University of Calgary

PRISM Repository **https://prism.ucalgary.ca**

The Vault Open Theses and Dissertations

2013-07-10

Contributions to Copula Modeling of Mixed Discrete-Continuous Outcomes

Beilei, Wu

Beilei, W. (2013). Contributions to Copula Modeling of Mixed Discrete-Continuous Outcomes (Doctoral thesis, University of Calgary, Calgary, Canada). Retrieved from https://prism.ucalgary.ca. doi:10.11575/PRISM/25449

http://hdl.handle.net/11023/791

Downloaded from PRISM Repository, University of Calgary

UNIVERSITY OF CALGARY

Contributions to Copula Modeling of Mixed Discrete-Continuous Outcomes

by

Beilei Wu

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

July, 2013

© Beilei Wu 2013

Abstract

This thesis includes three topics that are concerned with joint modeling and analysis of multiple correlated mixed discrete and continuous outcomes.

The first topic is concerned with the analysis of multiple correlated discrete and continuous outcomes that are observed on the same subjects over time in the case of longitudinal studies, or from clustered subjects in cross-sectional settings. Joint analysis of such disparate responses (i.e., mixed discrete and continuous outcomes) is problematic in practice due mainly to the difficulty of defining or constructing a joint model. Our proposed approach is based on a new generalized linear mixed model (GLMM) that accounts for associations between the outcomes (of the same or of different types) for the same subject at the same time point, and/or at different time points for the longitudinal data, or between mixed outcomes within clusters, including the intrinsic association between the mixed outcomes for the same subject, in clustered settings. A latent-variable approach is adopted to sidestep complications of direct application of copula models to discrete data. The approach yields regression parameters that are marginally meaningful, and permits the adoption of flexible non-Gaussian distributions for the mixed outcomes as well as for the random effects. Special cases of our model include conventional GLMMs previously proposed by a number of authors, among whom are Faes (2013), Gueorguieva (2013), and Lin *et al.* (2010). Full and pairwise likelihood estimation methods are implemented for the model using PROC NLMIXED in SAS. The proposed methodology is illustrated using individual panel data on the wages, work hours, and union memberships, and data on fetal malformation and weight in a developmental toxicity study on mice.

In the second topic, we adopt the "continuous-ation" approach of Machado and Santos Silva (2005) and Denuit and Lambert (2005) to construct a Gaussian copula joint model for mixed discrete and continuous outcomes. The joint model does not require a latent

variable formulation of the discrete outcomes, and does not suffer from the complications of directly using discrete margins in copula models (Genest and Nešlehovà, 2007). A surrogate likelihood approach to estimation is implemented for the model and empirical results concerning the relative bias and efficiency of the resulting estimates are reported. The proposed methodology is illustrated using data on burn injuries.

The third and final topic concerns a methodology for calculating the sample size in clinical trials with multiple mixed binary and continuous co-primary endpoints. The Gaussian copula joint model we proposed permits the adoption of flexible marginal distributions for the mixed endpoints, and includes the conditional grouped continuous model (CGCM) — a popular model for mixed endpoints based on the multivariate Gaussian distribution — as a special case. The proposed methodology adopts a latent variable description of the binary endpoints and makes use of tests on the latent means to test for differences in the binary proportions. This approach results in a simple and streamlined methodology akin to that for multiple continuous co-primary endpoints studied in Sozu *et al.* (2011). In addition, our approach is more powerful than that recently proposed by Sozu *et al.* (2012), in that it yields smaller sample sizes at powers comparable to those considered in Sozu *et al.* (2012). We report the results of empirical comparisons as well as a numerical illustration of our methodology.

Acknowledgements

My first acknowledgment must go to my supervisor, Dr. A. R. de Leon. I express my heartfelt gratitude for his invaluable assistance, support, and guidance during the course of my graduate studies at the University of Calgary, as well as his patience, encouragement, and excellent advice during the preparation and writing of this thesis.

I also thank the members of the thesis committee, Drs. Gemai Chen, Xuewen Lu, Tolulope Sajobi (Department of Community Health Sciences, University of Calgary) and Abdulkadir Hussein (Department of Mathematics and Statistics, University of Windsor). It is no easy task reviewing a thesis, and I am grateful for their thoughtful and detailed comments. I am especially grateful to Drs. Chen and Lu for their generous help and their practical advice during the course of my studies.

