-\rho\phi(\tilde{\eta}_{1i})\Phi\left((\tilde{\eta}_{2i} - \rho\tilde{\eta}_{1i})/\tilde{q}\right) - \phi(\tilde{\eta}_{2i})\Phi\left((\tilde{\eta}_{1i} - \rho\tilde{\eta}_{2i})/\tilde{q}\right)\right\}$, $\rho = \tilde{d}_{12}/\sqrt{\tilde{d}_{11}\tilde{d}_{22}}$, $\tilde{q} = \sqrt{1-\rho}$, $\mathcal{F} = \Phi_2\left(\tilde{\eta}_{1i}, \tilde{\eta}_{2i}; \mathbf{\Omega}\right)$ and $\mathbf{\Omega}$ is a 2×2 correlation matrix with 1s in the diagonals and ρ in the off-diagonals. The elements $\tilde{d}_{..}$ and $\tilde{l}_{..}$ correspond o the $(\cdot, \cdot)^{th}$ entry of $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{L}}$, respectively.

Step 6 Based on Trinh & Genz (2015), the bivariate normal approximation for trivariate normal probabilities can be written as follows

$$\Phi_3(\eta_{1i}, \eta_{2i}, \eta_{3i}; \mathbf{\Sigma}) \approx \Phi_2\left(\tilde{\eta}_{1i}, \tilde{\eta}_{2i}; \mathbf{\Omega}\right)\Phi(\tilde{\eta}_{3i}).$$

A.3 Geometric proof of the restriction on a correlation matrix

Geometric proofs of the restriction on a correlation matrix were first provided by Glass & Collins (1970) and Leung & Lam (1975). In what follows, we discuss a proof and show that the restriction on the values ϑ_{12} can assume when ϑ_{13} and ϑ_{23} are fixed. This is also displayed through spherical triangles.

Suppose that the n observations of the error term ε_{1i} are coordinates on the n-orthogonal axes of an n-dimensional space. Thus, the observations of ε_{1i} may be considered as corresponding to a vector, $\tilde{\boldsymbol{\varepsilon}}_1$, in the n-space. Similarly, two vectors corresponding to the n observations on ε_{2i} and ε_{3i} may be established in the n-space. By using the well known result, that the Pearson's coefficient is equivalent to the cosine of the angle between two vectors (Anderson et al., 1958, pp. 49-50), we re-express ϑ_{zk} as

$$\vartheta_{zk} \quad = \quad \cos(\boldsymbol{\varphi}_{zk}), \tag{A.6}$$

where $\boldsymbol{\varphi}_{zk}$ denotes the angle that separates $\tilde{\boldsymbol{\varepsilon}}_z$ and $\tilde{\boldsymbol{\varepsilon}}_k$. Now, consider vectors $\tilde{\boldsymbol{\varepsilon}}_1$, $\tilde{\boldsymbol{\varepsilon}}_2$ and $\tilde{\boldsymbol{\varepsilon}}_3$ in a three-dimensional subspace of the n-dimensional space. Let the angles separating $\tilde{\boldsymbol{\varepsilon}}_1$ and $\tilde{\boldsymbol{\varepsilon}}_2$, $\tilde{\boldsymbol{\varepsilon}}_1$ and $\tilde{\boldsymbol{\varepsilon}}_3$, and $\tilde{\boldsymbol{\varepsilon}}_2$ and $\tilde{\boldsymbol{\varepsilon}}_3$ be fixed at $\boldsymbol{\varphi}_{12}$, $\boldsymbol{\varphi}_{13}$ and $\boldsymbol{\varphi}_{23}$, respectively. Then, $\tilde{\boldsymbol{\varepsilon}}_1$, $\tilde{\boldsymbol{\varepsilon}}_2$ and $\tilde{\boldsymbol{\varepsilon}}_3$ form a spherical triangle on the surface of a sphere of radius equal to one, centred at the origin $\boldsymbol{O} = (0, 0, 0)$ with vertices $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ (Figure A.1). Planes $\mathcal{P}_2$ and $\mathcal{P}_3$, $\mathcal{P}_1$ and $\mathcal{P}_3$, and $\mathcal{P}_1$ and $\mathcal{P}_2$ form the dihedral angles $\angle\mathcal{CAB}$, $\angle\mathcal{CBA}$ and $\angle\mathcal{ACB}$ respectively. Suppose that $\angle\mathcal{CAB} = \mathfrak{a}$, $\angle\mathcal{CBA} = \mathfrak{b}$ and $\angle\mathcal{ACB} = \mathfrak{c}$ and assume that angles $\boldsymbol{\varphi}_{12}$, $\boldsymbol{\varphi}_{13}$, $\boldsymbol{\varphi}_{23}$, $\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{c}$ are between 0 and π radians. By using the spherical law of cosines for angles, we have the following three

equations

$$\cos\boldsymbol{\varphi}_{12} = \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23} + \sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23}\cos\mathfrak{c}, \quad (A.7)$$

$$\cos\boldsymbol{\varphi}_{13} = \cos\boldsymbol{\varphi}_{12}\cos\boldsymbol{\varphi}_{23} + \sin\boldsymbol{\varphi}_{12}\sin\boldsymbol{\varphi}_{23}\cos\mathfrak{b}, \quad (A.8)$$

$$\cos\boldsymbol{\varphi}_{23} = \cos\boldsymbol{\varphi}_{12}\cos\boldsymbol{\varphi}_{13} + \sin\boldsymbol{\varphi}_{12}\sin\boldsymbol{\varphi}_{13}\cos\mathfrak{a}. \quad (A.9)$$

Solving (A.7), (A.8) and (A.9) with respect to $\mathfrak{c}$, $\mathfrak{b}$ and $\mathfrak{a}$, respectively, it can be shown that the correlation parameters are restricted to a specific range. For instance, by solving equation (A.7) for $\cos\mathfrak{c}$ we have that $\cos\mathfrak{c} = (\cos\boldsymbol{\varphi}_{12} - \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23})/\sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23}$. Since $\cos\mathfrak{c} \in (-1,1)$ it follows that $-1 < (\cos\boldsymbol{\varphi}_{12} - \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23})/\sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23} < 1$, which implies that $-\sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23} < \cos\boldsymbol{\varphi}_{12} - \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23} < \sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23}$ and therefore

$$\cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23} - \sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23} < \cos\boldsymbol{\varphi}_{12} < \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23} + \sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23}. \quad (A.10)$$

Then, by using equation (A.6) and the trigonometric identity $\cos^2(\boldsymbol{\varphi}_{zk}) + \sin^2(\boldsymbol{\varphi}_{zk}) = 1 \implies \sin(\boldsymbol{\varphi}_{zk}) = \sqrt{1-\vartheta_{zk}^2}$, $\forall z = 1, 2, k = 3$, it follows that inequality (A.10) becomes

$$\vartheta_{13}\vartheta_{23} - \sqrt{1-\vartheta_{13}^2}\sqrt{1-\vartheta_{23}^2} < \vartheta_{12} < \vartheta_{13}\vartheta_{23} + \sqrt{1-\vartheta_{13}^2}\sqrt{1-\vartheta_{23}^2},$$

which is equal to (2.7). The interval for ϑ_{13} and ϑ_{23} is obtained by solving (A.9) and (A.8) respectively.

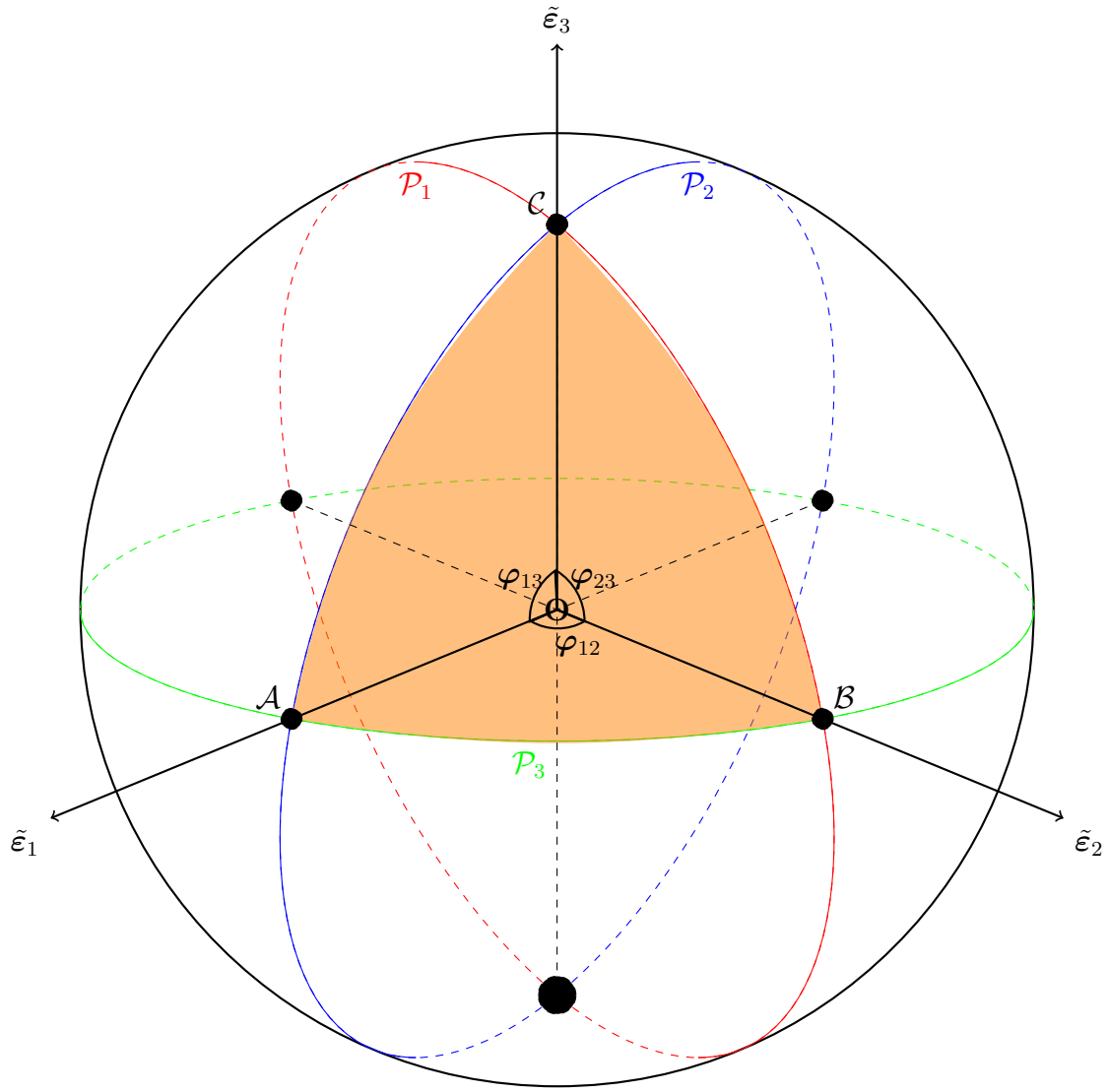

Figure A.1: Spherical representation of intercorrelations among the error terms $\tilde{\varepsilon}_1$, $\tilde{\varepsilon}_2$ and $\tilde{\varepsilon}_3$.

A.4 Proof of Propositions 2.3.2 and 2.3.3

The first-order derivatives of the log-likelihood function for a multivariate probit model are obtained as follows. First, we express the multivariate normal cdf Φ_M in terms of multivariate integrals. Then, by using conditional density distributions, we decompose ϕ_M into a product of two normal probability density functions (pdfs) and re-express Φ_M based on that decomposition. In doing so we proceed with the calculation of the two derivatives, where the derivative of Φ_M with respect to $\boldsymbol{\beta}_m$ is mainly based on a decomposition formula, while the derivative of Φ_M with respect to ϑ_{zk} has been derived by applying an idea by Plackett (1954).

The multivariate integrals

$$\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \;=\; \int_{-\infty}^{w_{M,i}} \dots \int_{-\infty}^{w_{1,i}} \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \tag{A.11}$$

can be written in a more convenient form by using the conditional distribution of the normal multivariate distribution. This can be achieved by partitioning both $\boldsymbol{l}_i$ and $\boldsymbol{\Upsilon}_i^*$ such that

$$\boldsymbol{l}_i \;=\; (\boldsymbol{l}_{1,i}, \boldsymbol{l}_{2,i})^\top,$$

and

$$\boldsymbol{\Upsilon}_i^* = \left(\begin{array}{c|c} \boldsymbol{\Theta}_{11,i}^* & \boldsymbol{\Theta}_{12,i}^* \\ \hline \boldsymbol{\Theta}_{21,i}^* & \boldsymbol{\Theta}_{22,i}^* \end{array}\right) = \left(\begin{array}{cccc|ccc} 1 & r_{12,i}^* & \dots & r_{1u,i}^* & r_{1,u+1,i}^* & \dots & r_{1,M,i}^* \\ r_{21,i}^* & 1 & \dots & r_{2u,i}^* & r_{2,u+1,i}^* & \dots & r_{2,M,i}^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ r_{u1,i}^* & r_{u2,i}^* & \dots & 1 & r_{u,u+1,i}^* & \dots & r_{u,M,i}^* \\ \hline r_{u+1,1,i}^* & r_{u+1,2,i}^* & \dots & r_{u+1,u,i}^* & 1 & \dots & r_{u+1,M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{M1,i}^* & r_{M2,i}^* & \dots & r_{Mu,i}^* & r_{M,u+1,i}^* & \dots & 1 \end{array}\right), \tag{A.12}$$

respectively, where $\boldsymbol{l}_{1,i} = (l_{1,i}, l_{2,i}, \dots, l_{u,i})^\top$, $\boldsymbol{l}_{2,i} = (l_{u+1,i}, l_{u+2,i}, \dots, l_{M,i})^\top$, $u =$

$1, \ldots, M-1$, $r^*_{zk,i} = \tanh(\vartheta^*_{zk})(2y_{zi}-1)(2y_{ki}-1)$, $\boldsymbol{\Theta}^*_{11,i}$ is a $u \times u$ matrix, $\boldsymbol{\Theta}^*_{22,i}$ is a $(M-u)\times(M-u)$ matrix and $\boldsymbol{\Theta}^*_{21,i} = \boldsymbol{\Theta}^{*\top}_{12,i}$. By using the chain rule for random variables and the partitioned vector $\boldsymbol{l}_i$ as well as the partitioned matrix $\boldsymbol{\Upsilon}^*_i$, the M-variate normal pdf $\phi_M(\boldsymbol{l}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}^*_i)$ can be expressed as the product of the conditional density function of $\boldsymbol{l}_{2,i}$ given $\boldsymbol{l}_{1,i}$ times the pdf of $\boldsymbol{l}_{1,i}$

$$\phi_M(\boldsymbol{l}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}^*_i) \;=\; \phi_{M-u}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i})\phi_u(\boldsymbol{l}_{1,i}), \tag{A.13}$$

where

$$\begin{aligned} \boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i} &\overset{iid}{\sim} \mathcal{N}_{M-u}(\mathbb{E}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}), \mathrm{Var}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i})) \\ &\overset{iid}{\sim} \mathcal{N}_{M-u}(\boldsymbol{\mu}_{\boldsymbol{l}_{2,i}} + \boldsymbol{\Theta}^*_{21,i}\boldsymbol{\Theta}^{*-1}_{11}(\boldsymbol{l}_{1,i} - \boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}), \boldsymbol{\Theta}^*_{22,i} - \boldsymbol{\Theta}^*_{21,i}\boldsymbol{\Theta}^{*-1}_{11,i}\boldsymbol{\Theta}^*_{12,i}), \end{aligned} \tag{A.14}$$

and

$$\begin{aligned} \boldsymbol{l}_{1,i} &\overset{iid}{\sim} \mathcal{N}_u(\mathbb{E}(\boldsymbol{l}_{1,i}), \mathrm{Var}(\boldsymbol{l}_{1,i})) \\ &\overset{iid}{\sim} \mathcal{N}_u(\boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}, \boldsymbol{\Theta}^*_{11,i}). \end{aligned} \tag{A.15}$$

$\boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}$ and $\boldsymbol{\mu}_{\boldsymbol{l}_{2,i}}$ stand for the mean of $\boldsymbol{l}_{1,i}$ and $\boldsymbol{l}_{2,i}$ respectively. It follows that the integrals (A.11) can be rewritten as

$$\begin{aligned} \Phi_M(\boldsymbol{w}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}^*_i) \;=\; & \int_{-\infty}^{w_{M,i}} \cdots \int_{-\infty}^{w_{1,i}} \phi_{M-u}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}; \boldsymbol{M}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}}, \boldsymbol{\Theta}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}})\phi_u(\boldsymbol{l}_{1,i}; \boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}, \boldsymbol{\Theta}^*_{11,i}) \\ & \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}, \end{aligned} \tag{A.16}$$

where $\boldsymbol{M}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}} = \boldsymbol{\mu}_{\boldsymbol{l}_{2,i}} + \boldsymbol{\Theta}^*_{21,i}\boldsymbol{\Theta}^{*-1}_{11}(\boldsymbol{l}_{1,i} - \boldsymbol{\mu}_{\boldsymbol{l}_{1,i}})$, $\boldsymbol{\Theta}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}} = \boldsymbol{\Theta}^*_{22,i} - \boldsymbol{\Theta}^*_{21,i}\boldsymbol{\Theta}^{*-1}_{11,i}\boldsymbol{\Theta}^*_{12,i}$, $\boldsymbol{\mu}_{\boldsymbol{l}_{1,i}} = \boldsymbol{\mu}_{\boldsymbol{l}_{2,i}} = \boldsymbol{0}$ and $\boldsymbol{\Theta}^*_{11,i}$ denotes the $u \times u$ sub-matrix of $\boldsymbol{\Upsilon}^*_i$.

A.4.1 Proof of Proposition 2.3.2

Proof. Consider formula (A.13) and let $u = 1$, such that

$$\phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \;=\; \phi_{M-1}(\boldsymbol{l}_{2,i}|l_{1,i})\phi(l_{1,i}).$$

By re-ordering matrix (A.12) we obtain

$$\boldsymbol{\Upsilon}_i^{*m} = \left(\begin{array}{c|c} \overbrace{\Theta_{11,i}^{*m}}^{1\times 1} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*m}}^{1\times(M-1)} \\ \hline \underbrace{\boldsymbol{\Theta}_{21,i}^{*m}}_{(M-1)\times 1} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*m}}_{(M-1)\times(M-1)} \end{array} \right),$$

