9-11-2017

Statistical Methods for Analyzing Bivariate Mixed Outcomes

Ved Deshpande
University of Connecticut - Storrs, ved.deshpande@uconn.edu

Statistical Methods for Analyzing Bivariate Mixed Outcomes

Ved Deshpande, Ph.D.
University of Connecticut, 2017

ABSTRACT

Multivariate outcomes are ubiquitous. Joint analysis of multivariate outcomes provides several benfits over separate analysis of each outcome. However, joint analysis of multivariate outcomes that are mixed, i.e., not on the same scale of measurement, can be challenging. This dissertation provides novel methods to analyze bivariate mixed outcomes, where we have exactly one continuous outcome and one binary outcome. A penalized generalized estimating equations framework to perform simultaneous estimation and variable selection for bivaraite mixed outcomes in the presence of a large number of covariates is provided. Next, fully Bayesian and empirical Bayes approaches to estimating the association between the two outcomes using a copula-based model are provided. Finally, methods for estimating and testing genomic effects in bivariate mixed secondary outcome models under case-control designs are presented.

Statistical Methods for Analyzing Bivariate Mixed Outcomes

Ved Deshpande

Integrated B.Sc. and M.Sc., Statistics and Informatics, Indian Institute of Technology,

Kharagpur, India, 2010

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2017

APPROVAL PAGE

Doctor of Philosophy Dissertation

Statistical Methods for Analyzing Bivariate Mixed Outcomes

Presented by

Ved Deshpande, Integrated B.Sc. and M.Sc., Statistics and Informatics

Co-Major Advisor _____

Dipak K. Dey

Co-Major Advisor _____

Elizabeth D. Schifano

Associate Advisor _____

Haim Bar

University of Connecticut

2017

Acknowledgments

I would like to thank both of my advisors, Dr. Dipak Dey and Dr. Elizabeth Schifano, without whose support I could not have completed this dissertation. Dr. Dey gave me much advice on how to be a good researcher, and taught me how to "connect the dots" to develop new ideas. Dr. Schifano was with me in the trenches, providing me with unwavering support and encouragement, especially during those times when things seemed absolutely hopeless. From going line-by-line through proofs, to catching tiny bugs in my code that escaped me for weeks, to providing detailed corrections on manuscripts, Dr. Schifano has helped me at every step of the way. I would also like to thank her for giving me my first taste of research, by offering me a research assistantship during my second year.

I am also very grateful to Dr. Ming-Hui Chen, for selecting me to be a part of UCONN's Statistical Consulting Services. Being a part of SCS has contributed tremendously towards my development as an applied statistician, and opened doors to a variety of career paths. I am certain that I would not have gotten the career opportunities that I have had without this experience. Dr. Chen has also been a role model of dedication and integrity. Seeing someone as incredibly busy as him never cut corners in even the smallest of tasks has inspired me to do the same, and I am better off for it.

I would like to thank Dr. Haim Bar for agreeing to be a part of my dissertation

committee, and for working with me on the research assistantship that I did in my second year. I learnt a great deal of R programming and applied statistics from him.

Next, I would like to thank every professor who has taught me at UCONN. They not only gave me an excellent education, but also infected me with their passion for statistics. In particular, I must thank Dr. Nitis Mukhopadhyay, who lit the first spark of my love for statistics; Dr. Rick Vitale, who showed me that measure theory was beautiful rather than terrifying; and Dr. Jun Yan, who helped me discover the power of the R programming language.

I have had a lot of fun during the last five years, and I owe that to no one more than my friends and colleagues in the department. Thank you for all the laughs, arguments, rants, philosophy, and motivation. I wish you all the best of luck in your endeavours.

Finally, and most importantly, I would like to thank my family for their constant support, warmth, and love. I dedicate this dissertation to you. Mom and Dad, you have always been supportive of my life choices, providing me with gentle guidance, and pulling me out of tight corners. Dooti, I have lost count of the innumerable number of things that you have done for me over these past eight years. Words utterly fail to capture my gratitude. Thank you–for everything.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The task of modeling multivariate outcomes on sets of covariates is becoming increasingly common across research disciplines. Multivariate outcomes that are measured from the same sampling unit are likely to be correlated. Joint modeling of correlated multivariate outcomes is preferable over separate modeling of the outcomes because we may be able to obtain more efficient parameter estimates through information sharing across correlated outcomes (Teixeira Pinto and Normand, 2009).

Multivariate outcomes are often *mixed*, i.e., they are measured on different scales of measurement. A common subcase of mixed multivariate outcomes is when exactly two outcomes per sampling unit are measured, with one outcome measured on a continuous scale and the other outcome measured on a binary scale. Joint modeling of such *bivariate mixed outcomes* is usually performed by specifying a joint probability model for the outcomes. However, specifying a joint model for mixed outcomes is challenging due to the lack of appropriate multivariate distributions for mixed outcomes. Likelihood-based approaches that aim to circumvent this problem include the factorization approach, in which the joint distribution of the outcomes is factorized into the marginal distribution

of one outcome and the conditional distribution of the other outcome given the first outcome, and the latent variable approach, in which unobserved shared latent variables account for the correlation between the outcomes. See Teixeira Pinto and Normand (2009) for a survey of these methods. A drawback of the factorization approach is that the model is not invariant to the choice of the conditioning outcome (Wu and de Leon, 2014). Disadvantages of the latent variable approach include sensitivity to misspecification of the covariance structure, and arbitrary and untestable distributional assumptions on the latent variables (Prentice and Zhao, 1991).

To overcome the drawbacks of direct likelihood-based approached to model bivariate mixed outcomes, a few indirect approaches have been proposed. One class of approaches utilizes generalized estimating equations (GEEs) (Prentice and Zhao, 1991; Rochon, 1996; Liu et al., 2009). GEEs are both convenient and robust; convenient because GEEs only require the specification of the first two moments of each outcome and an approximation to their correlation structure, and robust because GEEs consistently estimate the regression parameters even if the correlation structure is misspecified. GEEs are primarily used when the correlation between the outcomes is a nuisance parameter, and the marginal parameters are of primary interest. On the other hand, in many research studies, the association between the outcomes is of primary interest. Because GEEs are inefficient for estimating association parameters (Liang et al., 1992; Hall and Severini, 1998), they are not appropriate in such cases. In this context, copulas (Nelsen, 2007) are a convenient tool to efficiently model the association between correlated outcomes.

Like GEEs, copulas do not require direct specification of a joint distribution between the outcomes. Rather, they "glue" the marginal distributions of the outcomes together. With the ability to specify marginal distributions independently and the vast number of dependence structures available (Joe, 2014), copulas provide researchers great flexibility in modeling correlated outcomes.

We now provide a brief introduction to the framework of GEEs and copulas. Note that for the latter, we consider two-dimensional copulas only, which are relevant to this dissertation.

1.1 Generalized estimating equations

Liang and Zeger (1986) introduced GEEs as an extension of generalized linear models (McCullagh, 1984) to account for the correlation between longitudinal outcomes. More generally, GEEs can be used with any kind of clustered outcomes, not necessarily longitudinal. Suppose there are n clusters, and we observe outcomes $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{im_i})^T$, $i = 1, \ldots, n$. Each outcome y_{ij} has associated with it a p-dimensional covariate vector $\boldsymbol{x_{ij}}$, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. Observations from different clusters are assumed to be independent, but observations from the same cluster are assumed to be correlated. For notational simplicity, we assume that $m_i = m$, i.e., all clusters are of the same size.

GEEs require specification of the first two moments of the outcomes. Similar to generalized linear models, we specify link functions $g_{ij}(\mu_{ij}) = \boldsymbol{x_{ij}^T\beta}$, where $\mu_{ij} = \mathbb{E}(y_{ij})$.

Denote $\boldsymbol{\mu_i} = (\mu_{i1}, \ldots, \mu_{im})^T$. Next, we specify variance functions $v_{ij}(y_{ij}) = \psi h_{ij}(\mu_{ij})$. Note that we have assumed that within a cluster, a common set of regression coefficients $\boldsymbol{\beta}$ and a common dispersion parameter ψ apply to each outcome. This assumption will be relaxed in the chapters that follow. The GEEs are given by

$$\boldsymbol{S}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{D_i^T V_i^{-1}} (\boldsymbol{y_i} - \boldsymbol{\mu_i}) = \boldsymbol{0},$$

where $\boldsymbol{D_i^T} = \partial \boldsymbol{\mu_i}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}^T$, and $\boldsymbol{V_i}$ is the variance-covariance matrix of $\boldsymbol{y_i}$, given by $\boldsymbol{V_i} = \psi \boldsymbol{A_i^{1/2}} \boldsymbol{R} \boldsymbol{A_i^{1/2}}$, where $\boldsymbol{A_i} = \text{diag}(h_{i1}(\mu_{i1}), \ldots, h_{im}(\mu_{im}))$, and $\boldsymbol{R} \equiv ((\rho_{jj'}))$ is the *working correlation matrix* of $(y_{i1}, \ldots, y_{im_i})^T$, assumed to be the same for all $i = 1, \ldots, n$. Some commonly used working correlation structures include independence ($\rho_{jj'} = 0, \ \ j \neq j'$), exchangeable ($\rho_{jj'} = \rho, \ \ j \neq j'$), AR(1) ($\rho_{jj'} = \rho^{|j - j'|}$), and unstructured, among others. For simplicity, assume a single parameter ρ indexes $\boldsymbol{R}$.

In practice, the GEEs are solved by iterating between a Newton-Raphson type update for $\boldsymbol{\beta}$ and moment-based estimation of $(\psi, \rho)^T$. For more details on the moment-based estimators of ψ and ρ, see Liang and Zeger (1986). Given the current estimate $\hat{\boldsymbol{\beta}}^k$ and estimates $\hat{\psi}$ and $\hat{\rho}$, the regression coefficients $\boldsymbol{\beta}$ are updated as

$$\hat{\boldsymbol{\beta}}^{k+1} = \hat{\boldsymbol{\beta}}^k + \left(\sum_{i=1}^{n} \boldsymbol{D_i} (\hat{\boldsymbol{\beta}}^k)^T \tilde{\boldsymbol{V_i}}^{-1} \boldsymbol{D_i} (\hat{\boldsymbol{\beta}}^k) \right)^{-1} \left[\sum_{i=1}^{n} \boldsymbol{D_i} (\hat{\boldsymbol{\beta}}^k)^T \tilde{\boldsymbol{V_i}}^{-1} \boldsymbol{r_i} (\hat{\boldsymbol{\beta}}^k) \right],$$

with $\tilde{\boldsymbol{V_i}} \equiv \boldsymbol{V_i}(\hat{\boldsymbol{\beta}}^k, \hat{\psi}, \hat{\rho})$, and $\boldsymbol{r_i}(\hat{\boldsymbol{\beta}}^k) = \boldsymbol{y_i} - \boldsymbol{\mu_i}(\hat{\boldsymbol{\beta}}^k)$. Liang and Zeger (1986) showed that

solving the GEEs lead to consistent estimates of the regression parameters $\boldsymbol{\beta}$, even if the working correlation structure $\boldsymbol{R}$ is misspecified, which makes for convenient and robust inference.

1.2 Copulas

Sklar (1959) first introduced the term *copula* in the statistical literature to denote functions that joined marginal distributions together to form a joint distribution. The functions themselves, however, can be traced back to Hoeffding (1940). Sklar's theorem (Sklar, 1959) shows the relationship between the joint distribution function, the marginal distribution functions, and the copula.

Theorem 1.1 (Sklar's Theorem)**.** *Let H be a joint distribution function of random variables Y_1 and Y_2 with margins F_1 and F_2. Then there exists a copula C such that*

$$H(y_1, y_2) = C(F_1(y_1), F_2(y_2)), \quad \forall (y_1, y_2) \in \mathbb{R}^2. \tag{1.1}$$

If F_1 and F_2 are continuous, then C is unique. Otherwise, C is uniquely defined on Range(F_1) $\times$ Range(F_2).

Conversely, if C is a copula and F_1 and F_2 are distribution functions, then the function H defined by (1.1) is a joint distribution with margins F_1 and F_2.

Sklar's theorem is important because it ensures that irrespective of the choice of F_1,

F_2 and C, the resulting H function is a valid joint probability distribution function. This allows for great flexibility in constructing joint probability distributions.

We have informally denoted copulas as functions that join marginal probability distribution functions together to form a joint probability distribution. For the purposes of this dissertation, this is sufficient. Formal mathematical definitions of copulas can be found in Nelsen (2007).

1.2.1 Measures of dependence

With copulas, we can specify the type of dependence structure that we wish to model between random variables. Usually, parametric copula families are indexed by a parameter θ, which controls the strength of dependence between the random variables. Two important measures of dependence are Kendall's Tau and Spearman's Rho, defined as follows:

Definition 1.2 (Kendall's Tau)**.** *For random variables X_1 and Y_1, the population version of Kendall's Tau is*

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0],$$

where (X_2, Y_2) is an independent copy of (X_1, Y_1).

For continuous random variables X_1 and Y_1 with copula C, we have

$$\tau = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1. \tag{1.2}$$

Definition 1.3 (Spearman's Rho)**.** *For random variables X_1 and Y_1, the population version of Spearman's Rho is*

$$\rho = 3(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0],$$

where (X_1, Y_1), (X_2, Y_2), and (X_3, Y_3) are independent copies.

For continuous random variables X_1 and Y_1 with copula C, we have

$$\rho = 12 \int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3.$$

Another concept of dependence is tail dependence, which relates to the dependence between extreme values in the upper-right and the lower-left quadrant of a bivariate distribution.

Definition 1.4 (Tail dependence)**.** *Let Y_1 and Y_2 be continuous random variables with*

margins F_1 and F_2 and copula C. Then define

$$\lambda_L = \lim_{u \to 0^+} P(Y_2 \leq F_2^{-1}(u) | Y_1 \leq F_1^{-1}(u)) = \lim_{u \to 0^+} \frac{C(u,u)}{u},$$
$$\lambda_U = \lim_{u \to 1^-} P(Y_2 > F_2^{-1}(u) | Y_1 > F_1^{-1}(u)) = 2 - \lim_{u \to 1^-} \frac{1 - C(u,u)}{1-u}.$$

C has lower tail dependence if $\lambda_L \in (0,1]$, and lower tail independence if $\lambda_L = 0$, and similarly for upper tail dependence and λ_U.

1.2.2 Some commonly used copula families

Here we provide definitions and some useful properties of the copula families that are used in this dissertation. These copula families broadly belong to the elliptical and Archimedean classes of copulas. For details on these classes of copulas, see Nelsen (2007).

1. The Gaussian copula family belongs to the elliptical class of copulas. It is a symmetric copula that does not exhibit tail dependence. The two-dimensional Gaussian copula is given by

$$C(u_1, u_2 | \theta) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2) | \theta), \quad \theta \in (-1, 1),$$

where $\Phi_2(\cdot, \cdot | \theta)$ is the bivariate standard normal distribution function with correlation parameter θ, and Φ^{-1} is the inverse of the univariate standard normal distribution

function. For this copula, Kendall's Tau is given by $\tau = (2/\pi)\arcsin\theta$.

2. The Clayton copula family belongs to the Archimedean class of copulas. It is an asymmetric copula that exhibits lower tail dependence. The copula is given by

$$C(u_1, u_2|\theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{\frac{1}{\theta}}, \quad \theta \in (-1, \infty) \setminus \{0\}.$$

For this copula, Kendall's Tau is given by $\tau = \theta/(\theta + 2)$.

3. The Gumbel copula family belongs to the Archimedean class of copulas. It is an asymmetric copula that exhibits upper tail dependence. The copula is given by

$$C(u_1, u_2|\theta) = \exp\left[-\left\{(-\log u_1)^\theta + (-\log u_2)^\theta\right\}^{\frac{1}{\theta}}\right], \quad \theta \in [1, \infty).$$

For this copula, Kendall's Tau is given by $\tau = (\theta - 1)/\theta$.

4. The Frank copula family belongs to the Archimedean class of copulas. It is a symmetric copula that does not exhibit tail dependence. The copula is given by

$$C(u_1, u_2|\theta) = -\frac{1}{\theta}\log\left\{1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^\theta - 1}\right\}, \quad \theta \in (-\infty, \infty) \setminus \{0\}.$$

For this copula, Kendall's Tau is given by $\tau = 1 + (4/\theta)[D_1(\theta) - 1]$, where $D_1(\theta)$ is the Debye function of the first kind defined as $D_1(\theta) = (1/\theta)\int_0^\theta t/(e^t - 1)dt$.

1.2.3 Estimation and inference with copulas

We restrict our attention to fully parametric copula models, in which both the margins and the copula have parametric forms, which is relevant to this dissertation. The starting point of estimation is the joint probability density function of the random variables. Let Y_1 and Y_2 be continuous random variables with marginal probability distribution functions F_1 and F_2, respectively. Let the copula C specify the joint probability distribution of Y_1 and Y_2. Let β_1 and β_2 be the marginal parameters that index F_1 and F_2, respectively. Applying Sklar's theorem, we can obtain the joint probability density function as

$$f_{12}(y_1, y_2|\beta_1, \beta_2, \theta) = c(F_1(y_1|\beta_1), F_2(y_2|\beta_2)|\theta) f_1(y_1|\beta_1) f_2(y_2|\beta_2), \tag{1.3}$$

where $c(u, v) = \partial^2 C(u, v)/\partial u\, \partial v$ is the copula density function, and f_j is the marginal density function associated with F_j, $j = 1, 2$. Using (1.3), the likelihood function can be constructed. Frequentist estimation can be performed by jointly maximizing β_1, β_2, and θ. A computationally more convenient alternative is the method of Inference Function for Margins (IFM), in which estimation proceeds in two stages. In the first stage, the marginal parameters β_1 and β_2 are estimated by maximizing the marginal univariate likelihoods. In the second stage, the copula parameter θ is estimated conditional on the estimates of the marginal parameters. Although computationally convenient, IFM can be inefficient (Joe and Xu, 1996). Bayesian approaches offer an alternative solution

that can be simultaneously efficient and computationally convenient through the use of Markov Chain Monte Carlo (MCMC) sampling algorithms to conduct inference based on the complete posterior distribution of all the model parameters. Details on conducting Bayesian inference for models relevant to this dissertation are provided in Chapter 3.

1.3 Overview of dissertation

With recent advances in data collection and storage technologies, it is a common task to model outcomes on a large number of covariates. GEEs in their usual form do not perform regularization of parameter estimates or variable selection, which are important tasks to perform in the presence of a large number of covariates. In Chapter 2, we provide a framework to perform simultaneous estimation and variable selection for bivariate mixed outcomes with *penalized generalized estimating equations* (PGEEs). In the context of variable selection, controlling the false discovery rate (FDR) is also an important requirement. We also provide a method to estimate and control the FDR in the PGEE framework for bivariate mixed outcomes.

As mentioned previously, when the association between correlated outcomes is of primary interest, copula-based models are a better alternative than GEEs. In Chapter 3, we provide a fully Bayesian approach and an empirical Bayes approach to estimate a copula-based model for bivariate mixed outcomes. These methods use Hamiltonian Monte Carlo (HMC, Duane et al. (1987), Neal (2011)) to perform MCMC sampling, which makes

them extremely fast compared to equivalent methods based on the Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) or the Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) algorithms. We also investigate the ability of the fully Bayesian method to select the correct copula family, viz., the problem of copula selection.

In Chapter 4, we propose and compare two GEE-based methods to jointly model bivariate mixed *secondary phenotypes* in genome-wide association studies (GWASs) with a case-control sampling design. Because the case-control sampling design can distort the population association between the phenotypes and the genomic variables of interest, naive application of GEEs is not recommended. Our methods extend existing GEE-based solutions to analyzing secondary outcomes in case-control studies to the case of bivariate mixed outcomes.

Finally, in Chapter 5, we discuss limitations and directions for future work related to the methods proposed in this dissertation.

Chapter 2

Variable selection for correlated bivariate mixed outcomes using penalized generalized estimating equations

2.1 Introduction

As mentioned previously, in the presence of a large number of covariates, it is of interest to modify the usual generalized estimating equations (GEEs) to perform both estimation as well as variable selection. These can often be achieved simultaneously through penalized regression techniques, using penalties such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), the elastic net (EN) (Zou and Hastie, 2005), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), the minimax concave penalty (MCP) (Zhang, 2007), and others. To incorporate penalized

regression techniques in GEEs, Fu (2003) and Johnson et al. (2008) laid the framework for *penalized generalized estimating equations* (PGEEs), while Wang et al. (2012) gave the form of PGEEs for commensurate longitudinal outcomes. PGEEs perform simultaneous parameter estimation and variable selection through the incorporation of a sparsity-inducing penalty term in GEEs. In this chapter, we provide the framework to apply PGEEs in the (non-longitudinal) bivariate mixed outcome case. Through simulation studies, we show that gains can be made in both estimation and variable selection by using joint analysis rather than by separate marginal analyses of the outcomes. In the context of variable selection, controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995) is often of importance as well. Breheny (2009) and Yi et al. (2015) showed how to estimate and control the FDR for penalized regression. We generalize this method to the PGEE framework for bivariate mixed outcomes, and through simulations, demonstrate that our method is able to control the FDR at a desired level.

We illustrate the application of our PGEE framework and FDR control methodology to data from the Medical Expenditure Panel Survey (MEPS). MEPS provides a nationally representative sample of health care data at the individual level, and contains information on medical spending, health status, demographics, health conditions, access to care, health insurance coverage, income, and employment. Our analysis is inspired by the work done in Zimmerman (2013), who sought to jointly model annual drug spending (modeled as a continuous variable) and health status (modeled as a binary variable) for Medicare enrollees in 2004 and 2005, the two years before Medicare

Part D became active. While the primary goal of that analysis was to investigate the strength of association between these two outcomes, our goal is to identify important covariates that affect drug spending and health status. With our penalized GEE framework, we are able to consider a larger set of covariates than Zimmerman (2013). Then, by borrowing information from total drug spending, we are able to identify important covariates for health status that may not be detectable from a marginal analysis on the latter outcome. We also estimate the false discovery rate to reassure ourselves that we are detecting additional signal, rather than noise.

The rest of the chapter is organized as follows. In Section 2.2, we provide the framework for applying PGEEs to bivariate mixed outcomes. We also provide an iterative algorithm to solve the PGEEs and a method to control the FDR. Section 2.3 contains results from simulation experiments. In Section 2.4, we apply the PGEE framework to the MEPS data and discuss our findings.

2.2 Penalized generalized estimating equations for bivariate mixed outcomes

2.2.1 Notations

From the ith individual, we observe a continuous outcome y_{ic}, a binary outcome y_{ib}, a p-dimensional covariate vector $\boldsymbol{x_i}$ corresponding to the continuous outcome y_{ic}, and a

q-dimensional covariate vector $\boldsymbol{z_i}$ corresponding to the binary outcome y_{ib}, $i = 1 \ldots n$. It is common to assume $\boldsymbol{x_i} = \boldsymbol{z_i}$ (i.e., use the same set of covariates to model both outcomes), but this need not be so. Let $\boldsymbol{y_i} = (y_{ic}, y_{ib})^T$ denote the bivariate vector of outcomes from the ith individual. We assume that outcomes from the same individual are correlated, but outcomes from different individuals are independent.

We specify the link functions $g_c(\mu_{ic}) = \boldsymbol{x_i}^T\boldsymbol{\beta_c}$ and $g_b(\mu_{ib}) = \boldsymbol{z_i}^T\boldsymbol{\beta_b}$, where $\mu_{ic} = \mathrm{E}(y_{ic})$ and $\mu_{ib} = \mathrm{E}(y_{ib})$. Denote $\boldsymbol{\mu_i} = (\mu_{ic}, \mu_{ib})^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta_c}^T, \boldsymbol{\beta_b}^T)^T$. We specify the variance functions $v_c(y_{ic}) = \psi_c h_c(\mu_{ic})$ and $v_b(y_{ib}) = \psi_b h_b(\mu_{ib})$, where ψ_c and ψ_b are dispersion parameters. For illustration, we shall take $g_c(\cdot)$ to be the identity link and $g_b(\cdot)$ to be the logit link. For simplicity, we further assume that $\psi_c = \psi_b = 1$.