I am thankful to the Faculty of Graduate Studies, for awarding me a Queen Elizabeth II Graduate Scholarship, a Women's Club Graduate Scholarship, and a Chancellor's Challenge Graduate Scholarship. Thanks also go to the Department of Mathematics and Statistics, for its generous financial support through graduate assistantships and graduate research scholarships.

I also express my heartfelt thanks and gratitude to the following people who helped me, worked with me, and shared invaluable ideas with me during my nine years of studies at the University of Calgary:

Dr. Hyang Mi Kim, for kindly writing reference letters on my behalf on numerous occasions.

Yanmei Fei, for being such a nice, helpful, and efficient administrator. Her assistance at crucial times is much appreciated.

Niroshan Withanage, for being the best classmate and officemate one can ever ask for. I hope he becomes a wonderful professor like our supervisor.

Longlong Huang, for the good times we had sharing ideas on how to raise kids and for her kindness when I lived alone in this new city.

Mrs. Marguerite Fenyvesi and Dr. Alberto Nettel-Aguirre, for always remembering to send me useful information and for their reference letters when I was looking for jobs. Many thanks as well to Shenghua Cai for his valuable suggestions about my program codes.

Finally, I express my love and gratitude to my father (although he has left me, his spirit still sustains me to this day) and my mother, my husband Zhijian, and my lovely sons Zixi and Zirui, for their understanding and spiritual support during my time at the University of Calgary.

Table of Contents

Abstract . . . i
Acknowledgements . . . iii
Table of Contents . . . v
List of Tables . . . vii
List of Figures . . . ix
1 Introduction . . . 1
1.1 Background of thesis . . . 1
1.2 Review of literature . . . 4
1.2.1 Direct approaches to joint model construction . . . 4
1.2.2 Indirect approaches to joint model construction . . . 6
1.3 Overview of thesis . . . 11
2 Joint analysis of multiple mixed longitudinal outcomes via Gaussian copula mixed models . . . 13
2.1 Introduction . . . 13
2.2 Gaussian copula joint models . . . 14
2.3 Case of longitudinal data . . . 17
2.4 Associations . . . 20
2.4.1 Conditional dependence . . . 21
2.4.2 Marginal dependence . . . 22
2.5 Likelihood estimation . . . 23
2.5.1 Full likelihood estimation . . . 24
2.5.2 Pairwise likelihood estimation . . . 25
2.6 Application to wages-hours-union memberships data . . . 27
2.7 Discussion . . . 33
3 Gaussian copula mixed models for clustered mixed outcomes in developmental toxicology . . . 35
3.1 Introduction . . . 35
3.2 Gaussian copula mixed model . . . 36
3.2.1 Distribution of residual errors . . . 40
3.2.2 Marginal associations between outcomes . . . 41
3.2.3 Conditional assessment of outcomes . . . 43
3.3 Likelihood estimation . . . 44
3.4 Simulation study . . . 44
3.5 Developmental toxicity of ethylene glycol . . . 47
3.5.1 Conditional plots . . . 52
3.5.2 Quantitative risk assessment . . . 53
3.6 Discussion . . . 57
4 Gaussian copula joint models for mixed outcomes with “continued” binary variables . . . 61
4.1 Introduction . . . 61
4.2 Gaussian copula joint model with jittered binary outcomes . . . 63
4.3 Associations . . . 65

4.4 Likelihood estimation 67
4.5 Simulation study 68
4.6 Application to burn injury data 70
4.7 Discussion 75
5 Sample size determination in clinical trials with mixed co-primary endpoints via Gaussian copulas 77
5.1 Introduction 77
5.2 Gaussian copula joint model 79
5.3 Sample size calculation via latent-means tests 81
5.4 Empirical comparisons 84
5.5 Application to PREMIER Study 87
5.6 Discussion 91
6 Conclusion 95
6.1 Summary 95
6.2 Future research 98
A SAS codes for Gaussian copula mixed models 100
A.1 Logit-normal-normal Gaussian copula mixed model 100
A.2 Robit-t Gaussian copula mixed model 103
B Derivation of conditional mean (3.13) 105
Bibliography 107