$\forall m$, where $\Theta_{11,i}^{*m}$, $\boldsymbol{\Theta}_{12,i}^{*m}$, $\boldsymbol{\Theta}_{21,i}^{*m}$ and $\boldsymbol{\Theta}_{22,i}^{*m}$ are defined in Proposition 2.3.2, while the full matrix $\boldsymbol{\Upsilon}_i^{*m}$ can be found in Appendix A.5. Then the multivariate normal cdf (A.16) becomes

$$\begin{aligned} \Phi_M(\boldsymbol{w}_i; 0, \boldsymbol{\Upsilon}_i^{*m}) &= \int_{-\infty}^{w_{M,i}} \dots \int_{-\infty}^{w_{m,i}} \dots \int_{-\infty}^{w_{1,i}} \phi_M(\boldsymbol{l}_i; 0, \boldsymbol{\Upsilon}_i^{*m}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \\ &= \int_{\bar{\mathbf{C}}_i} \phi_{M-1}(\boldsymbol{l}_{-m,i}|l_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m}) \phi(l_{m,i}; \mu_{l_{m,i}}, \Theta_{11,i}^{*m}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}, \quad \text{(A.17)} \end{aligned}$$

for $\bar{\mathbf{C}}_i = \bar{C}_{1i} \times \bar{C}_{2i} \times \ldots \times \bar{C}_{Mi}$, where $\bar{C}_{mi}$ is the interval $[w_{m,i}, +\infty)$ if $y_{mi} = 1$ and the interval $(-\infty, w_{m,i}]$ if $y_{mi} = 0$. Vector $\boldsymbol{l}_{-m,i} = (l_{1,i}, \dots, l_{m-1,i}, l_{m+1,i}, \dots, l_{M,i})^\top$, where $l_{m,i}$ refers to the m^{th} element of vector $\boldsymbol{l}_i$. $\boldsymbol{M}_i^{*m}$ and $\boldsymbol{\Theta}_i^{*m}$, respectively, denote the mean and the variance of $\boldsymbol{l}_{-m,i}|l_{m,i}$, while $\mu_{l_{m,i}}$ and $\Theta_{11,i}^{*m}$ denote the mean and variance of $l_{m,i}$. Applying the properties of the conditional multivariate normal distribution, it follows that $\mathbb{E}(l_{m,i}) = \mu_{l_{m,i}} = 0$ and $\mathbb{E}(\boldsymbol{l}_{-m,i}) = \boldsymbol{\mu}_{\boldsymbol{l}_{-m,i}} = \mathbf{0}$. (Note that $\mathbb{E}(\boldsymbol{l}_{-m,i}|l_{m,i}) \neq \mathbf{0}$.) Hence, the distribution of $\boldsymbol{l}_{-m,i}|l_{m,i}$ and $l_{m,i}$ is equal to

$$\begin{aligned} \boldsymbol{l}_{-m,i}|l_{m,i} &\overset{iid}{\sim} \mathcal{N}_{M-1}\left(\mathbb{E}(\boldsymbol{l}_{-m,i}|l_{m,i}), \mathrm{Var}(\boldsymbol{l}_{-m,i}|l_{m,i})\right) \\ &\overset{iid}{\sim} \mathcal{N}_{M-1}\left(\boldsymbol{\mu}_{\boldsymbol{l}_{-m,i}} + \boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11}^{*m}\right)^{-1}(l_{m,i} - \mu_{l_{m,i}}), \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11,i}^{*m}\right)^{-1}\boldsymbol{\Theta}_{12,i}^{*m}\right) \\ &\overset{iid}{\sim} \mathcal{N}_{M-1}\left(\boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11}^{*m}\right)^{-1} l_{m,i}, \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11,i}^{*m}\right)^{-1}\boldsymbol{\Theta}_{12,i}^{*m}\right), \end{aligned}$$

and

$$\begin{aligned} l_{m,i} &\overset{iid}{\sim} \mathcal{N}(\mathbb{E}(l_{m,i}), \mathrm{Var}(l_{m,i})) \\ &\overset{iid}{\sim} \mathcal{N}(\mu_{l_{m,i}}, \Theta^{*m}_{11,i}) \\ &\overset{iid}{\sim} \mathcal{N}(0, \Theta^{*m}_{11,i}), \end{aligned}$$

respectively, where the sub-matrix $\Theta^{*m}_{11,i}$ in this case is equal to 1, $\forall m, i$. It follows that (A.17) becomes

$$\begin{aligned} \Phi_M(\boldsymbol{w}_i; 0, \boldsymbol{\Upsilon}^{*m}_i) &= \int_{\bar{\mathbf{C}}_i} \phi(l_{m,i}; 0, 1)\phi_{M-1}(\boldsymbol{l}_{-m,i}|l_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \\ &= \int_{-\infty}^{w_{m,i}} \phi(l_{m,i}; 0, 1) \left\{ \int_{\bar{\mathbf{C}}_{i,-m}} \phi_{M-1}(\boldsymbol{l}_{-m,i}|l_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) d\boldsymbol{l}_{-m,i} \right\} dl_{m,i} \\ &= \int_{-\infty}^{w_{m,i}} \phi(l_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{w}_{-m,i}|l_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) dl_{m,i}, \end{aligned} \tag{A.18}$$

where $\boldsymbol{w}_{-m,i} = (w_{1,i}, w_{2,i}, \ldots, w_{m-1,i}, w_{m+1,i}, \ldots, w_{M,i})^\top$ and $\bar{\mathbf{C}}_{i,-m} \in \{\bar{\mathbf{C}}_i\} \setminus \bar{C}_{mi}$. According to the properties of the conditional multivariate normal distribution, it follows that the expected value of $\boldsymbol{w}_{-m,i}|l_{m,i}$ is equal to $\boldsymbol{M}^{*m}_i = \boldsymbol{\Theta}^{*m}_{21,i} l_{m,i}$ while its variance-covariance matrix is expressed as $\boldsymbol{\Theta}^{*m}_i = \boldsymbol{\Theta}^{*m}_{22,i} - \boldsymbol{\Theta}^{*m}_{21,i}\boldsymbol{\Theta}^{*m}_{12,i}$. By using the chain rule as well as the fundamental theorem of calculus, it follows that the derivative of (A.18) with respect to $\boldsymbol{\beta}_m$ is equal to

$$\begin{aligned} \frac{\partial \Phi_M(\boldsymbol{w}_i; 0, \boldsymbol{\Upsilon}^{*m}_i)}{\partial \boldsymbol{\beta}_m} &= \frac{\partial \Phi_M(\boldsymbol{w}_i; 0, \boldsymbol{\Theta}^{*m}_i)}{\partial w_{m,i}} \frac{\partial w_{m,i}}{\partial \boldsymbol{\beta}_m} \\ &= \frac{\partial}{\partial w_{m,i}} \left\{ \int_{-\infty}^{w_{m,i}} \phi(l_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{w}_{-m,i}|l_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) dl_{m,i} \right\} \times \\ &\quad \left(\frac{\partial w_{m,i}}{\partial \boldsymbol{\beta}_m} \right) \\ &= \phi(w_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{w}_{-m,i}|w_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) \left(\frac{\partial w_{m,i}}{\partial \boldsymbol{\beta}_m} \right). \end{aligned}$$

Since $w_{m,i} = (2y_{mi} - 1)\mathbf{x}^\top_{mi}\boldsymbol{\beta}_m$, the derivative of $w_{m,i}$ with respect to $\boldsymbol{\beta}_m$ is equal to

$$\frac{\partial w_{m,i}}{\partial \boldsymbol{\beta}_m} = (2y_{mi} - 1)\mathbf{x}^\top_{mi},$$

and thus

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*m})}{\partial \boldsymbol{\beta}_m} = \phi(w_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{w}_{-m,i}|w_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m})(2y_{mi} - 1)\mathbf{x}_{mi}^{\top},$$

for $\boldsymbol{M}_i^{*m} = \boldsymbol{\Theta}_{21,i}^{*m} w_{m,i}$ and $\boldsymbol{\Theta}_i^{*m} = \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\boldsymbol{\Theta}_{12,i}^{*m}$, as required. □

A.4.2 Proof of Proposition 2.3.3

Proof. If we differentiate equation (A.11) with respect to the correlation coefficient ϑ_{zk}^*, we get the following

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*})}{\partial \vartheta_{zk}^*} = \frac{\partial}{\partial \vartheta_{zk}^*} \left\{ \int_{-\infty}^{w_{M,i}} \cdots \int_{-\infty}^{w_{1,i}} \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \right\},$$

and by using the chain rule

$$\begin{aligned} \frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*})}{\partial \vartheta_{zk,i}^*} &= \frac{\partial}{\partial r_{zk}^*} \left\{ \int_{\bar{\mathbf{C}}_i} \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*} \\ &= \left\{ \int_{\bar{\mathbf{C}}_i} \frac{\partial \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*})}{\partial r_{zk,i}^*} \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \end{aligned} \tag{A.19}$$

where $r_{zk,i}^*$ and region $\bar{\mathbf{C}}_i$ have been defined previously. By using the following differential equation derived by Plackett (1954)

$$\frac{\partial \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*})}{\partial r_{zk,i}^*} = \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*})}{\partial l_{z,i} \partial l_{k,i}},$$

equation (A.19) becomes

$$\begin{aligned} \frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*})}{\partial \vartheta_{zk,i}^*} &= \left\{ \int_{\bar{\mathbf{C}}_i} \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*})}{\partial l_{z,i} \partial l_{k,i}} \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*} \\ &= \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \left[\int_{\bar{\mathbf{C}}_{zk,i}} \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*})}{\partial l_{z,i} \partial l_{k,i}} d\boldsymbol{l}_{zk,i} \right] d\boldsymbol{l}_{-zk,i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \end{aligned} \tag{A.20}$$

where $\bar{\mathbf{C}}_{-zk,i} \in \bar{\mathbf{C}}_i \setminus \{\bar{C}_{zi}, \bar{C}_{ki}\}$, $\bar{\mathbf{C}}_{zk,i} = \bar{C}_{zi} \times \bar{C}_{ki}$, $\boldsymbol{l}_{zk,i} = (l_{z,i}, l_{k,i})^{\top}$ and $\boldsymbol{l}_{-zk,i} = (l_{1,i}, \ldots, l_{k-1,i},$

$l_{k+1,i}, \ldots, l_{z-1,i}, l_{z+1,i}, l_{M,i})^\top$. According to the fundamental theorem of calculus, the integral inside the brackets is equal to

$$\begin{aligned}
\int_{\bar{\mathbf{C}}_{zk,i}} \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial l_{z,i} \partial l_{k,i}} d\boldsymbol{l}_{zk,i} &= \frac{\partial^2}{\partial l_{z,i} \partial l_{k,i}} \left\{ \int_{\bar{\mathbf{C}}_{zk,i}} \phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) dl_{k,i} dl_{z,i} \right\} \\
&= \frac{\partial^2}{\partial l_{z,i} \partial l_{k,i}} \left\{ \int_{\bar{C}_{zi}} \int_{\bar{C}_{ki}} \phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) dl_{k,i} dl_{z,i} \right\} \\
&= \phi_M(l_{1,i}, \ldots, l_{z-1,i}, w_{z,i}, l_{z+1,i}, \ldots, l_{k-1,i}, w_{k,i}, l_{k+1,i}, \ldots, \\
&\quad l_{M,i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*).
\end{aligned}$$

Therefore, (A.20) can be expressed as

$$\begin{aligned}
\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk,i}^*} &= \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_M(l_{1,i}, \ldots, l_{z-1,i}, w_{z,i}, l_{z+1,i}, \ldots, l_{k-1,i}, w_{k,i}, l_{k+1,i}, \ldots, \right. \\
&\quad \left. l_{M,i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) d\boldsymbol{l}_{-zk,i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}.
\end{aligned}$$

The last expression can be written in a more convenient form by using the conditional distributions of the normal multivariate distribution. This can be done by imposing the special case $u = 2$ in equation (A.13), that is

$$\phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk}) = \phi_{M-2}(\boldsymbol{l}_{2,i} | \boldsymbol{l}_{1,i}) \phi_2(\boldsymbol{l}_{1,i}), \tag{A.21}$$

where $\boldsymbol{l}_{2,i}$ and $\boldsymbol{l}_{1,i}$ correspond to $\boldsymbol{l}_{-zk,i}$ and $\boldsymbol{w}_{zk,i}$, respectively, with $\boldsymbol{w}_{zk,i} = (w_{z,i}, w_{k,i})^\top$. Re-ordering matrix (A.12), we obtain

$$\boldsymbol{\Upsilon}_i^{*zk} = \begin{pmatrix} \overbrace{\boldsymbol{\Theta}_{11,i}^{*zk}}^{2\times 2} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*zk}}^{2\times(M-2)} \\ \underbrace{\boldsymbol{\Theta}_{21,i}^{*zk}}_{(M-2)\times 2} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*zk}}_{(M-2)\times(M-2)} \end{pmatrix}, \tag{A.22}$$

$\forall z = 1, \ldots, M-1, k = z+1, \ldots M$, where the sub-matrices $\boldsymbol{\Theta}_{11,i}^{*zk}$, $\boldsymbol{\Theta}_{12,i}^{*zk}$, $\boldsymbol{\Theta}_{21,i}^{*zk}$ and $\boldsymbol{\Theta}_{22,i}^{*zk}$ are defined in Proposition 2.3.3, while the full matrix $\boldsymbol{\Upsilon}_i^{*zk}$ can be found in

Appendix A.5. By using both (A.21) and (A.22), we have that

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk})}{\partial \vartheta_{zk,i}^*} = \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_{M-2}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk}) \times \right. \\ \left. \phi_2(\boldsymbol{w}_{zk,i}; \boldsymbol{\mu}_{\boldsymbol{w}_{zk,i}}, \boldsymbol{\Theta}_{11,i}^{*zk}) d\boldsymbol{l}_{-zk,i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \tag{A.23}$$

where $\boldsymbol{M}_i^{*-zk}$ and $\boldsymbol{\Theta}_i^{*-zk}$ refer to the mean and variance-covariance matrix of $\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}$, while $\boldsymbol{\mu}_{\boldsymbol{w}_{zk,i}}$ and $\boldsymbol{\Theta}_{11,i}^{*zk}$ denote the mean and variance-covariance of $\boldsymbol{w}_{zk,i}$. By using the properties of the conditional multivariate normal distribution, it follows that $\mathbb{E}(\boldsymbol{w}_{zk,i}) = \boldsymbol{\mu}_{\boldsymbol{w}_{zk,i}} = 0$ and $\mathbb{E}(\boldsymbol{l}_{-zk,i}) = \boldsymbol{\mu}_{\boldsymbol{l}_{-zk,i}} = 0$. (Note that $\mathbb{E}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}) \neq 0$.) Hence, according to (A.14) and (A.15)

$$\begin{aligned} \boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i} &\overset{iid}{\sim} \mathcal{N}_M(\mathbb{E}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}), \mathrm{Var}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i})) \\ &\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{\mu}_{\boldsymbol{l}_{-zk,i}} + \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11}^{*zk}\right)^{-1} (\boldsymbol{w}_{zk,i} - \boldsymbol{\mu}_{\boldsymbol{l}_{zk,i}}), \boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{\Theta}_{12,i}^{*zk}) \\ &\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{w}_{zk,i}, \boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{zk}\right)^{-1} \boldsymbol{\Theta}_{12,i}^{*zk}), \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{l}_{zk,i} &\overset{iid}{\sim} \mathcal{N}_M(\mathbb{E}(\boldsymbol{w}_{zk,i}), \mathrm{Var}(\boldsymbol{w}_{zk,i})) \\ &\overset{iid}{\sim} \mathcal{N}_M(\mu_{\boldsymbol{w}_{zk,i}}, \boldsymbol{\Theta}_{11,i}^{*zk}) \\ &\overset{iid}{\sim} \mathcal{N}_M(\mathbf{0}, \boldsymbol{\Theta}_{11,i}^{*zk}), \end{aligned}$$

where the sub-matrix $\boldsymbol{\Theta}_{11,i}^{*zk}$ is a 2×2 diagonal matrix with unit variances and correlations equal to $r_{zk,i}^*$. For simplicity, we will denote this matrix as $\boldsymbol{\Theta}_i^{*zk}$. Consequently, equation (A.23) can be expressed as

$$\frac{\partial \Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk})}{\partial \vartheta_{zk}^*} = \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_2(\boldsymbol{w}_{zk,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*zk}) \phi_{M-2}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk}) \right. \\ \left. d\boldsymbol{l}_{-zk,i} \right\} \times \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}.$$

Because only the term $\phi_{M-2}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i};\boldsymbol{M}_i^{*-zk},\boldsymbol{\Theta}_i^{*-zk})$ depends on $\boldsymbol{l}_{-zk,i}$, it follows that

$$\begin{aligned}\frac{\partial\Phi_M(\boldsymbol{w}_i;\boldsymbol{0},\boldsymbol{\Upsilon}_i^{*zk})}{\partial\vartheta_{zk}^*} &= \left\{\phi_2(\boldsymbol{w}_{zk,i};\boldsymbol{0},\boldsymbol{\Theta}_i^{*zk})\int_{\tilde{\mathbf{C}}_{-zk,i}}\phi_{M-2}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i};\boldsymbol{M}_i^{*-zk},\boldsymbol{\Theta}_i^{*-zk})\right.\\ &\quad \left.d\boldsymbol{l}_{-zk,i}\right\}\times\frac{\partial r_{zk,i}^*}{\partial\vartheta_{zk}^*}\\ &= \left\{\phi_2(\boldsymbol{w}_{zk,i};\boldsymbol{0},\boldsymbol{\Theta}_i^{*zk})\Phi_{M-2}(\boldsymbol{w}_{-zk,i}|\boldsymbol{w}_{zk,i};\boldsymbol{M}_i^{*-zk},\boldsymbol{\Theta}_i^{*-zk})\right\}\frac{\partial r_{zk,i}^*}{\partial\vartheta_{zk}^*},\end{aligned} \tag{A.24}$$

where the last term comes from basic results of the multivariate normal distribution function. In addition, $\boldsymbol{w}_{-zk,i}=(w_{1,i},w_{2,i},\ldots,w_{z-1,i},w_{z+1,i},\ldots,w_{k-1,i},w_{k+1,i},\ldots w_{M,i})^\top$, while the partial derivative $\partial r_{zk,i}^*/\partial\vartheta_{zk}^*$ is equal to

$$\begin{aligned}\frac{\partial r_{zk,i}^*}{\partial\vartheta_{zk}^*} &= \frac{\partial}{\partial\vartheta_{zk}^*}\left\{\tanh(\vartheta_{zk}^*)(2y_{z,i}-1)(2y_{k,i}-1)\right\}\\ &= (2y_{z,i}-1)(2y_{k,i}-1)\frac{\partial}{\partial\vartheta_{zk}^*}\left\{\tanh(\vartheta_{zk}^*)\right\}\\ &= (2y_{z,i}-1)(2y_{k,i}-1)\mathrm{sech}^2(\vartheta_{zk}^*)\\ &= (2y_{z,i}-1)(2y_{k,i}-1)\frac{1}{\cosh^2(\vartheta_{zk}^*)}\\ &= (2y_{z,i}-1)(2y_{k,i}-1)\frac{1}{\left(\frac{\exp(\vartheta_{zk}^*)+\exp(-\vartheta_{zk}^*)}{2}\right)^2}\\ &= (2y_{z,i}-1)(2y_{k,i}-1)\frac{4}{\left\{\exp(\vartheta_{zk}^*)+\exp(-\vartheta_{zk}^*)\right\}^2}\\ &= (2y_{z,i}-1)(2y_{k,i}-1)\frac{4e^{2\vartheta_{zk}^*}}{\left\{e^{2\vartheta_{zk}^*}+1\right\}^2},\end{aligned}$$

by using definitions and properties of the hyperbolic functions. Therefore, (A.24) becomes

$$\begin{aligned}\frac{\partial\Phi_M(\boldsymbol{w}_i;\boldsymbol{0},\boldsymbol{\Upsilon}_i^{*zk})}{\partial\vartheta_{*zk}} &= \phi_2(\boldsymbol{w}_{zk,i};\boldsymbol{0},\boldsymbol{\Theta}_i^{*zk})\Phi_{M-2}(\boldsymbol{w}_{-zk,i}|\boldsymbol{w}_{zk,i};\boldsymbol{M}_i^{*-zk},\boldsymbol{\Theta}_i^{*-zk})(2y_{z,i}-1)\times\\ &\quad (2y_{k,i}-1)\frac{4e^{2\vartheta_{zk}^*}}{\left\{e^{2\vartheta_{zk}^*}+1\right\}^2},\end{aligned}$$

for $\boldsymbol{w}_{zk,i} = (w_{z,i}, w_{k,i})^\top$, $\boldsymbol{w}_{-zk,i} = (w_{1,i}, w_{2,i}, \ldots, w_{z-1,i}, w_{z+1,i}, \ldots, w_{k-1,i}, w_{k+1,i}, \ldots w_{M,i})^\top$, $\boldsymbol{\Theta}_i^{*zk} = \boldsymbol{\Theta}_{11,i}^{*zk}$, $\boldsymbol{M}_i^{*-zk} = \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{w}_{zk,i}$ and $\boldsymbol{\Theta}_i^{*-zk} = \boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{\Theta}_{12,i}^{*zk}$, as required. □

A.5 Correlation matrices $\boldsymbol{\Upsilon}_i^{*m}$ and $\boldsymbol{\Upsilon}_i^{*zk}$

For $m = 1$, matrix $\boldsymbol{\Upsilon}_i^{*m}$ is equal to

$$\boldsymbol{\Upsilon}_i^{*1} = \left(\begin{array}{c|c} \Theta_{11,i}^{*1} & \boldsymbol{\Theta}_{12,i}^{*1} \\ \hline \boldsymbol{\Theta}_{21,i}^{*1} & \boldsymbol{\Theta}_{22,i}^{*1} \end{array}\right) = \left(\begin{array}{c|cccccc} 1 & r_{12,i}^* & r_{13,i}^* & \cdots & r_{1,M-1,i}^* & r_{1M,i}^* \\ \hline r_{12,i}^* & 1 & r_{23,i}^* & \cdots & r_{2,M-1,i}^* & r_{2M,i}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{1,M-1,i}^* & r_{2,M-1,i} & r_{3,M-1,i}^* & \cdots & 1 & r_{M-1,M,i}^* \\ r_{1M,i}^* & r_{2M,i}^* & r_{3M,i}^* & \cdots & r_{M-1,M,i}^* & 1 \end{array}\right),$$

while for $m \geq 2$

$$\begin{aligned} \boldsymbol{\Upsilon}_i^{*m} &= \left(\begin{array}{c|c} \Theta_{11,i}^{*m} & \boldsymbol{\Theta}_{12,i}^{*m} \\ \hline \boldsymbol{\Theta}_{21,i}^{*m} & \boldsymbol{\Theta}_{22,i}^{*m} \end{array}\right) \\ &= \left(\begin{array}{c|ccccccc} 1 & r_{m1,i}^* & r_{m2,i}^* & \cdots & r_{m,m-1,i}^* & r_{m,m+1,i}^* & \cdots & r_{mM,i}^* \\ \hline r_{m1,i}^* & 1 & r_{12,i}^* & \cdots & r_{1,m-1,i}^* & r_{1,m+1,i}^* & \cdots & r_{1M,i}^* \\ r_{m2,i}^* & r_{12,i}^* & 1 & \cdots & r_{2,m-1,i}^* & r_{2,m+1,i}^* & \cdots & r_{2M,i}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ r_{m,m-1,i}^* & r_{1,m-1,i} & r_{2,m-1,i}^* & \cdots & 1 & r_{m-1,m+1,i}^* & \cdots & r_{m-1,m,i}^* \\ r_{m,m+1,i}^* & r_{1,m+1,i}^* & r_{2,m+1,i}^* & \cdots & r_{m-1,m+1,i}^* & 1 & \cdots & r_{m+1,M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{mM,i}^* & r_{1M,i}^* & r_{2M,i}^* & \cdots & r_{m-1,M,i}^* & r_{m+1,M,i}^* & \cdots & 1 \end{array}\right). \end{aligned}$$