2.2.2 Generalized estimating equations for bivariate mixed outcomes

Rochon (1996) gave the setup for generalized estimating equations for bivariate mixed outcomes:

$$\boldsymbol{S}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\boldsymbol{D_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) = \boldsymbol{0}, \tag{2.1}$$

where

$$\boldsymbol{D_i}^T = \frac{\partial\boldsymbol{\mu_i}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T} = \begin{pmatrix} \partial\mu_{ic}/\partial\boldsymbol{\beta_c}^T & \boldsymbol{0} \\ \boldsymbol{0} & \partial\mu_{ib}/\partial\boldsymbol{\beta_b}^T \end{pmatrix}, \tag{2.2}$$

and $\boldsymbol{V_i}$ is the variance-covariance matrix of $\boldsymbol{y_i}$, given by $\boldsymbol{V_i} = \boldsymbol{A_i}^{1/2}\boldsymbol{R}\boldsymbol{A_i}^{1/2}$, where

$$\boldsymbol{A_i} = \begin{pmatrix} h_c(\mu_{ic}) & 0 \\ 0 & h_b(\mu_{ib}) \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Here $\boldsymbol{R}$ is a *working correlation matrix* and ρ measures the strength of association between the continuous and binary outcomes. Note that ρ, which we shall refer to as the association parameter, is assumed to be fixed across i.

Wang et al. (2012) showed that if the marginal density of each outcome can be assumed to come from a canonical exponential family, then $\boldsymbol{S}(\boldsymbol{\beta})$ in (2.1) can be simplified to

$$\boldsymbol{S}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \boldsymbol{X_i}^T \boldsymbol{A_i}^{1/2}(\boldsymbol{\beta})\hat{\boldsymbol{R}}^{-1}\boldsymbol{A_i}^{-1/2}(\boldsymbol{\beta})(\boldsymbol{y_i} - \boldsymbol{\mu_i}(\boldsymbol{\beta})), \tag{2.3}$$

where $\boldsymbol{X_i}$ is the covariate matrix for the ith individual. In the bivariate mixed outcome case, $\boldsymbol{X_i}$ reduces to the block-diagonal structure

$$\boldsymbol{X_i} = \begin{pmatrix} \boldsymbol{x_i}^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{z_i}^T \end{pmatrix}. \tag{2.4}$$

$\hat{\boldsymbol{R}}$ is the estimated working correlation matrix, in which the association parameter ρ is replaced by an estimate $\hat{\rho}$. We compute $\hat{\rho}$ using the biserial correlation between the binary outcomes and the residuals of the continuous outcomes.

2.2.3 Penalized generalized estimating equations for bivariate mixed outcomes

A sparsity-inducing penalty term can be incorporated into (2.3) if we wish to perform simultaneous estimation and variable selection with the GEEs. The PGEEs for bivariate mixed outcomes are given as

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{S}(\boldsymbol{\beta}) - \boldsymbol{q}_{\boldsymbol{\lambda}}(|\boldsymbol{\beta}|)sign(\boldsymbol{\beta}), \tag{2.5}$$

where $\boldsymbol{S}(\boldsymbol{\beta})$ is defined in (2.3),

$$\boldsymbol{q}_{\boldsymbol{\lambda}}(|\boldsymbol{\beta}|) = [q_{\lambda_c}(\beta_{c1}), q_{\lambda_c}(\beta_{c2}), \ldots, q_{\lambda_c}(\beta_{cp}), q_{\lambda_b}(\beta_{b1}), q_{\lambda_b}(\beta_{b2}), \ldots, q_{\lambda_b}(\beta_{bq})]^T \tag{2.6}$$

is a $(p+q)$-dimensional vector of the first derivatives of penalty functions, where λ_c and λ_b are the tuning parameters for the penalty functions associated with continuous regression coefficients and binary regression coefficients, respectively, and

$$sign(\boldsymbol{\beta}) = [sign(\beta_{c1}), \ldots, sign(\beta_{cp}), sign(\beta_{b1}), \ldots, sign(\beta_{bq})]^T, \tag{2.7}$$

where $sign(t) = I(t > 0) - I(t < 0)$. Note that the product of $\boldsymbol{q}_{\boldsymbol{\lambda}}(\cdot)$ and $sign(\cdot)$ in (2.5) is component-wise. Unlike previous frameworks for the PGEEs such as in Johnson et al. (2008) and Wang et al. (2012), we require two tuning parameters λ_c and λ_b, because

the continuous and binary outcomes are on fundamentally different scales. Restricting the model to a single tuning parameter would necessarily lead to over-penalization or under-penalization in at least one component of $(\boldsymbol{\beta_c}, \boldsymbol{\beta_b})$.

Although a variety of sparsity-inducing penalties can be chosen in (2.5), we restrict our attention to the SCAD penalty $q_\lambda(\theta) = \lambda\{I(\theta \leq \lambda) + (a-1)^{-1}\lambda^{-1}(a\lambda - \theta)_+ I(\theta > \lambda)\}$, for $\theta \geq 0$ and for fixed $a > 2$, where $(t)_+ = \max(t, 0)$. We fix $a = 3.7$ as recommended in Fan and Li (2001).

2.2.4 Algorithm to solve PGEEs

Analogous to Wang et al. (2012), we use a Newton-Raphson type iterative scheme to solve PGEEs for bivariate mixed outcomes:

$$\hat{\boldsymbol{\beta}}^{k+1} = \hat{\boldsymbol{\beta}}^k + [\boldsymbol{H}(\hat{\boldsymbol{\beta}}^k) + \boldsymbol{E}(\hat{\boldsymbol{\beta}}^k)]^{-1}[\boldsymbol{S}(\hat{\boldsymbol{\beta}}^k) - \boldsymbol{E}(\hat{\boldsymbol{\beta}}^k)\hat{\boldsymbol{\beta}}^k], \tag{2.8}$$

where

$$\boldsymbol{H}(\hat{\boldsymbol{\beta}}^k) = \sum_{i=1}^{n} \boldsymbol{X_i}^T \boldsymbol{A_i}^{1/2}(\hat{\boldsymbol{\beta}}^k)\hat{\boldsymbol{R}}^{-1}\boldsymbol{A_i}^{1/2}(\hat{\boldsymbol{\beta}}^k)\boldsymbol{X_i}, \tag{2.9}$$

$$\boldsymbol{E}(\hat{\boldsymbol{\beta}}^k) = \text{diag}\left\{\frac{q_{\lambda_c}(|\hat{\beta}^k_{c1}|_+)}{\varepsilon + |\hat{\beta}^k_{c1}|}, \ldots, \frac{q_{\lambda_c}(|\hat{\beta}^k_{cp}|_+)}{\varepsilon + |\hat{\beta}^k_{cp}|}, \frac{q_{\lambda_b}(|\hat{\beta}^k_{b1}|_+)}{\varepsilon + |\hat{\beta}^k_{b1}|}, \ldots, \frac{q_{\lambda_b}(|\hat{\beta}^k_{bq}|_+)}{\varepsilon + |\hat{\beta}^k_{bq}|}\right\}, \tag{2.10}$$

where ε is a small fixed positive number, which we set to 10^{-6}. This algorithm has close connections to the local quadratic approximation algorithm of Fan and Li (2001)

and the minorization-maximization (MM) algorithm of Hunter and Li (2005) for solving penalized regression problems.

The two tuning parameters λ_c and λ_b are chosen using four-fold cross-validation over a two-dimensional grid. The loss function used for the cross-validation is the sum of a squared error loss for the estimated continuous regression coefficient vector $\widehat{\boldsymbol{\beta_c}}$:

$$L_c(\boldsymbol{y_c}, \widehat{\boldsymbol{\eta_c}}) = \sum_{i=1}^{n} (y_{ic} - \widehat{\eta_{ic}})^2, \tag{2.11}$$

where $\widehat{\eta_{ic}} = \boldsymbol{x_i}^T \widehat{\boldsymbol{\beta_c}}$, and a deviance loss for the estimated binary regression coefficient vector $\widehat{\boldsymbol{\beta_b}}$:

$$L_b(\boldsymbol{y_b}, \widehat{\boldsymbol{\eta_b}}) = \frac{1}{\log(2)} \sum_{i=1}^{n} \log[1 + \exp\{-2\widehat{\eta_{ib}}(2y_{ib} - 1)\}], \tag{2.12}$$

where $\widehat{\eta_{ib}} = \boldsymbol{z_i}^T \widehat{\boldsymbol{\beta_b}}$. Note that $y_{ib} \in \{0, 1\}$. Convergence of the algorithm is declared if two conditions are satisfied: $\|\hat{\boldsymbol{\beta}}^{k+1} - \hat{\boldsymbol{\beta}}^k\|_1 < 10^{-6}$ and $\|\boldsymbol{U}(\hat{\boldsymbol{\beta}}^{k+1})\|_1 < 10^{-6}$, where $\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{n} |\theta_i|$ is the L_1-norm of an n-dimensional vector $\boldsymbol{\theta}$, and $\boldsymbol{U}(\boldsymbol{\beta})$ are the penalized estimating functions from (2.5).

From the Newton-Raphson scheme, analogous to Wang et al. (2012), we can obtain the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$, given by

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) \approx [\boldsymbol{H}(\widehat{\boldsymbol{\beta}}) + \boldsymbol{E}(\widehat{\boldsymbol{\beta}})]^{-1} \boldsymbol{M}(\widehat{\boldsymbol{\beta}}) [\boldsymbol{H}(\widehat{\boldsymbol{\beta}}) + \boldsymbol{E}(\widehat{\boldsymbol{\beta}})]^{-1}, \tag{2.13}$$

where $\boldsymbol{H}$ and $\boldsymbol{E}$ are defined in (2.9) and (2.10), and

$$\boldsymbol{M}(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \boldsymbol{X}_{\boldsymbol{i}}^T \boldsymbol{A}_{\boldsymbol{i}}^{1/2}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{R}}^{-1} [\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\beta}}) \boldsymbol{\varepsilon}_i^T(\widehat{\boldsymbol{\beta}})] \widehat{\boldsymbol{R}}^{-1} \boldsymbol{A}_{\boldsymbol{i}}^{1/2}(\widehat{\boldsymbol{\beta}}) \boldsymbol{X}_{\boldsymbol{i}}, \tag{2.14}$$

with $\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}_{\boldsymbol{i}}^{-1/2}(\widehat{\boldsymbol{\beta}})(\boldsymbol{y_i} - \boldsymbol{\mu_i}(\widehat{\boldsymbol{\beta}}))$.

2.2.5 Controlling the false discovery rate

In this section, we propose a method to estimate and thus control the false discovery rate (FDR) in the PGEE setting by selecting appropriate values for the penalty parameters λ_c and λ_b. Breheny (2009) and Yi et al. (2015) proposed such a method to control the FDR for penalized linear regression and penalized logistic regression. We generalize this method to PGEEs for mixed outcomes.

The FDR can be expressed as

$$\text{FDR} = \frac{\text{E(F)}}{\text{S}}, \tag{2.15}$$

where S is the total number of covariates selected by the variable selection procedure and F is the number of false discoveries. Under sparsity-inducing penalty functions like SCAD, the jth covariate is selected if its regression coefficient β_j is estimated as non-zero, i.e., $\hat{\beta}_j \neq 0$. We shall say that the jth covariate is *null* if $\beta_j = 0$. Thus, a false discovery is a null covariate that is selected by the variable selection procedure. Note

that since F is unknown in practice, it is replaced with its expectation in (2.15).

Next, letting $\alpha_j = P(\hat{\beta}_j \neq 0 | \beta_j = 0)$ be the probability of making a false discovery on the jth covariate, the numerator of (2.15) can be estimated by

$$\widehat{\mathrm{E(F)}} = \sum_{j=1}^{J} \alpha_j, \tag{2.16}$$

where J is the number of covariates being considered in the variable selection procedure. This approach to estimating the FDR is conservative (overestimates the FDR), since the sum in (2.16) is over all covariates and not just the null covariates. However, we do not know which covariates are null in practice.

We rewrite the estimating functions of the unpenalized GEEs from (2.1) as

$$\boldsymbol{S}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{D_i^T V_i^{-1}} (\boldsymbol{y_i - \mu_i}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{W_i^T r_i} = n^{-1} \boldsymbol{W^T r},$$

where $\boldsymbol{W_i^T} = \boldsymbol{D_i^T V_i^{-1}}$, $\boldsymbol{r_i} = \boldsymbol{(y_i - \mu_i)}$, $\boldsymbol{W^T} = [\boldsymbol{W_1^T}, \cdots, \boldsymbol{W_n^T}]$, $\boldsymbol{r} = [\boldsymbol{r_1^T}, \cdots, \boldsymbol{r_n^T}]$. Note that each $\boldsymbol{r_i} = [r_{ic}, r_{ib}]^T$ is a 2-dimensional vector; hence $\boldsymbol{r}$ is a $2n$-dimensional vector. Denoting $\boldsymbol{w^{(j)}}$ as the jth column vector of $\boldsymbol{W}$, $j = 1, \cdots, (p+q)$, we can express the jth component of $\boldsymbol{S}(\boldsymbol{\beta})$ as $S_j(\boldsymbol{\beta}) = n^{-1} \boldsymbol{w^{(j)T} r}$.

Wolfson (2011) mentions that although estimating equations may not correspond to the gradient of some (unknown) loss function, they can be obtained as the modification of such a gradient, and can be expected to have similar behavior as the gradient. Hence,

at the solution, the Karush-Kuhn-Tucker optimality conditions should hold, which give the following conditions for PGEEs:

$$n^{-1}\boldsymbol{w}^{(j)T}\boldsymbol{r} = \lambda_j sign(\hat{\beta}_j) \qquad \forall \hat{\beta}_j \neq 0, \tag{2.17a}$$

$$n^{-1}|\boldsymbol{w}^{(j)T}\boldsymbol{r}| \leq \lambda_j \qquad \forall \hat{\beta}_j = 0, \tag{2.17b}$$

where λ_j is λ_c or λ_b, depending on whether β_j corresponds to the continuous outcomes or to the binary outcomes, respectively. Note that the conditions in (2.17) are derived assuming the LASSO penalty, but as mentioned in Breheny (2009), the same conditions can be applied to the SCAD penalty, which we use.

We show in Appendix A that the conditions in (2.17) further imply the conditions

$$n^{-1}|\boldsymbol{w}^{(j)T}\boldsymbol{r}^{(-j)}| > \lambda_j \qquad \forall \hat{\beta}_j \neq 0, \tag{2.18a}$$

$$n^{-1}|\boldsymbol{w}^{(j)T}\boldsymbol{r}^{(-j)}| \leq \lambda_j \qquad \forall \hat{\beta}_j = 0, \tag{2.18b}$$

where the $-j$ superscript indicates quantities calculated without using the jth covariate. Hence, we have

$$\alpha_j = P(\hat{\beta}_j \neq 0 | \beta_j = 0) = P(n^{-1}|\boldsymbol{w}^{(j)T}\boldsymbol{r}^{(-j)}| > \lambda_j | \beta_j = 0). \tag{2.19}$$

In general, the distribution of the $\boldsymbol{r}^{(-j)}$'s is complex, hence obtaining an analytical expression for (2.19) is difficult. However, analogous to Breheny (2009), we can make

an approximation:

$$\boldsymbol{r}^{(-j)} \overset{approx}{\sim} \boldsymbol{N}_{2n}(\boldsymbol{0}, \widetilde{\boldsymbol{V}}), \tag{2.20}$$

where $\widetilde{\boldsymbol{V}} = diag(\boldsymbol{V}, \ldots, \boldsymbol{V})$, and

$$\boldsymbol{V} = \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_b \\ \rho\sigma_c\sigma_b & \sigma_b^2 \end{pmatrix}, \tag{2.21}$$

where the variance parameters σ_c^2 and σ_b^2 can be estimated from the data as $\hat{\sigma_c^2} = n^{-1}\|\boldsymbol{r_c}\|_2^2$, $\hat{\sigma_b^2} = n^{-1}\|\boldsymbol{r_b}\|_2^2$, with $\boldsymbol{r_c} = [r_{c1}, \ldots, r_{cn}]^T$ and $\boldsymbol{r_b} = [r_{b1}, \ldots, r_{bn}]^T$. The association parameter ρ is already estimated from the algorithm that solves the PGEEs. Note that the block-diagonal structure of the variance-covariance matrix of $\boldsymbol{r}^{(-j)}$ from (2.20) reflects the assumption that the bivariate outcomes from a single individual are correlated, but outcomes between individuals are independent.

Using (2.20), we can approximate (2.19) as:

$$\widehat{\alpha_j} = 2\Phi\left(\frac{-n\lambda_j}{\sqrt{\boldsymbol{w}^{(j)T}\widetilde{\boldsymbol{V}}\boldsymbol{w}^{(j)}}}\right). \tag{2.22}$$

To estimate the total FDR across both continuous and binary outcomes, we can use (2.15), (2.16), and (2.22), with $J = p + q$. Alternatively, we can estimate the FDR

separately for the continuous and the binary outcomes using:

$$\widehat{\mathrm{FDR}_c} = \frac{\widehat{\mathrm{E}(\mathrm{F}_c)}}{\mathrm{S}_c}, \quad \widehat{\mathrm{E}(\mathrm{F}_c)} = \sum_{j=1}^{p} \widehat{\alpha}_j, \qquad \widehat{\mathrm{FDR}_b} = \frac{\widehat{\mathrm{E}(\mathrm{F}_b)}}{\mathrm{S}_b}, \quad \widehat{\mathrm{E}(\mathrm{F}_b)} = \sum_{j=p+1}^{p+q} \widehat{\alpha}_j, \quad (2.23)$$

where S_c is the total number of continuous outcome covariates selected and S_b is the total number of binary outcome covariates selected.

Note that in general, there will be multiple pairs of tuning parameters (λ_c, λ_b) that can control the FDR at a desired level. Hence, in practice, we choose λ_c and λ_b as the pair with the lowest cross-validated error amongst all pairs that control the FDR at the desired level.

2.3 Simulation studies

We conducted simulation studies to compare our method of modeling the bivariate outcomes jointly versus modeling each outcome separately using two unrelated PGEEs. We also conducted a simulation study to investigate the effectiveness of our FDR control method.

2.3.1 Data generation

Comparing the joint PGEE method versus the separate PGEEs method

We generated 100 data sets, each consisting of $n = 500$ pairs of correlated bivariate mixed outcomes, with $p = q = 50$ covariates per outcome. Marginally, the continuous outcomes follow normal distributions with the identity link to covariates and the binary outcomes follow Bernoulli distributions with the logit link to covariates. Denote the covariate matrices as $\boldsymbol{X} = [\boldsymbol{x^{(1)}}, \boldsymbol{x^{(2)}}, \ldots, \boldsymbol{x^{(p)}}]$ and $\boldsymbol{Z} = [\boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}, \ldots, \boldsymbol{z^{(q)}}]$ for the continuous and the bivariate responses, respectively, where $\boldsymbol{x^{(j)}}$ is the jth column of $\boldsymbol{X}$, and similarly for $\boldsymbol{Z}$. We assumed intercepts for both outcomes, so $\boldsymbol{x^{(1)}} = \boldsymbol{z^{(1)}} = \boldsymbol{1_n}$, whose coefficients are not penalized. Covariates were generated from a multivariate normal distribution with a zero-mean vector, unit marginal variances, and an AR(1) correlation structure with a correlation of 0.25. Three situations for these covariate matrices were considered: (i) All covariates are shared (between the bivariate outcomes), i.e., $\boldsymbol{X} = \boldsymbol{Z}$, (ii) Some but not all covariates are shared, in which case we set $\boldsymbol{z^{(j)}} = \boldsymbol{x^{(j)}}$ for $j = 2, 3$, and (iii) No covariates are shared, in which case we generated $\boldsymbol{X}$ and $\boldsymbol{Z}$ independently.

Next, the true regression coefficient vectors are chosen as $\boldsymbol{\beta_{0c}} = (0.2, 2.0, 0, \ldots, 0, 3.0, -1.5, 2.0)^T$ and $\boldsymbol{\beta_{0b}} = (1.2, 0.8, 0.6, -0.4, 0, \ldots, 0)^T$. This setup lets us consider the case that when the two covariate matrices are identical, exactly one of the covariates is associated with both of the outcomes, while all other covariates are associated with at

most one of the outcomes. The correlated bivariate mixed responses are then generated as follows. For $i = 1, \ldots, n$:

$$
\begin{aligned}
(u_i, v_i) &\sim C(\cdot, \cdot|\theta), \\
y_{ic} &\sim \Phi^{-1}(u_i|\mu = \boldsymbol{x_i}^T \boldsymbol{\beta_{0c}}, \sigma = 1), \\
w_i &\sim F^{-1}(v_i|\mu = \boldsymbol{z_i}^T \boldsymbol{\beta_{0b}}, s = 1), \\
y_{ib} &= I(w_i > 0),
\end{aligned}
$$

where $C(\cdot, \cdot|\theta)$ is a two-dimensional Gaussian copula with correlation parameter θ, $\Phi^{-1}(\cdot|\mu, \sigma)$ is the inverse cumulative distribution function of a normal distribution with mean μ and standard deviation σ, and $F^{-1}(\cdot|\mu, s)$ is the inverse cumulative distribution function of the logistic distribution with location μ and scale s. Thus θ, the parameter of the copula, is the correlation between the continuous outcome and the latent variable that generates the binary outcome. We feel that specifying a correlation between the continuous outcome y_{ic} and the latent logistic variable w_i is more natural than specifying a direct correlation between the continuous outcome y_{ic} and the binary outcome y_{ib}. Note that the copula parameter, θ, and the association parameter in the PGEEs, ρ, are different quantities. Finally, note that to generate the binary response with a logistic link, we used the fact that generating $y \sim Bernoulli(p = e^{\phi}/(1 + e^{\phi}))$ is equivalent to generating $w \sim Logistic(\mu = \phi, s = 1)$, $y = I(w > 0)$. We considered scenarios with $\theta = 0.2$, $\theta = 0.4$, $\theta = 0.6$, and $\theta = 0.8$, corresponding to varying strengths of association

between the continuous and binary outcomes.

FDR control

We generated 500 data sets of correlated bivariate mixed outcomes. For brevity, we only considered the case when all covariates are shared between the bivariate outcomes. The design of the simulation is largely the same as the one described above, with the exception that we set both the continuous and binary regression coefficients to $(1, -1, 1, -1, 1, -1, 0, \ldots, 0)$, as in Breheny (2009).

2.3.2 Simulation results

Comparing the joint PGEE method versus the separate PGEEs method

Here, we compare the joint and the separate PGEE methods in terms of accuracy and variable selection metrics. For each of the 100 data sets generated under each scenario, we applied our iterative algorithm to solve the PGEEs and obtained estimates of the regression coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\boldsymbol{c}}^T, \boldsymbol{\beta}_{\boldsymbol{b}}^T)^T$. As described in Section 2.2.4, the tuning parameters λ_c and λ_b were selected using four-fold cross-validation over a two-dimensional grid, equally spaced on the log scale. We also applied separate PGEEs to the continuous and the binary outcomes and estimated the regression coefficients. For each of the separate estimations, the tuning parameter was selected using four-fold cross-validation over a one-dimensional grid, equally spaced on the log scale. To evaluate the accuracy of these estimates, we computed the mean squared error (MSE) as $(100)^{-1} \sum_{i=1}^{100} \|\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}_0\|_2^2$,

Table 1: Accuracy and variable selection metrics comparing the joint and the separate PGEE methods, for $\theta = 0.2, 0.4, 0.6, 0.8$, with *all covariates shared* between the continuous and binary outcomes. Maximum TP is 9.0 and maximum FP is 91.

(a) $\theta = 0.2$

Method	MSE	U	O	E	TP	FP
Joint	0.2416	0.52	0.19	0.29	8.29	1.28
Separate	0.2485	0.51	0.24	0.25	8.28	1.43

(b) $\theta = 0.4$

Method	MSE	U	O	E	TP	FP
Joint	0.1990	0.41	0.33	0.26	8.47	2.52
Separate	0.2227	0.45	0.27	0.28	8.39	1.48

(c) $\theta = 0.6$

Method	MSE	U	O	E	TP	FP
Joint	0.1962	0.40	0.24	0.36	8.48	1.77
Separate	0.2588	0.46	0.28	0.26	8.31	1.87

(d) $\theta = 0.8$

Method	MSE	U	O	E	TP	FP
Joint	0.1662	0.26	0.46	0.28	8.66	2.98
Separate	0.2172	0.46	0.27	0.27	8.41	1.83

where $\hat{\boldsymbol{\beta}}^{(i)}$ is the estimate for the true regression coefficient vector $\boldsymbol{\beta}_0$ from the ith data set. We also computed the absolute bias and the sandwich-formula based standard error for each true non-zero regression coefficient. To compare performance in variable selection, we computed the proportion of data sets in which the methods under-selected (U), over-selected (O) and exactly selected (E) the covariates with true non-zero regression coefficients. (A good variable selection method should have small U and O metrics, and

Table 2: Accuracy and variable selection metrics the comparing the joint and the separate PGEE methods, for $\theta = 0.2, 0.4, 0.6, 0.8$, with *some covariates shared* between the continuous and the binary outcomes. Maximum TP is 9.0 and maximum FP is 91.