List of Tables

2.1 Summary statistics for data on union membership, hourly wage and annual hours from 1980-1987 U. S. National Longitudinal Survey. 27

2.2 Full and pairwise likelihood estimates, their SEs, and p-values for wages-hours-union memberships data, using a logit-normal-normal Gaussian copula mixed model with shared random intercepts and slopes. 30

2.3 Estimated marginal biserial correlations between union membership indicator and log-hourly wage for time points $t = 1, \cdots, 8$, using full likelihood estimation. 31

2.4 Estimated marginal biserial correlations between union membership indicator and log-hourly wage for time points $t = 1, \cdots, 8$, using pairwise likelihood estimation. 31

3.1 Summary statistics for fetal malformation and fetal weight outcomes for EG mice data. 48

3.2 Estimates, their SEs and t-values for EG data using the logit-normal Gaussian copula mixed model with correlated random effects (REs), with only linear dose effects (Model 1) and with both linear and quadratic dose effects (Model 2), and using the robit-t Gaussian copula mixed model with correlated REs, with only linear dose effects (Model 3) and with both linear and quadratic dose effects (Model 4). 49

3.3 Estimates of overall risk probability $P(z)$, additional risk $r_a(z)$, and extra risk $r_e(z)$, for dose $z = 0, 0.75, 1.5, 3$g/kg, based on Models 1 to 4. 56

3.4 Benchmark dose BMD_q and lower effective dose LED_q for $100q = 1, 5, 10\%$, based on extra risk, estimated using Models 1 to 4. 58

4.1 MLEs (based on $M = 1000$ jitters), their SEs, and t-values for the Gaussian copula joint model (Model 1) with marginal models $\text{logit}(p_i) = \text{logit}\{E(X_i)\} = \alpha_1 + \alpha_2 \times \text{age}$ and $Y_i \sim \textit{normal}(\mu_i = \beta_1 + \beta_2 \times \text{age}, \sigma^2)$, $i = 1, \cdots, 981$. For comparison, we also included Song *et al.*'s (2009) result based on a Gaussian copula joint model with $X_i \sim \textit{Bernoulli}(p_i)$ used directly as a discrete margin and $Y_i \sim \textit{normal}(\mu_i, \sigma^2)$, and de Leon and Wu's (2011) result based on the robit-normal Gaussian copula joint model in Chapter 2 with a t-latent distribution (hence, a robit regression model for X_i) and $Y_i \sim \textit{normal}(\mu_i, \sigma^2)$. 71

5.1 Sample sizes $n = n_{T_0} = n_{T_1}$ for overall power $1 - \beta \approx 80\%$, $\alpha = 0.025$, $Q = C = 1$, and with fixed $z_2^\dagger / z_1^\dagger$ and $\delta^* = \mu_{T_1}^* - \mu_{T_0}^* = \Phi^{-1}(\pi_{T_1}) - \Phi^{-1}(\pi_{T_0})$. Marginal sample sizes E_1 and E_2 for the continuous and binary endpoints, respectively, were calculated with individual power of at least 80%. 85

5.2 Sample sizes $n = n_{T_0} = n_{T_1}$ for overall power $1 - \beta \approx 80\%$, $\alpha = 0.025$, $Q = C = 1$, and with fixed $\delta^* = \mu_{T_1}^* - \mu_{T_0}^* = \Phi^{-1}(\pi_{T_1}) - \Phi^{-1}(\pi_{T_0})$ and $\bar{\delta}$. Marginal sample sizes E_1 and E_2 for the continuous and binary endpoints, respectively, were calculated with individual power of at least 80%. 88

5.3 Sample sizes $n = n_{T_0} = n_{T_1}$ in Sozu *et al.* (2012) (obtained with approximate overall power of 80% using their approach) and corresponding power values $1 - \beta$ obtained using our approach, for $\alpha = 0.025$, $Q = C = 1$, and with fixed $\bar{\delta}$ (as determined by fixing $z_2^\dagger / z_1^\dagger$ in Table 5.1) and $\delta^* = \mu^*_{T_1} - \mu^*_{T_0} = \Phi^{-1}(\pi_{T_1}) - \Phi^{-1}(\pi_{T_0})$. The rows are grouped according to the fixed values $z_2^\dagger / z_1^\dagger = 1, 1.5$, and 3, in Table 5.1; note, however, that these are not the actual values of $z_2^\dagger / z_1^\dagger$ for the given $\bar{\delta}, \delta^*$, and n. 89