Matrix $\boldsymbol{\Upsilon}_i^{zk}$ is equal to

$$\boldsymbol{\Upsilon}_i^{*zk} = \left(\begin{array}{c|c} \boldsymbol{\Theta}_{11,i}^{*zk} & \boldsymbol{\Theta}_{12,i}^{*zk} \\ \hline \boldsymbol{\Theta}_{21,i}^{*zk} & \boldsymbol{\Theta}_{22,i}^{*zk} \end{array}\right) =$$

$$\left(\begin{array}{cc|ccccccccccc} 1 & r_{zk,i}^* & r_{z1,i}^* & r_{z2,i}^* & \cdots & r_{z,z-1,i}^* & r_{z,z+1,i}^* & \cdots & r_{z,k-1,i}^* & r_{z,k+1} & \cdots & r_{zM,i}^* \\ r_{zk,i}^* & 1 & r_{k1,i}^* & r_{k2,i}^* & \cdots & r_{k,z-1,i}^* & r_{k,z+1,i}^* & \cdots & r_{k,k-1,i}^* & r_{k,k+1,i}^* & \cdots & r_{kM,i}^* \\ \hline r_{z1,i}^* & r_{k1,i}^* & 1 & r_{12,i}^* & \cdots & r_{1,z-1,i}^* & r_{1,z+1,i}^* & \cdots & r_{1,k-1,i}^* & r_{1,k+1,i}^* & \cdots & r_{1M,i}^* \\ r_{z2,i}^* & r_{k2,i}^* & r_{12,i}^* & 1 & \cdots & r_{2,z-1,i}^* & r_{2,z+1,i}^* & \cdots & r_{2,k-1,i}^* & r_{2,k+1,i}^* & \cdots & r_{2M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{z,z-1,i}^* & r_{k,z-1,i}^* & r_{1,z-1,i}^* & r_{2,z-1,i}^* & \cdots & 1 & r_{z-1,z+1,i}^* & \cdots & r_{z-1,k-1,i}^* & r_{z-1,k+1,i}^* & \cdots & r_{z-1,M,i}^* \\ r_{z,z+1,i}^* & r_{k,z+1,i}^* & r_{1,z+1,i}^* & r_{2,z+1,i}^* & \cdots & r_{z-1,z+1,i}^* & 1 & \cdots & r_{z+1,k-1,i}^* & r_{z+1,k+1,i}^* & \cdots & r_{z+1,M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ r_{z,k-1,i}^* & r_{k,k-1,i}^* & r_{1,k-1,i}^* & r_{2,k-1,i}^* & \cdots & r_{z-1,k-1,i}^* & r_{z+1,k-1,i}^* & \cdots & 1 & r_{k-1,k+1,i}^* & \cdots & r_{k-1,M,i}^* \\ r_{z,k+1,i}^* & r_{k,k+1,i}^* & r_{1,k+1,i}^* & r_{2,k+1,i}^* & \cdots & r_{z-1,k+1,i}^* & r_{z+1,k+1,i}^* & \cdots & r_{k-1,k+1,i}^* & 1 & \cdots & r_{k+1,M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{z,M,i}^* & r_{k,M,i}^* & r_{1M,i}^* & r_{2M,i}^* & \cdots & r_{z-1,M,i}^* & r_{z+1,M,i}^* & \cdots & r_{k-1,M,i}^* & r_{k+1,M,i}^* & \cdots & 1 \end{array}\right)$$

A.6 Derivation of results in Section 2.3.2

A.6.1 Derivation of (2.11)

For notational convenience we denote $\boldsymbol{g}(\boldsymbol{\delta}^{[\cdot]})$ as $\boldsymbol{g}^{[\cdot]}$, $\boldsymbol{g}_p(\boldsymbol{\delta}^{[\cdot]})$ as $\boldsymbol{g}_p^{[\cdot]}$, $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}^{[\cdot]})$ as $\boldsymbol{\mathcal{H}}^{[\cdot]}$ and $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}^{[\cdot]})$ as $\boldsymbol{\mathcal{H}}_p^{[\cdot]}$.

By using the Taylor series expansion for $\boldsymbol{g}_p^{[\varkappa+1]}$ at $\boldsymbol{\delta}^{[\varkappa]}$ we have that $\mathbf{0} = \boldsymbol{g}_p^{[\varkappa+1]} \approx \boldsymbol{g}_p^{[\varkappa]} + \boldsymbol{\mathcal{H}}_p^{[\varkappa]}(\boldsymbol{\delta}^{[\varkappa+1]} - \boldsymbol{\delta}^{[\varkappa]})$, where $\boldsymbol{g}_p^{[\varkappa]} = \boldsymbol{g}^{[\varkappa]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}^{[\varkappa]}$ and $\boldsymbol{\mathcal{H}}_p^{[\varkappa]} = \boldsymbol{\mathcal{H}}^{[\varkappa]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}$. Suppose that $\boldsymbol{\mathcal{I}}^{[\varkappa]} = -\boldsymbol{\mathcal{H}}^{[\varkappa]}$; then we have that

$$\mathbf{0} \;=\; \boldsymbol{g}_p^{[\varkappa]} + \left(\boldsymbol{\delta}^{[\varkappa+1]} - \boldsymbol{\delta}^{[\varkappa]}\right)\left(-\boldsymbol{\mathcal{I}}^{[\varkappa]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right).$$

Re-arranging the above equation we get

$$\begin{aligned}
\boldsymbol{g}_p^{[\varkappa]} &= \left(\boldsymbol{\delta}^{[\varkappa+1]} - \boldsymbol{\delta}^{[\varkappa]}\right)\left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right) \implies \\
\implies \boldsymbol{g}^{[\varkappa]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}^{[\varkappa]} &= \boldsymbol{\delta}^{[\varkappa+1]}\left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right) - \boldsymbol{\delta}^{[\varkappa]}\boldsymbol{\mathcal{I}}^{[\varkappa]} - \boldsymbol{\delta}^{[\varkappa]}\tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}} \implies \\
\implies \boldsymbol{\delta}^{[\varkappa+1]}\left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right) &= \boldsymbol{g}^{[\varkappa]} + \boldsymbol{\delta}^{[\varkappa]}\boldsymbol{\mathcal{I}}^{[\varkappa]} \implies \\
\implies \boldsymbol{\delta}^{[\varkappa+1]} &= \left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\left(\sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\boldsymbol{\delta}^{[\varkappa]} + \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}^{-1}\boldsymbol{g}^{[\varkappa]}\right).
\end{aligned}$$

Therefore, the parameter estimator can be expressed as

$$\boldsymbol{\delta}^{[\varkappa+1]} \;=\; \left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\bar{\mathbf{z}}^{[\varkappa]},$$

where $\bar{\mathbf{z}}^{[\varkappa]} = \boldsymbol{\mu}_{\bar{\mathbf{z}}}^{[\varkappa]} + \bar{\boldsymbol{\epsilon}}^{[\varkappa]}$, $\boldsymbol{\mu}_{\bar{\mathbf{z}}}^{[\varkappa]} = \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\boldsymbol{\delta}^{[\varkappa]}$ and $\bar{\boldsymbol{\epsilon}}^{[\varkappa]} = \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}^{-1}\boldsymbol{g}^{[\varkappa]}$, as required.

A.6.2 Derivation of (2.12)

Based on the notation in Section 2.3.2, we have that

$$\begin{aligned}
\mathbb{E}(\|\boldsymbol{\mu}_{\bar{\mathbf{z}}} - \hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}\|^2) &= \mathbb{E}\left(\left\| (\bar{\mathbf{z}} - \bar{\boldsymbol{\epsilon}}) - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2\right) \\
&= \mathbb{E}\left(\left\| (\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}) - \bar{\boldsymbol{\epsilon}}\right\|^2\right) \\
&= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2 + \bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}} - 2\left\| (\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}})\,\bar{\boldsymbol{\epsilon}}\right\|\right) \\
&= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2 + \bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}} - 2\left\| \{\boldsymbol{\mu}_{\bar{\mathbf{z}}} + \bar{\boldsymbol{\epsilon}} - \mathbf{C}_{\boldsymbol{\lambda}}(\boldsymbol{\mu}_{\bar{\mathbf{z}}} + \bar{\boldsymbol{\epsilon}})\}\,\bar{\boldsymbol{\epsilon}}\right\|\right) \\
&= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2 + \bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}} - 2\left\|\boldsymbol{\mu}_{\bar{\mathbf{z}}}\bar{\boldsymbol{\epsilon}} + \bar{\boldsymbol{\epsilon}}^2 - \mathbf{C}_{\boldsymbol{\lambda}}\boldsymbol{\mu}_{\bar{\mathbf{z}}}\bar{\boldsymbol{\epsilon}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\boldsymbol{\epsilon}}^2)\right\|\right) \\
&= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2\right) + \mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}}\right) - 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\boldsymbol{\mu}_{\bar{\mathbf{z}}}\right) - \\
&\quad 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}}\right) + 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\boldsymbol{\mu}_{\bar{\mathbf{z}}}\right) + 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\bar{\boldsymbol{\epsilon}}\right) \\
&= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2\right) - \mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}}\right) - 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\boldsymbol{\mu}_{\bar{\mathbf{z}}}\right) + \\
&\quad 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\boldsymbol{\mu}_{\bar{\mathbf{z}}}\right) + 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\bar{\boldsymbol{\epsilon}}\right).
\end{aligned}$$

By using the following results (e.g., Wood, 2006, Section 1.8.5)

$$\begin{aligned}
\mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}}) &= \mathbb{E}\left(\sum_i \bar{\boldsymbol{\epsilon}}_i^2\right) = \tilde{n}\cdot 1 = \tilde{n}, \text{ for } \tilde{n} = 6n, \\
\mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top\boldsymbol{\mu}_{\bar{\mathbf{z}}}) &= \mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top)\boldsymbol{\mu}_{\bar{\mathbf{z}}} = \mathbf{0}, \\
\mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\bar{\boldsymbol{\epsilon}}) &= \mathbb{E}\left(\mathrm{tr}(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\bar{\boldsymbol{\epsilon}})\right), \text{ since a scalar is its own trace} \\
&= \mathrm{tr}\left(\mathbf{C}_{\boldsymbol{\lambda}}\mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}})\right) \\
&= \mathrm{tr}\left(\mathbf{C}_{\boldsymbol{\lambda}}\mathbf{I}\right)\cdot 1 \\
&= \mathrm{tr}\left(\mathbf{C}_{\boldsymbol{\lambda}}\right),
\end{aligned}$$

it follows that

$$\begin{aligned}\mathbb{E}(\|\boldsymbol{\mu}_{\bar{\mathbf{z}}} - \hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}\|^2) &= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2\right) - \tilde{n} - 2\cdot\mathbf{0} + 2\cdot\mathbf{0} + 2\mathrm{tr}\left(\mathbf{C}_{\hat{\boldsymbol{\lambda}}}\right) \\ &= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2\right) - \tilde{n} + 2\mathrm{tr}\left(\mathbf{C}_{\boldsymbol{\lambda}}\right),\end{aligned}$$

as required.

A.6.3 Equivalence of $\mathcal{V}(\boldsymbol{\lambda})$ and AIC

The AIC of a model can be defined as follows

$$\mathrm{AIC} = 2Q - 2\ell(\hat{\boldsymbol{\delta}}),$$

where Q is the number of estimated parameters in the model.

Consider a Taylor expansion of $-2\ell(\hat{\boldsymbol{\delta}})$ about $\boldsymbol{\delta}$

$$\begin{aligned}-2\ell(\hat{\boldsymbol{\delta}}) &\approx -2\ell(\boldsymbol{\delta}) + (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \nabla_{\boldsymbol{\delta}}\left\{-2\ell(\boldsymbol{\delta})\right\} + \frac{1}{2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \nabla\nabla_{\boldsymbol{\delta}}\left\{-2\ell(\boldsymbol{\delta})\right\}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \\ &\approx -2\ell(\boldsymbol{\delta}) - 2(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{g} - (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}), \qquad \text{(A.25)}\end{aligned}$$

where $\boldsymbol{g} := \boldsymbol{g}(\boldsymbol{\delta})$ and $\boldsymbol{\mathcal{H}} := \boldsymbol{\mathcal{H}}(\boldsymbol{\delta})$. By using $\boldsymbol{\mathcal{I}} = -\boldsymbol{\mathcal{H}}$, we have that

$$\begin{aligned}-(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) &= (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \\ &= \|\sqrt{\boldsymbol{\mathcal{I}}}\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right)\|^2 \\ &= \|\sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}} - \boldsymbol{\mathcal{I}}\boldsymbol{\delta}\|^2,\end{aligned}$$

and by applying $\bar{\mathbf{z}} = \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\delta} + \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}$ to the above expression we get

$$\begin{aligned}
-(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) &= \|\sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}} - \bar{\mathbf{z}} + \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 \\
&= \| - \left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}} - \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right) \|^2 \\
&= \| \left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\right) - \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 \qquad \text{(A.26)} \\
&= \left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}} \right\rangle - \\
&\quad 2\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \right\rangle + \left\langle \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \right\rangle \\
&= \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2 - 2\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \right\rangle + \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2, \qquad \text{(A.27)}
\end{aligned}$$

where (A.26) results from the fact that $\| - \tilde{\boldsymbol{\chi}}\|^2 = \|\tilde{\boldsymbol{\chi}}\|^2$, for any vector $\tilde{\boldsymbol{\chi}}$. Similarly, by using the expression for the pseudo-data vector $\bar{\mathbf{z}}$ in the second term in (A.25) we have

$$\begin{aligned}
(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{g} &= \left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right)^\top \sqrt{\boldsymbol{\mathcal{I}}}\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \\
&= \left(\sqrt{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \\
&= \left(\sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}} - \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\delta}\right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \\
&= \left(\sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}} - \bar{\mathbf{z}} + \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \\
&= -\left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}} - \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right)\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \\
&= -\left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} + \left(\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \\
&= -\left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\right)^\top \bullet \left(\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right) + \boldsymbol{g}^\top \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{g} \\
&= -\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g} \right\rangle + \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2. \qquad \text{(A.28)}
\end{aligned}$$

Substituting both (A.27) and (A.28) in (A.25), we obtain

$$\begin{aligned}
-2\ell(\hat{\boldsymbol{\delta}}) &\approx -2\ell(\boldsymbol{\delta}) - 2\left\{-\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle + \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2\right\} + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2 - \\
&\quad 2\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle + \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 \\
&\approx -2\ell(\boldsymbol{\delta}) + 2\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle - 2\|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2 - \\
&\quad 2\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle + \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 \\
&\approx -2\ell(\boldsymbol{\delta}) - \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2,
\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. It follows that

$$\begin{aligned}
\text{AIC} &\approx 2Q - 2\ell(\boldsymbol{\delta}) - \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2 \\
&\approx 2\text{tr}(\mathbf{C}_{\boldsymbol{\lambda}}) - 2\ell(\boldsymbol{\delta}) - \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2, \qquad \text{(A.29)}
\end{aligned}$$

where $\text{tr}(\mathbf{C}_{\boldsymbol{\lambda}})$ denotes the number of estimated parameters in the model and thus $Q = \text{tr}(\mathbf{C}_{\boldsymbol{\lambda}})$. Since we are interested in optimizing a criterion with respect to the smoothing parameter $\boldsymbol{\lambda}$, we drop any terms that are not affected by $\boldsymbol{\lambda}$, i.e., $-2\ell(\boldsymbol{\delta})$ and $-\|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2$. Therefore (A.29) becomes

$$\text{AIC} \approx 2\text{tr}(\mathbf{C}_{\lambda}) + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2,$$

as required.

A.7 Data generating processes used in the simulation study I

A.7.1 DGP1 & DGP2

Both DGP1 and DGP2 were based on the following trivariate system of equations

$$\begin{aligned} y_{1i}^* &= 1.6 + 0.9v_{1i} - 1.3z_{1i} + \varepsilon_{1i} \\ y_{2i}^* &= -1.0 - 1.4v_{1i} + 1.0z_{1i} + \varepsilon_{2i} \\ y_{3i}^* &= -1.4 + 2.0v_{1i} - 1.5z_{1i} + \varepsilon_{3i}, \end{aligned}$$

where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and v_{mi} and z_{mi}, $\forall m$, denote a binary regressor and a continuous covariate, respectively. DGP1 is fully parametric and was used for comparing `mvprobit()` and `SemiParTRIV()`/`gjrm()`. The correlation parameters were set as ($\vartheta_{12} = -0.8$, $\vartheta_{13} = -0.6$, $\vartheta_{23} = 0.8$). These values were obtained after fitting the trivariate probit model on the North Carolina data set used for the case study in Section 3.4. DGP2 was based on the following set of correlations ($\vartheta_{12} = -0.1$, $\vartheta_{13} = 0.3$, $\vartheta_{23} = 0.9$), which was selected while trying out different combinations of values; this choice seemed to be problematic from an estimation perspective as convergence was not achieved in most of the replicates used in the simulation study, and the estimates of the correlation parameters were not close to the true values. For each set-up, we generated 250 datasets with sample sizes equal to 1000 and 10000. Note that the responses were unbalanced (similarly as in the case study). Specifically, responses y_{1i}, y_{2i} and y_{3i} had typical observed value 1 proportions of 90.5%, 15.1% and 21.5%, respectively.

`STATA` and `R` code for DGP1

```
# Generate some data using STATA:
set obs 1000
```

```
matrix Er = (1,  -0.8, -0.6 \ -0.8, 1, 0.8 \ -0.6, 0.8, 1)
forvalues i = 1/250{
set seed ‘i’
drawnorm er1‘i’ er2‘i’ er3‘i’, corr(Er)
matrix C = (1, .5\ .5, 1)
drawnorm x1‘i’ x2‘i’, corr(C)
gen v1‘i’ = round(normal(x1‘i’))
gen z1‘i’ = normal(x2‘i’)
gen y1‘i’ = ( 1.6 + 0.9 * v1‘i’ - 1.3 * z1‘i’ + er1‘i’>0)
gen y2‘i’ = (-1.0 - 1.4 * v1‘i’ + 1.0 * z1‘i’ + er2‘i’>0)
gen y3‘i’ = (-1.4 + 2.0 * v1‘i’ - 1.5 * z1‘i’ + er3‘i’>0)
}

# Fit the model via the mvprobit routine in STATA:
ssc install mvprobit
forvalues i = 1/250{
capture mvprobit( y1‘i’ = v1‘i’ z1‘i’)(y2‘i’ = v1‘i’ z1‘i’)
(y3‘i’ = v1‘i’ z1‘i’)
matrix estparams‘i’ = e(b)
* creates a matrix of the parameter estimates
}

# Fit the model via the SemiParTRIV routine in R:
library(GJRM)
library(foreign)

SimsSTATADGP2 <- read.dta("SimsDGP2.dta") # extracts the simulation
                                          # STATA file

gamma11 <- gamma12 <- gamma13 <- gamma21 <- gamma22 <- gamma23 <- NULL
```

```
gamma31 <- gamma32 <- gamma33 <- theta12 <- theta13 <- theta23 <- NULL

n.rep <- 250 # number of replicates
n <- 1000
p <- 10 # number of variables generated in STATA, i.e., er1, er2, er3,
        # x1, x2, v1, z1, y1, y2, y3

eqn1 <- y1 ~ v1 + z1
eqn2 <- y2 ~ v1 + z1
eqn3 <- y3 ~ v1 + z1
f.l <- list(eqn1, eqn2, eqn3)

for(i in 1:n.rep){
  j <- ifelse(i>0, i+1, 1)
  DataSTATADGP2 <- SimsSTATADGP2[1:n, (i * p +1):(j * p)]
  v1 <- DataSTATADGP2[1:n, 6]
  z1 <- DataSTATADGP2[1:n, 7]
  y1 <- DataSTATADGP2[1:n, 8]
  y2 <- DataSTATADGP2[1:n, 9]
  y3 <- DataSTATADGP2[1:n, 10]
  data <- DataSTATADGP2
  out <- SemiParTRIV(f.l, data = data)
  X1.d2 <- out$X1.d2 # number of columns in the design matrix of first
                     # equation
  X2.d2 <- out$X2.d2 # number of columns in the design matrix of second
                     # equation
  X3.d2 <- out$X3.d2 # number of columns in the design matrix of third
                     # equation
  gamma11[i] <- out$fit$argument[1]
  gamma12[i] <- out$fit$argument[2]
```

```
  gamma13[i] <- out$fit$argument[X1.d2]
  gamma21[i] <- out$fit$argument[X1.d2 + 1]
  gamma22[i] <- out$fit$argument[X1.d2 + 2]
  gamma23[i] <- out$fit$argument[X1.d2 + X2.d2]
  gamma31[i] <- out$fit$argument[X1.d2 + X2.d2 + 1]
  gamma32[i] <- out$fit$argument[X1.d2 + X2.d2 + 2]
  gamma33[i] <- out$fit$argument[X1.d2 + X2.d2 + X3.d2]
  theta12[i] <- out$theta12
  theta13[i] <- out$theta13
  theta23[i] <- out$theta23
}

# Note: for sample size 10000 we replace set obs 1000 with set obs
# 10000 in the STATA code and n <- 1000 with n <- 10000 in the R code.
```