(a) $\theta = 0.2$

Method	MSE	U	O	E	TP	FP
Joint	0.1947	0.50	0.16	0.34	8.43	0.88
Separate	0.2120	0.57	0.16	0.27	8.35	1.09

(b) $\theta = 0.4$

Method	MSE	U	O	E	TP	FP
Joint	0.1595	0.45	0.16	0.39	8.51	0.84
Separate	0.1774	0.52	0.16	0.32	8.43	0.91

(c) $\theta = 0.6$

Method	MSE	U	O	E	TP	FP
Joint	0.1576	0.41	0.29	0.30	8.54	1.28
Separate	0.1982	0.51	0.23	0.26	8.41	1.17

(d) $\theta = 0.8$

Method	MSE	U	O	E	TP	FP
Joint	0.1271	0.28	0.31	0.41	8.69	1.78
Separate	0.1574	0.42	0.20	0.38	8.54	1.07

a large E metric). Finally, we calculated the average number of true positives per data set (TP) and the average number of false positives per data set (FP) for both the methods. Table 1 shows the MSE and variable selection metrics for the joint and the separate methods, where all covariates are shared between the bivariate outcomes. We observe that the joint method has smaller MSE than the separate method, with larger gains for the larger values of θ. Under-selection is usually considered worse than over-selection in

Table 3: Accuracy and variable selection metrics comparing the joint and the separate PGEE methods, for $\theta = 0.2, 0.4, 0.6, 0.8$, with *no covariates shared* between the continuous and the binary outcomes. Maximum TP is 9.0 and maximum FP is 91.

(a) $\theta = 0.2$

Method	MSE	U	O	E	TP	FP
Joint	0.2505	0.52	0.14	0.34	8.26	0.95
Separate	0.2475	0.51	0.15	0.34	8.28	1.16

(b) $\theta = 0.4$

Method	MSE	U	O	E	TP	FP
Joint	0.1891	0.40	0.21	0.39	8.48	1.28
Separate	0.2217	0.45	0.21	0.34	8.39	1.15

(c) $\theta = 0.6$

Method	MSE	U	O	E	TP	FP
Joint	0.1685	0.36	0.29	0.35	8.54	1.33
Separate	0.2160	0.46	0.21	0.33	8.41	0.99

(d) $\theta = 0.8$

Method	MSE	U	O	E	TP	FP
Joint	0.1226	0.20	0.40	0.40	8.75	1.69
Separate	0.2160	0.46	0.21	0.33	8.41	0.99

variable selection, and we observe that the joint method has a smaller U metric than the separate method for $\theta = 0.4, 0.6, 0.8$, while its U metric for $\theta = 0.2$ is almost equal to that of the separate method. The joint method also has a larger E metric than the separate method for $\theta = 0.2$ and $\theta = 0.6$, and a similar E metric to the separate method for $\theta = 0.4$ and $\theta = 0.8$. The joint method has a uniformly larger TP metric than the separate method. The tradeoff to this gain is a slightly larger FP metric for the joint

method, for $\theta = 0.4$ and $\theta = 0.8$. Table 2 shows the metrics for the scenario where some, but not all covariates are shared between the bivariate outcomes. The joint method has smaller MSE, smaller U, larger E, and larger TP than the separate method. It has smaller FP for $\theta = 0.2$ and $\theta = 0.4$, but has larger FP for the larger values of θ. Table 3 shows the metrics for the scenario where no covariates are shared between the bivariate outcomes. The $\theta = 0.2$ setting under this scenario is the only subcase where the separate method is generally superior than the joint method. However, for larger values of θ, we see similar trends as described previously.

Table 4 shows the absolute bias and the standard errors for the true non-zero coefficients for the scenarios corresponding to Table 1. We observe that the absolute bias and the standard errors under both methods are similar for $\theta = 0.2$, while the absolute bias is smaller for most of the binary outcome coefficients for $\theta = 0.6$. For the same covariate setting, but with $\theta = 0.4$ and $\theta = 0.8$, the standard errors are usually smaller for the joint method. For the other scenarios considered (Table 5 and Table 6), the joint and the separate methods perform comparably in terms of absolute bias and standard error for the continuous outcome coefficients, but the joint method is generally superior to the separate method for the binary outcome coefficients.

Overall, we see that the joint method makes gains over the separate method in estimation and variable selection metrics for the binary outcome coefficients, especially

Table 4: Absolute bias (AB) and sandwich-formula based standard errors (SE) of estimates of true non-zero regression coefficients for the joint and the separate PGEE methods, with *all covariates shared* between the continuous and binary outcomes.

(a) $\theta = 0.2$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.002	0.004	0.006	0.004	0.005	0.009	0.035	0.138	0.167
	Separate	0.002	0.004	0.005	0.005	0.003	0.008	0.036	0.144	0.164
SE	Joint	0.044	0.044	0.046	0.047	0.045	0.122	0.128	0.110	0.082
	Separate	0.044	0.044	0.046	0.047	0.046	0.122	0.128	0.110	0.084

(b) $\theta = 0.4$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.001	0.004	0.006	0.007	0.007	0.017	0.029	0.091	0.129
	Separate	0.002	0.004	0.005	0.005	0.002	0.014	0.028	0.108	0.140
SE	Joint	0.044	0.045	0.042	0.043	0.042	0.121	0.126	0.105	0.082
	Separate	0.044	0.045	0.046	0.047	0.046	0.122	0.128	0.113	0.088

(c) $\theta = 0.6$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.002	0.004	0.006	0.005	0.006	0.017	0.028	0.093	0.122
	Separate	0.001	0.003	0.005	0.006	0.001	0.013	0.039	0.135	0.138
SE	Joint	0.044	0.044	0.044	0.045	0.044	0.122	0.127	0.111	0.088
	Separate	0.044	0.044	0.046	0.047	0.046	0.122	0.128	0.110	0.088

(d) $\theta = 0.8$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.001	0.003	0.005	0.007	0.007	0.018	0.033	0.066	0.066
	Separate	0.001	0.003	0.005	0.007	0.001	0.015	0.034	0.102	0.130
SE	Joint	0.045	0.045	0.039	0.040	0.039	0.121	0.124	0.098	0.085
	Separate	0.044	0.044	0.046	0.047	0.046	0.122	0.128	0.114	0.090

for larger values of θ. Intuitively, this makes sense, as the binary outcome coefficients– which are harder to estimate due to the smaller information content of binary outcomes– benefit from *borrowing information* from the continuous outcomes. The benefit increases as the strength of association between the outcomes increases.

FDR control

For each of 500 data sets generated, we noted the estimated FDR and the true FDR over a 30×30 grid of (λ_c, λ_b) values. The smoothed average of these FDRs for $\theta = 0.2, 0.4, 0.6, 0.8$ are plotted in the contour plots in Figure 1. Each level of the contour plots shows the various (λ_c, λ_b) combinations which result in the same FDR.

For a particular level of the FDR, our method controls the FDR if the estimated FDR contour lies "above" the true FDR contour. We observe that for all the scenarios, our method is able to control the FDR, albeit a little conservatively for larger values of θ. For larger values of θ, in the lower right corner of the figures, we also observe that the estimated FDR contours curve upward. This means that for a fixed value of λ_b, if λ_c increases beyond a certain point, the estimated FDR increases. This phenomenon occurs because for overly large values of λ_c, the number of false discoveries in the *binary* coefficients increases, due to the correlation between the outcomes. In practice, this is not a concern, because we choose the optimal (λ_c, λ_b) pair using cross-validation, and our simulation results indicate that the optimal pair of tuning parameters usually does not lie in the lower right corner.

Table 5: Absolute bias (AB) and sandwich-formula based standard errors (SE) of estimates of true non-zero regression coefficients for the joint and the separate PGEE methods, with *some covariates shared* between the continuous and the binary outcomes.

(a) $\theta = 0.2$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.002	0.003	0.003	0.002	0.008	0.008	0.031	0.043	0.143
	Separate	0.002	0.003	0.004	0.002	0.009	0.005	0.030	0.043	0.176
SE	Joint	0.044	0.045	0.045	0.047	0.045	0.121	0.125	0.112	0.084
	Separate	0.044	0.045	0.046	0.047	0.046	0.121	0.125	0.113	0.079

(b) $\theta = 0.4$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.001	0.004	0.006	0.007	0.007	0.017	0.029	0.091	0.129
	Separate	0.001	0.004	0.005	0.006	0.001	0.013	0.039	0.135	0.139
SE	Joint	0.044	0.045	0.042	0.043	0.042	0.121	0.126	0.105	0.082
	Separate	0.044	0.044	0.046	0.047	0.046	0.122	0.128	0.110	0.088

(c) $\theta = 0.6$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.002	0.001	0.003	0.003	0.005	0.004	0.013	0.031	0.126
	Separate	0.001	0.001	0.004	0.000	0.009	0.004	0.013	0.041	0.162
SE	Joint	0.045	0.045	0.042	0.044	0.042	0.120	0.122	0.104	0.081
	Separate	0.044	0.045	0.046	0.048	0.046	0.120	0.124	0.113	0.081

(d) $\theta = 0.8$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.002	0.000	0.003	0.005	0.002	0.003	0.017	0.034	0.095
	Separate	0.001	0.001	0.004	0.001	0.008	0.007	0.021	0.021	0.118
SE	Joint	0.045	0.046	0.039	0.040	0.039	0.119	0.120	0.096	0.078
	Separate	0.044	0.045	0.046	0.048	0.046	0.121	0.124	0.117	0.089

Table 6: Absolute bias (AB) and sandwich-formula based standard errors (SE) of estimates of true non-zero regression coefficients for the joint and the separate PGEE methods, with *no covariates shared* between the continuous and the binary outcomes.

(a) $\theta = 0.2$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.001	0.002	0.001	0.008	0.003	0.007	0.033	0.149	0.167
	Separate	0.002	0.002	0.001	0.008	0.003	0.008	0.036	0.144	0.164
SE	Joint	0.044	0.044	0.046	0.047	0.046	0.121	0.126	0.108	0.082
	Separate	0.044	0.045	0.046	0.047	0.046	0.122	0.128	0.110	0.084

(b) $\theta = 0.4$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.001	0.002	0.000	0.008	0.003	0.014	0.022	0.094	0.124
	Separate	0.001	0.001	0.000	0.008	0.002	0.014	0.028	0.107	0.140
SE	Joint	0.044	0.043	0.045	0.045	0.044	0.121	0.123	0.111	0.088
	Separate	0.044	0.045	0.046	0.047	0.046	0.122	0.128	0.113	0.088

(c) $\theta = 0.6$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.001	0.001	0.001	0.008	0.002	0.015	0.021	0.077	0.112
	Separate	0.001	0.000	0.000	0.008	0.002	0.013	0.039	0.135	0.139
SE	Joint	0.044	0.041	0.042	0.043	0.042	0.121	0.116	0.107	0.085
	Separate	0.044	0.045	0.046	0.047	0.046	0.122	0.128	0.110	0.088

(d) $\theta = 0.8$

		Continuous Outcome					Binary Outcome			
Metric	Method	β_1	β_2	β_3	β_4	β_5	β_{51}	β_{52}	β_{53}	β_{54}
AB	Joint	0.000	0.003	0.004	0.007	0.001	0.017	0.023	0.050	0.054
	Separate	0.000	0.001	0.001	0.008	0.002	0.015	0.034	0.102	0.130
SE	Joint	0.045	0.038	0.039	0.040	0.039	0.120	0.107	0.100	0.086
	Separate	0.044	0.045	0.046	0.047	0.046	0.122	0.128	0.114	0.090

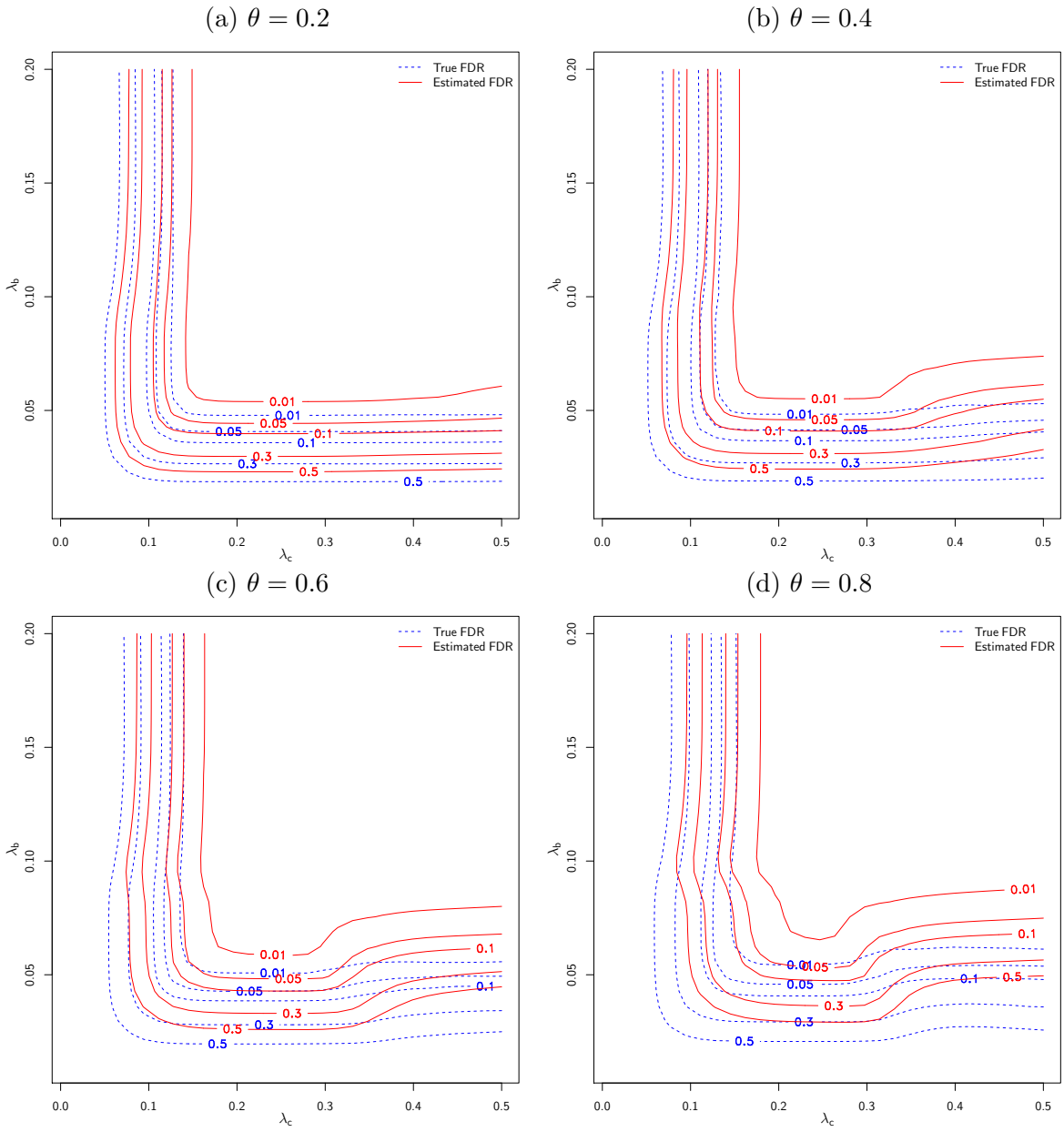

Figure 1: Contour plots of smoothed true and estimated FDRs. The continuous penalty parameter is on the horizontal axis and the binary penalty parameter is on the vertical axis. Each contour shows the combination of penalty parameters that result in the same true/estimated FDR.

2.4 MEPS data analysis

In this section, we demonstrate the application of our PGEE framework and FDR control methodology to data from the Medical Expenditure Panel Survey (MEPS) (`https://meps.ahrq.gov/`). Our goal is to identify covariates on demographics, medical conditions, income, employment, health insurance coverage, and access to care that are significantly associated with total annual drug spending and health status. We used the 2005 data and restricted attention to Medicare enrollees, 65 years of age and older, with an annual drug spending of $100 or more. We used the natural logarithm of total drug spending as our continuous outcome. As done in Zimmerman (2013), we dichotomized health status into *fair or poor* (1) and *better than fair* (0), which formed our binary outcome. We considered a total of 40 covariates, and we used the same set of covariates to model both total drug spending and health status. The complete list of covariates with descriptions can be found in Table 7. The data set also provides sampling weights for each observation, which we incorporated into the estimation methods. The final data set contains data for 2,953 individuals, who represent 30,146,029 individuals of the U.S. population. We applied both our joint PGEE method as well as separate PGEEs to the responses. Similar to the simulations, four-fold cross-validation was used to select the optimal tuning parameters. Table 8 shows the estimated regression coefficients under the joint method and under the separate method. Sandwich-formula based standard errors for the regression coefficients from the joint model can be found

in Table 9. For the continuous outcome–the logarithm of total drug spending–we observe that the joint and separate methods perform similarly in terms of both variable selection and estimation. For the joint model, the covariates with the largest coefficients are CARDIOVASCULAR, DIABETES, and MENTAL, which are binary indicators for the presence of a cardiovascular disease, some form of diabetes, and a mental disease, respectively. Intuitively this makes sense, as pre-existing medical conditions should have strong associations with drug spending.

For the binary outcome–the indicator of *fair or poor* health status–the joint model is able to detect signal from more covariates than the separate model. This is consistent with the results from our simulation studies, in which the gains in variable selection metrics through joint modeling are primarily made for the binary outcome coefficients. Of course, false discoveries could be a concern here. Hence, we estimated the FDR using the method described in Section 2.2.5 and found it to be 0.07. Interestingly, among all covariates selected by the joint method for the binary outcome, LANG_ENG has the largest coefficient in absolute value. The negative coefficient indicates that individuals whose language of comfort is English report better health status than other individuals. The moderate positive coefficient of DIFF_USC_TRAVEL indicates that individuals who find it difficult to travel to their Usual Source of Care (USC) provider report worse health statuses. Both the joint model and the separate model emphasize the importance of regular physical activity to good health, as seen in the large negative coefficient of PHYACT. In the joint model, the effect of dental health on health status can be seen

via the coefficients of DNUNAB (individual was unable to receive dental treatment when it was required) and DENTCK_LESS_ONEYR (frequency of dental checkups are less than once a year). Next, income and employment are positively associated with good health as seen through the negative coefficients of LN_INCOME and EMPLOYED. Interestingly, other than DIABETES, most of the variables related to prior medical conditions have relatively small coefficients. SEATBELT_NOT_ALWAYS has a moderate positive coefficient, indicating that some individuals may have experienced poor health status due to a motor vehicle accident.

Finally, our joint method estimated the association parameter, ρ, to be 0.13. Our simulations indicate that the difference between the copula parameter, θ, and ρ, is roughly 0.10, so θ may be roughly regarded as 0.23.

Table 7: Covariates used in analysis of MEPS data

Name	Type	Description
AGEX	Continuous	Age as of 12/31/2005
SEX	Binary	Male?
RACE_WHITE	Binary	White?
MARRIED	Binary	Currently married?
LN_INCOME	Continuous	Shifted natural logarithm of income
LOW_INC_FAM	Binary	Low income family?
LANG_ENG	Binary	Primary language spoken at home English?
TMTK_MORE_ONEHR	Binary	Takes > 1 hour to travel to USC?
DIFF_USC_TRAVEL	Binary	Difficult to travel to USC?
DIFF_USC_PHONE	Binary	Difficult to reach USC by phone?
MDUNAB	Binary	Did not receive medical treatment?
DNUNAB	Binary	Did not receive dental treatment?
PMUNAB	Binary	Did not receive prescription medication?
MDDLAY	Binary	Delay in receiving medical treatment?
DNDLAY	Binary	Delay in receiving dental treatment?
PMDLAY	Binary	Delay in receiving prescription medication?
MCDEV	Binary	Covered by Medicaid?
PRVEV	Binary	Covered by private insurance?
TRIEV	Binary	Covered by TRICARE?
DENTCK_LESS_ONEYR	Binary	Frequency of dental checkups < 1/year?
CHOLCK_MORE_5YR	Binary	> 5 years since last blood cholesterol check?
CHECK_MORE_1YR	Binary	> 1 year since routing medical checkup?
FLUSHT_MORE_1YR	Binary	> 1 year since last flu shot?
NOTEETH	Binary	Lost all natural teeth?
STOOL	Binary	Has had a blood stool test?
BOWEL	Binary	Has had sigmoidoscopy/colonoscopy?
PHYACT	Binary	Mod./vig. physical activity $>= 3$/week?
BMI	Continuous	Body Mass Index
SEATBELT_NOT_ALWAYS	Binary	Does not always wear a seatbelt?
CANCER	Binary	Has some form of cancer?
DIABETES	Binary	Has some form of diabetes?
COPD	Binary	Has chronic obstructive pulmonary disease?
CARDIOVASCULAR	Binary	Has a cardiovascular disease?
ARTHRITIS	Binary	Has arthritis?
ASTHAMA	Binary	Has asthma?
STOMACH_ULCERS	Binary	Has stomach ulcers?
MENTAL	Binary	Has a mental disease?
KIDNEY	Binary	Has a renal disease?
PRIO	Count	Number of “priority conditions”
EMPLOYED	Binary	Currently employed?

Table 8: Estimated regression coefficients for log(drug spending) and health status outcomes under the joint and the separate PGEE methods. A dot indicates that the covariate was not selected by that method, for that outcome. Covariates that are not selected by either method are not shown.

	log(drug spending)		health status: fair or poor	
Covariate	Joint method	Sep. method	Joint method	Sep. method
Intercept	6.330	6.274	5.346	0.053
SEX	-0.057	-0.065	0.259	.
RACE_WHITE	.	0.007	-0.387	.
MARRIED	-0.017	-0.021	.	.
LN_INCOME	.	.	-0.489	-0.004
LANG_ENG	.	.	-1.017	.
TMTK_MORE_ONEHR	.	-0.004	.	.
DIFF_USC_TRAVEL	.	.	0.541	.
MDUNAB	-0.127	-0.159	.	.
DNUNAB	.	.	0.891	.
MDDLAY	-0.002	.	.	.
MCDEV	0.028	0.049	.	.
PRVEV	.	.	-0.273	.
DENTCK_LESS_ONEYR	.	.	0.556	.
CHECK_MORE_1YR	-0.105	-0.132	.	.
FLUSHT_MORE_1YR	-0.111	-0.125	.	.
NOTEETH	0.006	0.009	.	.
BOWEL	0.044	0.051	.	.
PHYACT	-0.060	-0.066	-0.893	-1.156
BMI	0.007	0.009	.	.
SEATBELT_NOT_ALWAYS	.	.	0.573	.
CANCER	-0.001	-0.009	.	.
DIABETES	0.412	0.407	0.624	.
COPD	0.060	0.087	0.034	.
CARDIOVASCULAR	0.418	0.425	.	.
ASTHAMA	0.155	0.201	.	.
MENTAL	0.368	0.365	0.383	.
KIDNEY	0.090	0.124	0.020	.
PRIO	0.090	0.087	0.176	0.019
EMPLOYED	-0.033	-0.038	-0.646	.

Table 9: Estimated regression coefficients and sandwich formula-based standard errors for covariates selected by the joint method in the MEPS data analysis. A dot in the Coefficient column indicates that the covariate was not selected for that outcome.

	log(drug spending)		health status: fair or poor	
Covariate	Coefficient	Standard error	Coefficient	Standard Error
SEX	-0.057	0.017	0.259	0.090
RACE_WHITE	.	.	-0.387	0.125
MARRIED	-0.017	0.008	.	.
LN_INCOME	.	.	-0.489	0.118
LANG_ENG	.	.	-1.017	0.192
DIFF_USC_TRAVEL	.	.	0.541	0.159
MDUNAB	-0.127	0.058	.	.
DNUNAB	.	.	0.891	0.271
MDDLAY	-0.002	0.001	.	.
MCDEV	0.028	0.011	.	.
PRVEV	.	.	-0.273	0.091
DENTCK_LESS_ONEYR	.	.	0.556	0.089
CHECK_MORE_1YR	-0.105	0.028	.	.
FLUSHT_MORE_1YR	-0.111	0.025	.	.
NOTEETH	0.006	0.003	.	.
BOWEL	0.044	0.015	.	.
PHYACT	-0.060	0.017	-0.893	0.089
BMI	0.007	0.002	.	.
SEATBELT_NOT_ALWAYS	.	.	0.573	0.120
CANCER	-0.001	0.001	.	.
DIABETES	0.412	0.038	0.624	0.107
COPD	0.060	0.022	0.034	0.014
CARDIOVASCULAR	0.418	0.040	.	.
ASTHAMA	0.155	0.038	.	.
MENTAL	0.368	0.039	0.383	0.106
KIDNEY	0.090	0.026	0.020	0.008
PRIO	0.090	0.008	0.176	0.023
EMPLOYED	-0.033	0.014	-0.646	0.139

Chapter 3

Fully and empirical Bayes approaches to estimating copula-based models for bivariate mixed outcomes using Hamiltonian Monte Carlo

3.1 Introduction

As noted in Chapter 1, when measuring the association between correlated outcomes is of primary interest, copula-based models provide a better alternative to GEEs. However, care must be taken when specifying the copula structure for mixed outcomes. Sklar's Theorem (Sklar, 1959) ensures the uniqueness of a copula only when the marginals are

continuous. In the burn injury data of Fan and Gijbels (1996), it is of interest to assess the association between *total burn area* (continuous outcome) and *survival status* (binary outcome, either dead or survived). When any of the marginals are discrete, the copula is uniquely defined only on the Cartesian product of the ranges of the marginals. Another consequence of having discrete marginals is that dependence measures such as Kendall's Tau and Spearman's Rho may now be restricted by the marginal distributions. See Genest and Nešlehová (2007) for more details on the limitations of using copulas with discrete marginals. Song et al. (2009) developed a regression framework for bivariate mixed outcomes using Gaussian copulas and generalized linear models (GLMs) as marginal models. As they applied the copula directly on discrete marginals, their model suffers from the problems noted above. To avoid these problems, de Leon and Wu (2011) used a latent variable formulation, in which a continuous latent variable is dichotomized to generate the binary outcome. Under this approach, the copula is specified on the continuous outcome and the continuous latent variable, thus avoiding the issues accompanying copulas with discrete margins.