5.4 Sample sizes $n = n_{T_0} = n_{T_1}$ in Table 5.1 (obtained with approximate or asymptotic overall power of 80% using the normal-logistic Gaussian copula joint model) and corresponding power values $1 - \beta$ obtained using Monte Carlo approximation, for $\alpha = 0.025$ and with fixed $\bar{\delta}$ (as determined by fixing $z_2^\dagger / z_1^\dagger$ in Table 5.1) and $\delta^* = \mu^*_{T_1} - \mu^*_{T_0} = \Phi^{-1}(\pi_{T_1}) - \Phi^{-1}(\pi_{T_0})$. The rows are grouped according to the fixed values $z_2^\dagger / z_1^\dagger = 1, 1.5$, and 3, in Table 5.1. . . . 90

5.5 Sample sizes $n = n_{T_0} = n_{T_1}$ for endpoints mTSS (continuous) and ACR50 (binary), for overall power $1 - \beta \approx 80\%$ and $\alpha = 0.025$, with fixed $\bar{\delta}$ and δ^*. Sample sizes E_1 and E_2 for mTSS and ACR50, respectively, were calculated with individual power of at least 80%. 92

5.6 Sample sizes $n = n_{T_0} = n_{T_1}$ in Sozu *et al.* (2012) (obtained with approximate overall power of 80% using their approach) for endpoints mTSS (continuous) and ACR50 (binary), and corresponding power values $1 - \beta$ obtained using our approach, for $\alpha = 0.025$, with fixed $\bar{\delta}$ and δ^*. 93

List of Figures and Illustrations

2.1 Spaghetti plots for log(Hourly wage) and log(Annual hours). 28
2.2 Histograms of log(Hourly wage) at time $t = 1$ to time $t = 8$. 28
2.3 Histograms of log(Annual hours) at time $t = 1$ to time $t = 8$. 29

3.1 Contour plots for the joint density $f_{\varepsilon^*_{ih},\varepsilon_{ih}}(\cdot)$ of ε^*_{ih} and ε_{ih}, for various $\widetilde{\rho}$ values, with $\varepsilon^*_{ih} \sim$ *logistic*$(0,1)$ and $\varepsilon_{ih} \sim$ *normal*$(0,1)$. 41
3.2 Relative bias "$100 \times$(mean of estimates–true value)/true value" of MLEs and PLEs of (a) α_1, (b) α_2, (c) λ^*, (d) β_1, (e) β_2, (f) λ, (g) σ, and (h) $\widetilde{\rho}$ for robit-normal mixed model with shared litter-level normal random effects. . . 45
3.3 Relative efficiency "mean of SEs/empirical SD" of MLEs and PLEs of (a) α_1, (b) α_2, (c) λ^*, (d) β_1, (e) β_2, (f) λ, (g) σ, and (h) $\widetilde{\rho}$ for robit-normal mixed model with shared litter-level normal random effects. 46
3.4 LOESS-smoothed marginal residual plots for EG mice data using the logit-normal Gaussian copula mixed model with correlated random effects. 50
3.5 LOESS-smoothed marginal residual plots for EG mice data using the robit-t Gaussian copula mixed models with correlated random effects. 51
3.6 Plots of (a and d) $P(X_{ih} = 1|Y_{ih} = y)$ as function of dose z; (b and f) $E(Y_{ih}|X_{ih} = x)$ as function of dose z; and (c and g) $P(X_{ih} = 1|Y_{ih} = y)$ as function of y, for fixed dose z; $E(Y_{ih})$ and $P(X_{ih} = 1)$, as functions of dose z, are shown for comparison. Top panel obtained using Model 1, bottom panel using Model 3. 54
3.7 Plots of (a and d) $P(X_{ih} = 1|Y_{ih} = y)$ as function of dose z; (b and f) $E(Y_{ih}|X_{ih} = x)$ as function of dose z; and (c and g) $P(X_{ih} = 1|Y_{ih} = y)$ as function of y, for fixed dose z; $E(Y_{ih})$ and $P(X_{ih} = 1)$, as functions of dose z, are shown for comparison. Top panel obtained using Model 2, bottom panel using Model 4. 55