R code for DGP2

```
library(GJRM)

theta12.sim <- -0.1
theta13.sim <-  0.3
theta23.sim <-  0.9
n.rep <- 250
n <- 1000 # then n <- 10000
Sigma.er <- matrix( c( 1, theta12.sim, theta13.sim,
                       theta12.sim, 1, theta23.sim,
                       theta13.sim, theta23.sim, 1), 3 , 3)
theta.cov <- 0.5
SigmaCov       <- matrix(theta.cov, 2, 2)
diag(SigmaCov) <- 1
f.l <- list(y1 ~ v1 + z1, y2 ~ v1 + z1, y3 ~ v1 + z1 )
```

```
gamma11 <- gamma12 <- gamma13 <- gamma21 <- gamma22 <- gamma23 <- NULL
gamma31 <- gamma32 <- gamma33 <- theta12 <- theta13 <- theta23 <- NULL


for(i in 1:n.rep){
  set.seed(i)
  er  <- rMVN(n, rep(0,3), Sigma.er)
  cov <- rMVN(n, rep(0,2), SigmaCov)
  cov <- pnorm(cov)
  v1 <- round(cov[,1])
  z1 <- cov[,2]
  y1 <- ifelse(1.6  + 0.9 * v1 - 1.3 * z1 + er[,1] > 0, 1, 0)
  y2 <- ifelse(-1.0 - 1.4 * v1 + 1.0 * z1 + er[,2] > 0, 1, 0)
  y3 <- ifelse(-1.4 + 2.0 * v1 - 1.5 * z1 + er[,3] > 0, 1, 0)
  dataSim <- data.frame(y1, y2, y3, v1, z1)
  out <- SemiParTRIV(f.l, data = dataSim) # penCor = "lasso"
                                          # or penCor = "ridge"
                                          # or penCor = "alasso"
                                          # for penalized correlation
                                          #study
  X1.d2 <- out$X1.d2
  X2.d2 <- out$X2.d2
  X3.d2 <- out$X3.d2
  gamma11[i] <- coef(out)[1]
  gamma12[i] <- coef(out)[2]
  gamma13[i] <- coef(out)[X1.d2]
  gamma21[i] <- coef(out)[X1.d2 + 1]
  gamma22[i] <- coef(out)[X1.d2 + 2]
  gamma23[i] <- coef(out)[X1.d2 + X2.d2]
  gamma31[i] <- coef(out)[X1.d2 + X2.d2 + 1]
  gamma32[i] <- coef(out)[X1.d2 + X2.d2 + 2]
```

```
  gamma33[i] <- coef(out)[X1.d2 + X2.d2 + X3.d2]
  theta12[i] <- out$theta12
  theta13[i] <- out$theta13
  theta23[i] <- out$theta23
}
```

Appendix B

Complements to Chapter 3

B.1 Correlation-based penalty

B.1.1 The penalty functions

$$
\begin{aligned}
\textbf{Lasso: } \mathcal{P}^{\mathrm{L}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &= \mathcal{P}^{\mathrm{L}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_1) \\
&= \lambda_{\vartheta^*}\|\mathbf{R}_q\boldsymbol{\delta}\|_1 \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}|\mathbf{R}_q\boldsymbol{\delta}| \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\left\{\left(\mathbf{R}_q\boldsymbol{\delta}\right)^\top\mathbf{R}_q\boldsymbol{\delta}\right\}^{1/2} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\left\{\left(\boldsymbol{e}_q^\top\boldsymbol{\delta}\right)^2\right\}^{1/2} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}|\boldsymbol{e}_q^\top\boldsymbol{\delta}| \\
&= \lambda_{\vartheta^*}\left\{|\boldsymbol{e}_{Q-2}^\top\boldsymbol{\delta}| + |\boldsymbol{e}_{Q-1}^\top\boldsymbol{\delta}| + |\boldsymbol{e}_{Q}^\top\boldsymbol{\delta}|\right\} \\
&= \lambda_{\vartheta^*}\left(|\vartheta^*_{12}| + |\vartheta^*_{13}| + |\vartheta^*_{23}|\right),
\end{aligned}
$$

where $\boldsymbol{e}_q = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ with a one at the q^{th} position, $\forall q$.

$$
\begin{aligned}
\textbf{Ridge: } \mathcal{P}^{\mathrm{R}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &= \frac{1}{2}\mathcal{P}^{\mathrm{R}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_2^2) \\
&= \frac{1}{2}\lambda_{\vartheta^*}\|\mathbf{R}_q\boldsymbol{\delta}\|_2^2 \\
&= \frac{1}{2}\lambda_{\vartheta^*}\left\{\left[\sum_{q=Q-2}^{Q}\left((\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta}\right)\right]^{1/2}\right\}^2 \\
&= \frac{1}{2}\lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\left((\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta}\right) \\
&= \frac{1}{2}\lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\left(\boldsymbol{e}_q^\top\boldsymbol{\delta}\right)^2 \\
&= \frac{1}{2}\lambda_{\vartheta^*}\left\{(\boldsymbol{e}_{Q-2}^\top\boldsymbol{\delta})^2 + (\boldsymbol{e}_{Q-1}^\top\boldsymbol{\delta})^2 + (\boldsymbol{e}_{Q}^\top\boldsymbol{\delta})^2\right\} \\
&= \frac{1}{2}\lambda_{\vartheta^*}\left(\vartheta_{12}^{*2} + \vartheta_{13}^{*2} + \vartheta_{23}^{*2}\right).
\end{aligned}
$$

$$
\begin{aligned}
\textbf{Ad. Lasso: } \mathcal{P}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &= \mathcal{P}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_1) \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\frac{|\mathbf{R}_q\boldsymbol{\delta}|}{|\mathbf{R}_q\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}|^{\bar{\gamma}}} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\frac{\left\{\left(\mathbf{R}_q\boldsymbol{\delta}\right)^\top\mathbf{R}_q\boldsymbol{\delta}\right\}^{1/2}}{\left\{\left(\mathbf{R}_q\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}\right)^\top\mathbf{R}_q\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}\right\}^{\bar{\gamma}/2}} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\frac{\left\{\left(\boldsymbol{e}_q^\top\boldsymbol{\delta}\right)^2\right\}^{1/2}}{\left\{\left(\boldsymbol{e}_q^\top\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}\right)^2\right\}^{\bar{\gamma}/2}} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\frac{|\boldsymbol{e}_q^\top\boldsymbol{\delta}|}{|\boldsymbol{e}_q^\top\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}|^{\bar{\gamma}}} \\
&= \lambda_{\vartheta^*}\left\{\frac{|\boldsymbol{e}_{Q-2}^\top\boldsymbol{\delta}|}{|\boldsymbol{a}_{Q-2}^\top\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}|^{\bar{\gamma}}} + \frac{|\boldsymbol{e}_{Q-1}^\top\boldsymbol{\delta}|}{|\boldsymbol{e}_{Q-1}^\top\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}|^{\bar{\gamma}}} + \frac{|\boldsymbol{e}_{Q}^\top\boldsymbol{\delta}|}{|\boldsymbol{a}_{Q}^\top\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}|^{\bar{\gamma}}}\right\} \\
&= \lambda_{\vartheta^*}\left(\frac{|\vartheta_{12}^*|}{|\hat{\vartheta}_{12}^{*\mathrm{MLE}}|^{\bar{\gamma}}} + \frac{|\vartheta_{13}^*|}{|\hat{\vartheta}_{13}^{*\mathrm{MLE}}|^{\bar{\gamma}}} + \frac{|\vartheta_{23}^*|}{|\hat{\vartheta}_{23}^{*\mathrm{MLE}}|^{\bar{\gamma}}}\right).
\end{aligned}
$$

B.1.2 LQA of the penalty function $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$

The approximated penalty functions for both Lasso and Adaptive Lasso belong to the L_1-type family. Based on (3.5) and by applying the chain rule, it follows that $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ can be written as

$$\begin{aligned}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \nabla_{\tilde{\boldsymbol{\delta}}}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})^\top(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}) \\ &\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \frac{\partial \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})}{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\cdot\frac{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}{\partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}}\cdot\frac{\partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\partial\tilde{\boldsymbol{\delta}}^\top}\cdot\left(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right). \qquad \text{(B.1)}\end{aligned}$$

By using the local approximation $(\mathbf{R}_q\boldsymbol{\delta})^\top / \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top \approx 1$ for $\tilde{\boldsymbol{\delta}} \approx \boldsymbol{\delta}$ (Fan & Li, 2001) as well as the following approximation (Ulbricht, 2010)

$$\begin{aligned}(\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\left(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right) &= (\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta} - (\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}} \\ &= \frac{1}{2}\left\{(\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta} - 2\,(\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}} + \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right\} + \\ &\quad \frac{1}{2}\left\{(\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta} - \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right\} \\ &= \frac{1}{2}\left(\mathbf{R}_q^\top(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}})^\top(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}})\mathbf{R}_q\right) + \\ &\quad \frac{1}{2}\left((\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta} - \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right) \\ &\approx \frac{1}{2}(\boldsymbol{\delta}^\top\mathbf{R}_q^\top\mathbf{R}_q\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}^\top\mathbf{R}_q^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}),\end{aligned}$$

equation (B.1) becomes

$$\begin{aligned}
\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top} \cdot \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top \cdot \mathbf{R}_q \cdot \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right) \\
&\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top} \cdot \frac{1}{2}(\boldsymbol{\delta}^\top\mathbf{R}_q^\top\mathbf{R}_q\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}^\top\mathbf{R}_q^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}) \\
&\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \frac{1}{2}\boldsymbol{\delta}^\top\left\{\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top}\mathbf{R}_q\mathbf{R}_q^\top\right\}\boldsymbol{\delta} - \\
&\qquad \frac{1}{2}\tilde{\boldsymbol{\delta}}^\top\left\{\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top}\mathbf{R}_q\mathbf{R}_q^\top\right\}\tilde{\boldsymbol{\delta}},
\end{aligned}$$

where $\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) = \partial\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})/\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1$, $\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}}) = \partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1/\partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}$ and $\mathbf{R}_q = \partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}/\partial\tilde{\boldsymbol{\delta}}^\top$. The constant terms do not affect (3.1) and hence can be eliminated. Therefore $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ can be locally approximated (except for a constant term) by

$$\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) \approx \frac{1}{2}\boldsymbol{\delta}^\top\left\{\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top}\mathbf{R}_q\mathbf{R}_q^\top\right\}\boldsymbol{\delta}.$$

B.1.3 Derivation of $\boldsymbol{\Lambda}^{\mathrm{L}}_{\lambda_{\vartheta^*}}$ and $\boldsymbol{\Lambda}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}$

Based on the approximation derived in Appendix B.1.2, we have that the penalty matrix $\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ is equal to

$$\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}} = \nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top}\mathbf{R}_q\mathbf{R}_q^\top.$$

Quantity $\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})$ for Lasso and Adaptive Lasso, respectively, is equal to

$$\begin{aligned}
\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathrm{L}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) &= \frac{\partial\mathcal{P}^{\mathrm{L}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})}{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} = \frac{\partial\left(\lambda_{\vartheta^*}\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1\right)}{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} = \lambda_{\vartheta^*}, \\
\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) &= \frac{\partial\mathcal{P}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})}{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} = \frac{\partial\left(\lambda_{\vartheta^*}\sum_q w_q|\mathbf{R}_q\tilde{\boldsymbol{\delta}}|\right)}{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} = \lambda_{\vartheta^*}w_q.
\end{aligned}$$

Derivative $\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})$ is equal to

$$\begin{aligned}
\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}}) &= \frac{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}{\partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}} \\
&= \frac{\partial}{\partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}}\sum_{q=Q-2}^{Q}\left\{\left(\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^{1/2}\right\} \\
&= \sum_{q=Q-2}^{Q}\left\{\frac{1}{2}\left(\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^{-1/2}\cdot 2\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right\} \\
&= \sum_{q=Q-2}^{Q}\frac{\mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\sqrt{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}}} \\
&\approx \sum_{q=Q-2}^{Q}\frac{\mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\sqrt{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}+\bar{c}}},
\end{aligned}$$

where the denominator was approximated by $\left(\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}\right)^{-1/2}$ which allows for $\tilde{\boldsymbol{\delta}} = \mathbf{0}$. It follows that $\boldsymbol{\Lambda}^{\mathrm{L}}_{\lambda_{\vartheta^*}}$ can be expressed as

$$\begin{aligned}
\boldsymbol{\Lambda}^{\mathrm{L}}_{\lambda_{\vartheta^*}} &= \sum_{q=Q-2}^{Q} \left\{ \frac{\lambda_{\vartheta^*}}{\mathbf{R}_q\tilde{\boldsymbol{\delta}}} \cdot \frac{\mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\sqrt{(\mathbf{R}_q\tilde{\boldsymbol{\delta}})^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}}} \cdot \mathbf{R}_q\mathbf{R}_q^\top \right\} \\
&= \sum_{q=Q-2}^{Q} \left\{ \frac{\lambda_{\vartheta^*}}{\sqrt{(\mathbf{R}_q\tilde{\boldsymbol{\delta}})^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}}} \cdot \mathbf{R}_q\mathbf{R}_q^\top \right\} \\
&= \lambda_{\vartheta^*} \left\{ \frac{1}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 1, 0, 0) + \right. \\
&\quad \frac{1}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 0, 1, 0) + \\
&\quad \left. \frac{1}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 0, 0, 1) \right\} \\
&= \lambda_{\vartheta^*} \operatorname{diag}\left(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, \frac{1}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}}, \frac{1}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}}, \frac{1}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \right),
\end{aligned}$$

while $\boldsymbol{\Lambda}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}$ is equal to

$$\begin{aligned}
\boldsymbol{\Lambda}^{\mathrm{AL}}_{\lambda_{\vartheta^*}} &= \sum_{q=Q-2}^{Q} \left\{ \frac{\lambda_{\vartheta^*}\mathrm{w}_q}{\mathbf{R}_q\tilde{\boldsymbol{\delta}}} \cdot \frac{\mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\sqrt{(\mathbf{R}_q\tilde{\boldsymbol{\delta}})^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}}} \cdot \mathbf{R}_q\mathbf{R}_q^\top \right\} \\
&= \sum_{q=Q-2}^{Q} \left\{ \frac{\lambda_{\vartheta^*}\mathrm{w}_q}{\sqrt{(\mathbf{R}_q\tilde{\boldsymbol{\delta}})^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}}} \cdot \mathbf{R}_q\mathbf{R}_q^\top \right\} \\
&= \lambda_{\vartheta^*} \left\{ \frac{\mathrm{w}_{12}}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 1, 0, 0) + \right. \\
&\quad \frac{\mathrm{w}_{13}}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 0, 1, 0) + \\
&\quad \left. \frac{\mathrm{w}_{23}}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 0, 0, 1) \right\} \\
&= \lambda_{\vartheta^*} \operatorname{diag}\left(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{12}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{13}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{23}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \right).
\end{aligned}$$

B.2 Data generating process used in the simulation study II

B.2.1 DGP3

The trivariate system of equations was based on

$$\begin{aligned} y_{1i}^* &= 1.05 + 0.90v_{1i} + s_1(z_{1i}) + \varepsilon_{1i} \\ y_{2i}^* &= -1.45 - 1.40v_{1i} + s_2(z_{1i}) + \varepsilon_{2i} \\ y_{3i}^* &= -1.60 + 2.00v_{1i} + s_3(z_{1i}) + \varepsilon_{3i} \end{aligned}$$

where $\boldsymbol{\varepsilon}_i \sim (\mathbf{0}, \boldsymbol{\Sigma})$ and s_m, for all m, corresponds to the smooth component which was represented using penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives. The correlation parameters were set to the same values as those used for DGP2, while the smooth functions are given by $s_1(z_{1i}) = 0.5\cos(2\pi z_{1i})$, $s_2(z_{1i}) = z_{1i} + \exp\{-30(z_{1i} - 0.5)^2\}$ and $s_3(z_{1i}) = -0.5(z_{1i} + 3z_{1i}^3)$. The other settings are similar to those described in Appendix A.7.1. For each replicate and fitted model the estimated smooth functions were evaluated at 200 fixed values in the ranges of the respective covariates. Parameter estimation was carried out using a Lasso-type penalty for the correlations, i.e. $\boldsymbol{\Gamma}_{\boldsymbol{\lambda}}^{\mathrm{L}} = \tilde{\boldsymbol{S}}_{\boldsymbol{\lambda}} + \boldsymbol{\Lambda}_{\lambda_{\vartheta^*}}^{\mathrm{L}}$; using Ridge and Adaptive Lasso did led to virtually identical results.

R code for DGP3

```
library(GJRM)


# Simulate some data:
n     <- 1000 # then n <- 10000
n.rep <- 250
theta12.sim <- -0.1
```

```
theta13.sim <-  0.3
theta23.sim <-  0.9
Sigma.er <- matrix( c( 1, theta12.sim, theta13.sim,
                       theta12.sim, 1, theta23.sim,
                       theta13.sim, theta23.sim, 1 ), 3 , 3)
SigmaCov <- matrix(0.5, 2, 2)
diag(SigmaCov) <- 1
f.l <- list(y1 ~ v1 + s(z1), y2 ~ v1 + s(z1), y3 ~ v1 + s(z1) )
F1 <- F2 <- F3 <- matrix(NA, 200, n.rep)
theta12 <- theta13 <- theta23 <- NULL
# smooth functions
f1 <- function(x) 0.5*cos(pi*2*x)
f2 <- function(x) x+exp(-30*(x-0.5)^2)
f3 <- function(x) -0.5*(x+3*x^3)
xt <- seq(0.0000001, 0.9999999, length.out = 200) # grid to evaluate
smooth functions
dt <- data.frame(z = xt)
f1t <- f1(xt) - mean(f1(xt))
f2t <- f2(xt) - mean(f2(xt))
f3t <- f3(xt) - mean(f3(xt))

for(i in 1:n.rep){
set.seed(i)
u   <- rMVN(n, rep(0,3), Sigma.er)
cov <- rMVN(n, rep(0,2), SigmaCov)
cov <- pnorm(cov)
v1  <- round(cov[, 1])
z1  <- cov[, 2]
y1 <- ifelse( 1.05 + 0.9*v1  + f1(z1) + u[,1]  > 0, 1, 0)
y2 <- ifelse(-1.45 - 1.4*v1  + f2(z1) + u[,2]  > 0, 1, 0)
```

```
y3 <- ifelse(-1.6  + 2.0*v1  + f3(z1) + u[,3]  > 0, 1, 0)
dataSim <- data.frame(y1, y2, y3, v1, z1)
out <- SemiParTRIV(f.l, data = dataSim, penCor = "lasso")
X1 <- PredictMat( out$gam1$smooth[[1]], dt )
X2 <- PredictMat( out$gam2$smooth[[1]], dt )
X3 <- PredictMat( out$gam3$smooth[[1]], dt )
lg1 <- length(coef(out$gam1))
lg2 <- length(coef(out$gam2))
F1[,i] <- X1%*%
coef(out)[(out$gam1$smooth[[1]]$first.para:out$gam1$smooth[[1]]$
last.para)]
F2[,i] <- X2%*%
coef(out)[lg1 + (out$gam2$smooth[[1]]$first.para:out$gam2$smooth[[1]]$
last.para)]
F3[,i] <- X3%*%
coef(out)[lg1 + lg2 + (out$gam3$smooth[[1]]$first.para:out$gam3
$smooth[[1]]$last.para)]
F1[,i] <- F1[,i] - mean(F1[,i])
F2[,i] <- F2[,i] - mean(F2[,i])
F3[,i] <- F3[,i] - mean(F3[,i])
theta12[i] <- out$theta12
theta13[i] <- out$theta13
theta23[i] <- out$theta23
}
```

B.3 Some theoretical aspects

B.3.1 Proof of Theorem 3.3.1

Proof. By definition, the gradient of the log-likelihood function at $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ is equal to zero, that is $\boldsymbol{g}(\hat{\boldsymbol{\delta}}^{\text{MLE}}) = \mathbf{0}$. If $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ is close to $\boldsymbol{\delta}_0$, then $\boldsymbol{g}(\hat{\boldsymbol{\delta}}^{\text{MLE}})$ can be approximated by a Taylor series around the true parameter $\boldsymbol{\delta}_0$. We apply the mean value theorem in order to truncate the Taylor series at the second term, that is

$$\boldsymbol{g}(\hat{\boldsymbol{\delta}}^{\text{MLE}}) \approx \boldsymbol{g}(\boldsymbol{\delta}_0) + \boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) = \mathbf{0}.$$

Multiplying both sides by $\sqrt{n}$ and rearranging, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) \approx \{-\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\}^{-1} \left\{\sqrt{n}\boldsymbol{g}(\boldsymbol{\delta}_0)\right\},$$

and by dividing $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ and $\boldsymbol{g}(\boldsymbol{\delta}_0)$ by n we obtain

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) \approx \left\{-\frac{1}{n}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1} \left\{\sqrt{n}\frac{\boldsymbol{g}(\boldsymbol{\delta}_0)}{n}\right\}.$$

Since $\boldsymbol{g}(\boldsymbol{\delta}_0)/n$ is the mean of a random sample, we may apply the Central Limit Theorem (CLT) to $\sqrt{n}\boldsymbol{g}(\boldsymbol{\delta}_0)/n$. According to the theorem and given that $\mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\delta}_0)) = \mathbf{0}$ (as $\boldsymbol{\delta}_0$ is the maximizer of $\ell(\boldsymbol{\delta}_0)$, $\forall i$) we have that

$$\sqrt{n}\left\{\frac{1}{n}\boldsymbol{g}(\boldsymbol{\delta}_0) - \cancelto{0}{\mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\delta}_0))}\right\} \to \mathcal{N}\left(\mathbf{0}, \mathbf{Cov}\left(\boldsymbol{g}_i(\boldsymbol{\delta}_0)\right)\right),$$

where

$$\mathbf{Cov}\left(\boldsymbol{g}_i(\boldsymbol{\delta}_0)\right) = \mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\delta}_0)\boldsymbol{g}_i(\boldsymbol{\delta}_0)^\top) = \{-\mathbb{E}\left(\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0)\right)\} = -\mathbb{E}\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0) = -\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0).$$

It follows that

$$\sqrt{n}\left\{\frac{1}{n}\boldsymbol{g}(\boldsymbol{\delta}_0)\right\} \to \mathcal{N}\left(\mathbf{0}, -\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right).$$