The methods from Song et al. (2009) and de Leon and Wu (2011) are frequentist approaches that use maximum likelihood based techniques for parameter estimation. Indeed, most applications of copulas involve frequentist estimation. However recent advances in Markov Chain Monte Carlo (MCMC) sampling algorithms have enabled Bayesian methods for copula estimation to gain traction. See Silva and Lopes (2008) for an overview of Bayesian copula model estimation and model selection. In the case of

mixed outcomes, Smith and Khaled (2012) provided a Bayesian framework for copula estimation of mixed outcomes based on data augmentation with latent variables, and provided sampling schemes to perform inference using MCMC. Craiu and Sabeti (2012) extended the latent variable formulation of de Leon and Wu (2011) to allow the copula parameter to vary as a function of the covariates using splines, and provided an adaptive MCMC scheme for sampling and inference.

The MCMC sampling algorithms of these methods are based on the Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) and Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) algorithms. Due to the random walk nature of these algorithms, they often take an unacceptably long time to converge to the target distribution (Neal, 1993). On the other hand, Hamiltonian Monte Carlo (HMC) (Neal, 2011; Duane et al., 1987) avoids random walk behavior by introducing auxiliary variables to simulate Hamiltonian dynamics. This enables HMC to produce distant proposals with high acceptance probability, resulting in fast exploration of the target density. The increased efficiency of HMC comes with two additional costs. First, HMC requires computation of the gradient of the log-density of the target distribution. Second, HMC's performance is highly sensitive to the choice of at least two tuning parameters: the step size ϵ and the number of steps L. Fortunately, the Stan software (Carpenter et al., 2016) efficiently computes gradients using automatic differentiation (Griewank and Walther, 2008), and finds optimal values for the tuning parameters based on the No-U-turn Sampler (NUTS) (Hoffman and Gelman, 2014). With Stan, we only need

to specify the Bayesian model in Stan's modeling language, and the software returns samples from the target density.

As noted in Silva and Lopes (2008), one aspect of Bayesian estimation of copula models involves specifying a prior for the copula parameter. Specifying a noninformative prior for the copula parameter is challenging. The main problem here is that a prior that is non-informative on the space of the copula parameter may be strongly informative on the space of Kendall's Tau. Also, an appropriate prior distribution may not be obvious, given the support of the copula parameter. For example, the Clayton copula's parameter is $(-1, \infty) \setminus \{0\}$, and the choice of prior distribution here is not obvious. A possible solution for this second issue may be to transform the copula parameter to bring it's support to the real line, and specify a diffuse prior on the transformed parameter. However, this still induces a strongly informative prior on Kendall's Tau. Specifying a uniform prior on Kendall's Tau simply reverses the problem. Furthermore, an expression for Kendall's Tau is not available in closed form for some copula families, which results in an increased computational burden during MCMC sampling. For these reasons, finding a way to circumvent the specification of a prior for the copula parameter may be of interest. Recently, Roy et al. (2016) proposed an empirical Bayes approach for estimating parameters in Bayesian models for which prior specification is difficult, and showed that their empirical Bayes approach resulted in parameter estimates with less bias than a corresponding fully Bayesian approach. It is of interest to compare the performance of the empirical Bayes approach and the fully Bayesian approach for

correlated mixed outcomes.

In this chapter, we propose fully Bayesian HMC-based estimation of a copula model for bivariate mixed outcomes. We also extend the empirical Bayes approach to estimate the copula parameter for our copula model. Section 3.2 provides details on the aforementioned copula model for bivariate mixed outcomes, and the estimation methods for the model are explained in Section 3.3. Section 3.4 provides results from simulation studies that compare the performance of the fully Bayesian and the empirical Bayes approaches, as well as assess the ability of the fully Bayesian method to select the correct copula family. In Section 3.5 we illustrate the application of the fully Bayesian method on the burn injury data set.

3.2 Models and notations

From the ith individual, we observe a continuous outcome y_{ic}, a binary outcome y_{ib}, a p-dimensional covariate vector $\boldsymbol{x_i}$ corresponding to the continuous outcome y_{ic}, and a q-dimensional covariate vector $\boldsymbol{z_i}$ corresponding to the binary outcome y_{ib}, $\quad i = 1 \ldots n$. It is common to assume $\boldsymbol{x_i} = \boldsymbol{z_i}$ (i.e., use the same set of covariates to model both outcomes), but this need not be so. Let $\boldsymbol{y_i} = (y_{ic}, y_{ib})^T$ denote the bivariate vector of outcomes from the ith individual. We assume that outcomes from the same individual are correlated, but outcomes from different individuals are independent.

We shall assume that the continuous outcome marginally follows a normal distribution with the identity link to its covariates and that the binary outcome follows a Bernoulli distribution with the logit link to its covariates. To model the dependency between the outcomes, we shall assume that a copula specifies the dependency between the continuous outcome and a continuous latent variable that generates the binary outcome. The generative model for the data can be described as follows. For $i = 1, \ldots, n$:

$$
\begin{aligned}
(u_i, v_i) &\sim C(\cdot, \cdot|\theta), \\
y_{ic} &= F_c^{-1}(u_i|\mu = \boldsymbol{x_i}^T\boldsymbol{\beta_c}, \sigma = \sigma_c), \\
w_i &= F_w^{-1}(v_i|\mu = \boldsymbol{z_i}^T\boldsymbol{\beta_b}, s = 1), \\
y_{ib} &= I(w_i > 0),
\end{aligned}
$$

where $C(\cdot, \cdot|\theta)$ is a two-dimensional copula with parameter θ, $F_c^{-1}(\cdot|\mu, \sigma) \equiv \Phi^{-1}(\cdot|\mu, \sigma)$ is the inverse cumulative distribution function of the normal distribution with mean μ and standard deviation σ, and $F_w^{-1}(\cdot|\mu, s)$ is the inverse cumulative distribution function of the logistic distribution with location μ and scale s. Also, $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ are the continuous latent variables that generate the observed binary outcomes $\boldsymbol{y_b}$. Note that generating a Bernoulli random variable with logit link is equivalent to generating a random variable following the logistic distribution and dichotomizing it.

The use of a continuous latent variable ensures a unique copula structure (Braeken et al., 2007). The choice of the logistic distribution for the latent variable results in a

model similar to the logit copula model from Nikoloulopoulos and Karlis (2008). We could also have chosen other distributions for the latent variable. For example, choosing a normal distribution results in a probit-binary model similar to Catalano and Ryan (1992), while choosing a generalized t-distribution (Kotz and Nadarajah, 2004) leads to a robit regression model (Liu, 2004) for the binary outcome.

3.3 Estimation methods and model selection

3.3.1 Fully Bayesian approach

Let $\boldsymbol{\psi} = (\boldsymbol{\beta_c}, \boldsymbol{\beta_b}, \sigma_c)^T$ denote the marginal parameters. To conduct estimation and inference on $(\boldsymbol{\psi}, \theta)$, we would like to draw samples from the posterior density

$$\pi(\boldsymbol{\psi}, \theta | \boldsymbol{y_c}, \boldsymbol{y_b}) \propto f(\boldsymbol{y_c}, \boldsymbol{y_b} | \boldsymbol{\psi}, \theta) h(\boldsymbol{\psi}) h(\theta). \tag{3.1}$$

Craiu and Sabeti (2012) derived the contribution of the ith observation to the likelihood for this model:

$$\begin{aligned} f(y_{ic}, y_{ib} | \boldsymbol{\psi}, \theta) =& \frac{1}{\sigma_c} \phi \left(\frac{y_{ic} - \boldsymbol{x_i}^T \boldsymbol{\beta_c}}{\sigma_c} \right) \times \left[c^{(0,1)} \left\{ \frac{1}{1 + \exp(\boldsymbol{x_i}^T \boldsymbol{\beta_c})}, \Phi \left(\frac{y_{ic} - \boldsymbol{x_i}^T \boldsymbol{\beta_c}}{\sigma_c} \right) \middle| \theta \right\} \right]^{1 - y_{ib}} \\ & \times \left[1 - c^{(0,1)} \left\{ \frac{1}{1 + \exp(\boldsymbol{x_i}^T \boldsymbol{\beta_c})}, \Phi \left(\frac{y_{ic} - \boldsymbol{x_i}^T \boldsymbol{\beta_c}}{\sigma_c} \right) \middle| \theta \right\} \right]^{y_{ib}}, \end{aligned} \tag{3.2}$$

where $c^{(a,b)}(u, v|\theta) = \partial^{a+b}C(u, v|\theta)/\partial u^a \partial v^b$, for $0 \leq a, b \leq 1$. Schepsmeier and Stöber (2014) provide closed form expressions for $c^{(a,b)}(\cdot, \cdot|\theta)$ for a number of common copula families.

The model specification is completed by specifying the prior distributions $h(\boldsymbol{\psi})$ and $h(\theta)$. We specify diffuse normal priors for the regression parameters $\boldsymbol{\beta_c}$ and $\boldsymbol{\beta_b}$, and a half-Cauchy prior for σ_c, as recommended in Gelman (2006). As for the copula parameter θ, there are a few ways to specify the prior. One way is to not specify a prior on θ directly, but rather on Kendall's Tau τ, which in turn will induce a prior on θ. However, as mentioned previously, unless an expression for Kendall's Tau (as a function of the copula parameter) exists in closed form, the computational burden significantly increases. Alternatively, we could specify priors on θ directly, while taking care that the support of the prior distribution matches the support of θ. Silva and Lopes (2008) suggest the uniform distribution for the Gaussian copula parameter, and a gamma distribution with mean one and variance 10^6 for the Clayton copula parameter. Yet another way is to apply a monotone transformation $g(\cdot)$ to θ to bring the support of the transformed parameter $g(\theta)$ to the real line, and then specify a prior on $g(\theta)$. In our simulations and real data analysis, we adopt this approach. Specifically, for the Gaussian, Clayton, Gumbel, and Frank copulas, we use the transformations $g(\theta) = \tanh^{-1}(\theta)$, $g(\theta) = \log(1 + \theta)$, $g(\theta) = \log(\theta - 1)$, and $g(x) = x$, respectively, where $\tanh^{-1}(\cdot)$ is the inverse hyperbolic tangent function, also known as Fisher's Z-transformation.

3.3.2 Empirical Bayes approach

As an alternative to estimating the copula parameter θ jointly with the marginal parameters, we may first try to obtain a point estimate of θ, and then perform fully Bayesian inference on the marginal parameters conditional on the estimate of θ. To do this, we adapt the method of Roy et al. (2016) to the bivariate mixed outcomes case.

Denote $\boldsymbol{y} = (\boldsymbol{y_c}, \boldsymbol{y_b})$. Then denote $L_\theta(\boldsymbol{\psi}|\boldsymbol{y}) \equiv L(\boldsymbol{\psi}, \theta|\boldsymbol{y})$ as the likelihood function. Let $\pi(\boldsymbol{\psi})$ denote the prior of the marginal parameters. Then for fixed θ, the posterior density of $\boldsymbol{\psi}$ is

$$\pi_\theta(\boldsymbol{\psi}|\boldsymbol{y}) = \frac{L_\theta(\boldsymbol{\psi}|\boldsymbol{y})\pi(\boldsymbol{\psi})}{m_\theta(\boldsymbol{y})}, \tag{3.3}$$

where $m_\theta(\boldsymbol{y})$ is the marginal density of the data conditional on θ, given by

$$m_\theta(\boldsymbol{y}) = \int L_\theta(\boldsymbol{\psi}|\boldsymbol{y})\pi(\boldsymbol{\psi})d\boldsymbol{\psi}. \tag{3.4}$$

The empirical Bayes estimate of θ is found by maximizing $m_\theta(\boldsymbol{y})$. The details of how to perform this estimation follow.

Empirical Bayes estimation of the copula parameter

We now proceed to describe the computational algorithm to obtain the empirical Bayes estimate of θ. Rather than maximize $m_\theta(\boldsymbol{y})$, which is usually difficult to estimate directly, we maximize the Bayes factor $B_{\theta,\theta_1} = m_\theta/m_{\theta_1}$, where θ_1 is a fixed value of θ. Usually, B_{θ,θ_1} itself does not have a closed form expression. However, as mentioned in

Roy et al. (2016), it can be consistently estimated by

$$S^{-1}\sum_{s=1}^{S}\frac{f(\boldsymbol{y}|\boldsymbol{\psi}^{(\boldsymbol{s})},\theta)}{f(\boldsymbol{y}|\boldsymbol{\psi}^{(\boldsymbol{s})},\theta_1)}, \tag{3.5}$$

where $\{\boldsymbol{\psi}^{(\boldsymbol{s})}\}_{s=1}^{S}$ is a Markov chain with stationary density $\pi_{\theta_1}(\boldsymbol{\psi}|\boldsymbol{y})$. Thus, an empirical Bayes estimate of θ can be found by maximizing (3.5). However, as noted in Roy et al. (2016), this estimator is often unstable and highly sensitive to the choice of θ_1. An improved estimator of θ is based on the idea of replacing the single value θ_1 with a grid of *skeleton points* $\theta_1,\ldots,\theta_k$. Denote $\{\boldsymbol{\psi}_{\boldsymbol{j}}^{(\boldsymbol{s})}\}_{s=1}^{S_j}$ as a Markov chain with stationary density $\pi_{\theta_j}(\boldsymbol{\psi}|\boldsymbol{y})$ for $j=1,\ldots,k$. A consistent estimator of B_{θ,θ_1} based on the grid $\theta_1,\ldots,\theta_k$ is

$$\sum_{j=1}^{k}\sum_{s=1}^{S_j}\frac{f(\boldsymbol{y}|\boldsymbol{\psi}_{\boldsymbol{j}}^{(\boldsymbol{s})},\theta)}{\sum_{l=1}^{k}S_l f(\boldsymbol{y}|\boldsymbol{\psi}_{\boldsymbol{j}}^{(\boldsymbol{s})},\theta_l)/r_l}, \tag{3.6}$$

where $r_l = m_{\theta_l}/m_{\theta_1},\ \ l=2,\ldots,k$, with $r_1=1$. Since $r_1,\ldots,r_k$ are unknown in general, we replace them with consistent estimates $\hat{r_1},\ldots,\hat{r_k}$. These estimates can be found using the "reverse logistic regression" procedure from Geyer (1994), described briefly below. The final empirical Bayes estimate of θ is found by maximizing (3.6) with respect to θ.

Reverse logistic regression to estimate $r_j's$

Reverse logistic regression (Geyer, 1994) is a method to estimate ratios of normalizing constants of posterior densities. Denote $\{\boldsymbol{X_{sj}}\}_{s=1}^{S_j} \equiv \{\boldsymbol{\psi}_{\boldsymbol{j}}^{(\boldsymbol{s})}\}_{s=1}^{S_j}$ as posterior samples from $\pi_{\theta_j}(\boldsymbol{\psi})$ for $j=1,\ldots,k$. In practice, $\boldsymbol{X_{sj}}$ are generated using the unnormalized density

$h_j \equiv c_j \cdot \pi_{\theta_j}$, where c_j is the normalizing constant. To estimate $r_j = c_j/c_1 = m_{\theta_j}/m_{\theta_1}$, $j = 1, \ldots, k$, we maximize

$$l(\boldsymbol{r}) = \sum_{j=1}^{k} \sum_{s=1}^{S_j} \log p_j(\boldsymbol{X_{sj}}, \boldsymbol{r}),$$

where

$$p_j(\boldsymbol{x}, \boldsymbol{r}) = \frac{h_j(\boldsymbol{x})}{\sum_{l=1}^{k} h_l(\boldsymbol{x}) \times \left\{ \frac{r_j}{r_l} \frac{S_l}{S_j} \right\}}.$$

Note that $r_1 = 1$ by construction, and hence we maximize over the reduced set $(r_2, \cdots, r_k)$.

Estimation of marginal parameters

We now fix the value of the copula parameter θ at its empirical Bayes estimate, $\hat{\theta}$. We generate new MCMC samples $\{\boldsymbol{\psi^{(s)}}\}_{s \geq 1}$ from $\pi_{\hat{\theta}}(\boldsymbol{\psi})$, and conduct inference using these samples.

3.3.3 Hamiltonian Monte Carlo

Here, we briefly describe the mechanics of HMC, which is our algorithm of choice for generating MCMC samples for both the fully Bayesian and the empirical Bayes methods. HMC accelerates convergence to the target distribution by simulating Hamiltonian

dynamics. The interested reader is referred to Neal (2011) for a detailed technical explanation of HMC, and to Betancourt (2017) for an excellent exposition on the intuition behind HMC.

For each variable γ_d in the probability model, $d = 1, \cdots, D$, HMC introduces an auxiliary momentum variable r_d. The momentum variables are usually drawn independently from the standard normal distribution. This yields the joint probability density function

$$p(\boldsymbol{\gamma}, \boldsymbol{r}) \propto \exp\left\{ l(\boldsymbol{\gamma}) - \frac{1}{2}\boldsymbol{r}^T\boldsymbol{r} \right\},$$

where l is the logarithm of the unnormalized joint density of $\boldsymbol{\gamma}$. The joint density p can be interpreted as a Hamiltonian system where $\boldsymbol{\gamma}$ represents a particle's position in D-dimensional space, r_d represents the particle's momentum in the dth dimension, l is a negative potential energy function, $\boldsymbol{r}^T\boldsymbol{r}/2$ is the kinetic energy of the particle, and log $p(\boldsymbol{\gamma}, \boldsymbol{r})$ is the negative energy of the particle.

Hamiltonian dynamics describe a particle's motion in continuous time. In order to simulate these dynamics on a computer, we need to discretize time. This is done by using the "leapfrog" integrator. The leapfrog integrator updates the momentum and position of the particle sequentially, first by simulating the momentum dynamics over a small time period $\epsilon/2$, then by simulating the position dynamics over the longer time period ϵ, and finally by completing the momentum dynamics over time period $\epsilon/2$. This sequence of steps ensures that the momentum and position of the particle end up at the

same point in time. A single leapfrog update is given as

$$\begin{aligned}
\boldsymbol{r}^{t+\epsilon/2} &= \boldsymbol{r}^{t} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\gamma}} l(\boldsymbol{\gamma}^{t}), \\
\boldsymbol{\gamma}^{t+\epsilon} &= \boldsymbol{\gamma}^{t} + \epsilon \boldsymbol{r}^{t+\epsilon/2}, \\
\boldsymbol{r}^{t+\epsilon} &= \boldsymbol{r}^{t+\epsilon/2} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\gamma}} l(\boldsymbol{\gamma}^{t+\epsilon}),
\end{aligned}$$

where $\boldsymbol{r}^{t}$ and $\boldsymbol{\gamma}^{t}$ are the momentum and the position variables at time t, and $\nabla_{\boldsymbol{\gamma}}$ is the gradient with respect to $\boldsymbol{\gamma}$.

Once the momentum variables have been sampled, L leapfrog updates are applied to the position and the momentum variables, generating a proposal position-momentum pair $\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{r}}$. The proposal pair is accepted according to a Metropolis reject step with probability

$$\min\left\{1, \frac{\exp\left\{l(\widetilde{\boldsymbol{\gamma}}) - \frac{1}{2}\widetilde{\boldsymbol{r}}^{T}\widetilde{\boldsymbol{r}}\right\}}{\exp\left\{l(\boldsymbol{\gamma^{t-1}}) - \frac{1}{2}\boldsymbol{r^{0}}^{T}\boldsymbol{r^{0}}\right\}}\right\},$$

where $\boldsymbol{r^{0}}$ are the sampled momentum variables before they are put through the leapfrog integrator.

The efficiency of HMC comes from the fact that using Hamiltonian dynamics with many leapfrog steps leads to proposals for $\boldsymbol{\gamma}$ that have a high acceptance probability even through they are distant from the previous sample. This results in fast exploration of the target density. As mentioned previously, the performance of HMC is sensitive to choosing suitable values of ϵ and L. The No-U-Turn Sampler (NUTS)

(Hoffman and Gelman, 2014) automatically adapts L during sampling. Stan implements NUTS and also sets ϵ optimally using the dual averaging algorithm of Nesterov (2009). Details on Stan's implementation of HMC/NUTS can be found in the Stan Language Manual (`https://github.com/stan-dev/stan/releases/download/v2.15.0/stan-reference-2.15.0.pdf`).

3.3.4 Model selection

We consider the problem of selecting the optimal copula family to model the dependence between the bivariate mixed outcomes. To simplify matters, we shall assume that the families of the marginal distributions of the outcomes are fixed. By doing so, we can recast the copula selection problem as a model selection problem, as the only difference between competing models is the choice of copula family. We consider two Bayesian model selection criteria: the deviance information criterion (DIC, Spiegelhalter et al. (2002)), and the conditional predictive ordinate (CPO, Geisser (1993), Gelfand et al. (1992)). As mentioned in Chen et al. (2008), the DIC and the CPO are popular criteria for Bayesian model selection because they are well defined under improper priors, as long as the posterior distribution is proper, which gives them an advantage over Bayes factors. Silva and Lopes (2008) mention another desirable property of the DIC: it is invariant (i.e., chooses the same model) under monotone increasing transformations of the marginal distributions.

The model deviance $\Delta(\boldsymbol{\omega})$, defined as

$$\Delta(\boldsymbol{\omega}) = -2\log f(\boldsymbol{y}|\boldsymbol{\omega}, \mathcal{M}),$$

is twice the negative observed-data log-likelihood for model $\mathcal{M}$ with parameters $\boldsymbol{\omega} = (\boldsymbol{\psi}, \theta)$. Define the effective number of parameters as

$$p_\Delta = \overline{\Delta(\boldsymbol{\omega})} - \Delta(\bar{\boldsymbol{\omega}}),$$

where the "bar" superscript denotes the expectation taken with respect to the posterior distribution of $\boldsymbol{\omega}$, i.e. $\mathbb{E}_{\boldsymbol{\omega}|\boldsymbol{y}}$. Then the DIC is defined as

$$\begin{aligned} \mathrm{DIC}(\mathcal{M}) &= \Delta(\bar{\boldsymbol{\omega}}) + 2p_\Delta \\ &= \overline{\Delta(\boldsymbol{\omega})} + p_\Delta \\ &= -4\mathbb{E}_{\boldsymbol{\omega}|\boldsymbol{y}}[\log f(\boldsymbol{y}|\boldsymbol{\omega}, \mathcal{M})] + 2\log f(\boldsymbol{y}|\mathbb{E}_{\boldsymbol{\omega}|\boldsymbol{y}}[\boldsymbol{\omega}], \mathcal{M}). \end{aligned} \tag{3.7}$$

Neither of the integrals in (3.7) is usually available in closed form. However, they can be estimated using MCMC samples. The estimated DIC is given by:

$$\widehat{\mathrm{DIC}}(\mathcal{M}) = -4\left(\frac{1}{S}\sum_{s=1}^{S}\log f(\boldsymbol{y}|\boldsymbol{\omega}^{(s)}, \mathcal{M})\right) + 2\log f(\boldsymbol{y}|\left(\frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\omega}^{(s)}\right), \mathcal{M}), \tag{3.8}$$

where $\{\boldsymbol{\omega}^{(s)}\}_{s=1}^{S}$ are MCMC samples from the posterior distribution. Models with

smaller DIC values are preferred.