4.1 Relationship between normal correlation $\widetilde{\rho}$ vs. the corresponding Kendall's tau $\tau\{\Phi^{-1}(F_{Y_i^*}(Y_i^*)), \Phi^{-1}(F_{Y_i}(Y_i))\} = \tau(Y_i^*, Y_i)$. 66
4.2 Plots (a), (b), (c), (d), (e), and (f) represent the relative bias of MLEs of α_1, α_2, β_1, β_2, σ, and $\widetilde{\rho}$, respectively, based on $R = 500$ repeats and $M = 1000$ jitters for a Gaussian copula joint model with logit link for binary outcome. . 69
4.3 Plots (a), (b), (c), (d), (e), and (f) represent the relative efficiency of MLEs of α_1, α_2, β_1, β_2, σ, and $\widetilde{\rho}$, respectively, based on $R = 500$ repeats and $M = 1000$ jitters for a Gaussian copula joint model with logit link for binary outcome. . 70
4.4 Index plot (a) and LOESS-smoothed plot (b) of marginal residuals against predicted $Y_i = \log(\text{burn area}+1)$ for burn injury data, using a Gaussian copula joint model with marginal models $\text{logit}(p_i) = \text{logit}\{E(X_i)\} = \alpha_1 + \alpha_2 \times \text{age}$ and $Y_i \sim$ *normal*$(\mu_i = \beta_1 + \beta_2 \times \text{age}, \sigma^2)$ (Model 1). 72

4.5 Plots of (a) the conditional probability of death as a function of age z at fixed $y = \log(\text{burn area} + 1) = 6.1, 6.9, 7.7$, and (b) the conditional mean of $y = \log(\text{burn area} + 1)$, given the disposition of death $x = 1$, as a function of age z. 74

5.1 Overall power as a function of ρ for $Q = C = 1$, with the same set-up considered in Sozu *et al.* (2012). 83

Chapter 1

Introduction

1.1 Background of thesis

Multivariate mixed discrete (i.e., nominal, ordinal, or binary) and continuous outcomes are commonplace in health, medicine and many other fields (e.g., applied econometrics, actuarial science). One example of mixed outcomes in a study on burn injuries includes the disposition of death of a patient, a binary outcome, and the patient's total burn area, a continuous outcome. An example from individual panel data on 545 men during the period 1980–1987 (Vella and Verbeek, 1998) includes the workers' wages and work hours, which are continuous variables, and their union membership, which is binary variable. Joint analysis of such outcomes entails specification of flexible models that take into account the associations among the outcomes. Such joint models for multivariate mixed outcomes allow analysts to simultaneously account for relationships between outcomes, assess the joint influence of predictors/covariates on them, and characterize within-cluster (or subject) associations for clustered data (or longitudinal data). From a statistical standpoint, joint analysis of such outcomes avoids multiple testing and naturally leads to global tests, thus resulting in increased power and better control of Type I error rates (de Leon and Carrière Chough, 2010; de Leon and Wu, 2011). Significant efficiency gains over separate univariate analysis have also been reported when dealing with multivariate mixed outcomes with missing values (McCulloch, 2008). de Leon and Carrière Chough (2013) provides a book length treatment of mixed data analysis.

Methods for continuous multivariate responses are well developed and traditional approaches for joint analysis, e.g., generalized linear mixed models (GLMMs), can be used for inference. However, joint analysis of mixed discrete and continuous data often leads to

complications in practice due to the relative lack of available models, mainly due to difficulties in constructing a mixed-outcome joint model. One approach is to convert one type of outcomes to another (e.g., discretization of continuous variables), then to adopt an appropriate method of analysis. While simple and straightforward, this approach is clearly not efficient since, for example, the crude coding of continuous variables is not satisfactory in many applications. We need to develop alternative methods of constructing or specifying, either directly or indirectly, a joint model or joint distribution for mixed outcomes.