By using the limiting distribution $\mathbb{P}\left(\lim\left(-1/n\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right)\right) = -1/n\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ we have that

$$\left\{-\frac{1}{n}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\left\{\sqrt{n}\frac{\boldsymbol{g}(\boldsymbol{\delta}_0)}{n}\right\} \;\rightarrow\; \mathcal{N}\left(\mathbf{0}, \bar{\boldsymbol{\Gamma}}\right),$$

for

$$\bar{\boldsymbol{\Gamma}} = \left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1} = \left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}.$$

Equivalently

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\mathrm{MLE}} - \boldsymbol{\delta}_0) \;\rightarrow\; \mathcal{N}\left(\mathbf{0}, \left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\right),$$

or

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\mathrm{MLE}} - \boldsymbol{\delta}_0) \;\rightarrow\; \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1}\right),$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) = -\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ and $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)$ denotes the Fisher information matrix, as required. □

B.3.2 Proof of Theorem 3.3.2

Proof. The first-order Taylor expansion of $\boldsymbol{g}_p(\cdot)$ around $\boldsymbol{\delta}_0$ is as follows

$$\boldsymbol{g}_p(\hat{\boldsymbol{\delta}}) \;\approx\; \boldsymbol{g}_p(\boldsymbol{\delta}_0) + \boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0). \tag{B.2}$$

By using the fact that $\boldsymbol{g}_p(\hat{\boldsymbol{\delta}}) = \mathbf{0}$ and multiplying all terms by $\sqrt{n}$ leads to

$$\sqrt{n}\boldsymbol{g}_p(\boldsymbol{\delta}_0) + \sqrt{n}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \;=\; 0.$$

Inverting the above series results to

$$\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \;=\; -\left\{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{\sqrt{n}\boldsymbol{g}_p(\boldsymbol{\delta}_0)\right\}.$$

We then divide both $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)$ and $\boldsymbol{g}_p(\boldsymbol{\delta}_0)$ by n, that is

$$\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \;=\; -\left\{\frac{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\sqrt{n}\frac{\boldsymbol{g}_p(\boldsymbol{\delta}_0)}{n}\right\}.$$

By using the CLT on $\sqrt{n}\boldsymbol{g}_p(\boldsymbol{\delta}_0)/n$ we obtain the following

$$\sqrt{n}\left\{\frac{\boldsymbol{g}_p(\boldsymbol{\delta})}{n} - \mathbb{E}(\boldsymbol{g}_{pi}(\boldsymbol{\delta}_0))\right\} \;\to\; \mathcal{N}\left(\mathbf{0}, \mathbf{Cov}(\boldsymbol{g}_{pi}(\boldsymbol{\delta}_0))\right), \tag{B.3}$$

where $\mathbb{E}(\boldsymbol{g}_{pi}(\boldsymbol{\delta}_0)) = 1/n\mathbb{E}\left(\boldsymbol{g}_p(\boldsymbol{\delta}_0)\right) = 1/n\mathbb{E}\left(\boldsymbol{g}(\boldsymbol{\delta}_0)\right) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0) = 1/n\left[\mathbb{E}\left(\boldsymbol{g}(\boldsymbol{\delta}_0)\right) - \mathbb{E}\left(\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right)\right] = 1/n\left[\mathbf{0} - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] = -1/n\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0$ and $\mathbf{Cov}(\boldsymbol{g}_{pi}(\boldsymbol{\delta}_0)) = \mathbf{Cov}(\boldsymbol{g}_i(\boldsymbol{\delta}_0) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0) = \mathbf{Cov}(\boldsymbol{g}_i(\boldsymbol{\delta}_0)) = -\mathbb{E}\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0) = -1/n\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$. Therefore, (B.3) can be re-expressed as

$$\sqrt{n}\left\{\frac{\boldsymbol{g}_p(\boldsymbol{\delta}) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{n}\right\} \;\to\; \mathcal{N}\left(\mathbf{0}, -1/n\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right),$$

and thus

$$\sqrt{n}\frac{\boldsymbol{g}_p(\boldsymbol{\delta})}{n} \;\to\; \mathcal{N}\left(-\frac{\sqrt{n}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{n}, -1/n\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right).$$

Next we use the law of large numbers that says that the observed information converges to the expected Fisher information as the sample size increases. That is, $-\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0) \to -\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)$. Therefore

$$-\left\{\frac{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\sqrt{n}\frac{\boldsymbol{g}_p(\boldsymbol{\delta}_0)}{n}\right\} \;\to\; \mathcal{N}\left(\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{\sqrt{n}}\right\}, \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{n}\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\right),$$

which implies that

$$\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \;\to\; \mathcal{N}\left(\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{\sqrt{n}}\right\}, \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{n}\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\right), \tag{B.4}$$

From the above result we can calculate the bias of the estimator $\hat{\boldsymbol{\delta}}$, that is

$$\begin{aligned}
\mathbf{Bias}(\hat{\boldsymbol{\delta}}) &= \mathbb{E}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \\
&\approx \frac{1}{\sqrt{n}}\mathbb{E}\left[\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)\right] \\
&\approx \frac{1}{\sqrt{n}}\left[\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{\sqrt{n}}\right\}\right] \\
&\approx -\frac{1}{\sqrt{n}}n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{\sqrt{n}}\right\} \\
&\approx -\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0 \\
&\approx -\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0 \\
&\approx -\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0.
\end{aligned}$$

as well as its asymptotic covariance matrix

$$\begin{aligned}
\mathbf{Cov}(\hat{\boldsymbol{\delta}}) &\approx \frac{1}{n}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{n}\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1} \\
&\approx \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1} \\
&\approx \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \\
&\approx \left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}.
\end{aligned}$$

Rearranging (B.4) leads to

$$\sqrt{n}\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}\left[(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) + \left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] \;\rightarrow\; \mathcal{N}\left(\mathbf{0}, n\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right),$$

which results from the following: expression (B.4) can be re-written as

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \;\rightarrow\; \mathcal{N}\big(&-\sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\}, \\
&n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\big),
\end{aligned}$$

or

$$\sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \;\rightarrow\; \mathcal{N}\left(-\sqrt{n}\left\{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\}, n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\right),$$

and therefore,

$$\begin{aligned}
\sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)+\sqrt{n}\left\{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\} &\rightarrow \mathcal{N}\left(\mathbf{0}, n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\right)\\
\Longrightarrow \sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}\left[\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0+\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] &\rightarrow \mathcal{N}\left(\mathbf{0}, n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\right)\\
\Longrightarrow \sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}\left[\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0+\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] &\rightarrow \mathcal{N}\left(\mathbf{0}, n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\right)\\
\Longrightarrow \sqrt{n}\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}\left[(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)+\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] &\rightarrow \mathcal{N}\left(\mathbf{0}, n\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right).
\end{aligned}$$

□

B.3.3 Asymptotic order of $\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0$, $\mathrm{Cov}(\hat{\boldsymbol{\delta}})$ and $\mathrm{Bias}(\hat{\boldsymbol{\delta}})$

Proof of the asymptotic order of $\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0$

Rearranging equation (B.2) leads to

$$\begin{aligned}
\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0 &= -\left\{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\boldsymbol{g}_p(\boldsymbol{\delta}_0)+\ldots\\
&= -\left\{\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)-\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\left\{\boldsymbol{g}(\boldsymbol{\delta}_0)-\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\}+\ldots\\
&= -\left\{\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)-\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\left\{\boldsymbol{g}(\boldsymbol{\delta}_0)-\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\}+\ldots,
\end{aligned}$$

and by applying assumptions (i)-(iv) we have that

$$\begin{aligned}
\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0 &= -\left\{\mathcal{O}_P(n^{1/2})+\mathcal{O}(n)-o(n^{1/2})\right\}^{-1}\left\{\mathcal{O}_P(n^{1/2})-o(n^{1/2})\right\}\\
&= \left\{\mathcal{O}_P(n)\right\}^{-1}\left\{\mathcal{O}_P(n^{1/2})\right\}\\
&= \mathcal{O}_P(n^{-1})\mathcal{O}_P(n^{1/2})\\
&= \mathcal{O}_P(n^{-1/2}).
\end{aligned}$$

Proof of the asymptotic order of $\mathbf{Cov}(\hat{\boldsymbol{\delta}})$

The asymptotic covariance of $\hat{\boldsymbol{\delta}}$ is of order

$$\begin{aligned}
\mathbf{Cov}(\hat{\boldsymbol{\delta}}) &\approx \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\} \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \\
&= \left\{\mathcal{O}(n) + o(n^{1/2})\right\}^{-1} \left\{\mathcal{O}(n)\right\} \left\{\mathcal{O}(n) + o(n^{1/2})\right\}^{-1} \\
&= \left\{\mathcal{O}(n)\right\}^{-1} \left\{\mathcal{O}(n)\right\} \left\{\mathcal{O}(n)\right\}^{-1} \\
&= \mathcal{O}(n^{-1}).
\end{aligned}$$

Proof of the asymptotic order of $\mathbf{Bias}(\hat{\boldsymbol{\delta}})$

The asymptotic bias of $\hat{\boldsymbol{\delta}}$ is of order

$$\begin{aligned}
\mathbf{Bias}(\hat{\boldsymbol{\delta}}) &\approx -\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0 \\
&= -\left\{-\mathcal{O}(n) + o(n^{1/2})\right\}^{-1} o(n^{1/2}) \\
&= \left\{\mathcal{O}(n)\right\}^{-1} o(n^{1/2}) \\
&= \mathcal{O}(n^{-1})\, o(n^{1/2}) \\
&= o(n^{-1/2}).
\end{aligned}$$

B.3.4 Proof of Theorem 3.3.3

Proof. If $\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0| = o(n^{1/2})$ and $\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}| = o(n^{1/2})$, then as $n \to \infty$ we have that $1/\sqrt{n}\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0| \to \mathbf{0}$ and $1/\sqrt{n}\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}| \to \mathbf{0}$. Given these two conditions, it follows that

$$\begin{aligned}
\mathbb{E}\left(\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)\right) &= \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{n}\right\}^{-1} \left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}\boldsymbol{\delta}_0}}{\sqrt{n}}\right\} \\
&\to \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{n}\right\}^{-1} \cdot \mathbf{0} \\
&\to \mathbf{0},
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{Cov}\left(\sqrt{n}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)\right) &= \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{n}\right\}^{-1}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{n}\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{n}\right\}^{-1} \\
&= n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \\
&= \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{\sqrt{n}}\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{\sqrt{n}}\right\}^{-1} \\
&\to \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{\sqrt{n}}+\mathbf{0}\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{\sqrt{n}}+\mathbf{0}\right\}^{-1} \\
&\to \sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\sqrt{n} \\
&\to n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1} \\
&\to n\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1} \\
&\to \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta})\right\}^{-1},
\end{aligned}$$

and thus

$$\sqrt{n}\left(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0\right) \sim \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1}\right).$$

□

B.3.5 Proof of Theorem 3.3.4

Proof. If $\hat{\boldsymbol{\delta}}$ minimizes $-\ell_p(\boldsymbol{\delta})$, then it also minimizes $-\ell_p(\boldsymbol{\delta})/n$. Similarly, $\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}$ minimizes $-\ell(\boldsymbol{\delta})$ as well as $-\ell(\boldsymbol{\delta})/n$. Because $\bar{\boldsymbol{\lambda}}$ is fixed, we have that $-\ell_p(\hat{\boldsymbol{\delta}})/n \to -\ell(\hat{\boldsymbol{\delta}}^{\mathrm{MLE}})/n$ and $-\ell_p(\hat{\boldsymbol{\delta}})/n \to -\ell(\hat{\boldsymbol{\delta}})/n$; thus $-\ell(\hat{\boldsymbol{\delta}})/n \to -\ell(\hat{\boldsymbol{\delta}}^{\mathrm{MLE}})/n$ hold as well. Since $\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}$ is a unique minimizer of $-\ell(\boldsymbol{\delta})/n$ and $-\ell(\boldsymbol{\delta})/n$ is convex, it follows that $\hat{\boldsymbol{\delta}} \to \hat{\boldsymbol{\delta}}^{\mathrm{MLE}}$. The consistency of $\hat{\boldsymbol{\delta}}$ follows from the consistency of $\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}$. □

Appendix C

Complements to Chapter 5

C.1 Proof of Lemma 5.2.1

Proof. For convenience we ignore index $\tilde{k}$ and term $\mathcal{Y}_{i\tilde{k}}$. By definition,

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{y}_i;\boldsymbol{\delta}) &= \mathbb{P}(-\tilde{y}_{1i}y^*_{1i}\leq 0,\ldots,-\tilde{y}_{Mi}y^*_{Mi}\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}(\eta_{1i}+\varepsilon_{1i})\leq 0,\ldots,-\tilde{y}_{Mi}(\eta_{Mi}+\varepsilon_{Mi})\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}(\Phi^{-1}(\mathrm{F}_1(\eta_{1i}))+\varepsilon_{1i})\leq 0,\ldots,-\tilde{y}_{Mi}(\Phi^{-1}(\mathrm{F}_M(\eta_{Mi}))+\varepsilon_{Mi})\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}\Phi^{-1}(\mathrm{F}_1(\eta_{1i}))-\tilde{y}_{1i}\varepsilon_{1i}\leq 0,\ldots,-\tilde{y}_{Mi}\Phi^{-1}(\mathrm{F}_M(\eta_{Mi}))-\tilde{y}_{Mi}\varepsilon_{Mi}\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}\varepsilon_{1i}\leq \tilde{y}_{1i}\Phi^{-1}(\mathrm{F}_1(\eta_{1i})),\ldots,-\tilde{y}_{Mi}\varepsilon_{Mi}\leq \tilde{y}_{Mi}\Phi^{-1}(\mathrm{F}_M(\eta_{Mi})))\\
&= \Phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\mathbf{H}_i;\mathbf{0},\boldsymbol{\Sigma})\\
&= \int_{-\infty}^{\tilde{y}_{Mi}\Phi^{-1}(\mathrm{F}_M(\eta_{Mi}))}\cdots\int_{-\infty}^{\tilde{y}_{1i}\Phi^{-1}(\mathrm{F}_1(\eta_{1i}))}\phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\Sigma})\prod_{\tilde{c}=1}^{M}dl_{\tilde{c},i}. \qquad \text{(C.1)}
\end{aligned}$$

Since $\tilde{y}_{mi}$ is either equal to -1 or 1, it follows that $\boldsymbol{\mathcal{B}}_i=\boldsymbol{\mathcal{B}}_i^{-1}$ and $|\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i|=|\boldsymbol{\Sigma}|$. In addition, the pdf of a multivariate normal vector $-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\Sigma}$ can be re-expressed as the pdf of a multivariate normal vector

$\boldsymbol{\varepsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i$, that is

$$\begin{aligned}\phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(-\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i)^\top(\boldsymbol{\Sigma})^{-1}(-\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i)\right\}\\ &= |2\pi(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{l}_i^\top(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)^{-1}\boldsymbol{l}_i\right\}\\ &= \phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i).\end{aligned}$$

Therefore, equation (C.1) can be written as

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{y}_i;\boldsymbol{\delta}) &= \int_{-\infty}^{\tilde{y}_{Mi}\Phi^{-1}(\mathrm{F}_M(\eta_{Mi}))}\cdots\int_{-\infty}^{\tilde{y}_{1i}\Phi^{-1}(\mathrm{F}_1(\eta_{1i}))}\phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)\prod_{\tilde{c}=1}^{M}dl_{\tilde{c},i}\\ &= \Phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\mathbf{H}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)\\ &= \Phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{W}}_i;\mathbf{0},\boldsymbol{\Upsilon}_i^*),\end{aligned}$$

where

$$\boldsymbol{\Upsilon}_i^* = \begin{pmatrix}1 & r_{12,i}^* & \dots & r_{1M,i}^*\\ r_{12,i}^* & 1 & \dots & r_{2M,i}^*\\ \vdots & \vdots & \ddots & \vdots\\ r_{1M,i}^* & r_{2M,i}^* & \dots & 1\end{pmatrix},$$

for $r_{zk,i}^* = \tanh(\vartheta_{zk}^*)(2y_{zi}-1)(2y_{ki}-1)$, $\forall z,k,i$. Note that the above derivation applies to all $\tilde{k}$s, thus the likelihood $\mathcal{L}_{i\tilde{k}}$ is equal to

$$\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i;\boldsymbol{\delta}) = \{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{\mathcal{W}}_i)_{\tilde{k}};\mathbf{0},(\boldsymbol{\Upsilon}_i)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}^*},$$

as required. □

C.2 Proof of Propositions 5.2.2 and 5.2.3

The first-order derivatives of the log-likelihood function for a multivariate probit model are obtained as follows. First, we express the multivariate normal cdf Φ_M in terms of multivariate integrals. Then, by using conditional density distributions, we decompose ϕ_M into a product of two normal probability density functions (pdfs) and re-express Φ_M based on that decomposition. In doing so we proceed with the calculation of the two derivatives, where the derivative of Φ_M with respect to $\boldsymbol{\beta}_m$ is mainly based on a decomposition formula, while the derivative of Φ_M with respect to ϑ_{zk} has been derived by applying an idea by Plackett (1954).

The multivariate integrals

$$\Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \;=\; \int_{-\infty}^{\mathcal{W}_{M,i}} \dots \int_{-\infty}^{\mathcal{W}_{1,i}} \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \tag{C.2}$$

can be written in a more convenient form by using the conditional distribution of the normal multivariate distribution. This can be achieved by partitioning both $\boldsymbol{l}_i$ and $\boldsymbol{\Upsilon}_i^*$ such that

$$\boldsymbol{l}_i \;=\; (\boldsymbol{l}_{1,i}, \boldsymbol{l}_{2,i})^\top,$$

and

$$\boldsymbol{\Upsilon}_i^* = \left(\begin{array}{c|c} \boldsymbol{\Theta}_{11,i}^* & \boldsymbol{\Theta}_{12,i}^* \\ \hline \boldsymbol{\Theta}_{21,i}^* & \boldsymbol{\Theta}_{22,i}^* \end{array}\right) = \left(\begin{array}{cccc|ccc} 1 & r_{12,i}^* & \dots & r_{1u,i}^* & r_{1,u+1,i}^* & \dots & r_{1,M,i}^* \\ r_{21,i}^* & 1 & \dots & r_{2u,i}^* & r_{2,u+1,i}^* & \dots & r_{2,M,i}^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ r_{u1,i}^* & r_{u2,i}^* & \dots & 1 & r_{u,u+1,i}^* & \dots & r_{u,M,i}^* \\ \hline r_{u+1,1,i}^* & r_{u+1,2,i}^* & \dots & r_{u+1,u,i}^* & 1 & \dots & r_{u+1,M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{M1,i}^* & r_{M2,i}^* & \dots & r_{Mu,i}^* & r_{M,u+1,i}^* & \dots & 1 \end{array}\right), \tag{C.3}$$

respectively, where $\boldsymbol{l}_{1,i} = (l_{1,i}, l_{2,i}, \dots, l_{u,i})^\top$, $\boldsymbol{l}_{2,i} = (l_{u+1,i}, l_{u+2,i}, \dots, l_{M,i})^\top$, $u =$

$1, \ldots, M-1$, $r^*_{zk,i} = \tanh(\vartheta^*_{zk})(2y_{zi}-1)(2y_{ki}-1)$, $\boldsymbol{\Theta}^*_{11,i}$ is a $u \times u$ matrix, $\boldsymbol{\Theta}^*_{22,i}$ is a $(M-u)\times(M-u)$ matrix and $\boldsymbol{\Theta}^*_{21,i} = \boldsymbol{\Theta}^{*\top}_{12,i}$. By using the chain rule for random variables and the partitioned vector $\boldsymbol{l}_i$ as well as the partitioned matrix $\boldsymbol{\Upsilon}^*_i$, the M-variate normal pdf $\phi_M(\boldsymbol{l}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}^*_i)$ can be expressed as the product of the conditional density function of $\boldsymbol{l}_{2,i}$ given $\boldsymbol{l}_{1,i}$ times the pdf of $\boldsymbol{l}_{1,i}$

$$\phi_M(\boldsymbol{l}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}^*_i) \;=\; \phi_{M-u}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i})\phi_u(\boldsymbol{l}_{1,i}), \tag{C.4}$$

where

$$\begin{aligned} \boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i} &\overset{iid}{\sim} \mathcal{N}_{M-u}(\mathbb{E}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}), \mathrm{Var}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i})) \\ &\overset{iid}{\sim} \mathcal{N}_{M-u}(\boldsymbol{\mu}_{\boldsymbol{l}_{2,i}} + \boldsymbol{\Theta}^*_{21,i}\boldsymbol{\Theta}^{*-1}_{11}(\boldsymbol{l}_{1,i} - \boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}), \boldsymbol{\Theta}^*_{22,i} - \boldsymbol{\Theta}^*_{21,i}\boldsymbol{\Theta}^{*-1}_{11,i}\boldsymbol{\Theta}^*_{12,i}), \end{aligned} \tag{C.5}$$

and

$$\begin{aligned} \boldsymbol{l}_{1,i} &\overset{iid}{\sim} \mathcal{N}_u(\mathbb{E}(\boldsymbol{l}_{1,i}), \mathrm{Var}(\boldsymbol{l}_{1,i})) \\ &\overset{iid}{\sim} \mathcal{N}_u(\boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}, \boldsymbol{\Theta}^*_{11,i}). \end{aligned} \tag{C.6}$$

$\boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}$ and $\boldsymbol{\mu}_{\boldsymbol{l}_{2,i}}$ stand for the mean of $\boldsymbol{l}_{1,i}$ and $\boldsymbol{l}_{2,i}$ respectively. It follows that the integrals (C.2) can be rewritten as

$$\begin{aligned} \Phi_M(\boldsymbol{\mathcal{W}}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}^*_i) \;=\; & \int_{-\infty}^{\mathcal{W}_{M,i}} \cdots \int_{-\infty}^{\mathcal{W}_{1,i}} \phi_{M-u}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}; \boldsymbol{M}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}}, \boldsymbol{\Theta}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}}) \\ & \phi_u(\boldsymbol{l}_{1,i}; \boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}, \boldsymbol{\Theta}^*_{11,i}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}, \end{aligned} \tag{C.7}$$

where $\boldsymbol{M}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}} = \boldsymbol{\mu}_{\boldsymbol{l}_{2,i}} + \boldsymbol{\Theta}^*_{21,i}\boldsymbol{\Theta}^{*-1}_{11}(\boldsymbol{l}_{1,i} - \boldsymbol{\mu}_{\boldsymbol{l}_{1,i}})$, $\boldsymbol{\Theta}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}} = \boldsymbol{\Theta}^*_{22,i} - \boldsymbol{\Theta}^*_{21,i}\boldsymbol{\Theta}^{*-1}_{11,i}\boldsymbol{\Theta}^*_{12,i}$, $\boldsymbol{\mu}_{\boldsymbol{l}_{1,i}} = \boldsymbol{\mu}_{\boldsymbol{l}_{2,i}} = \boldsymbol{0}$ and $\boldsymbol{\Theta}^*_{11,i}$ denotes the $u \times u$ sub-matrix of $\boldsymbol{\Upsilon}^*_i$.