Under model $\mathcal{M}$, the CPO for the ith observation is defined as

$$\text{CPO}_i = f(\boldsymbol{y_i}|\boldsymbol{y_{(-i)}}, \mathcal{M}) = \int f(\boldsymbol{y_i}|\boldsymbol{\omega}, \mathcal{M})\pi(\boldsymbol{\omega}|\boldsymbol{y_{(-i)}}, \mathcal{M})d\boldsymbol{\omega}, \tag{3.9}$$

where $\boldsymbol{y_{(-i)}} = (\boldsymbol{y_1}, \cdots, \boldsymbol{y_{i-1}}, \boldsymbol{y_{i+1}} \cdots, \boldsymbol{y_n})$, and $\pi(\boldsymbol{\omega}|\boldsymbol{y_{(-i)}}, \mathcal{M})$ is the posterior density under model $\mathcal{M}$ based on $\boldsymbol{y_{(-i)}}$. From (3.9), we see that CPO_i is the marginal posterior predictive density of $\boldsymbol{y_i}$ given $\boldsymbol{y_{(-i)}}$. Large values of CPO_i indicate a better fit of the model. Gelfand and Dey (1994) provide a way to estimate CPO_i using MCMC samples:

$$\widehat{\text{CPO}}_i = S\left(\sum_{s=1}^{S} \frac{1}{f(\boldsymbol{y_i}|\boldsymbol{\omega^{(s)}}, \mathcal{M})}\right)^{-1}, \tag{3.10}$$

where $\{\boldsymbol{\omega^{(s)}}\}_{s=1}^{S}$ are MCMC samples from the posterior distribution.

A natural way to summarize the CPO_i over all observations is the logarithm of the pseudomarginal likelihood (LPML, Ibrahim et al. (2005)), defined as

$$\text{LPML} = \sum_{i=1}^{n} \log(\text{CPO}_i). \tag{3.11}$$

Models with larger LPML values are preferred.

3.4 Simulation studies

3.4.1 Comparison between the fully Bayesian and the empirical Bayes approaches

We conducted extensive simulation studies to assess the performance of the fully Bayesian approach and the empirical Bayes approach. Our simulations consisted of 100 replications under each of 24 scenarios that varied the sample size, the copula family used to generate the data, and the strength of association between the bivariate outcomes. We considered sample sizes of $n = 500$ and $n = 1000$ to assess the effect of varying the signal-to-noise ratio on the performance of the two methods. For the copula family, we used the Gaussian, Clayton, Gumbel, and Frank copulas, which reflect different types of tail dependence between the correlated variables. The Gaussian and the Frank copulas do not model tail dependence, while the Clayton and the Gumbel copulas can model lower tail dependence and upper tail dependence, respectively. Under each copula family, we varied the copula parameter to correspond to a Kendall's Tau τ of $0.1, 0.3,$ and 0.6, to capture weak, medium, and strong dependence between the outcomes.

The data were generated according to the scheme explained in Section 3.2. A single common covariate $x_i \sim \text{Normal}(0, 1)$ is considered along with an intercept for both the continuous outcome and the binary outcome, so that $\boldsymbol{x_i}^T = \boldsymbol{z_i}^T = (1, x_i)$. We set the true regression coefficients $\boldsymbol{\beta_c} \equiv (\beta_{c0}, \beta_{c1})^T = (1.2, 2.0)$, $\boldsymbol{\beta_b} \equiv (\beta_{b0}, \beta_{b1})^T = (1.2, 0.8)$, and we set the marginal variance of the continuous outcome $\sigma_c^2 = 1$.

For each data set generated, we applied the fully Bayesian method and the empirical Bayes method to estimate the parameters. For these simulations, we assumed that the true copula family was known, i.e., we did not perform model selection. A simulation study addressing model selection performance is described in Section 3.4.2. For the fully Bayesian approach, we specified zero-mean Normal priors on the transformed copula parameter $g(\theta)$, which was defined in Section 3.3.1. It is common practice to make this prior diffuse by specifying a large prior variance. However, we have observed that a diffuse prior on $g(\theta)$ can lead to divergent transitions for the HMC sampling algorithm, which in turn leads to biased estimates of the parameters (Betancourt, 2016). One possible reason for the poor performance of HMC with diffuse priors on $g(\theta)$ is that these priors induce strongly informative and often nonsensical prior distributions on θ. Consequently, we decided to consider priors on $g(\theta)$ that induced plausible prior distributions on θ, i.e., weakly informative priors, in the language of Gelman et al. (2008). We found a unit prior variance to be a sensible choice for the prior distribution of $g(\theta)$ for the Gaussian, Clayton, and Gumbel copula families, and a variance of 10 to be a good choice for the Frank copula family. With these prior choices, none of the HMC runs led to divergent transitions, under any of scenarios considered. We used diffuse normal priors for the regression coefficients, and a half-Cauchy prior for the σ_c, as recommended by Gelman (2006). For the empirical Bayes approach, we specified the grid of copula parameter points as $g^{-1}(g(\theta_{true}) \cdot (0.5, 0.8, 1.0, 1.2, 1.5))$, similar to the approach in Roy et al. (2016), where θ_{true} is the true value of the copula parameter. For

the fully Bayesian approach, we run two MCMC chains for 2000 iterations each. The first 1000 iterations from each chain are discarded as warmup, leaving us with a total of 2000 iterations across the chains. The empirical Bayes method requires three sets of MCMC chains: one set to estimate the r_l's from (3.6), one set to compute the empirical Bayes estimate of the copula parameter, $\hat{\theta}$, and another set to obtain posterior samples for the marginal parameters, given $\hat{\theta}$. To ensure a fair comparison of both methods, the final set of MCMC iterations for the empirical Bayes method is run with the same specifications as the chains for the fully Bayesian method. Convergence of the chains was assessed using the potential scale reduction statistic, $\hat{R}$ (Gelman and Rubin, 1992). Furthermore, the robustness of the HMC exploration of the posterior distribution was assessed by checking for divergent transitions, as well as by using the energy diagnostic of Betancourt (2017).

We assessed the performance of the two methods on estimation, inference and computational efficiency. First, we compared the accuracy of the two methods by computing the bias and the root mean square error (RMSE) of the point estimates of the model parameters. For all parameters except θ under the empirical Bayes method, we used the posterior mean as the point estimate. Next, we computed the coverage of the 95% credible intervals for the marginal parameters, as well as the coverage of the 95% credible intervals for θ under the fully Bayesian method. Interval estimates for θ are unavailable under the empirical Bayes method. We also compared the width of these intervals under both methods. Note that as a metric, interval width should be considered only if the

coverage of the interval is sufficient. Table 10 displays these metrics for the various scenarios considered.

We first observe that both methods produced unbiased estimates for the marginal parameters. The RMSE for the marginal parameters is similar under both methods, and is generally small compared to the true values of the parameters. Both methods exhibit small bias and RMSE for the copula parameter θ under the Gaussian, Clayton, and Gumbel copula families. Under the Frank copula family, both methods exhibit larger, but similar, bias and RMSE for the copula parameter, with the exception of one case– when Kendall's Tau is set to 0.6, and $n = 1000$ observations are generated, then the fully Bayesian method performs substantially worse than the empirical Bayes method. We investigated this scenario further by specifying more informative priors on the copula parameter, but this did not improve the performance of the fully Bayesian method. The convergence diagnostics did not reveal anything anomalous either, so we are currently unable to explain the poor performance of the fully Bayesian method for this scenario. The coverage and width metrics for the marginal parameters are similar under the two methods. We also note that the coverage is generally above 90%, with the exception of the single scenario mentioned above for the fully Bayesian method, where the coverage for the copula parameter is only 57%, which is consistent with the large bias observed. Finally, as expected, the RMSE and width metrics are usually smaller for the larger sample size, compared to the analogous scenario with smaller sample size.

To compare the performance in estimating the dependence across copula families, we

computed these metrics for our estimates of Kendall's Tau. For the copula families that we have considered, there is a one-to-one relationship between the copula parameter and Kendall's Tau. Thus, under the fully Bayesian method, posterior samples for Kendall's Tau can be obtained by simply transforming the posterior samples of the copula parameter. Under the empirical Bayes method, we simply transform the empirical Bayes estimate of the copula parameter to obtain an empirical Bayes estimate of Kendall's Tau. These metrics are displayed in Table 11. We observe that the performance of each methods to estimate Kendall's Tau is similar across copula families. Note that Kendall's Tau scales down the bias and RMSE metrics, which is why the differences between the two methods appear small, even for the scenario mentioned above. Also note that because of the one-to-one relationship between θ and τ, the coverage for τ is exactly the same as the coverage for θ.

Finally, Table 12 displays the average difference in computation time, in hours, between the two methods. The difference in computation time for a single data set is computed as the empirical Bayes method's time to complete minus fully Bayesian method's time to complete. Hence, a positive number means that the empirical Bayes method took longer to complete, on average. We see that the fully Bayesian method is the clear winner here. The average difference ranges from 1.69 hours, in the scenario where $n = 500, \tau = 0.1$, and the Frank copula is used, to 10.35 hours, in the scenario where $n = 1000, \tau = 0.3$, and the Gumbel copula is used. The empirical Bayes method is slower than the fully Bayesian method for a two reasons. First, multiple MCMC chains

need to be run for the three stages of estimation. However, with HMC, these additional chains can be generated fairly quickly. The bulk of computation time for the empirical Bayes method goes into the optimization involved in estimating the weights from reverse logistic regression procedure. We observe that the average time difference seems to grow linearly with increasing sample size, which indicates that this method is better suited for smaller data sets. Although the computation time for the empirical Bayes method could be reduced by either reducing the size of the grid for the copula parameter or by running the MCMC chains for a smaller number of iterations, we have observed that this comes with a significant decrease in estimation accuracy.

Table 10: Bias, root mean square error, coverage of 95% credible intervals, and average width of 95% credible intervals for model parameters, under the Fully Bayesian method (FB) and the empirical Bayes method (EB).

(a) Gaussian copula

Parameter	τ	n	Bias		RMSE		Coverage		Width	
			FB	EB	FB	EB	FB	EB	FB	EB
θ	0.1	500	0.01	0.01	0.06	0.07	0.93	-	0.23	-
		1000	0.01	0.01	0.05	0.05	0.93	-	0.17	-
	0.3	500	0.00	0.00	0.06	0.06	0.95	-	0.20	-
		1000	0.00	0.00	0.04	0.04	0.95	-	0.14	-
	0.6	500	0.01	0.00	0.03	0.03	0.95	-	0.11	-
		1000	0.00	0.00	0.02	0.02	0.94	-	0.08	-
β_{c0}	0.1	500	0.00	0.00	0.04	0.04	0.95	0.95	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.92	0.93	0.12	0.12
	0.3	500	0.00	0.00	0.05	0.05	0.89	0.88	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.92	0.92	0.12	0.12
	0.6	500	0.00	0.00	0.04	0.04	0.94	0.92	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.97	0.97	0.12	0.12
β_{c1}	0.1	500	0.00	0.00	0.05	0.05	0.88	0.89	0.17	0.17
		1000	0.00	0.00	0.04	0.04	0.90	0.89	0.12	0.12
	0.3	500	0.00	0.00	0.04	0.04	0.96	0.96	0.17	0.17
		1000	0.01	0.01	0.03	0.03	0.96	0.96	0.12	0.12
	0.6	500	0.00	0.00	0.04	0.04	0.96	0.95	0.17	0.17
		1000	0.00	0.00	0.04	0.04	0.91	0.88	0.12	0.12
β_{b0}	0.1	500	0.01	0.01	0.14	0.14	0.88	0.87	0.43	0.43
		1000	0.00	0.00	0.08	0.08	0.93	0.92	0.31	0.31
	0.3	500	0.00	0.00	0.12	0.12	0.91	0.90	0.44	0.43
		1000	0.01	0.01	0.08	0.08	0.95	0.96	0.31	0.31
	0.6	500	0.01	0.01	0.14	0.14	0.86	0.86	0.42	0.42
		1000	0.02	0.02	0.07	0.07	0.97	0.97	0.30	0.30
β_{b1}	0.1	500	0.01	0.01	0.13	0.13	0.95	0.95	0.45	0.45
		1000	0.00	0.00	0.08	0.08	0.95	0.96	0.33	0.32
	0.3	500	0.02	0.02	0.13	0.13	0.94	0.93	0.45	0.45
		1000	0.03	0.03	0.09	0.09	0.92	0.92	0.33	0.33
	0.6	500	0.00	0.00	0.11	0.11	0.95	0.94	0.42	0.42
		1000	0.01	0.01	0.08	0.08	0.96	0.94	0.30	0.30
σ_c	0.1	500	0.00	0.00	0.03	0.03	0.93	0.92	0.12	0.12
		1000	0.00	0.00	0.02	0.02	0.98	0.98	0.09	0.09
	0.3	500	0.00	0.00	0.03	0.03	0.95	0.93	0.12	0.12
		1000	0.00	0.00	0.02	0.02	0.94	0.95	0.09	0.08
	0.6	500	0.00	0.00	0.03	0.03	0.95	0.94	0.12	0.11
		1000	0.00	0.00	0.02	0.02	0.92	0.92	0.09	0.08

(b) Clayton copula

Parameter	τ	n	Bias		RMSE		Coverage		Width	
			FB	EB	FB	EB	FB	EB	FB	EB
	0.1	500	0.00	0.00	0.07	0.07	0.94	-	0.28	-
		1000	0.02	0.01	0.15	0.06	0.95	-	0.24	-
θ	0.3	500	0.01	0.01	0.14	0.13	0.95	-	0.51	-
		1000	0.00	0.00	0.08	0.08	0.98	-	0.36	-
	0.6	500	0.01	0.00	0.35	0.35	0.97	-	1.45	-
		1000	0.03	0.03	0.29	0.30	0.91	-	1.04	-
	0.1	500	0.00	0.00	0.05	0.05	0.94	0.94	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.91	0.91	0.12	0.12
β_{c0}	0.3	500	0.01	0.01	0.05	0.05	0.92	0.91	0.17	0.16
		1000	0.00	0.00	0.03	0.03	0.96	0.96	0.12	0.12
	0.6	500	0.00	0.00	0.05	0.05	0.91	0.90	0.17	0.16
		1000	0.00	0.00	0.03	0.03	0.95	0.95	0.12	0.12
	0.1	500	0.00	0.00	0.04	0.04	0.97	0.97	0.17	0.17
		1000	0.02	0.00	0.17	0.03	0.95	0.95	0.15	0.12
β_{c1}	0.3	500	0.00	0.00	0.05	0.05	0.94	0.92	0.16	0.16
		1000	0.00	0.00	0.03	0.03	0.97	0.96	0.12	0.12
	0.6	500	0.00	0.00	0.04	0.04	0.95	0.95	0.16	0.16
		1000	0.00	0.00	0.03	0.03	0.93	0.93	0.12	0.11
	0.1	500	0.02	0.02	0.13	0.13	0.91	0.92	0.44	0.44
		1000	0.00	0.01	0.15	0.09	0.92	0.90	0.33	0.31
β_{b0}	0.3	500	0.01	0.01	0.12	0.12	0.91	0.91	0.43	0.42
		1000	0.01	0.01	0.07	0.07	0.96	0.96	0.30	0.30
	0.6	500	0.01	0.01	0.11	0.11	0.92	0.84	0.40	0.38
		1000	0.01	0.01	0.07	0.07	0.96	0.94	0.29	0.27
	0.1	500	0.02	0.02	0.12	0.12	0.96	0.97	0.46	0.46
		1000	0.01	0.01	0.09	0.09	0.95	0.95	0.33	0.33
β_{b1}	0.3	500	0.02	0.02	0.11	0.11	0.97	0.94	0.44	0.43
		1000	0.00	0.00	0.08	0.08	0.96	0.96	0.31	0.31
	0.6	500	0.01	0.00	0.10	0.10	0.92	0.92	0.39	0.39
		1000	0.02	0.02	0.08	0.08	0.94	0.93	0.28	0.28
	0.1	500	0.01	0.01	0.03	0.03	0.94	0.94	0.12	0.12
		1000	0.00	0.00	0.03	0.02	0.94	0.95	0.09	0.08
σ_c	0.3	500	0.01	0.01	0.03	0.03	0.93	0.92	0.12	0.11
		1000	0.00	0.00	0.02	0.02	0.95	0.95	0.09	0.08
	0.6	500	0.00	0.00	0.03	0.03	0.95	0.92	0.12	0.11
		1000	0.00	0.00	0.02	0.02	0.90	0.89	0.08	0.08

(c) Gumbel copula

Parameter	τ	n	Bias		RMSE		Coverage		Width	
			FB	EB	FB	EB	FB	EB	FB	EB
	0.1	500	0.02	0.01	0.05	0.05	0.96	-	0.20	-
		1000	0.00	0.00	0.04	0.04	0.95	-	0.14	-
θ	0.3	500	0.00	0.01	0.09	0.09	0.94	-	0.32	-
		1000	0.01	0.01	0.06	0.06	0.94	-	0.23	-
	0.6	500	0.00	0.00	0.14	0.14	0.98	-	0.76	-
		1000	0.00	0.00	0.13	0.13	0.94	-	0.53	-
	0.1	500	0.01	0.01	0.04	0.04	0.96	0.96	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.96	0.96	0.12	0.12
β_{c0}	0.3	500	0.00	0.00	0.05	0.05	0.91	0.94	0.17	0.17
		1000	0.01	0.01	0.03	0.03	0.94	0.95	0.12	0.12
	0.6	500	0.00	0.00	0.05	0.05	0.91	0.90	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.94	0.95	0.12	0.12
	0.1	500	0.00	0.00	0.04	0.04	0.96	0.96	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.94	0.94	0.12	0.12
β_{c1}	0.3	500	0.00	0.00	0.05	0.05	0.92	0.92	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.96	0.96	0.12	0.12
	0.6	500	0.01	0.01	0.05	0.05	0.94	0.94	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.95	0.94	0.12	0.12
	0.1	500	0.00	0.00	0.12	0.12	0.94	0.94	0.44	0.44
		1000	0.00	0.00	0.08	0.08	0.94	0.95	0.31	0.31
β_{b0}	0.3	500	0.00	0.00	0.13	0.13	0.92	0.92	0.43	0.43
		1000	0.01	0.02	0.08	0.08	0.93	0.94	0.31	0.31
	0.6	500	0.01	0.01	0.12	0.12	0.93	0.93	0.43	0.43
		1000	0.00	0.00	0.08	0.08	0.99	0.99	0.30	0.30
	0.1	500	0.02	0.02	0.12	0.12	0.94	0.93	0.46	0.46
		1000	0.00	0.00	0.09	0.09	0.95	0.94	0.33	0.33
β_{b1}	0.3	500	0.02	0.02	0.13	0.13	0.93	0.93	0.45	0.45
		1000	0.00	0.01	0.07	0.07	0.94	0.94	0.32	0.32
	0.6	500	0.03	0.03	0.11	0.11	0.94	0.92	0.43	0.43
		1000	0.01	0.01	0.08	0.08	0.95	0.95	0.31	0.31
	0.1	500	0.00	0.00	0.03	0.03	0.96	0.95	0.12	0.12
		1000	0.00	0.00	0.02	0.02	0.98	0.98	0.09	0.09
σ_c	0.3	500	0.01	0.01	0.03	0.03	0.94	0.94	0.12	0.12
		1000	0.00	0.00	0.02	0.02	0.96	0.96	0.09	0.08
	0.6	500	0.01	0.01	0.03	0.03	0.92	0.92	0.12	0.12
		1000	0.00	0.00	0.02	0.02	0.93	0.90	0.09	0.08

(d) Frank copula

Parameter	τ	n	Bias		RMSE		Coverage		Width	
			FB	EB	FB	EB	FB	EB	FB	EB
	0.1	500	0.05	0.05	0.35	0.35	0.96	-	1.44	-
		1000	0.00	0.02	0.25	0.26	0.96	-	1.02	-
θ	0.3	500	0.10	0.08	0.43	0.42	0.95	-	1.69	-
		1000	0.11	0.00	0.31	0.31	0.96	-	1.16	-
	0.6	500	0.07	0.03	1.03	1.03	0.90	-	3.17	-
		1000	0.87	0.06	0.96	0.56	0.57	-	1.87	-
	0.1	500	0.01	0.01	0.04	0.04	0.98	0.98	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.95	0.95	0.12	0.12
β_{c0}	0.3	500	0.01	0.01	0.05	0.05	0.92	0.91	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.97	0.98	0.12	0.12
	0.6	500	0.00	0.00	0.04	0.04	0.94	0.94	0.16	0.16
		1000	0.00	0.00	0.03	0.03	0.93	0.93	0.12	0.12
	0.1	500	0.01	0.01	0.04	0.04	0.98	0.98	0.17	0.17
		1000	0.00	0.00	0.03	0.03	0.94	0.95	0.12	0.12
β_{c1}	0.3	500	0.00	0.00	0.04	0.04	0.95	0.96	0.17	0.17
		1000	0.00	0.00	0.04	0.04	0.93	0.93	0.12	0.12
	0.6	500	0.00	0.00	0.05	0.05	0.93	0.93	0.16	0.16
		1000	0.00	0.00	0.03	0.03	0.96	0.97	0.12	0.12
	0.1	500	0.01	0.01	0.12	0.12	0.94	0.94	0.43	0.43
		1000	0.00	0.00	0.07	0.07	0.95	0.94	0.31	0.31
β_{b0}	0.3	500	0.02	0.02	0.12	0.12	0.92	0.92	0.43	0.44
		1000	0.00	0.00	0.08	0.08	0.95	0.95	0.31	0.31
	0.6	500	0.02	0.02	0.10	0.10	0.97	0.97	0.42	0.42
		1000	0.00	0.01	0.07	0.07	0.96	0.95	0.30	0.30
	0.1	500	0.03	0.03	0.13	0.13	0.92	0.92	0.46	0.46
		1000	0.01	0.01	0.08	0.08	0.95	0.95	0.33	0.33
β_{b1}	0.3	500	0.03	0.03	0.13	0.13	0.93	0.93	0.46	0.46
		1000	0.01	0.01	0.08	0.08	0.93	0.94	0.32	0.32
	0.6	500	0.02	0.02	0.11	0.11	0.96	0.96	0.43	0.43
		1000	0.00	0.00	0.07	0.07	0.96	0.96	0.31	0.30
	0.1	500	0.00	0.00	0.03	0.03	0.94	0.94	0.12	0.12
		1000	0.01	0.01	0.02	0.02	0.94	0.93	0.09	0.09
σ_c	0.3	500	0.00	0.00	0.03	0.03	0.95	0.93	0.12	0.12
		1000	0.00	0.00	0.02	0.02	0.94	0.93	0.09	0.08
	0.6	500	0.00	0.00	0.03	0.03	0.92	0.92	0.12	0.12
		1000	0.00	0.00	0.02	0.02	0.92	0.90	0.08	0.08

Table 11: Bias, root mean square error, coverage of 95% credible intervals, and average width of 95% credible intervals for the estimate of Kendall's Tau τ, under the Fully Bayesian method (FB) and the empirical Bayes method (EB). Data sets are generated with true $\tau \in \{0.1, 0.3, 0.6\}$.

Copula	τ	n	Bias		RMSE		Coverage		Width	
			FB	EB	FB	EB	FB	EB	FB	EB
Gaussian	0.1	500	0.00	0.00	0.04	0.04	0.93	-	0.15	-
		1000	0.01	0.01	0.03	0.03	0.93	-	0.11	-
	0.3	500	0.00	0.00	0.04	0.04	0.95	-	0.14	-
		1000	0.00	0.00	0.03	0.03	0.95	-	0.10	-
	0.6	500	0.00	0.00	0.03	0.03	0.95	-	0.11	-
		1000	0.00	0.00	0.02	0.02	0.94	-	0.08	-
Clayton	0.1	500	0.00	0.00	0.03	0.03	0.94	-	0.11	-
		1000	0.00	0.00	0.03	0.02	0.95	-	0.09	-
	0.3	500	0.01	0.00	0.03	0.03	0.95	-	0.13	-
		1000	0.00	0.00	0.02	0.02	0.98	-	0.09	-
	0.6	500	0.00	0.00	0.03	0.03	0.97	-	0.12	-
		1000	0.00	0.00	0.02	0.02	0.91	-	0.08	-
Gumbel	0.1	500	0.01	0.00	0.04	0.04	0.96	-	0.15	-
		1000	0.00	0.00	0.03	0.03	0.95	-	0.11	-
	0.3	500	0.01	0.00	0.04	0.04	0.94	-	0.16	-
		1000	0.00	0.01	0.03	0.03	0.94	-	0.11	-
	0.6	500	0.00	0.00	0.02	0.02	0.98	-	0.12	-
		1000	0.00	0.00	0.02	0.02	0.94	-	0.09	-
Frank	0.1	500	0.01	0.01	0.04	0.04	0.96	-	0.16	-
		1000	0.00	0.00	0.03	0.03	0.96	-	0.11	-
	0.3	500	0.01	0.01	0.04	0.04	0.95	-	0.15	-
		1000	0.01	0.00	0.03	0.03	0.96	-	0.10	-
	0.6	500	0.00	0.00	0.04	0.04	0.90	-	0.12	-
		1000	0.04	0.00	0.04	0.02	0.57	-	0.08	-

Table 12: Average time differences in hours to complete parameter estimation between the fully Bayesian method (FB) and the empirical Bayes method (EB) Δ_{time}. The difference in computation time for a single data set is computed as the time for EB minus the time for FB.