A variety of approaches to joint model construction for mixed discrete and continuous outcomes have been investigated by a number of authors. One direct approach employs the factorization method, which specifies the joint distribution as a product of a marginal distribution for one set of outcomes and a conditional distribution for the other. Olkin and Tate (1961) introduced so-called general location models (GLOMs), which are defined as the product of a multinomial distribution for the discrete outcomes, and a conditional multivariate normal distribution for the continuous outcomes. Recent applications of GLOMs are illustrated by Fitzmaurice and Laird (1997,1995), Hirakawa (2012), and Kang and Yang (2013), among others. Alternatively, a joint distribution can be formed from the product of a marginal distribution for the continuous outcomes and a conditional distribution for the discrete outcomes, given the continuous outcomes; an example of such a model is the conditional grouped continuous model (CGCM) (see, e.g., de Leon, 2005). A number of refinements and extensions of both GLOMs and CGCMs have been studied by several authors; for example, de Leon and Carrière (2007) recently introduced the general mixed-data model (GMDM), a hybrid of GLOMs and CGCMs that can be used to model the joint distribution of mixed nominal, ordinal, and continuous data. Even though factorization provides a general route for directly specifying mixed-outcome joint distributions, the resulting joint model is not invariant to the direction of conditioning taken or the factorization adopted; that is, factorization models are not unique. As a consequence, different factorization models may

yield very different estimates and inferences, especially of associations. In addition, they induce a hierarchy in the data which asymmetrically treats conditioned variables as primary responses, and conditioning variables as intermediate outcomes. Furthermore, these models become intractable in high-dimensional data settings, since the number of possible factorizations increases with the number of outcomes. For example, there are 2 possible factorizations for bivariate data, 6 for trivariate data, and 12 for data with 4 outcomes.

An indirect approach to joint model construction introduces shared or correlated random effects to incorporate associations between outcomes in the joint model. This approach overcomes the drawbacks of the factorization models by treating the outcomes symmetrically. In addition, the hierarchical structure of the resulting joint model can account for different measurement levels, delineate various associations in the data, incorporate covariate effects, and can be readily extended to longitudinal and clustered data settings. Although random effects models are attractive in practice, they also have their drawbacks. For one, the correlations between outcomes may be restricted within narrow ranges; for another, computation may become infeasible in high-dimensional problems. See de Leon and Carrière Chough (2010) and McCulloch (2008) for more details.

A recent alternative indirect strategy for jointly modeling mixed outcomes that has attracted the attention of researchers involves using copula functions. Applications of this approach are found in de Leon and Wu (2011), Song et al. (2009), and Zimmer and Trivedi (2006), among others. Trivedi and Zimmer (2007) provide general discussions of copula functions and their statistical applications. The theoretical details in copula construction and discussions of important methodological issues are given in Song (2007) and Joe (1997).

In the next section, the above methodologies and strategies are further reviewed and a number of issues concerning their applications to joint modeling of mixed correlated outcomes are discussed. These issues involve model specification as well as ensuing inference based on such models. Particular attention is given to advantages and disadvantages of various

approaches. The chapter concludes with a brief description and overview of the thesis.

1.2 Review of literature

Data from statistical applications in clustered and longitudinal settings are usually combinations of mixed discrete and continuous outcomes. To analyze such correlated outcomes jointly, researchers usually emphasize how to determine the mixed outcomes' joint distribution, then from this joint models, how to obtain associations in the data. Many different approaches of joint model specifications for mixed outcomes have been proposed in the literature. To fix notations, let the data be represented by vectors $\mathbf{X}$ and $\mathbf{Y}$ of discrete and continuous outcomes, respectively, where $\mathbf{X}$ can comprise nominal, ordinal or binary outcomes. Joint analysis of $\mathbf{X}$ and $\mathbf{Y}$ requires either direct or indirect specification of their joint density $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$. This is illustrated in the subsequent sections.