C.2.1 Proof of Proposition 5.2.2

Proof. Consider formula (C.4) and let $u = 1$, such that

$$\phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \;=\; \phi_{M-1}(\boldsymbol{l}_{2,i}|l_{1,i})\phi(l_{1,i}).$$

By re-ordering matrix (C.3) we obtain

$$\boldsymbol{\Upsilon}_i^{*m} = \begin{pmatrix} \overbrace{\Theta_{11,i}^{*m}}^{1\times 1} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*m}}^{1\times(M-1)} \\ \underbrace{\boldsymbol{\Theta}_{21,i}^{*m}}_{(M-1)\times 1} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*m}}_{(M-1)\times(M-1)} \end{pmatrix},$$

$\forall m$, where $\Theta_{11,i}^{*m}$, $\boldsymbol{\Theta}_{12,i}^{*m}$, $\boldsymbol{\Theta}_{21,i}^{*m}$ and $\boldsymbol{\Theta}_{22,i}^{*m}$ are defined in Proposition 5.2.2. Then the multivariate normal cdf (C.7) becomes

$$\begin{aligned} \Phi_M(\boldsymbol{\mathcal{W}}_i; 0, \boldsymbol{\Upsilon}_i^{*m}) &= \int_{-\infty}^{\mathcal{W}_{M,i}} \dots \int_{-\infty}^{\mathcal{W}_{m,i}} \dots \int_{-\infty}^{\mathcal{W}_{1,i}} \phi_M(\boldsymbol{l}_i; 0, \boldsymbol{\Upsilon}_i^{*m}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \\ &= \int_{\bar{\mathbf{C}}_i} \phi_{M-1}(\boldsymbol{l}_{-m,i}|l_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m}) \phi(l_{m,i}; \mu_{l_{m,i}}, \Theta_{11,i}^{*m}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}, \end{aligned} \tag{C.8}$$

for $\bar{\mathbf{C}}_i = \bar{C}_{1i} \times \bar{C}_{2i} \times \ldots \times \bar{C}_{Mi}$, where $\bar{C}_{mi}$ is the interval $[\mathcal{W}_{m,i}, +\infty)$ if $y_{mi} = 1$ and the interval $(-\infty, \mathcal{W}_{m,i}]$ if $y_{mi} = 0$. Vector $\boldsymbol{l}_{-m,i} = (l_{1,i}, \ldots, l_{m-1,i}, l_{m+1,i}, \ldots, l_{M,i})^\top$, where $l_{m,i}$ refers to the m^{th} element of vector $\boldsymbol{l}_i$. $\boldsymbol{M}_i^{*m}$ and $\boldsymbol{\Theta}_i^{*m}$, respectively, denote the mean and the variance of $\boldsymbol{l}_{-m,i}|l_{m,i}$, while $\mu_{l_{m,i}}$ and $\Theta_{11,i}^{*m}$ denote the mean and variance of $l_{m,i}$. Applying the properties of the conditional multivariate normal distribution, it follows that $\mathbb{E}(l_{m,i}) = \mu_{l_{m,i}} = 0$ and $\mathbb{E}(\boldsymbol{l}_{-m,i}) = \boldsymbol{\mu}_{\boldsymbol{l}_{-m,i}} = \mathbf{0}$. (Note that $\mathbb{E}(\boldsymbol{l}_{-m,i}|l_{m,i}) \neq \mathbf{0}$.) Hence, the distribution of $\boldsymbol{l}_{-m,i}|l_{m,i}$ and $l_{m,i}$ is equal to

$$\begin{aligned} \boldsymbol{l}_{-m,i}|l_{m,i} &\overset{iid}{\sim} \mathcal{N}_{M-1}\left(\mathbb{E}(\boldsymbol{l}_{-m,i}|l_{m,i}), \mathrm{Var}(\boldsymbol{l}_{-m,i}|l_{m,i})\right) \\ &\overset{iid}{\sim} \mathcal{N}_{M-1}\left(\boldsymbol{\mu}_{\boldsymbol{l}_{-m,i}} + \boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11}^{*m}\right)^{-1}(l_{m,i} - \mu_{l_{m,i}}), \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11,i}^{*m}\right)^{-1}\boldsymbol{\Theta}_{12,i}^{*m}\right) \\ &\overset{iid}{\sim} \mathcal{N}_{M-1}\left(\boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11}^{*m}\right)^{-1} l_{m,i}, \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11,i}^{*m}\right)^{-1}\boldsymbol{\Theta}_{12,i}^{*m}\right), \end{aligned}$$

and

$$\begin{aligned} l_{m,i} &\overset{iid}{\sim} \mathcal{N}(\mathbb{E}(l_{m,i}), \mathrm{Var}(l_{m,i})) \\ &\overset{iid}{\sim} \mathcal{N}(\mu_{l_{m,i}}, \Theta^{*m}_{11,i}) \\ &\overset{iid}{\sim} \mathcal{N}(0, \Theta^{*m}_{11,i}), \end{aligned}$$

respectively, where the sub-matrix $\Theta^{*m}_{11,i}$ in this case is equal to 1, $\forall m, i$. It follows that (C.8) becomes

$$\begin{aligned} \Phi_M(\boldsymbol{\mathcal{W}}_i; 0, \boldsymbol{\Upsilon}^{*m}_i) &= \int_{\bar{\mathbf{C}}_i} \phi(l_{m,i}; 0, 1)\phi_{M-1}(\boldsymbol{l}_{-m,i}|l_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \\ &= \int_{-\infty}^{\mathcal{W}_{m,i}} \phi(l_{m,i}; 0, 1) \left\{ \int_{\bar{\mathbf{C}}_{i,-m}} \phi_{M-1}(\boldsymbol{l}_{-m,i}|l_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) d\boldsymbol{l}_{-m,i} \right\} dl_{m,i} \\ &= \int_{-\infty}^{\mathcal{W}_{m,i}} \phi(l_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{\mathcal{W}}_{-m,i}|l_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) dl_{m,i}, \end{aligned} \tag{C.9}$$

where $\boldsymbol{\mathcal{W}}_{-m,i} = (\mathcal{W}_{1,i}, \mathcal{W}_{2,i}, \ldots, \mathcal{W}_{m-1,i}, \mathcal{W}_{m+1,i}, \ldots, \mathcal{W}_{M,i})^\top$ and $\bar{\mathbf{C}}_{i,-m} \in \{\bar{\mathbf{C}}_i\} \setminus \bar{C}_{mi}$. According to the properties of the conditional multivariate normal distribution, it follows that the expected value of $\boldsymbol{\mathcal{W}}_{-m,i}|l_{m,i}$ is equal to $\boldsymbol{M}^{*m}_i = \boldsymbol{\Theta}^{*m}_{21,i} l_{m,i}$ while its variance-covariance matrix is expressed as $\boldsymbol{\Theta}^{*m}_i = \boldsymbol{\Theta}^{*m}_{22,i} - \boldsymbol{\Theta}^{*m}_{21,i}\boldsymbol{\Theta}^{*m}_{12,i}$. By using the chain rule as well as the fundamental theorem of calculus, it follows that the derivative of (C.9) with respect to $\boldsymbol{\beta}_m$ is equal to

$$\begin{aligned} \frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; 0, \boldsymbol{\Upsilon}^{*m}_i)}{\partial \boldsymbol{\beta}_m} &= \frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; 0, \boldsymbol{\Theta}^{*m}_i)}{\partial \mathcal{W}_{m,i}} \frac{\partial \mathcal{W}_{m,i}}{\partial \boldsymbol{\beta}_m} \\ &= \frac{\partial}{\partial \mathcal{W}_{m,i}} \left\{ \int_{-\infty}^{\mathcal{W}_{m,i}} \phi(l_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{\mathcal{W}}_{-m,i}|l_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) dl_{m,i} \right\} \times \\ &\quad \left(\frac{\partial \mathcal{W}_{m,i}}{\partial \boldsymbol{\beta}_m} \right) \\ &= \phi(\mathcal{W}_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{\mathcal{W}}_{-m,i}|\mathcal{W}_{m,i}; \boldsymbol{M}^{*m}_i, \boldsymbol{\Theta}^{*m}_i) \left(\frac{\partial \mathcal{W}_{m,i}}{\partial \boldsymbol{\beta}_m} \right). \end{aligned}$$

Since $\mathcal{W}_{m,i} = (2y_{mi} - 1)\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))$, then the derivative of $\mathcal{W}_{m,i}$ with respect to $\boldsymbol{\beta}_m$ can be expressed as

$$\frac{\partial \mathcal{W}_{m,i}}{\partial \boldsymbol{\beta}_m} = \frac{\partial \mathcal{W}_{m,i}}{\partial \mathrm{F}_m(\eta_{mi})} \frac{\partial \mathrm{F}_m(\eta_{mi})}{\partial \eta_{mi}} \frac{\partial \eta_{mi}}{\partial \boldsymbol{\beta}_m},$$

where $\partial \mathcal{W}_{m,i}/\partial \mathrm{F}_m(\eta_{mi}) = (2y_{mi}-1)/\phi(\Phi^{-1}(\mathrm{F}_m(\eta_{mi})))$ (based on the inverse function theorem), $\partial \mathrm{F}_m(\eta_{mi})/\partial \eta_{mi} = f_m(\eta_{mi})$ and $\partial \eta_{mi}/\partial \boldsymbol{\beta}_m = \mathbf{x}_{mi}^\top$. Therefore, we have that

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \boldsymbol{\beta}_m} &= \phi(\mathcal{W}_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{\mathcal{W}}_{-m,i} | \mathcal{W}_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m}) \frac{f_m(\eta_{mi})}{\phi\left(\Phi^{-1}(\mathrm{F}_m(\eta_{mi}))\right)} \times \\ &\quad (2y_{mi} - 1)\mathbf{x}_{mi}^\top\end{aligned}$$

for $\boldsymbol{M}_i^{*m} = \boldsymbol{\Theta}_{21,i}^{*m}\mathcal{W}_{m,i}$ and $\boldsymbol{\Theta}_i^{*m} = \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\boldsymbol{\Theta}_{12,i}^{*m}$, as required. □

C.2.2 Proof of Proposition 5.2.3

Proof. If we differentiate equation (C.2) with respect to the correlation coefficient ϑ_{zk}^*, we get the following

$$\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_{\boldsymbol{i}}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk}^*} = \frac{\partial}{\partial \vartheta_{zk}^*} \left\{ \int_{-\infty}^{\mathcal{W}_{M,i}} \cdots \int_{-\infty}^{\mathcal{W}_{1,i}} \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \right\},$$

and by using the chain rule

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_{\boldsymbol{i}}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk,i}^*} &= \frac{\partial}{\partial r_{zk}^*} \left\{ \int_{\bar{\mathbf{C}}_i} \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*} \\ &= \left\{ \int_{\bar{\mathbf{C}}_i} \frac{\partial \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial r_{zk,i}^*} \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \qquad \text{(C.10)}\end{aligned}$$

where $r_{zk,i}^*$ and region $\bar{\mathbf{C}}_i$ have been defined previously. By using the following differential equation derived by Plackett (1954)

$$\frac{\partial \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial r_{zk,i}^*} = \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial l_{z,i} \partial l_{k,i}},$$

equation (C.10) becomes

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk,i}^*} &= \left\{ \int_{\bar{\mathbf{C}}_i} \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial l_{z,i} \partial l_{k,i}} \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*} \\ &= \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \left[\int_{\bar{\mathbf{C}}_{zk,i}} \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial l_{z,i} \partial l_{k,i}} d\boldsymbol{l}_{zk,i} \right] d\boldsymbol{l}_{-zk,i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \end{aligned} \tag{C.11}$$

where $\bar{\mathbf{C}}_{-zk,i} \in \bar{\mathbf{C}}_i \setminus \{\bar{C}_{zi}, \bar{C}_{ki}\}$, $\bar{\mathbf{C}}_{zk,i} = \bar{C}_{zi} \times \bar{C}_{ki}$, $\boldsymbol{l}_{zk,i} = (l_{z,i}, l_{k,i})^\top$ and $\boldsymbol{l}_{-zk,i} = (l_{1,i}, \ldots, l_{k-1,i}, l_{k+1,i}, \ldots, l_{z-1,i}, l_{z+1,i}, l_{M,i})^\top$. According to the fundamental theorem of calculus, the integral inside the brackets is equal to

$$\begin{aligned}\int_{\bar{\mathbf{C}}_{zk,i}} \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial l_{z,i} \partial l_{k,i}} d\boldsymbol{l}_{zk,i} &= \frac{\partial^2}{\partial l_{z,i} \partial l_{k,i}} \left\{ \int_{\bar{\mathbf{C}}_{zk,i}} \phi_M(\boldsymbol{\mathcal{W}_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) dl_{k,i} dl_{z,i} \right\} \\ &= \frac{\partial^2}{\partial l_{z,i} \partial l_{k,i}} \left\{ \int_{\bar{C}_{zi}} \int_{\bar{C}_{ki}} \phi_M(\boldsymbol{\mathcal{W}_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) dl_{k,i} dl_{z,i} \right\} \\ &= \phi_M(l_{1,i}, \ldots, l_{z-1,i}, \mathcal{W}_{z,i}, l_{z+1,i}, \ldots, l_{k-1,i}, \mathcal{W}_{k,i}, l_{k+1,i}, \ldots, \\ &\quad l_{M,i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*). \end{aligned}$$

Therefore, (C.11) can be expressed as

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk,i}^*} &= \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_M(l_{1,i}, \ldots, l_{z-1,i}, \mathcal{W}_{z,i}, l_{z+1,i}, \ldots, l_{k-1,i}, \mathcal{W}_{k,i}, l_{k+1,i}, \ldots, \right. \\ &\quad \left. l_{M,i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) d\boldsymbol{l}_{-zk,i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}. \end{aligned}$$

The last expression can be written in a more convenient form by using the conditional distributions of the normal multivariate distribution. This can be done by imposing the special case $u = 2$ in equation (C.4), that is

$$\phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk}) = \phi_{M-2}(\boldsymbol{l}_{2,i} | \boldsymbol{l}_{1,i}) \phi_2(\boldsymbol{l}_{1,i}), \tag{C.12}$$

where $\boldsymbol{l}_{2,i}$ and $\boldsymbol{l}_{1,i}$ correspond to $\boldsymbol{l}_{-zk,i}$ and $\boldsymbol{\mathcal{W}}_{zk,i}$, respectively, with $\boldsymbol{\mathcal{W}}_{zk,i} = (\mathcal{W}_{z,i}, \mathcal{W}_{k,i})^\top$. Re-ordering matrix (C.3), we obtain

$$\boldsymbol{\Upsilon}_i^{*zk} = \left(\begin{array}{c:c} \overbrace{\boldsymbol{\Theta}_{11,i}^{*zk}}^{2\times 2} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*zk}}^{2\times(M-2)} \\ \hdashline \underbrace{\boldsymbol{\Theta}_{21,i}^{*zk}}_{(M-2)\times 2} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*zk}}_{(M-2)\times(M-2)} \end{array} \right), \tag{C.13}$$

$\forall z = 1, \ldots, M-1, k = z+1, \ldots M$, where the sub-matrices $\boldsymbol{\Theta}_{11,i}^{*zk}$, $\boldsymbol{\Theta}_{12,i}^{*zk}$, $\boldsymbol{\Theta}_{21,i}^{*zk}$ and $\boldsymbol{\Theta}_{22,i}^{*zk}$ are defined in Proposition 5.2.3. By using both (C.12) and (C.13), we have that

$$\begin{aligned} \frac{\partial \Phi_M(\boldsymbol{\mathcal{W}_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk})}{\partial \vartheta_{zk,i}^*} &= \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_{M-2}(\boldsymbol{l}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk}) \phi_2(\boldsymbol{\mathcal{W}}_{zk,i}; \boldsymbol{\mu}_{\boldsymbol{\mathcal{W}}_{zk,i}}, \right. \\ & \left. \boldsymbol{\Theta}_{11,i}^{*zk}) d\boldsymbol{l}_{-zk,i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \end{aligned} \tag{C.14}$$

where $\boldsymbol{M}_i^{*-zk}$ and $\boldsymbol{\Theta}_i^{*-zk}$ refer to the mean and variance-covariance matrix of $\boldsymbol{l}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i}$, while $\boldsymbol{\mu}_{\boldsymbol{\mathcal{W}}_{zk,i}}$ and $\boldsymbol{\Theta}_{11,i}^{*zk}$ denote the mean and variance-covariance of $\boldsymbol{\mathcal{W}}_{zk,i}$. By using the properties of the conditional multivariate normal distribution, it follows that $\mathbb{E}(\boldsymbol{\mathcal{W}}_{zk,i}) = \boldsymbol{\mu}_{\boldsymbol{\mathcal{W}}_{zk,i}} = 0$ and $\mathbb{E}(\boldsymbol{l}_{-zk,i}) = \boldsymbol{\mu}_{\boldsymbol{l}_{-zk,i}} = 0$. (Note that $\mathbb{E}(\boldsymbol{l}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i}) \neq 0$.) Hence, according to (C.5) and (C.6)

$$\begin{aligned} \boldsymbol{l}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i} &\overset{iid}{\sim} \mathcal{N}_M(\mathbb{E}(\boldsymbol{l}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i}), \mathrm{Var}(\boldsymbol{l}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i})) \\ &\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{\mu}_{\boldsymbol{l}_{-zk,i}} + \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11}^{*zk}\right)^{-1} (\boldsymbol{\mathcal{W}}_{zk,i} - \boldsymbol{\mu}_{\boldsymbol{l}_{zk,i}}), \boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{\Theta}_{12,i}^{*zk}) \\ &\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{\mathcal{W}}_{zk,i}, \boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{zk}\right)^{-1} \boldsymbol{\Theta}_{12,i}^{*zk}), \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{l}_{zk,i} &\overset{iid}{\sim} \mathcal{N}_M(\mathbb{E}(\boldsymbol{\mathcal{W}}_{zk,i}), \mathrm{Var}(\boldsymbol{\mathcal{W}}_{zk,i})) \\ &\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{\mathcal{W}}_{zk,i}, \boldsymbol{\Theta}_{11,i}^{*zk}) \\ &\overset{iid}{\sim} \mathcal{N}_M(\mathbf{0}, \boldsymbol{\Theta}_{11,i}^{*zk}), \end{aligned}$$

where the sub-matrix $\boldsymbol{\Theta}_{11,i}^{*zk}$ is a 2×2 diagonal matrix with unit variances and correlations equal to $r_{zk,i}^*$. For simplicity, we will denote this matrix as $\boldsymbol{\Theta}_i^{*zk}$. Consequently,

equation (C.14) can be expressed as

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk})}{\partial \vartheta_{zk}^*} &= \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_2(\boldsymbol{\mathcal{W}}_{zk,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*zk}) \phi_{M-2}(\boldsymbol{l}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk}) \right. \\ &\quad \left. d\boldsymbol{l}_{-zk,i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}.\end{aligned}$$

Because only the term $\phi_{M-2}(\boldsymbol{l}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk})$ depends on $\boldsymbol{l}_{-zk,i}$, it follows that