Copula	τ	n	Δ_{time}
Gaussian	0.1	500	1.90
		1000	4.12
	0.3	500	2.38
		1000	4.49
	0.6	500	2.53
		1000	4.26
Clayton	0.1	500	2.59
		1000	4.97
	0.3	500	2.89
		1000	5.42
	0.6	500	2.77
		1000	4.31
Gumbel	0.1	500	5.23
		1000	9.73
	0.3	500	5.35
		1000	10.35
	0.6	500	4.53
		1000	8.35
Frank	0.1	500	1.69
		1000	3.93
	0.3	500	2.16
		1000	3.79
	0.6	500	2.25
		1000	3.84

3.4.2 Model selection

We conducted simulation studies to assess the fully Bayesian method's ability to detect the correct copula family. As noted in Section 3.3.4, by fixing the marginal distributions of the outcomes, the problem of selecting the copula family can be recast as a model selection problem. We only consider the fully Bayesian method here, as the simulation studies from the preceding subsection show that the methods offer similar performance in terms of estimation and inference for most of the scenarios considered, but the fully Bayesian method is significantly faster.

The simulation design and the scenarios considered are the same as those considered previously, except that we generate 500 replicated data sets per scenario. For each simulated data set, we fit models using the Gaussian, Clayton, Gumbel, and Frank copulas. Using the posterior samples, we computed the DIC and the LPML for the models corresponding to each copula. Table 13 shows the fraction of times each copula family was identified as best, using DIC and LPML.

We observe that the ability of the fully Bayesian method to select the correct copula family improves with increasing sample size and larger values of Kendall's Tau. Furthermore, it is easier to identify the Clayton and the Gumbel copulas, due to the unique tail dependence that they produce. These trends are consistent to those observed by Silva and Lopes (2008). Finally, we note that both DIC and LPML are consistent in these trends, and that LPML has a slightly better ability to detect the correct copula family.

Table 13: Fraction of 500 replications where a copula family used for estimation is selected by the model selection metric, for varying values of Kendall's Tau τ, sample size n, and copula family used for data generation. The model selection metrics considered are the Deviance Information Criterion (DIC) and the Logarithm of Pseudo Marginal Likelihood (LPML).

(a) Model selection metric: DIC

Generated copula	τ	n	Estimated copula			
			Gaussian	Clayton	Gumbel	Frank
Gaussian	0.1	500	0.19	0.29	0.29	0.23
		1000	0.26	0.21	0.31	0.23
	0.3	500	0.51	0.08	0.24	0.16
		1000	0.71	0.03	0.18	0.08
	0.6	500	0.75	0.03	0.20	0.02
		1000	0.86	0.00	0.13	0.01
Clayton	0.1	500	0.12	0.72	0.08	0.09
		1000	0.10	0.83	0.03	0.04
	0.3	500	0.03	0.97	0.00	0.00
		1000	0.02	0.98	0.00	0.00
	0.6	500	0.05	0.95	0.00	0.00
		1000	0.01	0.99	0.00	0.00
Gumbel	0.1	500	0.13	0.17	0.46	0.24
		1000	0.14	0.09	0.51	0.26
	0.3	500	0.22	0.00	0.59	0.19
		1000	0.18	0.00	0.68	0.14
	0.6	500	0.23	0.00	0.59	0.18
		1000	0.13	0.00	0.74	0.12
Frank	0.1	500	0.12	0.19	0.29	0.39
		1000	0.14	0.12	0.26	0.48
	0.3	500	0.15	0.02	0.19	0.64
		1000	0.10	0.00	0.12	0.77
	0.6	500	0.05	0.00	0.19	0.76
		1000	0.02	0.00	0.15	0.83

(b) Model selection metric: LPML

Generated copula	τ	n	Estimated copula			
			Gaussian	Clayton	Gumbel	Frank
Gaussian	0.1	500	0.21	0.29	0.28	0.23
		1000	0.27	0.20	0.30	0.23
	0.3	500	0.52	0.08	0.24	0.17
		1000	0.72	0.03	0.17	0.08
	0.6	500	0.74	0.03	0.21	0.02
		1000	0.86	0.00	0.13	0.01
Clayton	0.1	500	0.12	0.72	0.07	0.09
		1000	0.10	0.84	0.03	0.04
	0.3	500	0.03	0.97	0.00	0.00
		1000	0.02	0.98	0.00	0.00
	0.6	500	0.05	0.95	0.00	0.00
		1000	0.01	0.99	0.00	0.00
Gumbel	0.1	500	0.14	0.17	0.44	0.25
		1000	0.15	0.09	0.50	0.27
	0.3	500	0.22	0.00	0.58	0.19
		1000	0.18	0.00	0.67	0.15
	0.6	500	0.23	0.00	0.58	0.18
		1000	0.13	0.00	0.74	0.12
Frank	0.1	500	0.13	0.19	0.28	0.40
		1000	0.15	0.12	0.25	0.48
	0.3	500	0.14	0.02	0.18	0.65
		1000	0.10	0.00	0.12	0.78
	0.6	500	0.05	0.00	0.19	0.76
		1000	0.02	0.00	0.15	0.84

3.5 Application to burn injury data

The burn injury data (Fan and Gijbels, 1996) contains information on $n = 981$ patients of varying ages who suffered burn injuries. For each patient, we have two outcomes: the total burn area, and the survival status. It is naturally of interest to quantify the strength of association between total burn area and survival status. Furthermore, it is also of interest to estimate the association between the age of the patient and the two outcomes. We considered $y_c = \log(\text{total burn area} + 1)$ to be the continuous response, and $y_b =$ survival status to be the binary response (1 for dead and 0 for survived). We took $x =$ age as the common covariate to both outcomes.

We modeled the data using the copula model described in Section 3.2 with the fully Bayesian approach. To improve the efficiency of HMC, we standardized the covariate prior to performing estimation. The marginal models for the outcomes are given by

$$y_{ic} \sim \text{Normal}(\mu_{ic}, \sigma_c^2)$$

$$y_{ib} \sim \text{Bernoulli}(p_{ib}),$$

where

$$\mu_{ic} = \beta_{c0} + \beta_{c1} x_i, \quad p_{ib} = \frac{\exp(\beta_{b0} + \beta_{b1} x_i)}{1 + \exp(\beta_{b0} + \beta_{b1} x_i)}.$$

We fit four models corresponding to four choices of the copula family: Gaussian, Clayton, Gumbel, and Frank. For each model, two parallel chains were used to draw 2000

Table 14: DIC and LPML values for the models fitted on the burn injury data. Models with smaller values of DIC and larger values of LPML are preferred.

Metric	Copula family			
	Gaussian	Clayton	Gumbel	Frank
DIC	3621	3689	3607	3654
LPML	-1805	-1839	-1798	-1821

samples each from the posterior distribution. The first 1000 samples from each chain were discarded as warmup, leaving a total of 2000 samples. Convergence of the chains and robustness of the HMC algorithm's ability to explore the posterior distribution were assessed using the diagnostics described in Section 3.4.1. Autocorrelation plots of the chains indicated that no thinning was necessary. The Gaussian, Clayton, Gumbel, and Frank copula models took 26 seconds, 29 seconds, 44 seconds, and 27 seconds, respectively, to fit on an Intel Core i5-6200U CPU (2.30GHz) with four cores and 8 GB RAM.

Using DIC and LPML as our model selection criteria, we found that the Gumbel copula provided the best fit to the data, followed respectively by the Gaussian copula, the Frank copula, and finally the Clayton copula. Table 14 reports the DIC and the LPML values for the four copula families.

Table 15 reports the posterior means, the standard errors and the 95% posterior credible intervals for the parameters of the Gumbel copula model. Note that all reported quantities for regression coefficients are on the original scale of the covariate age. The 95% credible intervals for the coefficients of age indicate that it is significantly associated

Table 15: Posterior means, standard deviations, and 95% credible intervals for the parameters of the Gumbel copula model applied to the burn injury data. The top row shows the derived metrics for Kendall's Tau τ.

Parameter	Mean	SD	95% Credible interval
τ	0.612	0.027	(0.559, 0.661)
θ	2.588	0.179	(2.266, 2.950)
β_{c0}	6.661	0.067	(6.530, 6.795)
β_{c1}	0.005	0.002	(0.002, 0.008)
β_{b0}	-3.241	0.181	(-3.584,-2.885)
β_{b1}	0.041	0.004	(0.034, 0.049)
σ_c	1.250	0.028	(1.197, 1.305)

with both total burn area and survival status. The expected total burn area increases slowly with age. For a year's increase in age, the odds of death increase by a factor of $\exp(0.0411) = 1.042$. Perhaps more usefully, for a ten year increase in age, the odds of death increase by a factor of $\exp(10 \times 0.0411) \approx 1.5$. Finally, we observe a strong association between the logarithm of total burn area and survival status, as indicated by the statistically significant estimate of 0.61 for Kendall's Tau. To interpret this estimate, it is important to note that 0.61 is the estimate of Kendall's Tau between the logarithm of total burn area and the continuous latent variable assumed to underlie the binary outcome, survival status. Nonetheless, a strong dependence between the continuous outcome and the latent variable implies a strong dependence between the continuous outcome and the binary outcome. The estimates of the regression coefficients and Kendall's Tau from our model are similar to the corresponding estimates from the analyses conducted in Song et al. (2009) and de Leon and Wu (2011).

Chapter 4

Analyzing bivariate mixed secondary phenotypes in case-control genome-wide association studies using generalized estimating equations

4.1 Introduction

Genome-wide association studies (GWAS) are becoming increasingly popular for identifying associations between genetic variants such as single-nucleotide polymorphisms (SNPs) and disease phenotypes. Many GWASs use a case-control design, in which disease-affected individuals (the cases) are sampled separately from disease-free individuals (the controls). For example, the Environment and Genetics in Lung Cancer

Etiology (EAGLE) study (Landi et al., 2008) is a large case-control study conducted to investigate the genetic and environmental determinants of lung cancer and smoking persistence. The data from the EAGLE study contains genotypic and phenotypic information on approximately 2000 newly diagnosed lung cancer cases from the Lombardi region of Italy, and approximately 2000 age-, gender-, and region- matched controls. As is common practice in many GWASs, during the EAGLE study, information on phenotypes other than the disease status was collected. Because per-individual genotyping and sequencing costs are still considerable, there is substantial interest in extracting additional information out of GWAS datasets by analyzing these *secondary* phenotypes (Monsees et al., 2009). In the EAGLE study, the Fagerstrom Test Score for Nicotine Dependence (FTND) is a continuous phenotype that measures the intensity of physical addition to nicotine, while smoking status (current smoker or former smoker) is a binary phenotype. Both FTND and smoking status are measures of smoking persistence, and quantifying the association of each of these secondary phenotypes with the SNPs is of substantial interest.

A common strategy to identify associations between secondary phenotypes and SNPs is to fit separate regression models for each phenotype, and construct separate tests of association between the SNP and each secondary phenotype. This strategy could fail to detect some important SNPs, because the separate tests may be underpowered, so that they fail to reach the genome-wide level of significance (Schifano et al., 2013). On the other hand, if the secondary outcomes measure different aspects of the same underlying

physiological condition (such as smoking persistence, in the EAGLE study), then they are likely to be correlated. A simultaneous test of association between a genetic variant and the secondary outcomes could reach the genome-wide level of significance, where the separate tests failed. Simultaneous tests, which usually have higher power than the corresponding (multiple-comparisons-corrected) separate tests, can further benefit from the information sharing that occurs between the correlated phenotypes.

In the EAGLE study, FTND is a continuous outcome, whereas smoking status is a binary outcome. Constructing joint regression models for such mixed outcomes can be challenging. Several approaches have been proposed to jointly model a continuous outcome and a binary outcome (see Teixeira Pinto and Normand (2009) for an overview). Of these, the approach of using generalized estimating equations (GEEs) is one of the most convenient, in that GEEs only require specification of the first two moments of each outcome and an approximation to their correlation structure. Furthermore, parameter estimates from GEEs are robust to misspecification of the correlation structure, which is often the case in practice.

However, care must be taken while applying GEEs to secondary outcomes from case-control studies. In case-control designs, the cases are often oversampled. Consequently, the case-control sample is not a random sample from the population under study, and the population association between the secondary outcomes and the SNPs can be distorted in the case-control sample. The naive application of estimation methods such as GEEs to secondary outcomes from a case-control study could lead to biased estimates

and inflated Type I error rates (Monsees et al., 2009). Strategies to conduct valid inference on secondary outcomes under case-control designs include inverse probability weighted regression (Monsees et al., 2009; Schifano et al., 2013; Xing et al., 2016), use of retrospective likelihoods (Lin and Zeng, 2009; Wei et al., 2013; Ghosh et al., 2013) and conditional methods (Chen et al., 2013; Tchetgen, 2013). Of these methods, only SMAT, the method from Schifano et al. (2013) was developed for the joint analysis of multiple secondary phenotypes in case-control designs. However, their model was designed for multiple continuous outcomes measuring the same underlying trait. SMAT accounts for the case-control sampling design by weighting observations by the inverse of the subject's probability of being included in the case-control sample. Recently, Xing et al. (2016) showed that their method, which weights observations by the inverse of the subject's probability of being a case (or a control), conditional on their genetic and phenotypic information, performs better than the single-outcome equivalent of SMAT, in terms of estimation accuracy and power. However, this method was developed to handle a single secondary continuous outcome only, and not mixed secondary outcomes as found in the EAGLE study.

In this chapter, we propose two GEE-based methods to jointly model bivariate mixed secondary phenotypes, i.e., one continuous secondary phenotype and one binary secondary phenotype, in case-control designs. These methods extend the methods of Schifano et al. (2013) and Xing et al. (2016), respectively, to handle bivariate mixed outcomes. Under each proposed method, we show how a simultaneous test of association

between a SNP and the secondary phenotypes can be constructed. Through simulation studies, we compare both methods in terms of estimation accuracy, Type I error rate control, and power. We then apply our methods to the EAGLE data set and identify SNPs associated with smoking persistence and nicotine addiction.

4.2 Notations and generalized estimating equations for bivariate mixed outcomes

Assume that our case-control data set contains n_0 controls and n_1 cases. For all notation that follows, let the i subscript denote quantities for the ith individual, $i = 1, \ldots, n$, with $n = n_0 + n_1$. Then denote d_i as the indicator for disease (case-control) status, which is the primary outcome. We set $d_i = 1$ for the cases, and $d_i = 0$ for the controls. Next, denote the secondary continuous outcome as y_{ic}, and the secondary binary outcome as y_{ib}. Denote g_i as a SNP genotypic value, which is the independent variable of interest. For simplicity, we assume that g_i is biallelic, and is measured as the number of minor alleles at the SNP location. Further, denote $\boldsymbol{x_i}$ as a p-dimensional covariate vector corresponding to y_{ic}, and denote $\boldsymbol{z_i}$ a q-dimensional covariate vector corresponding to y_{ib}. For notational simplicity, we assume the common situation that the same set of covariates are used to model both of the secondary outcomes, i.e., $\boldsymbol{x_i} = \boldsymbol{z_i}$, although separate sets of covariates for each secondary outcome can easily be accommodated. Let $\boldsymbol{y_i} = (y_{ic}, y_{ib})^T$ denote the bivariate vector of secondary outcomes from the ith individual.

We assume that outcomes from the same individual are correlated, but outcomes from different individuals are independent.

We use GEEs to account for the correlation between the secondary outcomes y_{ic} and y_{ib}. Consequently, the marginal means of the secondary outcomes can be specified via link functions as

$$\mu_{ic} = \boldsymbol{x_i}^T \boldsymbol{\beta_c} + g_i \tau_c, \tag{4.1}$$

$$\log\left(\frac{\mu_{ib}}{1-\mu_{ib}}\right) = \boldsymbol{x_i}^T \boldsymbol{\beta_b} + g_i \tau_b, \tag{4.2}$$

where $\mu_{ic} = \mathbb{E}(y_{ic})$, $\mu_{ib} = \mathbb{E}(y_{ib})$, $\boldsymbol{\beta_c}$ and $\boldsymbol{\beta_b}$ are the covariate effects, and τ_c and τ_b are the SNP effects for y_{ic} and y_{ib}, respectively. Denote $\boldsymbol{\mu_i} = (\mu_{ic}, \mu_{ib})^T$, $\boldsymbol{\gamma_c} = (\boldsymbol{\beta_c}^T, \tau_c)^T$, $\boldsymbol{\gamma_b} = (\boldsymbol{\beta_b}^T, \tau_b)^T$, and $\boldsymbol{\gamma} = (\boldsymbol{\gamma_c}^T, \boldsymbol{\gamma_b}^T)^T$. We specify the variance functions $v_c(y_{ic}) = \psi_c h_c(\mu_{ic})$ and $v_b(y_{ib}) = \psi_b h_b(\mu_{ib})$, where ψ_c and ψ_b are dispersion parameters. For illustration, we have chosen the logit link for the secondary binary outcome y_{ib}, though other choices exist, such as the probit link, or the robit link (Liu, 2004), the latter of which is resistant to outliers. For simplicity of notation, we further assume that $\psi_c = \psi_b = 1$.

As given in Chapter 2, the specification of GEEs for bivariate mixed outcomes is given by

$$\boldsymbol{S}(\boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{S_i} = \boldsymbol{0}, \tag{4.3}$$

where $\boldsymbol{S_i} = \boldsymbol{D_i^T V_i^{-1}}(\boldsymbol{y_i} - \boldsymbol{\mu_i})$, where

$$\boldsymbol{D_i^T} = \frac{\partial \boldsymbol{\mu_i}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^T} = \begin{pmatrix} \partial \mu_{ic}/\partial \boldsymbol{\gamma_c^T} & \mathbf{0} \\ \mathbf{0} & \partial \mu_{ib}/\partial \boldsymbol{\gamma_b^T} \end{pmatrix},$$

and $\boldsymbol{V_i}$ is the variance-covariance matrix of $\boldsymbol{y_i}$, given by $\boldsymbol{V_i} = \boldsymbol{A_i^{1/2} R A_i^{1/2}}$, where

$$\boldsymbol{A_i} = \begin{pmatrix} h_c(\mu_{ic}) & 0 \\ 0 & h_b(\mu_{ib}) \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Here $\boldsymbol{R}$ is the working correlation matrix and ρ measures the strength of association between the continuous and binary outcomes. Note that ρ, which we shall refer to as the association parameter, is assumed to be fixed across i.

4.3 Adapting generalized estimating equations for bivariate mixed outcomes to case-control designs

In case-control studies, the cases are often oversampled. Hence, the sample proportion of cases will not match the population proportion of cases. Consequently, any analysis that ignores the sampling design in a case-control study can lead to biased estimates of the population effects. This problem is exacerbated if disease status is correlated with the secondary outcomes and the genomic variable (Monsees et al., 2009). We now

describe two methods to estimate and test the SNP effects α_c and α_b that account for the case-control sampling design. The methods are extensions of the methods provided in Schifano et al. (2013) and Xing et al. (2016), respectively, to bivariate mixed outcomes. Both of these methods correct the case-control sampling bias through the incorporation of observation *weights*. They differ in the way the weights are constructed, and in how the final population effect estimates are constructed.

4.3.1 Inverse-probability-of-sampling weighted generalized estimating equations

The method of inverse-probability-of-sampling weighted (IPSW) GEEs is based on Schifano et al. (2013). Under this method, the population effects are estimated by solving the weighted GEEs

$$\boldsymbol{S}_{\text{IPSW}}(\boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^{n} w_i \boldsymbol{S_i} = \mathbf{0}, \tag{4.4}$$

where the observation weights w_i are given by

$$w_i = \begin{cases} \dfrac{\pi}{p_n} & \text{if } d_i = 1 \\ \dfrac{1-\pi}{1-p_n} & \text{if } d_i = 0 \end{cases}, \tag{4.5}$$

where π is the disease prevalence in the population, and $p_n = n_1/n$ is the sample proportion of cases. Denote $I_S(i)$ as the indicator of individual i being included in the

sample of case-control data from a cohort of size N. Using Bayes theorem, for the cases, we have

$$\begin{aligned} Pr(I_S(i) = 1|d_i = 1) &= \frac{Pr(d_i = 1|I_S(i) = 1)Pr(I_S(i) = 1)}{Pr(d_i = 1)} \\ &= \frac{(n_1/n) \cdot (n/N)}{\pi} \\ &= (n/N)\frac{(p_n)}{\pi} \\ &= (n/N)(1/w_i). \end{aligned}$$

Similarly, for the controls, we have

$$\begin{aligned} Pr(I_S(i) = 1|d_i = 0) &= \frac{Pr(d_i = 0|I_S(i) = 1)Pr(I_S(i) = 1)}{Pr(d_i = 0)} \\ &= \frac{(n_0/n) \cdot (n/N)}{1 - \pi} \\ &= (n/N)\frac{(1 - p_n)}{1 - \pi} \\ &= (n/N)(1/w_i). \end{aligned}$$

Thus, w_i is proportional to the inverse probability that individual i was sampled from the population, given disease status. Usually, the case-control design ensures that $\pi << p_n$, and hence the weights serve to downweight the influence of the cases, and upweight the influence of the controls. Notice that under this scheme, all the cases have the same weight, and all the controls have the same weight.

Schifano et al. (2013) showed that the weighted GEEs in (4.4) are unbiased for arbitrary working correlation matrices in the context of multiple continuous secondary outcomes. The proof of this result easily generalizes to multiple mixed secondary outcomes. To conduct inference on α_c and α_b, we note that a sandwich estimator for the variance-covariance matrix of estimates the secondary outcome parameters $\hat{\boldsymbol{\gamma}}$, say $\hat{V}_{\boldsymbol{\gamma}}$, can be obtained after solving the weighted GEEs. The variance-covariance matrix of α_c and α_b can be obtained by selecting the relevant submatrix of $\hat{V}_{\boldsymbol{\gamma}}$. We can then perform hypothesis testing based on the asymptotic multivariate normality of $(\alpha_c, \alpha_b)^T$.

4.3.2 Inverse-probability-weighted generalized estimating equations

The method of inverse-probability weighted (IPW) GEEs is based on Xing et al. (2016). Estimation under this method begins by considering separate weighted estimating equations for the cases and the controls:

$$\boldsymbol{S}^{(\mathbf{0})}_{\text{IPW}}(\boldsymbol{\gamma}) = n_0^{-1} \sum_{i=1}^{n_0} \frac{\boldsymbol{S_i}}{1 - p_{d_i}} = \mathbf{0}, \tag{4.6}$$

$$\boldsymbol{S}^{(\mathbf{1})}_{\text{IPW}}(\boldsymbol{\gamma}) = n_1^{-1} \sum_{i=n_0+1}^{n_0+n_1} \frac{\boldsymbol{S_i}}{p_{d_i}} = \mathbf{0}, \tag{4.7}$$

where p_{d_i} denotes the probability of the ith individual being a case given y_{ic}, y_{ib}, g_i, and $\boldsymbol{x_i}$, in the population. The probability p_{d_i} is an unknown population quantity, but it

can be estimated from the case-control data, as detailed in Section 4.3.2. Xing et al. (2016) showed that solving (4.6) and (4.7) will each lead to consistent estimates for the SNP effects, in the context of a single continuous secondary outcome. The proof of this result easily generalizes to multiple mixed secondary outcomes.