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk})}{\partial \vartheta_{zk}^*} &= \left\{ \phi_2(\boldsymbol{\mathcal{W}}_{zk,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*zk}) \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_{M-2}(\boldsymbol{l}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk}) \right. \\ &\quad \left. d\boldsymbol{l}_{-zk,i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*} \\ &= \left\{ \phi_2(\boldsymbol{\mathcal{W}}_{zk,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*zk}) \Phi_{M-2}(\boldsymbol{\mathcal{W}}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk}) \right\} \times \\ &\quad \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \end{aligned} \tag{C.15}$$

where the last term comes from basic results of the multivariate normal distribution function. In addition, $\boldsymbol{\mathcal{W}}_{-zk,i} = (\mathcal{W}_{1,i}, \mathcal{W}_{2,i}, \ldots, \mathcal{W}_{z-1,i}, \mathcal{W}_{z+1,i}, \ldots, \mathcal{W}_{k-1,i}, \mathcal{W}_{k+1,i}, \ldots, \mathcal{W}_{M,i})^\top$, while the partial derivative $\partial r_{zk,i}^* / \partial \vartheta_{zk}^*$ is equal to

$$\begin{aligned}\frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*} &= \frac{\partial}{\partial \vartheta_{zk}^*} \left\{ \tanh(\vartheta_{zk}^*)(2y_{z,i} - 1)(2y_{k,i} - 1) \right\} \\ &= (2y_{z,i} - 1)(2y_{k,i} - 1) \frac{\partial}{\partial \vartheta_{zk}^*} \left\{ \tanh(\vartheta_{zk}^*) \right\} \\ &= (2y_{z,i} - 1)(2y_{k,i} - 1) \operatorname{sech}^2(\vartheta_{zk}^*) \\ &= (2y_{z,i} - 1)(2y_{k,i} - 1) \frac{1}{\cosh^2(\vartheta_{zk}^*)} \\ &= (2y_{z,i} - 1)(2y_{k,i} - 1) \frac{1}{\left(\frac{\exp(\vartheta_{zk}^*) + \exp(-\vartheta_{zk}^*)}{2} \right)^2} \\ &= (2y_{z,i} - 1)(2y_{k,i} - 1) \frac{4}{\left\{ \exp(\vartheta_{zk}^*) + \exp(-\vartheta_{zk}^*) \right\}^2} \\ &= (2y_{z,i} - 1)(2y_{k,i} - 1) \frac{4e^{2\vartheta_{zk}^*}}{\left\{ e^{2\vartheta_{zk}^*} + 1 \right\}^2},\end{aligned}$$

by using definitions and properties of the hyperbolic functions. Therefore, (C.15) becomes

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk})}{\partial \vartheta_{*zk}} &= \phi_2(\boldsymbol{\mathcal{W}}_{zk,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*zk}) \Phi_{M-2}(\boldsymbol{\mathcal{W}}_{-zk,i} | \boldsymbol{\mathcal{W}}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk}) \\ &\quad (2y_{z,i}-1)(2y_{k,i}-1) \frac{4e^{2\vartheta_{zk}^*}}{\left\{e^{2\vartheta_{zk}^*}+1\right\}^2},\end{aligned}$$

for $\boldsymbol{\mathcal{W}}_{zk,i} = (\mathcal{W}_{z,i}, \mathcal{W}_{k,i})^\top$, $\boldsymbol{\mathcal{W}}_{-zk,i} = (\mathcal{W}_{1,i}, \mathcal{W}_{2,i}, \ldots, \mathcal{W}_{z-1,i}, \mathcal{W}_{z+1,i}, \ldots, \mathcal{W}_{k-1,i}, \mathcal{W}_{k+1,i}, \ldots \mathcal{W}_{M,i})^\top$, $\boldsymbol{\Theta}_i^{*zk} = \boldsymbol{\Theta}_{11,i}^{*zk}$, $\boldsymbol{M}_i^{*-zk} = \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{\mathcal{W}}_{zk,i}$ and $\boldsymbol{\Theta}_i^{*-zk} = \boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk} \left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1} \boldsymbol{\Theta}_{12,i}^{*zk}$, as required. □

Appendix D

Complements to Chapter 6

D.1 Proof of Lemma 6.3.1

Proof. For convenience we ignore index $\tilde{k}$ and term $\mathcal{Y}_{i\tilde{k}}$. By definition,

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{y}_i;\boldsymbol{\delta}) &= \mathbb{P}(-\tilde{y}_{1i}y^*_{1i}\leq 0,\ldots,-\tilde{y}_{Mi}y^*_{Mi}\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}(\eta_{1i}+\varepsilon_{1i})\leq 0,\ldots,-\tilde{y}_{Mi}(\eta_{Mi}+\varepsilon_{Mi})\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}(\Phi^{-1}(F_1(\eta_{1i}))+\varepsilon_{1i})\leq 0,\ldots,-\tilde{y}_{Mi}(\Phi^{-1}(F_M(\eta_{Mi}))+\varepsilon_{Mi})\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}\Phi^{-1}(F_1(\eta_{1i}))-\tilde{y}_{1i}\varepsilon_{1i}\leq 0,\ldots,-\tilde{y}_{Mi}\Phi^{-1}(F_M(\eta_{Mi}))-\tilde{y}_{Mi}\varepsilon_{Mi}\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}\varepsilon_{1i}\leq \tilde{y}_{1i}\Phi^{-1}(F_1(\eta_{1i})),\ldots,-\tilde{y}_{Mi}\varepsilon_{Mi}\leq \tilde{y}_{Mi}\Phi^{-1}(F_M(\eta_{Mi})))\\
&= \Phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\mathbf{H}_i;\mathbf{0},\boldsymbol{\Sigma}_i)\\
&= \int_{-\infty}^{\tilde{y}_{Mi}\Phi^{-1}(F_M(\eta_{Mi}))}\cdots\int_{-\infty}^{\tilde{y}_{1i}\Phi^{-1}(F_1(\eta_{1i}))}\phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\Sigma}_i)\prod_{\tilde{c}=1}^{M}dl_{\tilde{c},i}. \qquad \text{(D.1)}
\end{aligned}$$

where $\boldsymbol{\varepsilon}_i=(\varepsilon_{1i},\ldots,\varepsilon_{Mi})^\top$ corresponds ot the error term of the M-variate Gaussian binary model. Since $\tilde{y}_{mi}$ is either equal to -1 or 1, it follows that $\boldsymbol{\mathcal{B}}_i=\boldsymbol{\mathcal{B}}_i^{-1}$ and $|\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i|=|\boldsymbol{\Sigma}_i|$. In addition, the pdf of a multivariate normal vector $-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\Sigma}_i$ can be re-expressed as the pdf of a multivariate

normal vector $\boldsymbol{\varepsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i$, that is

$$\begin{aligned}
\phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\Sigma}_i) &= |2\pi\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(-\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i)^\top(\boldsymbol{\Sigma}_i)^{-1}(-\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i)\right\}\\
&= |2\pi(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i)|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{l}_i^\top(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i)^{-1}\boldsymbol{l}_i\right\}\\
&= \phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i)\\
&= \phi_M(\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i).
\end{aligned}$$

where in the last expression we ignored index $\boldsymbol{\varepsilon}_i$ for convenience. Therefore, equation (D.1) can be written as

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{y}_i;\boldsymbol{\delta}) &= \int_{-\infty}^{\tilde{y}_{Mi}\Phi^{-1}(F_M(\eta_{Mi}))}\cdots\int_{-\infty}^{\tilde{y}_{1i}\Phi^{-1}(F_1(\eta_{1i}))}\phi_M(\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i)\prod_{\tilde{c}=1}^{M}dl_{\tilde{c},i}\\
&= \Phi_M(\boldsymbol{\mathcal{B}}_i\mathbf{H}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}_i\boldsymbol{\mathcal{B}}_i)\\
&= \Phi_M(\boldsymbol{\mathcal{W}}_i;\mathbf{0},\boldsymbol{\Upsilon}_i^*),
\end{aligned}$$

where

$$\boldsymbol{\Upsilon}_i^* = \begin{pmatrix} 1 & r_{12,i}^* & \dots & r_{1M,i}^* \\ r_{12,i}^* & 1 & \dots & r_{2M,i}^* \\ \vdots & \vdots & \ddots & \vdots \\ r_{1M,i}^* & r_{2M,i}^* & \dots & 1 \end{pmatrix},$$

for $r_{zk,i}^* = \mathrm{t}_{zz,i}\mathrm{t}_{kk,i}\bar{\sigma}_{zk,i}^*(2y_{zi}-1)(2y_{ki}-1)$, where $\mathrm{t}_{zz,i}$ and $\mathrm{t}_{kk,i}$ denote the $(z,z)^{th}$ and $(k,k)^{th}$ element of matrix $\mathbf{T}_i$, respectively, $\forall z,k,i$. Note that the above derivation applies to all $\tilde{k}$s, thus the likelihood $\mathcal{L}_{i\tilde{k}}$ is equal to

$$\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i;\boldsymbol{\delta}) = \{\Phi_M((\boldsymbol{\mathcal{W}}_i)_{\tilde{k}};\mathbf{0},(\boldsymbol{\Upsilon}_i)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}^*},$$

as required. □

D.2 Matrices $\mathbf{T}_i$ and $\bar{\mathbf{\Sigma}}_i^*$

Matrix $\mathbf{T}_i$ is equal to

$$\mathbf{T}_i = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & \left(1+\eta_{12,i}^2\right)^{-1/2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \left(1+\eta_{12,i}^2+\eta_{13,i}^2+\ldots+\eta_{M-1,M,i}^2\right)^{-1/2} \end{pmatrix},$$

while matrix $\bar{\mathbf{\Sigma}}_i^*$ is defined as

$$\bar{\mathbf{\Sigma}}^* = \begin{pmatrix} 1 & \eta_{12,i} & \dots & \eta_{1M,i} \\ \eta_{12,i} & 1+\eta_{12,i} & \dots & \eta_{1M,i}\eta_{12,i}+\eta_{2M,i} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{1M,i} & \eta_{1M,i}\eta_{12,i}+\eta_{2M,i} & \dots & 1+\eta_{1M,i}^2+\ldots+\eta_{M-1,M,i}^2 \end{pmatrix}.$$

D.3 Proof of Proposition 6.3.3

Proof. Since the correlation parameter $r^*_{zk,i}$ in matrix $\boldsymbol{\Upsilon}^*_i$ is defined as $r^*_{zk,i} = \mathrm{t}_{zz}\mathrm{t}_{kk}\bar{\sigma}^*_{zk}(2y_{zi}-1)(2y_{ki}-1)$, it follows that $r^*_{zk,i}$ may not only depend on $\eta_{zk,i}$ but may also depend on $\boldsymbol{\eta}_{-zk,i}$, where $\boldsymbol{\eta}_{-zk,i} \in \tilde{\boldsymbol{\eta}}_i \setminus \eta_{zk,i}$, for $\tilde{\boldsymbol{\eta}}_i = (\eta_{12,i}, \ldots, \eta_{M-1,M,i})^\top$. In order to account for these dependencies, we employ the multivariate chain rule

$$\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)}{\partial \boldsymbol{\beta}_{zk}} = \frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)}{\partial \boldsymbol{r}^*_i} \frac{\partial \boldsymbol{r}^*_i}{\partial \eta_{zk,i}} \frac{\partial \eta_{zk,i}}{\partial \boldsymbol{\beta}_{zk}}, \tag{D.2}$$

where $\boldsymbol{r}^*_i = \left(r^*_{12,i}, \ldots, r^*_{M-1,M,i}\right)^\top$ and $\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)/\partial \boldsymbol{r}^*_i = (\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)/\partial r^*_{12,i}, \ldots, \partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)/\partial r^*_{M-1,M,i}\big)$, $\forall i$. Based on result (C.15) in Appendix C.2.2, we have that

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}^*_i)}{\partial \boldsymbol{r}^*_i} = \big(&\phi_2(\boldsymbol{\mathcal{W}}_{12,i}; \mathbf{0}, \boldsymbol{\Theta}^{*12}_i)\Phi_{M-2}(\boldsymbol{\mathcal{W}}_{-12,i}|\boldsymbol{\mathcal{W}}_{12,i}; \boldsymbol{M}^{*-12}_i, \boldsymbol{\Theta}^{*-12}_i), \ldots, \\ &\phi_2(\boldsymbol{\mathcal{W}}_{M-1,M,i}; \mathbf{0}, \boldsymbol{\Theta}^{*M-1,M}_i)\Phi_{M-2}(\boldsymbol{\mathcal{W}}_{-M-1,M,i}|\boldsymbol{\mathcal{W}}_{M-1,M,i}; \\ &\boldsymbol{M}^{*-M-1,M}_i, \boldsymbol{\Theta}^{*-M-1,M}_i)\Big).\end{aligned} \tag{D.3}$$

where the notation is the same as in the previous chapters. The term $\partial \boldsymbol{r}^*_i/\partial \vartheta^*_{zk,i}$ can be obtained via the Jacobian matrix

$$\mathbf{J} = \frac{\partial \boldsymbol{r}^*_i}{\partial \tilde{\boldsymbol{\eta}}_i} = \left(\frac{\partial \boldsymbol{r}^*_i}{\partial \eta_{12,i}}; \ldots; \frac{\partial \boldsymbol{r}^*_i}{\partial \eta_{M-1,M,i}}\right)^\top = \begin{pmatrix} \frac{\partial r^*_{12,i}}{\partial \eta_{12,i}} & \cdots & \frac{\partial r^*_{12,i}}{\partial \eta_{M-1,M,i}} \\ \vdots & \ddots & \vdots \\ \frac{\partial r^*_{M-1,M,i}}{\partial \eta_{12,i}} & \cdots & \frac{\partial r^*_{M-1,M,i}}{\partial \eta_{M-1,M,i}} \end{pmatrix}, \tag{D.4}$$

as

$$\frac{\partial \boldsymbol{r}^*_i}{\partial \eta_{zk,i}} = \left(\frac{\partial r^*_{12,i}}{\partial \eta_{zk,i}}, \ldots, \frac{\partial r^*_{M-1,M,i}}{\partial \eta_{zk,i}}\right)^\top,$$

while $\partial \eta_{zk,i}/\partial \boldsymbol{\beta}_{zk}$ is equal to

$$\frac{\partial \eta_{zk,i}}{\partial \boldsymbol{\beta}_{zk}} = \mathbf{x}^\top_{zk,i}, \tag{D.5}$$

since $\eta_{zk,i} = \mathbf{x}_{zk,i}^{\top}\boldsymbol{\beta}_{zk}$. Based on the results (D.3), (D.4) and (D.5), it follows that (D.2) becomes

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{\mathcal{W}}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \boldsymbol{\beta}_{zk}} &= \big(\phi_2(\boldsymbol{\mathcal{W}}_{12,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*12})\Phi_{M-2}(\boldsymbol{\mathcal{W}}_{-12,i}|\boldsymbol{\mathcal{W}}_{12,i}; \boldsymbol{M}_i^{*-12}, \boldsymbol{\Theta}_i^{*-12}), \ldots, \\ &\quad \phi_2(\boldsymbol{\mathcal{W}}_{M-1,M,i}; \mathbf{0}, \boldsymbol{\Theta}_i^{*M-1,M})\Phi_{M-2}(\boldsymbol{\mathcal{W}}_{-M-1,M,i}|\boldsymbol{\mathcal{W}}_{M-1,M,i}; \\ &\quad \boldsymbol{M}_i^{*-M-1,M}, \boldsymbol{\Theta}_i^{*-M-1,M})\big)\left(\frac{\partial r_{12,i}^*}{\partial \eta_{zk,i}}, \ldots, \frac{\partial r_{M-1,M,i}^*}{\partial \eta_{zk,i}}\right)^{\top}\mathbf{x}_{zk,i}^{\top},\end{aligned}$$

for $\boldsymbol{\mathcal{W}}_{zk,i} = (\mathcal{W}_{z,i}, \mathcal{W}_{k,i})^{\top}$, $\boldsymbol{\mathcal{W}}_{-zk,i} = (\mathcal{W}_{1,i}, \mathcal{W}_{2,i}, \ldots, \mathcal{W}_{z-1,i}, \mathcal{W}_{z+1,i}, \ldots, \mathcal{W}_{k-1,i}, \mathcal{W}_{k+1,i}, \ldots \mathcal{W}_{M,i})^{\top}$, $\boldsymbol{\Theta}_i^{*zk} = \boldsymbol{\Theta}_{11,i}^{*zk}$, $\boldsymbol{M}_i^{*-zk} = \boldsymbol{\Theta}_{21,i}^{*zk}\left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1}\boldsymbol{\mathcal{W}}_{zk,i}$ and $\boldsymbol{\Theta}_i^{*-zk} = \boldsymbol{\Theta}_{22,i}^{*zk} - \boldsymbol{\Theta}_{21,i}^{*zk}\left(\boldsymbol{\Theta}_{11,i}^{*zk}\right)^{-1}\boldsymbol{\Theta}_{12,i}^{*zk}$, $\forall z, k$, as required. □

D.4 Data generating process used in the simulation study

D.4.1 DGP4

DGP4 was based on the following system of three equations

$$\begin{aligned} y_{1i}^{*} &= -0.55 + 0.90v_{1i} + s_1(z_{1i}) + \varepsilon_{1i} \\ y_{2i}^{*} &= -0.45 - 1.40v_{1i} + s_2(z_{1i}) + \varepsilon_{2i} \\ y_{3i}^{*} &= -0.60 + 2.00v_{1i} + s_3(z_{1i}) + \varepsilon_{3i}, \end{aligned}$$

while the additive predictors $\eta_{12,i}$, $\eta_{13,i}$ and $\eta_{23,i}$ were defined as

$$\begin{aligned} \eta_{12,i} &= 0.20 + 0.70v_{1i} + s_{\vartheta_{12}}(z_{1i}) \\ \eta_{13,i} &= -0.80 - 0.15v_{1i} + s_{\vartheta_{13}}(z_{1i}) \\ \eta_{23,i} &= -0.50 + 0.90v_{1i} + s_{\vartheta_{23}}(z_{1i}), \end{aligned}$$

where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, v_{1i} is a binary regressor, and s_m and $s_{\vartheta_{zk}}$ correspond to the smooth components which were represented using penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives, $\forall m, z, k, i$. The smooth functions are given by $s_1(z_{1i}) = 0.5\cos(2\pi z_{1i})$, $s_2(z_{1i}) = z_{1i} + \exp\{-30(z_{1i} - 0.5)^2\}$, $s_3(z_{1i}) = -0.5(z_{1i} + 3z_{1i}^3)$, $s_{\vartheta_{12}}(z_{1i}) = -2(0.25 \exp(z_{1i}) - z_{1i}^3)$, $s_{\vartheta_{13}}(z_{1i}) = z_{1i}^{5/2} + \exp(-3(z_{1i} - 0.45)^2)$ and $s_{\vartheta_{23}}(z_{1i}) = -2z_{1i}$. Sample sizes were set to 1000 and 3000 and the number of replicates to 1000.

R code for DGP4

```
library(GJRM)

n     <- 1000 # then n <- 3000
```

```
n.rep <- 1000

SigmaCov <- matrix(0.5, 2, 2);diag(SigmaCov) <- 1
f1 <- function(x) 0.5*cos(pi*2*x)
f2 <- function(x) x+exp(-30*(x-0.5)^2)
f3 <- function(x) -0.5*(x+3*x^3)
f4 <- function(x) (-2 * (0.25 * exp(x) - x^3))
f5 <- function(x) ((x^(5/2) + exp(-3*(x-0.45)^2)))
f6 <- function(x) (-2*x)

xt <- seq(0.0000001, 0.9999999, length.out = 200)
dt <- data.frame(z = xt)

f1t <- f1(xt) - mean(f1(xt))
f2t <- f2(xt) - mean(f2(xt))
f3t <- f3(xt) - mean(f3(xt))
f4t <- f4(xt) - mean(f4(xt))
f5t <- f5(xt) - mean(f5(xt))
f6t <- f6(xt) - mean(f6(xt))



gamma11 <- gamma12 <- gamma21 <- gamma22 <- gamma31 <- NULL
gamma32 <- theta121 <- theta122 <- theta131 <- NULL
theta132 <- theta231 <- theta232 <-  NULL
F1 <- F2 <- F3 <- F4 <- F5 <- F6 <- matrix(NA, 200, n.rep)

for(i in 1:n.rep){
  set.seed(i)
  data.gen <- function(SigmaCov, f1, f2, f3, f4, f5, f6){
  Mvdcov   <- mvdc(copula = normalCopula(0.5), margins = c("logis",
```

```
"norm"), paramMargins = list( list(location = 0, scale = 1),
 list(mean = 0, sd = 1)) )
cov <- rMvdc(1, Mvdcov)
v1 <- round(mm(plogis(cov[, 1])))
z1 <- mm(pnorm(cov[, 2]))
eta_theta12 <-    0.2 + 0.70*v1 + f4(z1)
eta_theta13 <-  - 0.8 - 0.15*v1 + f5(z1)
eta_theta23 <-  - 0.5 + 1.00*v1 + f6(z1)
Sigma.er <- matrix( c( 1, eta_theta12, eta_theta13,
                       eta_theta12,  1, eta_theta23,
                       eta_theta13, eta_theta23, 1 ), 3 , 3)
# Check if Sigma.er is positive-definite:
eS <- eigen(Sigma.er)
check.eigen <- any(eS$values < 0)
if (check.eigen == TRUE) {
  C <- matrix(c(1, 0, 0, eta_theta12, 1, 0, eta_theta13, eta_theta23,
  1), nrow = 3, byrow = TRUE)
  Sigma.star <- C %*% t(C)
  T <- diag(1/sqrt(diag(Sigma.star)))
  Sigma.er <- T %*% Sigma.star %*% T
} else Sigma.er <- Sigma.er
eta_theta12 <- Sigma.er[1, 2]; eta_theta13 <- Sigma.er[1, 3];
eta_theta23 <- Sigma.er[2, 3]

norm.copu  <- normalCopula( c(eta_theta12, eta_theta13, eta_theta23),
dim = 3, dispstr = "un")
Mvdu <- mvdc(copula = norm.copu, margins = c("logis", "gumbel", "norm"),
paramMargins = list( list(location = 0, scale = 1),
                                        list(location = 0, scale = 1),
                                        list(mean = 0, sd = 1)) )
```

```
  u <- rMvdc(1, Mvdu)

   y1 <- ifelse(-0.55 + 0.9*v1  + f1(z1) + u[,1]  > 0, 1, 0)
  y2 <- ifelse(-0.45 - 1.4*v1  + f2(z1) + u[,2]  > 0, 1, 0)
  y3 <- ifelse(-0.60  + 2.0*v1  + f3(z1) + u[,3]  > 0, 1, 0)
  c(y1, y2, y3, v1,z1)
}
  dataSim <- matrix(NA, nrow = n, ncol = 5)
  for(j in 1:n) dataSim[j,] <- data.gen(SigmaCov, f1, f2, f3, f4, f5, f6)
 dataSim <- data.frame(y1, y2, y3, v1, z1)
  s=mgcv::s
  f.l <- list(y1 ~ v1 + s(z1),
              y2 ~ v1 + s(z1),
              y3 ~ v1 + s(z1),
              ~ v1 + s(z1),
              ~ v1 + s(z1),
              ~ v1 + s(z1))

  out  <- try( SemiParTRIV(f.l, margins = c("logit", "cloglog", "probit"),
  data = dataSim, Chol = TRUE) )

  X1 <- PredictMat( out$gam1$smooth[[1]], dt )
  X2 <- PredictMat( out$gam2$smooth[[1]], dt )
  X3 <- PredictMat( out$gam3$smooth[[1]], dt )
  X4 <- PredictMat( out$gam4$smooth[[1]], dt )
  X5 <- PredictMat( out$gam5$smooth[[1]], dt )
  X6 <- PredictMat( out$gam6$smooth[[1]], dt )

  lg1 <- length(coef(out$gam1))
  lg2 <- length(coef(out$gam2))
```

```
lg3 <- length(coef(out$gam3))
lg4 <- length(coef(out$gam4))
lg5 <- length(coef(out$gam5))

F1[,i] <- X1%*%coef(out)[(out$gam1$smooth[[1]]$first.para:
out$gam1$smooth[[1]]$last.para)]
F2[,i] <- X2%*%coef(out)[lg1 + (out$gam2$smooth[[1]]$first.para:
out$gam2$smooth[[1]]$last.para)]
F3[,i] <- X3%*%coef(out)[lg1+ lg2+(out$gam3$smooth[[1]]$first.para:
out$gam3$smooth[[1]]$last.para)]
F4[,i] <- X4%*%coef(out)[lg1+lg2+lg3+(out$gam4$smooth[[1]]$first.para:
out$gam4$smooth[[1]]$last.para)]
F5[,i] <- X5%*%coef(out)[lg1+lg2+lg3+lg4+(out$gam5$smooth[[1]]$
first.para:out$gam5$smooth[[1]]$last.para)]
F6[,i] <- X6%*%coef(out)[lg1+lg2+lg3+lg4+lg5+(out$gam6$smooth[[1]]$
first.para:out$gam6$smooth[[1]]$last.para)]

F1[,i] <- F1[,i] - mean(F1[,i])
F2[,i] <- F2[,i] - mean(F2[,i])
F3[,i] <- F3[,i] - mean(F3[,i])
F4[,i] <- F4[,i] - mean(F4[,i])
F5[,i] <- F5[,i] - mean(F5[,i])
F6[,i] <- F6[,i] - mean(F6[,i])

gamma11[i] <- coef(out)[1]
gamma12[i] <- coef(out)[2]
gamma21[i] <- coef(out)[out$X1.d2+1]
gamma22[i] <- coef(out)[out$X1.d2+2]
gamma31[i] <- coef(out)[out$X1.d2+out$X2.d2+1]
gamma32[i] <- coef(out)[out$X1.d2+out$X2.d2+2]
```

```
theta121[i] <- coef(out)[out$X1.d2+out$X2.d2+out$X3.d2+1]
theta122[i] <- coef(out)[out$X1.d2+out$X2.d2+out$X3.d2+2]
theta131[i] <- coef(out)[out$X1.d2+out$X2.d2+out$X3.d2+out$X4.d2+1]
theta132[i] <- coef(out)[out$X1.d2+out$X2.d2+out$X3.d2+out$X4.d2+2]
theta231[i] <- coef(out)[out$X1.d2+out$X2.d2+out$X3.d2+out$X4.d2+
out$X5.d2+1]
theta232[i] <- coef(out)[out$X1.d2+out$X2.d2+out$X3.d2+out$X4.d2+
out$X5.d2+2]

 }
```

.

Bibliography

Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2), 182–198.

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

AIHW (2006). *Chronic Diseases and Associated Risk Factors in Australia*. Australian Institute of Health and Welfare.

Aldrich, J. H. & Nelson, F. D. (1984). *Linear probability, logit, and probit models*. Sage Publications.

Amos, A. F., McCarty, D. J., & Zimmet, P. (1997). The rising global burden of diabetes and its complications: estimates and projections to the year 2010. *Diabetic medicine*, 14(S5), S7–S85.

Andersen, E. W. (2004). Composite likelihood and two-stage estimation in family studies. *Biostatistics*, 5(1), 15–30.

Anderson, T. W., Anderson, T. W., Anderson, T. W., & Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. Wiley New York.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434), 444–455.

Ashford, J. & Sowden, R. (1970). Multi-variate probit analysis. *Biometrics*, 26, 535–546.

Azzalini, M. A. (2016). *mnormt: The Multivariate Normal and t Distributions*. R package version 1.5-5.

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., & Canning, D. (2011). Correcting hiv prevalence estimates for survey nonparticipation using heckman-type selection models. *Epidemiology*, 22(1), 27–35.

Bastida, E. & Pagán, J. A. (2002). The impact of diabetes on adult employment and earnings of mexican americans: findings from a community based study. *Health economics*, 11(5), 403–413.

Beck, C. A., Penrod, J., Gyorkos, T. W., Shapiro, S., & Pilote, L. (2003). Does aggressive care following acute myocardial infarction reduce mortality? analysis with instrumental variables to compare effectiveness in canadian and united states patient populations. *Health services research*, 38(6 Pt 1), 1423–1440.

Bhattacharya, J., Goldman, D., & McCaffrey, D. (2006). Estimating probit models with self-selected treatments. *Statistics in medicine*, 25(3), 389–413.

Black, S. A., Ray, L. A., & Markides, K. S. (1999). The prevalence and health burden of self-reported diabetes in older mexican americans: findings from the hispanic established populations for epidemiologic studies of the elderly. *American Journal of Public Health*, 89(4), 546–552.

Blondel, B., Kogan, M. D., Alexander, G. R., Dattani, N., Kramer, M. S., Macfarlane, A., & Wen, S. W. (2002). The impact of the increasing number of multiple births on the rates of preterm birth and low birthweight: an international study. *American Journal of Public Health*, 92(8), 1323–1330.

Buscha, F. & Conte, A. (2014). The impact of truancy on educational attainment during compulsory schooling: a bivariate ordered probit estimator with mixed effects. *The Manchester School*, 82(1), 103–127.

Butler, A. S., Behrman, R. E., et al. (2007). *Preterm Birth: Causes, Consequences, and Prevention*. National Academies Press.

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

Cappellari, L. & Jenkins, S. P. (2003). Multivariate probit regression using simulated maximum likelihood. *The Stata Journal*, 3(3), 278–294.

Chen, Z. & Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4), 762–769.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2), 347–361.

Chib, S., Greenberg, E., & Jeliazkov, I. (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics*, 18(2), 321–348.

Clarke, P. S. & Windmeijer, F. (2012). Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107(500), 1638–1652.

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.

Colchero, M. & Sosa-Rubí, S. (2012). Heterogeneity of income and lifestyle determinants of body weight among adult women in mexico, 2006. *Social Science & Medicine*, 75(1), 120–128.

Connors, R. D., Hess, S., & Daly, A. (2014). Analytic approximations for computing probit choice probabilities. *Transportmetrica A: Transport Science*, 10(2), 119–139.

Cox, D. & Barndorff-Nielsen, O. (1994). *Inference and Asymptotics*. CRC Press.

Czado, C. (2010). Pair-copula constructions of multivariate copulas. In P. Jaworski, F. Durante, W. K. Härdle, & T. Rychlik (Eds.), *Copula Theory and Its Applications* (pp. 93–109). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dall'Aglio, G., Kotz, S., & Salinetti, G. (2012). *Advances in probability distributions with given marginals: beyond the copulas*, volume 67. Springer Science & Business Media.

Donat, F. & Marra, G. (2017). Semi-parametric bivariate polychotomous ordinal regression. *Statistics and Computing*, 27(1), 283–299.

Duchon, J. (1977). *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*. Springer.

Eilers, P. H. & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2), 89–102.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

Frank, L. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109–135.

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

Freedman, D. A. & Sekhon, J. S. (2010). Endogeneity in probit response models. *Political Analysis*, 18(2), 138–150.

Genest, C., Nikoloulopoulos, A. K., Rivest, L.-P., & Fortin, M. (2013). Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. *Brazilian Journal of Probability and Statistics*, 27(3), 265–284.

Genz, A. (1991). An adaptive numerical integration algorithm for simplices. *Computing in the 90's*, 507, 279–285.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2), 141–149.

Genz, A. & Bretz, F. (2002). Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, 11(4), 950–971.

Genz, A. & Kass, R. E. (1997). Subregion-adaptive integration of functions having a dominant peak. *Journal of Computational and Graphical Statistics*, 6(1), 92–111.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities.

Gilbert, P. & Varadhan, R. (2016). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.

Glass, G. V. & Collins, J. R. (1970). Geometric proof of the restriction on the possible values of rxy when ryz are fixed. *Educational and Psychological Measurement*, 30, 37–39.

Greene, W. (2003). *Econometric Analysis*. Prentice Hall, New York.

Gronau, R. (1973). Wage comparisons–a selectivity bias. *Journal of Political Economy*.

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

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

Hajivassiliou, V. A. & McFadden, D. (1991). *The method of simulated scores for the estimation of LDV models with an application to external debt crises*. Yale University, Cowles Foundation for Research in Economics.

Harris, A. (2009). Diabetes, cardiovascular disease and labour force participation in australia: An endogenous multivariate probit analysis of clinical prevalence data. *Economic Record*, 85(271), 472–484.

Harris, M. I. (1998). Diabetes in america: epidemiology and scope of the problem. *Diabetes care*, 21(Supplement 3), C11–C14.

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

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

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. *Annals of Economic and Social Measurement*, 5(4), 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.

Heitjan, D. F. & Basu, S. (1996). Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3), 207–213.

Henningsen, A. (2015). *mvProbit: Multivariate Probit Models*. R package version 0.1-8.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

Hubert, L. J. (1972). A note on the restriction of range for pearson product-moment correlation coefficients. *Educational and Psychological Measurement*, 32, 767–770.

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

Joe, H. (1996). Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, 28, 120–141.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Joe, H. & Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, 100(4), 670–685.

Kasteridis, P. P., Munkin, M. K., Yen, S. T., et al. (2010). A binary-ordered probit model of cigarette demand. *Applied Economics*, 42(4), 413–426.

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, 49(1), 169–186.

Keane, M. P. (1990). *Four essays in empirical macro and labor economics*. Ph.D. dissertation, Brown University.

Keay, M.-J. (2016). Partial copula methods for models with multiple discrete endogenous explanatory variables and sample selection. *Economics Letters*, 144, 85–87.

Kessler, R. C., Maclean, J. R., Petukhova, M., Sarawate, C. A., Short, L., Li, T. T., & Stang, P. E. (2008). The effects of rheumatoid arthritis on labor force participation, work performance, and healthcare costs in two workplace samples. *Journal of Occupational and Environmental Medicine*, 50(1), 88–98.

Kiely, J. L. (1998). What is the population-based risk of preterm birth among twins and other multiples? *Clinical Obstetrics and Gynecology*, 41(1), 3–11.

Kim, Y., Kwon, S., & Song, S. H. (2006). Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. *Computational Statistics & Data Analysis*, 51(3), 1643–1655.

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

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

Koch, I. (1996). On the asymptotic performance of median smoothers in image analysis and nonparametric regression. *The Annals of Statistics*, 24(4), 1648–1666.

Król, A., Ferrer, L., Pignon, J.-P., Proust-Lima, C., Ducreux, M., Bouché, O., Michiels, S., & Rondeau, V. (2016). Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the ffcd 2000–05 trial. *Biometrics*, 72(3), 907–916.

Kuk, A. Y. & Nott, D. J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics & Probability Letters*, 47(4), 329–335.

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

Leigh, J. P. & Schembr, M. (2004). Instrumental variables technique: cigarette price provided better estimate of effects of smoking on sf-12. *Journal of clinical epidemiology*, 57(3), 284–293.

Leung, C.-K. & Lam, K. (1975). A note on the geometric representation of the correlation coefficients. *The American Statistician*, 29(3), 128–130.

Li, C., Poskitt, D., Zhao, X., et al. (2016). *The Bivariate Probit Model, Maximum Likelihood Estimation, Pseudo True Parameters and Partial Identification*. Technical report, Monash University, Department of Econometrics and Business Statistics.

Li, P. (2011). Estimation of sample selection models with two selection mechanisms. *Computational Statistics & Data Analysis*, 55(2), 1099–1108.

Li, R. & Sudjianto, A. (2005). Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, 47(2), 111–120.

Lie, H.-P. & Gardner, S. (2016). The interrelationship between smoking and depression in indonesia. *Health Policy and Technology*, 5(1), 26–31.

Lindenl, A. & Adams, J. L. (2006). Evaluating disease management programme effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice*, 12(2), 148–154.

Little, R. J. (1985). A note about models for selectivity bias. *Econometrica: Journal of the Econometric Society*, 53(6), 1469–1474.

Loureiro, M. L., Sanz-de Galdeano, A., & Vuri, D. (2010). Smoking habits: like father, like son, like mother, like daughter? *Oxford Bulletin of Economics and Statistics*, 72(6), 717–743.

LP, S. C. (2017). *Stata Statistical Software Release 15*. Stata Press Publication.

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

Marius Hofert, Ivan Kojadinovic, M. M. & Yan, J. (2017). *copula: Multivariate Dependence with Copulas*. R package version 0.999-18.

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

Marra, G. & Radice, R. (2017). *GJRM: Generalised Joint Regression Modelling*. R package version 0.1-1.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., McGovern, M. E., et al. (2017). A simultaneous equation approach to estimating hiv prevalence with non-ignorable missing responses. *Journal of the American Statistical Association*, 112(518), 484–496.

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

Marra, G. & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1), 53–74.

Marshall, A. W. & Olkin, I. (1988). Families of multivariate distributions. *Journal of the American statistical association*, 83(403), 834–841.

Martin, J. A., Park, M. M., et al. (1999). Trends in twin and triplet births: 1980–97. *National vital statistics reports*, 47(24), 1–16.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

McNeil, A. J. & Nešlehová, J. (2009). Multivariate archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. *The Annals of Statistics*, 37(5B), 3059–3097.

Meng, G. (2010). *Social and spatial determinants of adverse birth outcome inequalities in socially advanced societies*. Ph.D. dissertation, University of Waterloo.

Minor, T. (2011). The effect of diabetes on female labor force decisions: new evidence from the national health interview survey. *Health economics*, 20(12), 1468–1486.

Miranda, M. L., Maxson, P., & Edwards, S. (2009). Environmental contributions to disparities in pregnancy outcomes. *Epidemiologic Reviews*, 31(1), 67–83.

Mohanty, M. S. (2001). Testing for the specification of the wage equation: double selection approach or single selection approach. *Applied Economics Letters*, 8(8), 525–529.

Neelon, B., Anthopolos, R., & Miranda, M. L. (2014). A spatial bivariate probit model for correlated binary data with application to adverse birth outcomes. *Statistical Methods in Medical Research*, 23(2), 119–133.

Nikoloulopoulos, A. K. (2015). A vine copula mixed effect model for trivariate meta-analysis of diagnostic test accuracy studies accounting for disease prevalence. *Statistical Methods in Medical Research*, 25(2), 988–991.

Nikoloulopoulos, A. K. (2016). Efficient estimation of high-dimensional multivariate normal copula models with discrete spatial responses. *Stochastic Environmental Research and Risk Assessment*, 30(2), 493–505.

Nocedal, J. & Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Noh, H., Ghouch, A. E., & Bouezmarni, T. (2013). Copula-based regression estimation and inference. *Journal of the American Statistical Association*, 108(502), 676–688.

Oelker, M.-R., Gertheiss, J., & Tutz, G. (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling*, 14(2), 157–177.

Oelker, M.-R. & Tutz, G. (2013). A general family of penalties for combining differing types of penalties in generalized structured models. *Technical Report Number 139, 2013, Department of Statistics, University of Munich.*

Panagiotelis, A., Czado, C., & Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499), 1063–1072.

Paneth, N. S. (1995). The problem of low birth weight. *The Future of Children*, 5(1), 19–34.

Park, M. Y. & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B*, 69(4), 659–677.

Plackett, R. L. (1954). A reduction formula for normal multivariate integrals. *Biometrika*, 41(3/4), 351–360.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3), 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87(2), 425–435.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance–correlation parameters. *Biometrika*, 94(4), 1006–1013.

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

Radice, R., Zanin, L., & Marra, G. (2013). On the effect of obesity on employment in the presence of observed and unobserved confounding. *Statistica Neerlandica*, 67(4), 436–455.

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.

Rippe, R. C., Meulman, J. J., & Eilers, P. H. (2012). Visualization of genomic changes by segmented smoothing using an l_0 penalty. *PloS One*, 7(6), 1–14.

Rossi, P. (2015). *Bayesian Inference for Marketing/Micro-Econometrics*. R package version Version 3.0-2.

Rous, J. J., Jewell, R. T., & Brown, R. W. (2004). The effect of prenatal care on birthweight: a full-information maximum likelihood approach. *Health Economics*, 13(3), 251–264.

Rousseeuw, P. J. & Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics–Theory and Methods*, 22(4), 965–984.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields*. New Haven: Chapman & Hall/CRC, Boca Raton, FL.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge university press.

Sajaia, Z. (2008). Maximum likelihood estimation of a bivariate ordered probit model: implementation and monte carlo simulations. *The Stata Journal*, 4(2), 1–18.

Sharma, A., Siciliani, L., & Harris, A. (2013). Waiting times and socioeconomic status: does sample selection matter? *Economic Modelling*, 33, 659–667.

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

South, A. P., Jones, D. E., Hall, E. S., Huo, S., Meinzen-Derr, J., Liu, L., & Greenberg, J. M. (2012). Spatial analysis of preterm birth demonstrates opportunities for targeted intervention. *Maternal and Child Health Journal*, 16(2), 470–478.

Stanley, J. C. & Wang, M. D. (1969). Restrictions on the possible values of r12 given r13 and r23. *Educational and Psychological Measurement*, 29, 579–581.

Team, R. C. & contributors worldwide (2015). *The R Stats Package*. R package version 3.1.3.

Team, R. D. C. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Terracol, A. (2002). Triprobit and the ghk simulator: a short note.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1), 267–288.

Trinh, G. & Genz, A. (2015). Bivariate conditioning approximations for multivariate normal probabilities. *Statistics and Computing*, 25(5), 989–996.

Trivedi, P. K. & Zimmer, D. M. (2007). *Copula modeling: an introduction for practitioners*. Now Publishers Inc.

Tunceli, K., Bradley, C. J., Nerenz, D., Williams, L. K., Pladevall, M., & Lafata, J. E. (2005). The impact of diabetes on employment and work productivity. *Diabetes care*, 28(11), 2662–2667.

Ulbricht, J. (2010). *Variable selection in generalized linear models*. Verlag Dr. Hut.

Van de Ven, W. P. & Van Praag, B. M. (1981). The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of econometrics*, 17(2), 229–252.

Ward, B. W. (2015). Multiple chronic conditions and labor force outcomes: A population study of us adults. *American journal of industrial medicine*, 58(9), 943–954.

Ward, B. W. & Schiller, J. S. (2013). Prevalence of multiple chronic conditions among us adults: Estimates from the national health interview survey, 2010. *Preventing chronic disease*, 10.

Ward, B. W., Schiller, J. S., & Goodman, R. A. (2014). Multiple chronic conditions among us adults: a 2012 update. *Preventing Chronic Disease*, 11.

Whiting, D. R., Guariguata, L., Weil, C., & Shaw, J. (2011). Idf diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes research and clinical practice*, 94(3), 311–321.

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, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economics letters*, 69(3), 309–312.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B*, 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. (2006). *Generalized Additive Models: An Introduction With R*. Chapman & Hall/CRC, London.

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society Series B*, 70(3), 495–518.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B*, 73(1), 3–36.

Wood, S. N. (2013a). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

Wood, S. N. (2013b). A simple test for random effects in regression models. *Biometrika*, 100(4), 1005–1010.

Wooldridge, J. (2002). *Econometrics analysis of cross section and panel Data*. Cambridge: MIT Press.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models*. R package version 0.9-7.

Yoshida, T. & Naito, K. (2014). Asymptotics for penalised splines in generalised additive models. *Journal of Nonparametric Statistics*, 26(2), 269–289.

Zhang, R., Inder, B. A., & Zhang, X. (2015). Bayesian estimation of a discrete response model with double rules of sample selection. *Computational Statistics & Data Analysis*, 86, 81–96.

Zhang, X., Zhao, X., & Harris, A. (2009). Chronic diseases and labour force participation in australia. *Journal of Health Economics*, 28(1), 91–108.

Zhao, Y. & Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, 33(3), 335–356.

Zhong, W., Koopmeiners, J. S., & Carlin, B. P. (2012). A trivariate continual reassessment method for phase i/ii trials of toxicity, efficacy, and surrogate efficacy. *Statistics in Medicine*, 31(29), 3885–3895.

Zimmer, D. M. & Trivedi, P. K. (2006). Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business & Economic Statistics*, 24(1), 63–76.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2), 301–320.