Estimating p_d with case-control data

Xing et al. (2016) suggested that the probabilities p_{d_i} can be estimated using logistic regression:

$$p_{d_i} = Pr(d_i = 1|y_{ic}, y_{ib}, g_i, \boldsymbol{x_i}) = \frac{\exp(\theta_0 + \theta_c y_{ic} + \theta_b y_{ib} + \theta_g g_i + \boldsymbol{x_i}^T \boldsymbol{\theta_x})}{1 + \exp(\theta_0 + \theta_c y_{ic} + \theta_b y_{ib} + \theta_g g_i + \boldsymbol{x_i}^T \boldsymbol{\theta_x})}. \tag{4.8}$$

With case-control data, all the regression coefficients $\boldsymbol{\theta} = (\theta_0, \theta_c, \theta_b, \theta_g, \boldsymbol{\theta_x}^T)^T$ can be estimated unbiasedly, with the exception of the intercept, θ_0. If the disease prevalence is known (say π), θ_0 can be estimated by

$$\hat{\theta_0} = \theta_0^* + \log\left(\frac{n_0}{n_1}\right) + \log\left(\frac{\pi}{1-\pi}\right), \tag{4.9}$$

where $\hat{\theta_0^*}$ is the estimated intercept from the logistic regression applied to the case-control data. Having obtained valid estimates of all the logistic regression coefficients, we can obtain estimates of p_{d_i}.

Xing et al. (2016) also provided a method to estimate p_{d_i} when the disease prevalence is unknown. However, at the sample size considered in our simulations, we observed that

the performance of this method for bivariate mixed secondary outcomes is unreliable in terms of controlling the Type I error rate. Hence, we only use the method that assumes that the disease prevalence is known.

Combining estimators and variance-covariance matrix estimation

Solving (4.6) and (4.7) will lead to two consistent estimates for each of the SNP effects τ_c and τ_b. Let $\hat{\tau_c}^{(0)}$ and $\hat{\tau_b}^{(0)}$ be the estimates from the controls, and let $\hat{\tau_c}^{(1)}$ and $\hat{\tau_b}^{(1)}$ be the estimates from the cases. We can obtain more efficient estimates of τ_c and τ_b by combining the estimators from the controls and the cases using weighted sums, i.e.,

$$\hat{\tau_c} = a_0\hat{\tau_c}^{(0)} + (1 - a_0)\hat{\tau_c}^{(1)}, \tag{4.10}$$

$$\hat{\tau_b} = b_0\hat{\tau_b}^{(0)} + (1 - b_0)\hat{\tau_b}^{(1)}, \tag{4.11}$$

where the weights $0 \leq a_0, b_0 \leq 1$ are chosen to minimize the variances of $\hat{\tau_c}$ and $\hat{\tau_b}$, respectively. From Xing et al. (2016), the weights that minimize the variances of $\hat{\tau_c}$ and $\hat{\tau_b}$, are given by

$$\boldsymbol{a}^T \equiv (a_0, 1 - a_0) = \frac{\mathbf{1}^T\boldsymbol{V_{01}^{(c)}}}{\mathbf{1}^T\boldsymbol{V_{01}^{(c)}}\mathbf{1}},$$
$$\boldsymbol{b}^T \equiv (b_0, 1 - b_0) = \frac{\mathbf{1}^T\boldsymbol{V_{01}^{(b)}}}{\mathbf{1}^T\boldsymbol{V_{01}^{(b)}}\mathbf{1}},$$

where $\mathbf{1}^T = (1, 1)$, $\boldsymbol{V}_{\mathbf{01}}^{(\boldsymbol{c})}$ is the variance-covariance matrix of $\hat{\tau}_c{}^{(0)}$ and $\hat{\tau}_c{}^{(1)}$, and $\boldsymbol{V}_{\mathbf{01}}^{(\boldsymbol{b})}$ is the variance-covariance matrix of $\hat{\tau}_b{}^{(0)}$ and $\hat{\tau}_b{}^{(1)}$, to be discussed more below. Although the estimators $\hat{\tau}_c{}^{(0)}$ and $\hat{\tau}_c{}^{(1)}$ are estimated from separate data, they are correlated because they both use common parameters for p_{d_i}. To account for this correlation, consider the joint estimation of the primary outcome coefficients $\boldsymbol{\theta}$ and the secondary outcome coefficients $\boldsymbol{\gamma}$, performed by "stacking" all the relevant estimating functions:

$$\boldsymbol{U_i} \equiv \begin{pmatrix} \boldsymbol{T_i} \\ I(d_i = 0)\dfrac{\boldsymbol{S_i}}{1 - p_i} \\ I(d_i = 1)\dfrac{\boldsymbol{S_i}}{p_i} \end{pmatrix}, \tag{4.12}$$

where $\boldsymbol{T_i}$ denotes the contribution of the ith individual to the score function of the logistic regression (4.8). Based on (4.12), we can obtain the sandwich estimator $\hat{\boldsymbol{V}}_{\boldsymbol{\omega}}$ of the variance-covariance matrix of all the estimated parameters $\hat{\boldsymbol{\omega}} = (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\gamma}}^{(0)T}, \hat{\boldsymbol{\gamma}}^{(1)T})^T$, where $\hat{\boldsymbol{\gamma}}^{(0)}$ and $\hat{\boldsymbol{\gamma}}^{(1)}$ are the estimates of all the secondary outcome regression coefficients, from the controls and the cases, respectively. Let $\boldsymbol{A}(\hat{\boldsymbol{\omega}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \boldsymbol{U_i}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}}$ and $\boldsymbol{B}(\hat{\boldsymbol{\omega}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U_i}(\hat{\boldsymbol{\omega}})\boldsymbol{U_i}(\hat{\boldsymbol{\omega}})^T$. Then $\hat{\boldsymbol{V}}_{\boldsymbol{\omega}} = n^{-1}\boldsymbol{A}(\hat{\boldsymbol{\omega}})^{-1}\boldsymbol{B}(\hat{\boldsymbol{\omega}})[\boldsymbol{A}(\hat{\boldsymbol{\omega}})^{-1}]^T$. The required variance-covariance matrices $\boldsymbol{V}_{\mathbf{01}}^{(\boldsymbol{c})}$ and $\boldsymbol{V}_{\mathbf{01}}^{(\boldsymbol{b})}$ can be obtained by selecting the elements of $\hat{\boldsymbol{V}}_{\boldsymbol{\omega}}$ that correspond to the pairs $(\hat{\tau}_c{}^{(0)}, \hat{\tau}_c{}^{(1)})$, and $(\hat{\tau}_b{}^{(0)}, \hat{\tau}_b{}^{(1)})$, respectively. Furthermore, we can also extract the variance-covariance matrix of all the SNP coefficient estimates $(\hat{\tau}_c{}^{(0)}, \hat{\tau}_c{}^{(1)}, \hat{\tau}_b{}^{(0)}, \hat{\tau}_b{}^{(1)})$, which will be needed to conduct a joint test on τ_c and τ_b. Testing can be performed based on the asymptotic normality property of $\hat{\boldsymbol{\omega}}$, which was proved

in Xing et al. (2016) in the context of a single continuous secondary outcome, but which trivially extends to multiple mixed secondary outcomes.

4.4 Simulation studies

We conducted simulation studies to compare the performance of three methods to estimate regression coefficients for bivariate mixed secondary outcomes under a case-control design: 1) the naive, unweighted GEEs for bivariate mixed outcomes (NAÏVE), 2) the IPSW GEEs approach, and 3) the IPW GEEs approach. We considered scenarios where the disease was either rare or common. However, in either case, we assumed that the disease prevalence was known, or an estimate was available. We simulated case-control data prospectively through the following scheme:

1. Generate g, x_c, x_b. The genotypic variables s are generated from a Binomial(2, 0.3) distribution, corresponding to a biallelic SNP in Hardy-Weinberg equilibrium with a minor allele frequency of 0.3. We consider one continuous covariate x_c, generated from the standard normal distribution, and one binary covariate x_b, generated form Bernoulli(0.5). We also include an intercept term. Thus, the ith row of the design matrix is $(1, g_i, x_{ic}, x_{ib})$.

2. Generate $(y_c, y_b|s, x_c, x_b)$. We use a copula to specify the dependency between y_c and y_b. Specifically, we assume a continuous latent variable w to underlie y_b, and specify a Gaussian copula between y_c and w. The generative process can be

described as follows. For $i = 1, \ldots, n$:

$$
\begin{aligned}
(u_i, v_i) &\sim C(\cdot, \cdot|\tilde{\rho}), \\
y_{ic} &= F_c^{-1}(u_i|\mu = \beta_{0c} + \beta_{1c}x_{ic} + \beta_{2c}x_{ib} + g_i\tau_c, \sigma = \sigma_c), \\
w_i &= F_w^{-1}(v_i|\mu = \beta_{0b} + \beta_{1b}x_{ic} + \beta_{2b}x_{ib} + g_i\tau_b, s = 1), \\
y_{ib} &= I(w_i > 0),
\end{aligned}
$$

where $C(\cdot, \cdot|\tilde{\rho})$ is a two-dimensional Gaussian copula with correlation parameter $\tilde{\rho}$, $F_c^{-1}(\cdot|\mu, \sigma) \equiv \Phi^{-1}(\cdot|\mu, \sigma)$ is the inverse cumulative distribution function of the normal distribution with mean μ and standard deviation σ, and $F_w^{-1}(\cdot|\mu, s)$ is the inverse cumulative distribution function of the logistic distribution with location μ and scale s. Also, $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ are the continuous latent variables that generate the observed binary outcomes $\boldsymbol{y_b} = (y_{1b}, \ldots, y_{nb})^T$. Note that the copula parameter $\tilde{\rho}$ used to generate the data is different from the association parameter ρ of working correlation matrix of the GEEs. We set $\tau_c = \tau_b = 0$ under the null hypothesis and $\tau_c = \tau_b = \log(1.2)$ under the alternative hypothesis. The remaining coefficients are set as $\beta_{0c} = 2.3$, $\beta_{1c} = -0.2$, $\beta_{2c} = 1.0$, $\beta_{0b} = 0.5$, $\beta_{1b} = -0.3$ and $\beta_{2b} = -0.6$. To consider different strengths of association between the secondary outcomes, we let $\tilde{\rho} \in \{0.1, 0.3, 0.5\}$.

3. Generate $(d|y_c, y_b, g, x_c, x_b)$. The primary outcomes d are generated according to

the logistic regression in (4.8):

$$\begin{aligned} p_{d_i} &= Pr(d_i = 1 | y_{ic}, y_{ib}, g_i, x_{ic}, x_{ib}) \\ &= \frac{\exp(\theta_0 + \theta_c y_{ic} + \theta_b y_{ib} + \theta_g g_i + \theta_{xc} x_{ic} + \theta_{xb} x_{ib})}{1 + \exp(\theta_0 + \theta_c y_{ic} + \theta_b y_{ib} + \theta_g g_i + \theta_{xc} x_{ic} + \theta_{xb} x_{ib})}. \end{aligned}$$

We set $\theta_g \in \{0, \log(1.2)\}$, $\theta_c = 1.2$, $\theta_b = \log(1.5)$, $\theta_{xc} = 0.2$, and $\theta_{xb} = -0.4$. Given these parameters, we set the intercept θ_0 to calibrate our target population disease prevalences. We considered a target disease prevalence $\pi = 0.01$ for common diseases, and a target disease prevalence $\pi = 0.001$ for rare diseases.

4. Repeat steps 1-3 until we obtain $n_0 = 1000$ controls and $n_1 = 1000$ cases.

The values of the regression coefficients for the simulation studies are reflective of the effect sizes of the corresponding variables in the EAGLE study data. We simulated 10^6 data sets for each scenario with $\tau_c = \tau_b = 0$, and 10^5 data sets for each scenario corresponding to $\tau_c = \tau_b = \log(1.2)$.

For each scenario considered, we computed the root mean square error (RMSE) and the bias of the estimates for τ_c and τ_b, under the three methods. We also computed the Type I error rate and the power of the three methods for the simultaneous test $H_0 : \tau_c = \tau_b = 0$. The Type I error rate was computed at two levels of significance: $\alpha = 10^{-4}$ and $\alpha = 10^{-3}$. The power was computed at level of significance $\alpha = 10^{-3}$. Table 16a shows the RMSE and the bias for the three methods. As we might expect, the NAÏVE method has smaller RMSE and bias when the SNP effect in the primary model

θ_g is zero. However, when $\theta_g \neq 0$, the NAÏVE method is biased, whereas the IPSW and IPW methods are unbiased for most of the scenarios considered. When the disease is common and the correlation between the secondary outcomes is large ($\tilde{\rho} = 0.5$), the latter two methods show bias, although smaller than the bias of the NAÏVE method in the case where $\theta_g \neq 0$. Table 16b shows the Type I error and power for the three methods, for the simultaneous test. The NAÏVE method does not always control the Type I error at the required level of significance, especially when $\theta_g \neq 0$. The IPSW and the IPW methods control the Type I error rate across all the scenarios considered. Of these methods, the IPW method has higher power than the IPSW method. This can be explained by the fact that the IPW method takes all genotypic and phenotypic information into account while constructing its weights, which makes it more efficient than the IPSW method. In summary, the IPW method strikes the best balance between reliability and power, and it is the method we recommend to analyze bivariate mixed secondary outcomes under a case-control sampling design. Table 17 shows the Type I error and the power for the IPW method, for the separate tests $H_0 : \tau_c = 0$ and $H_0 : \tau_b = 0$. The method controls the Bonferroni-corrected Type I error rate for both tests. The observed powers to test $H_0 : \tau_c = 0$ are similar to the powers for the simultaneous test, while the observed powers to test $H_0 : \tau_b = 0$ are quite low. This reflects the difference of information contained in continuous and binary data, at the effect sizes considered in our simulation studies. We discuss the possible consequences of other effect sizes on the power of the simultaneous test versus the power of the joint test in Chapter 5.

Table 16: Root mean square error (RMSE), bias, Type I error, and power for the NAÏVE, IPSW, and IPW methods.

(a) RMSE and bias

π	θ_g	$\tilde{\rho}$	RMSE			Bias		
			NAÏVE	IPSW	IPW	NAÏVE	IPSW	IPW
0.001	0	0.1	0.07	0.10	0.10	0.01	0.01	0.00
		0.3	0.09	0.12	0.10	0.00	0.00	0.01
		0.5	0.09	0.12	0.11	0.00	0.01	0.01
	log(1.2)	0.1	0.10	0.12	0.11	0.06	0.01	0.01
		0.3	0.11	0.11	0.10	0.08	0.01	0.01
		0.5	0.12	0.11	0.10	0.08	0.00	0.01
0.01	0	0.1	0.09	0.12	0.11	0.04	0.03	0.04
		0.3	0.08	0.10	0.10	0.00	0.01	0.01
		0.5	0.08	0.14	0.12	0.02	0.09	0.06
	log(1.2)	0.1	0.10	0.10	0.09	0.05	0.00	0.00
		0.3	0.09	0.11	0.10	0.05	0.01	0.00
		0.5	0.12	0.13	0.12	0.09	0.04	0.05

(b) Type I error and power for the simultaneous test $H_0 : \tau_c = \tau_b = 0$.

π	θ_g	$\tilde{\rho}$	Type I Error						Power		
			$\alpha = 10^{-4}$			$\alpha = 10^{-3}$			$\alpha = 10^{-3}$		
			NAÏVE	IPSW	IPW	NAÏVE	IPSW	IPW	NAÏVE	IPSW	IPW
0.0001	0	0.1	0.00000	0.00000	0.00000	0.00003	0.00000	0.00007	1.00	0.70	0.79
		0.3	0.00000	0.00000	0.00000	0.00000	0.00001	0.00002	1.00	0.64	0.71
		0.5	0.00000	0.00000	0.00001	0.00461	0.00004	0.00003	1.00	0.50	0.61
	log(1.2)	0.1	0.00695	0.00000	0.00000	0.00937	0.00017	0.00035	1.00	0.70	0.79
		0.3	0.00444	0.00000	0.00000	0.02698	0.00003	0.00008	1.00	0.67	0.75
		0.5	0.00700	0.00000	0.00000	0.06807	0.00002	0.00016	1.00	0.60	0.71
0.01	0	0.1	0.00000	0.00000	0.00000	0.00004	0.00000	0.00001	1.00	0.78	0.81
		0.3	0.00000	0.00000	0.00001	0.00006	0.00001	0.00003	1.00	0.61	0.71
		0.5	0.00000	0.00000	0.00000	0.00000	0.00001	0.00000	0.99	0.54	0.64
	log(1.2)	0.1	0.00016	0.00001	0.00000	0.00029	0.00002	0.00001	1.00	0.64	0.76
		0.3	0.00014	0.00000	0.00000	0.00138	0.00000	0.00015	1.00	0.71	0.76
		0.5	0.00028	0.00001	0.00000	0.01047	0.00004	0.00023	1.00	0.64	0.64

Table 17: Type I error and power for the IPW method, for the separate tests $H_0 : \tau_c = 0$ and $H_0 : \tau_b = 0$.

π	θ_g	ρ	$H : \tau_c = 0$			$H : \tau_b = 0$		
			Type I Error		Power	Type I Error		Power
			$\alpha = 0.5 \times 10^{-4}$	$\alpha = 0.5 \times 10^{-3}$	$\alpha = 0.5 \times 10^{-3}$	$\alpha = 0.5 \times 10^{-4}$	$\alpha = 0.5 \times 10^{-3}$	$\alpha = 0.5 \times 10^{-3}$
0.0001	0	0.1	0.00000	0.00001	0.76	0.00000	0.00004	0.07
		0.3	0.00000	0.00000	0.70	0.00000	0.00000	0.09
		0.5	0.00000	0.00001	0.65	0.00001	0.00001	0.07
	log(1.2)	0.1	0.00000	0.00009	0.72	0.00000	0.00010	0.05
		0.3	0.00000	0.00001	0.72	0.00000	0.00002	0.07
		0.5	0.00000	0.00001	0.73	0.00000	0.00014	0.11
0.01	0	0.1	0.00000	0.00000	0.76	0.00000	0.00000	0.09
		0.3	0.00001	0.00003	0.70	0.00000	0.00000	0.13
		0.5	0.00000	0.00000	0.67	0.00000	0.00001	0.06
	log(1.2)	0.1	0.00000	0.00032	0.72	0.00000	0.00003	0.12
		0.3	0.00000	0.00015	0.74	0.00000	0.00000	0.06
		0.5	0.00000	0.00020	0.65	0.00000	0.00000	0.04

4.5 Case study: EAGLE

In this section we demonstrate the application of our methods on the data from the EAGLE study. We considered log(FTND + 1) as our continuous secondary phenotype, and SMOKING STATUS (1=former smoker, 0=current smoker) as our binary secondary phenotype. The covariates used in this analysis are SEX (1=male, 0=female) and AGE. AGE, which represents the age at diagnosis for cases and the age at study entry for controls, is actually an ordered categorical variable (0="59 or less", 1="60-64",...,7="90 or more"). However, for simplicity, during the analysis, we treated AGE as a continuous covariate. Finally, we assumed an additive genetic model for each SNP.

Prior to analysis, we performed data processing to restrict our attention to useful SNPs. We filtered out SNPs with a minor allele frequency of less than 0.05. We also filtered out SNPS whose estimated squared correlation between the estimated allele dosage and the true allele dosage was less than 0.3. Additionally, we dropped observations with missing phenotypic or covariate information.

Because the simulation studies show that the IPW GEEs method provides the highest power while controlling the Type I error rate, we identified the top SNPs using this method. Specifically, we conducted the simultaneous test of association between a SNP and the secondary outcomes, $H : \tau_c = \tau_b = 0$, for each SNP, and selected the SNPs that attained the genome-wide level of significance 5×10^{-8}. For these SNPs, we also computed Bonferroni-corrected p-values for the separate association tests $H : \tau_c = 0$

Table 18: Top SNPs from the EAGLE study reaching genome-wide level of significance 5×10^{-8} for the test $H : \tau_c = \tau_b = 0$. The corresponding Bonferroni-corrected p-values for the tests $H : \tau_c = 0$ and $H : \tau_b = 0$ are also shown. Corrected p-values larger than 1 are truncated to 1 and marked with an asterisk (*).

SNP	MAF	Chr.	Gene	p-value		
				$H : \tau_c = \tau_b = 0$	$H : \tau_c = 0$	$H : \tau_b = 0$
rs12548690	0.0866	8	CSMD1	3.91×10^{-9}	3.74×10^{-6}	3.89×10^{-6}
rs938760	0.0549	16	RBFOX1	5.54×10^{-9}	3.91×10^{-9}	1.00×10^{0}*
rs1878540	0.0561	16	RBFOX1	8.01×10^{-9}	5.66×10^{-9}	1.00×10^{0}*
rs4786675	0.0530	16	RBFOX1	2.07×10^{-8}	2.67×10^{-8}	1.00×10^{0}*

and $H : \tau_b = 0$. Table 18 shows the top SNPs and the corresponding p-values. We now observe the benefit performing the simultaneous test: SNP rs12548690 attains the genome-wide level of significance for the simultaneous test, but fails to do so for either of separate tests (even without the Bonferroni correction). This SNP is located on gene CSMD1. The other three SNPs that attain the genome-wide level of significance are all located on the same gene, RBFOX1. CSMD1 has been previously identified as associated with the ability to quit smoking (Uhl et al., 2008b), as well as with addiction to cannabis (Sherva et al., 2016), cocaine (Drgonova et al., 2015), and methamphetamine (Uhl et al., 2008a). RBFOX1 has been found to be associated with nicotine dependence and the ability to quit smoking (Chen et al., 2016; Zhong et al., 2015).

Chapter 5

Discussion

In this chapter, we summarize the contributions of this dissertation to analyzing bivariate mixed outcomes. We also discuss some limitations of our methods, directions for future work, and the associated challenges.

In Chapter 2, we provided a framework to perform simultaneous estimation and variable selection with correlated bivariate mixed outcomes using PGEEs. The simulation experiments and the MEPS data analysis indicate that the major gains in estimation and variable selection when outcomes are analyzed jointly occur in the binary outcome coefficients. Binary outcome regression coefficients are generally harder to estimate due to the smaller information content in binary data. Thus, by borrowing strength from the continuous outcomes through the correlation, joint estimation is able to outperform separate estimation for the binary outcome regression coefficients, while providing equivalent or better performance for the continuous outcome coefficients. We also provided a method to estimate and control the FDR in the PGEE framework for bivariate mixed outcomes. A useful extension of our method would be to allow for more than two outcomes. Another useful extension would be to allow for longitudinal data, for

each outcome. The challenge in each of these extensions lies in the estimation of the correlation structure. In the latter case, the iterative algorithm to solve the PGEEs would have to be modified as well.

In Chapter 3, we provided two methods to perform Bayesian estimation of parameters from a copula model for bivariate mixed outcomes. Our simulation studies show that the fully Bayesian method and the empirical Bayes method have equivalent performance in terms of parameter estimation and inference for most the scenarios considered. The fully Bayesian method, however, is much more computationally efficient. The fully Bayesian method also shows good performance in its ability to select the correct copula family. Once again, it would be of interest to extend the copula model to allow for more than two outcomes. However, the copula model with more than two outcomes necessitates the use of latent variables to represent the observed outcomes, and the likelihood will no longer have a closed form expression. Consequently, sampling from the posterior distribution becomes more challenging, possibly requiring the use of data augmentation techniques. Model selection becomes challenging as well, because metrics such as DIC and CPO require the likelihood in closed form. Extensions to DIC that use latent variables exist (Celeux et al., 2006), but their performance to select the correct copula model must be investigated. Finally, it would be of interest to extend the conditional copula model of Craiu and Sabeti (2012) to multivariate outcomes as well. For just two outcomes, this model involves a heavy computational burden, so an efficient HMC-based implementation of the multivariate extension of the conditional copula model would be

very useful.

In Chapter 4, we provided two methods to analyze bivariate mixed secondary outcomes in a case-control study. The simulation studies show that the methods can accurately estimate the SNP effect in the secondary outcome model, as well as provide powerful tests of association with the secondary outcomes, while controlling the Type I error rate. We observed that for our simulation settings, the power of the simultaneous test of association was similar to the power of the Bonferroni-corrected separate test of association between the SNP and the continuous secondary outcome. This result is likely because the stronger signal obtained from the continuous outcome "drowns out" the smaller signal from the binary outcome. It would be of interest to lower the signal from the continuous outcome relative to that of the binary outcome, and then compare the power of the simultaneous test and the separate test. We believe that in such a scenario, the simultaneous test would show a larger benefit over the separate test, possibly pushing the SNP below the required level of significance, where the separate tests might fail. In fact, this is exactly what we observed in the EAGLE case study analysis. We are currently investigating this problem through more simulation studies. Another point worth noting is that both of our methods assume that the disease prevalence is either known, or that an estimate is available. For rare diseases, obtaining the disease prevalence may be difficult. As mentioned previously, Xing et al. (2016) showed how to perform the estimation for rare diseases with unknown disease prevalence. We extended this method to the case of bivariate mixed secondary outcomes, but we observed that

the method did not control the Type I error rate if the SNP was not associated with the primary outcome. Further investigation showed that this failure was due to an inadequate sample size; for $n0 = n1 = 3000$, the method controlled the Type I error. We hypothesize that an absence of the disease prevalence constitutes a substantial loss in information provided to the model, which can be made up with the information provided by a larger number of samples. Finally, the methods we have proposed require fitting separate models for each SNP, and using a small genome-wide level of significance to minimize that chance of false positives. Extending the PGEEs from Chapter 2 to case-control studies would likely prove to be a more powerful approach than our current methods. However, this would require some theoretical work to show that unbiased weights can be correctly constructed under penalized estimation.

Appendix A

Proof of conditions in (2.18) from Section 2.2.5

We suppose that the conditions in (2.17) hold, i.e., for $j = 1, \ldots, p+q$,

$$n^{-1}\boldsymbol{w}^{(j)T}\boldsymbol{r} = \lambda_j sign(\hat{\beta}_j) \qquad \forall \hat{\beta}_j \neq 0, \tag{A.1a}$$

$$n^{-1}|\boldsymbol{w}^{(j)T}\boldsymbol{r}| \leq \lambda_j \qquad \forall \hat{\beta}_j = 0. \tag{A.1b}$$

We shall prove that the conditions in (2.18) hold, i.e., for $j = 1, \ldots, p+q$,

$$n^{-1}|\boldsymbol{w}^{(j)T}\boldsymbol{r}^{(-j)}| > \lambda_j \qquad \forall \hat{\beta}_j \neq 0, \tag{A.2a}$$

$$n^{-1}|\boldsymbol{w}^{(j)T}\boldsymbol{r}^{(-j)}| \leq \lambda_j \qquad \forall \hat{\beta}_j = 0. \tag{A.2b}$$

In what follows, quantities that are functions of $\boldsymbol{\beta}$ such as $\boldsymbol{w}^{(j)}, \boldsymbol{r}$ and $\boldsymbol{r}^{(-j)}$ are understood to be calculated at the solution $\hat{\boldsymbol{\beta}}$.

Proof of (A.2a): Assume $\hat{\beta}_j \neq 0$. Recall that

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\eta}), \tag{A.3}$$

where $\boldsymbol{\mu}(\boldsymbol{\eta}) = [\boldsymbol{\mu}(\boldsymbol{\eta}_1)^T, \ldots, \boldsymbol{\mu}(\boldsymbol{\eta}_n)^T]^T$, $\boldsymbol{\mu}(\boldsymbol{\eta}_i) = [\mu_c(\eta_{ic}), \mu_b(\eta_{ib})]^T$, $\eta_{ic} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$, and $\eta_{ib} = \boldsymbol{z}_i^T \hat{\boldsymbol{\beta}}_{\boldsymbol{b}}$. Here, $\mu_c(\cdot)$ and $\mu_b(\cdot)$ are the continuous and the binary inverse link functions respectively, i.e., $\mu_c \equiv g_c^{-1}$ and $\mu_b \equiv g_b^{-1}$ (see Section 2.2.1). Similarly, we have

$$\boldsymbol{r}^{(-\boldsymbol{j})} = \boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\eta}^{(-j)}), \tag{A.4}$$

where $\boldsymbol{\eta}^{(-j)}$ is analogous to $\boldsymbol{\eta}$, but is computed without the jth covariate. In the bivariate mixed outcomes setting, we mean the jth covariate to be the jth column of the full design matrix

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_1} \\ \vdots \\ \boldsymbol{X_n} \end{bmatrix}^{2n \times (p+q)}, \quad \text{where } \boldsymbol{X_i} = \begin{pmatrix} \boldsymbol{x}_i^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{z}_i^T \end{pmatrix}^{2 \times (p+q)}.$$

We assume that no column of $\boldsymbol{X}$ is identically equal to $\boldsymbol{0}$.

Denote the jth column of $\boldsymbol{X_i}$ as the 2-dimensional vector $\boldsymbol{c_i^{(j)}} \equiv [c_{ic}^{(j)}, c_{ib}^{(j)}]^T$. Note

that for fixed i, for $j \in \{1, \ldots, p\}$,

$$c_{ib}^{(j)} = 0, \tag{A.5a}$$

$$\eta_{ic} = \eta_{ic}^{(-j)} + c_{ic}^{(j)} \hat{\beta}_j, \tag{A.5b}$$

$$\eta_{ib} = \eta_{ib}^{(-j)}. \tag{A.5c}$$

Similarly, for fixed i, for $j \in \{p+1, \ldots, p+q\}$,

$$c_{ic}^{(j)} = 0, \tag{A.6a}$$

$$\eta_{ic} = \eta_{ic}^{(-j)}, \tag{A.6b}$$

$$\eta_{ib} = \eta_{ib}^{(-j)} + c_{ib}^{(j)} \hat{\beta}_j. \tag{A.6c}$$

From (A.3) and (A.4),

$$\begin{aligned} \boldsymbol{r}^{(-j)} &= \boldsymbol{r} + \boldsymbol{\mu}(\boldsymbol{\eta}) - \boldsymbol{\mu}(\boldsymbol{\eta}^{(-j)}), \\ \implies n^{-1}\boldsymbol{w}^{(j)T}\boldsymbol{r}^{(-j)} &= n^{-1}\boldsymbol{w}^{(j)T}\boldsymbol{r} + n^{-1}\boldsymbol{w}^{(j)T}[\boldsymbol{\mu}(\boldsymbol{\eta}) - \boldsymbol{\mu}(\boldsymbol{\eta}^{(-j)})]. \end{aligned} \tag{A.7}$$

Using (A.1a),

$$\begin{aligned} n^{-1}\boldsymbol{w^{(j)T}r^{(-j)}} &= \lambda_j sign(\hat{\beta}_j) + n^{-1}\boldsymbol{w^{(j)T}}[\boldsymbol{\mu(\eta)} - \boldsymbol{\mu(\eta^{(-j)})}], \\ &= \lambda_j sign(\hat{\beta}_j) + n^{-1}\sum_{i=1}^{n} w_{ic}^{(j)}(\mu_c(\eta_{ic}) - \mu_c(\eta_{ic}^{(-j)})) + n^{-1}\sum_{i=1}^{n} w_{ib}^{(j)}(\mu_b(\eta_{ib}) - \mu_b(\eta_{ib}^{(-j)})), \end{aligned} \tag{A.8}$$

where $\boldsymbol{w^{(j)}} \equiv [w_{1c}^{(j)}, w_{1b}^{(j)}, \ldots, w_{nc}^{(j)}, w_{nb}^{(j)}]^T$. Based on (A.8), to prove (A.2a), it is sufficient to show that for $\hat{\beta}_j > 0$,

$$w_{ic}^{(j)}(\mu_c(\eta_{ic}) - \mu_c(\eta_{ic}^{(-j)})) \geq 0 \quad \forall i, \tag{A.9a}$$

$$w_{ib}^{(j)}(\mu_b(\eta_{ib}) - \mu_b(\eta_{ib}^{(-j)})) \geq 0 \quad \forall i, \tag{A.9b}$$

and for $\hat{\beta}_j < 0$,

$$w_{ic}^{(j)}(\mu_c(\eta_{ic}) - \mu_c(\eta_{ic}^{(-j)})) \leq 0 \quad \forall i, \tag{A.10a}$$

$$w_{ib}^{(j)}(\mu_b(\eta_{ib}) - \mu_b(\eta_{ib}^{(-j)})) \leq 0 \quad \forall i. \tag{A.10b}$$

We shall prove (A.9). The proof for (A.10) follows analogously.

Assume $\hat{\beta}_j > 0$. Fix i. Now, $\boldsymbol{V_i^{-1}D_i} = \boldsymbol{A_i^{-1/2}\hat{R}^{-1}A_i^{1/2}X_i}$ (see (2.1) and (2.3)).

Hence, the jth column of $\boldsymbol{V_i}^{-1}\boldsymbol{D_i}$ is

$$\boldsymbol{A_i}^{-1/2}\hat{\boldsymbol{R}}^{-1}\boldsymbol{A_i}^{1/2}\boldsymbol{c_i^{(j)}} = \begin{pmatrix} 1/\sqrt{h_{ic}} & 0 \\ 0 & 1/\sqrt{h_{ib}} \end{pmatrix} \begin{pmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \sqrt{h_{ic}} & 0 \\ 0 & \sqrt{h_{ib}} \end{pmatrix} \begin{pmatrix} c_{ic}^{(j)} \\ c_{ib}^{(j)} \end{pmatrix}, \quad \text{(A.11)}$$

where $h_{ic} \equiv h_c(\mu_{ic}) \equiv h_c(\mu_c(\eta_{ic}))$ and $h_{ib} \equiv h_b(\mu_{ib}) \equiv h_b(\mu_b(\eta_{ib}))$ are the continuous and the binary variance functions, assuming unit dispersion parameters. Although we assume unit dispersion parameters to keep notation consistent with the main paper, the proof holds under general dispersion parameters. Continuing from (A.11), the jth column of $\boldsymbol{V_i}^{-1}\boldsymbol{D_i}$ is

$$\frac{1}{1-\hat{\rho}^2} \begin{pmatrix} c_{ic}^{(j)} - \hat{\rho}\sqrt{\frac{h_{ib}}{h_{ic}}}c_{ib}^{(j)} \\ c_{ib}^{(j)} - \hat{\rho}\sqrt{\frac{h_{ic}}{h_{ib}}}c_{ic}^{(j)} \end{pmatrix}.$$

By definition (see Section 2.2.5), $w_{ic}^{(j)}$ and $w_{ib}^{(j)}$ are the first and the second elements respectively of the jth column of $\boldsymbol{V_i}^{-1}\boldsymbol{D_i}$. Hence,

$$w_{ic}^{(j)} = \frac{1}{1-\hat{\rho}^2}\left(c_{ic}^{(j)} - \hat{\rho}\sqrt{\frac{h_{ib}}{h_{ic}}}c_{ib}^{(j)}\right), \quad \text{(A.12a)}$$

$$w_{ib}^{(j)} = \frac{1}{1-\hat{\rho}^2}\left(c_{ib}^{(j)} - \hat{\rho}\sqrt{\frac{h_{ic}}{h_{ib}}}c_{ic}^{(j)}\right). \quad \text{(A.12b)}$$

Subcase 1: $j \in \{1, \ldots, p\}$.

From (A.5a) and (A.12a),

$$w_{ic}^{(j)} = \frac{1}{1-\hat{\rho}^2}c_{ic}^{(j)}. \quad \text{(A.13)}$$

We assume that the mean functions $\mu_c(\cdot)$ and $\mu_b(\cdot)$ are monotonically increasing. This holds true for the identity link, which we have assumed as the link function for the continuous outcomes, and for practically all link functions commonly used for binary outcomes such as logit, probit and complimentary log-log.

With this assumption, the assumption that $\hat{\beta}_j > 0$, and (A.5b), we conclude that $(\mu_c(\eta_{ic}) - \mu_c(\eta_{ic}^{(-j)}))$ and $c_{ic}^{(j)}$ have the same sign. By (A.13), this is also the sign of $w_{ic}^{(j)}$ (assuming $\hat{\rho}^2 < 1$, which our estimation procedure ensures). This proves (A.9a). Also, under this subcase, from (A.5c), $\mu_b(\eta_{ib}) - \mu_b(\eta_{ib}^{(-j)}) = 0$, and (A.9b) holds as an equality.

Subcase 2: $j \in \{p+1, \ldots, p+q\}$.

From (A.6a) and (A.12b),

$$w_{ib}^{(j)} = \frac{1}{1-\hat{\rho}^2} c_{ib}^{(j)}. \tag{A.14}$$

Along the same line of reasoning as in Subcase 1, with (A.6c), we conclude that $(\mu_b(\eta_{ib}) - \mu_b(\eta_{ib}^{(-j)}))$ and $c_{ib}^{(j)}$ have the same sign, which, by (A.14), is the same sign as $w_{ib}^{(j)}$. This proves (A.9b). Also, under this subcase, from (A.6b), $\mu_c(\eta_{ib}) - \mu_c(\eta_{ic}^{(-j)}) = 0$, and (A.9a) holds as an equality.

Note that technically, (A.9) and (A.10) prove (A.2a) with equality:

$$n^{-1}|\boldsymbol{w}^{(j)T}\boldsymbol{r}^{(-j)}| \geq \lambda_j \qquad \forall \hat{\beta}_j \neq 0. \tag{A.15}$$

However, from (A.8), (A.9), and (A.10), equality holds if and only if

$$w_{ic}^{(j)}(\mu_c(\eta_{ic}) - \mu_c(\eta_{ic}^{(-j)})) = 0 \quad \forall i,$$

$$w_{ib}^{(j)}(\mu_b(\eta_{ib}) - \mu_b(\eta_{ib}^{(-j)})) = 0 \quad \forall i.$$

This, in turn, with (A.5), (A.6), (A.13), and (A.14), leaves two possibilities:

$$\hat{\beta}_j = 0,$$

or

$$c_{ic}^{(j)} = 0 \quad \forall i, \quad \text{if } j \in \{1, \ldots, p\}, \tag{A.17a}$$

$$c_{ib}^{(j)} = 0 \quad \forall i, \quad \text{if } j \in \{p+1, \ldots, p+q\}. \tag{A.17b}$$

We have assumed that $\hat{\beta}_j \neq 0$. Then, (A.17) means that a column of the full design matrix $\boldsymbol{X}$ must be identically equal to $\mathbf{0}$, which we have assumed is not the case. Hence, the equality in (A.15) cannot hold, which gives us (A.2a).

Proof of (A.2b): Assume $\hat{\beta}_j = 0$. Hence, $\boldsymbol{\eta} = \boldsymbol{\eta}^{(-j)}$. Then, from (A.7), $n^{-1}\boldsymbol{w}^{(j)T}\boldsymbol{r}^{(-j)} = n^{-1}\boldsymbol{w}^{(j)T}\boldsymbol{r}$, which, with (A.1b), proves (A.2b).

Bibliography

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

Betancourt, M. (2016), "Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo," *arXiv preprint arXiv:1604.00695*.

— (2017), "A conceptual introduction to Hamiltonian Monte Carlo," *arXiv preprint arXiv:1701.02434*.

Braeken, J., Tuerlinckx, F., and De Boeck, P. (2007), "Copula functions for residual dependency," *Psychometrika*, 72, 393–411.

Breheny, P. J. (2009), "Regularized methods for high-dimensional and bi-level variable selection," Ph.D. thesis, University of Iowa, Iowa City, IA.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2016), "Stan: A probabilistic programming language," *Journal of Statistical Software*, 20.

Catalano, P. J. and Ryan, L. M. (1992), "Bivariate latent variable models for clustered discrete and continuous outcomes," *Journal of the American Statistical Association*, 87, 651–658.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., et al. (2006), "Deviance information criteria for missing data models," *Bayesian Analysis*, 1, 651–673.

Chen, H. Y., Kittles, R., and Zhang, W. (2013), "Bias correction to secondary trait analysis with case-control design," *Statistics in Medicine*, 32, 1494–1508.

Chen, J., Bacanu, S.-A., Yu, H., Zhao, Z., Jia, P., Kendler, K. S., Kranzler, H. R., Gelernter, J., Farrer, L., Minica, C., et al. (2016), "Genetic relationship between schizophrenia and nicotine dependence," *Scientific Reports*, 6.

Chen, M.-H., Huang, L., Ibrahim, J. G., and Kim, S. (2008), "Bayesian variable selection and computation for generalized linear models with conjugate priors," *Bayesian Analysis*, 3, 585.

Craiu, R. V. and Sabeti, A. (2012), "In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes," *Journal of Multivariate Analysis*, 110, 106–120.

de Leon, A. R. and Wu, B. (2011), "Copula-based regression models for a bivariate mixed discrete and continuous outcome," *Statistics in Medicine*, 30, 175–185.

Drgonova, J., Walther, D., Singhal, S., Johnson, K., Kessler, B., Troncoso, J., and Uhl, G. R. (2015), "Altered CSMD1 expression alters cocaine-conditioned place preference: mutual support for a complex locus from human and mouse models," *PLOS ONE*, 10.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987), "Hybrid monte carlo," *Physics Letters B*, 195, 216–222.

Fan, J. and Gijbels, I. (1996), *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66, CRC Press.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fu, W. J. (2003), "Penalized estimating equations," *Biometrics*, 59, 126–132.

Geisser, S. (1993), *Predictive Inference*, vol. 55, CRC Press.

Gelfand, A. E. and Dey, D. K. (1994), "Bayesian model choice: asymptotics and exact calculations," *Journal of the Royal Statistical Society: Series B (Methodological)*, 501–514.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), "Model determination using predictive distributions with implementation via sampling-based methods," in *Bayesian Statistics 4*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford University Press, pp. 147–167.

Gelfand, A. E. and Smith, A. F. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 1, 515–534.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008), "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, 1360–1383.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 457–472.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 721–741.

Genest, C. and Nešlehová, J. (2007), "A primer on copulas for count data," *Astin Bulletin*, 37, 475–515.

Geyer, C. J. (1994), "Estimating Normalizing Constants and Reweighting Mixtures in Markov chain Monte Carlo," Tech. Rep. 568, School of Statistics, University of Minnesota, Minneapolis, MN.

Ghosh, A., Wright, F. A., and Zou, F. (2013), "Unified analysis of secondary traits in case-control association studies," *Journal of the American Statistical Association*, 108, 566–576.

Griewank, A. and Walther, A. (2008), *Evaluating derivatives: principles and techniques of algorithmic differentiation*, SIAM.

Hall, D. B. and Severini, T. A. (1998), "Extended generalized estimating equations for clustered data," *Journal of the American Statistical Association*, 93, 1365–1375.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Hoeffding, W. (1940), "Massstabinvariante Korrelationstheorie," *Schriften des Mathematischen Seminars und des Instituts für Angewandte Mathematik der Universität Berlin*, 5, 181–233.

Hoffman, M. D. and Gelman, A. (2014), "The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15, 1593–1623.

Hunter, D. R. and Li, R. (2005), "Variable selection using MM algorithms," *Annals of Statistics*, 33, 1617–1642.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2005), *Bayesian survival analysis*, Wiley Online Library.

Joe, H. (2014), *Dependence modeling with copulas*, CRC Press.

Joe, H. and Xu, J. J. (1996), "The estimation method of Inference Functions for Margins for multivariate models," Tech. rep., Department of Statistics, University of British Columbia.

Johnson, B. A., Lin, D., and Zeng, D. (2008), "Penalized estimating functions and variable selection in semiparametric regression models," *Journal of the American Statistical Association*, 103, 672–680.

Kotz, S. and Nadarajah, S. (2004), *Multivariate t-distributions and their applications*, Cambridge University Press.

Landi, M. T., Consonni, D., Rotunno, M., Bergen, A. W., Goldstein, A. M., Lubin, J. H., Goldin, L., Alavanja, M., Morgan, G., Subar, A. F., et al. (2008), "Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer," *BMC Public Health*, 8, 203.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 13–22.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992), "Multivariate regression analyses for categorical data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–40.

Lin, D. and Zeng, D. (2009), "Proper analysis of secondary phenotype data in case-control association studies," *Genetic Epidemiology*, 33, 256–265.

Liu, C. (2004), "Robit regression: a simple robust alternative to logistic and probit regression," in *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, eds. Gelman, A. and Meng, X.-L., Wiley: London, chap. 21, pp. 227–238.

Liu, J., Pei, Y., Papasian, C. J., and Deng, H.-W. (2009), "Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations," *Genetic Epidemiology*, 33, 217–227.

McCullagh, P. (1984), "Generalized linear models," *European Journal of Operational Research*, 16, 285–292.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, 21, 1087–1092.

Monsees, G. M., Tamimi, R. M., and Kraft, P. (2009), "Genome-wide association scans for secondary traits using case-control samples," *Genetic Epidemiology*, 33, 717–728.

Neal, R. M. (1993), "Probabilistic inference using Markov chain Monte Carlo methods," .

— (2011), "MCMC using Hamiltonian dynamics," in *Handbook of Markov chain Monte Carlo*, eds. Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., CRC Press, pp. 113–162.

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

Nesterov, Y. (2009), "Primal-dual subgradient methods for convex problems," *Mathematical Programming*, 120, 221–259.

Nikoloulopoulos, A. K. and Karlis, D. (2008), "Multivariate logit copula model with an application to dental data," *Statistics in Medicine*, 27, 6393–6406.

Prentice, R. L. and Zhao, L. P. (1991), "Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses," *Biometrics*, 47, 825–839.

Rochon, J. (1996), "Analyzing bivariate repeated measures for discrete and continuous outcome variables," *Biometrics*, 52, 740–750.

Roy, V., Evangelou, E., and Zhu, Z. (2016), "Efficient estimation and prediction for the Bayesian binary spatial model with flexible link functions," *Biometrics*, 72, 289–298.

Schepsmeier, U. and Stöber, J. (2014), "Derivatives and Fisher information of bivariate copulas," *Statistical Papers*, 55, 525–542.

Schifano, E. D., Li, L., Christiani, D. C., and Lin, X. (2013), "Genome-wide association analysis for multiple continuous secondary phenotypes," *The American Journal of Human Genetics*, 92, 744–759.

Sherva, R., Wang, Q., Kranzler, H., Zhao, H., Koesterer, R., Herman, A., Farrer, L. A., and Gelernter, J. (2016), "Genome-wide association study of cannabis dependence severity, novel risk variants, and shared genetic risks," *JAMA psychiatry*, 73, 472–480.

Silva, R. d. S. and Lopes, H. F. (2008), "Copula, marginal distributions and model selection: a Bayesian note," *Statistics and Computing*, 18, 313–320.

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

Smith, M. S. and Khaled, M. A. (2012), "Estimation of copula models with discrete margins via Bayesian data augmentation," *Journal of the American Statistical Association*, 107, 290–303.

Song, P. X.-K., Li, M., and Yuan, Y. (2009), "Joint regression analysis of correlated data using Gaussian copulas," *Biometrics*, 65, 60–68.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.

Tchetgen, E. J. T. (2013), "A general regression framework for a secondary outcome in case-control studies," *Biostatistics*, 117–128.

Teixeira Pinto, A. and Normand, S.-L. T. (2009), "Correlated bivariate continuous and binary outcomes: issues and applications," *Statistics in Medicine*, 28, 1753–1773.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodology)*, 58, 267–288.

Uhl, G. R., Drgon, T., Liu, Q.-R., Johnson, C., Walther, D., Komiyama, T., Harano, M., Sekine, Y., Inada, T., Ozaki, N., et al. (2008a), "Genome-wide association for methamphetamine dependence: convergent results from 2 samples," *Archives of General Psychiatry*, 65, 345–355.

Uhl, G. R., Liu, Q.-R., Drgon, T., Johnson, C., Walther, D., Rose, J. E., David, S. P., Niaura, R., and Lerman, C. (2008b), "Molecular genetics of successful smoking cessation: convergent genome-wide association study results," *Archives of General Psychiatry*, 65, 683–693.

Wang, L., Zhou, J., and Qu, A. (2012), "Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis," *Biometrics*, 68, 353–360.

Wei, J., Carroll, R. J., Müller, U. U., Keilegom, I. V., and Chatterjee, N. (2013), "Robust estimation for homoscedastic regression in the secondary analysis of case-control data," *Journal of the Royal Statistical Society: Series B (Methodology)*, 75, 185–206.

Wolfson, J. (2011), "EEBoost: A general method for prediction and variable selection based on estimating equations," *Journal of the American Statistical Association*, 106, 296–305.

Wu, B. and de Leon, A. R. (2014), "Gaussian copula mixed models for clustered mixed outcomes, with application in developmental toxicology," *Journal of Agricultural, Biological, and Environmental Statistics*, 19, 39–56.

Xing, C., McCarthy, J. M., Dupuis, J., Cupples, A. L., Meigs, J. B., Lin, X., and Allen, A. S. (2016), "Robust analysis of secondary phenotypes in case-control genetic association studies," *Statistics in Medicine*, 35, 4226–4237.

Yi, H., Breheny, P. J., Imam, N., Liu, Y., and Hoeschele, I. (2015), "Penalized multimarker vs. single-marker regression methods for genome-wide association studies of quantitative traits," *Genetics*, 199, 205–222.

Zhang, C. H. (2007), "Penalized linear unbiased selection," Tech. Rep. 2007–003, Department of Statistics and Bioinformatics, Rutgers University, Piscataway, NJ.

Zhong, X., Drgonova, J., Li, C.-Y., and Uhl, G. R. (2015), "Human cell adhesion molecules: annotated functional subtypes and overrepresentation of addiction-associated genes," *Annals of the New York Academy of Sciences*, 1349, 83–95.

Zimmerman, D. (2013), "Analysis of mixed outcomes in econometrics: Applications in health economics," in *Analysis of mixed data: methods & applications*, eds. de Leon, A. R. and Chough, K. C., CRC Press, chap. 11, pp. 157–172.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Methodology)*, 67, 301–320.