1.2.1 Direct approaches to joint model construction

As discussed in Section 1.1, factorization is commonly used in the literature as a direct approach in constructing a joint model for mixed outcomes. This approach entails factorizing the joint density $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$ into a conditional density of one set of outcomes and a marginal density of the other. As a consequence, this approach suggests two formulations of mixed-outcome joint models. The first factorizes the joint density of $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$ as $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y})f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$, where $f_{\mathbf{Y}}(\mathbf{y})$ and $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ denote the marginal density of $\mathbf{Y}$ and conditional density of $\mathbf{X}$ given $\mathbf{Y}$, respectively. The second uses the reverse factorization $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$, where $f_{\mathbf{X}}(\mathbf{x})$ represents the marginal density of $\mathbf{X}$ and $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ represents the conditional density of $\mathbf{Y}$ given $\mathbf{X}$.

In the former factorization model, it can be supposed that a thresholded latent vector $\mathbf{Y}^*$ describes the vector $\mathbf{X}$ and $\mathbf{Y}$ and $\mathbf{Y}^*$ have a joint multivariate normal distribution, resulting in the so-called CGCM, with a conditional probit regression model for $\mathbf{X}$, given $\mathbf{Y}$;

see Catalano and Ryan (1992), Catalano (1997) and de Leon (2005), for recent extensions and applications of this model. Note that $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y})f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ is specified through $f_{\mathbf{Y}^*,\mathbf{Y}}(\mathbf{y}^*,\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y})f_{\mathbf{Y}^*|\mathbf{Y}}(\mathbf{y}^*|\mathbf{y})$, where $f_{\mathbf{Y}^*,\mathbf{Y}}(\mathbf{y}^*,\mathbf{y})$ is conveniently formed as multivariate normal distribution, so that $f_{\mathbf{Y}}(\mathbf{y})$ and $f_{\mathbf{Y}^*|\mathbf{Y}}(\mathbf{y}^*|\mathbf{y})$ are modeled as normal densities. However, the conditional mean model $\boldsymbol{\mu}_{\mathbf{X}|\mathbf{Y}} = E(\mathbf{X}|\mathbf{Y})$ treats $\mathbf{Y}$ as the primary response and $\mathbf{X}$ as an intermediate outcome. Fedorov *et al.* (2012) recently applied this model to study optimality in dose-finding clinical studies with mixed categorical and continuous responses.

The latter factorization model, referred to as GLOM, is obtained by reversing the direction of the above factorization for the CGCM. The GLOM was introduced by Olkin and Tate (1961) and was later generalized by Fitzmaurice and Laird (1995) to incorporate regression models for mixed bivariate data. In their model, $\mathbf{X}$ is treated as the primary response and $\mathbf{Y}$ as the intermediate outcome; the resulting estimates, especially of the association between the binary and continuous responses, are more often than not quite different from those obtained from the CGCM. An application of the GLOM to quantitative risk assessment in developmental toxicology is studied by George *et al.* (2007) while its extension to mixed longitudinal count and continuous outcomes with possibly missing data is discussed in Kang and Yang (2013). Further refinements and extensions of CGCM and GLOM have been investigated by de Leon and Carrière (2007), who unified and synthesized GLOM and CGCM into the GMDM, a hybrid of the two that includes both as special cases.

Despite the fact that factorization models are straightforward to construct, they are inadequate, and hence inappropriate, in many applications for several reasons. First, it is important to decide the direction of conditioning when joint models are constructed. This is because different factorizations may yield very different estimates and inferences, especially of the associations among outcomes. The choice of the conditioning outcomes in practice, however, is made mainly for statistical convenience; as a consequence, the resulting models using different factorizations are not comparable. Second, factorization models fail

to properly account for different levels of measurement in the data, e.g., ordinal or count data in the case of GLOM and nominal binary variables in the case of CGCM. Third, these models may become intractable and computationally infeasible in applications with high-dimensional data. Because of these, a more symmetrical treatment of outcomes is needed in practice. Teixeira-Pinto and Normand (2009) recently provided an extensive discussion of these issues in the context of clustered mixed binary and continuous outcomes.

1.2.2 Indirect approaches to joint model construction

Indirect approaches of constructing correlated mixed-outcome joint models have also been previously studied by many authors; see, for example, Gueorguieva and Agresti (2001), Faes *et al.* (2008), and Faes (2013), among others. One such indirect approach adopts GLMMs, which employs random effects, either shared or correlated, to accommodate associations between the mixed outcomes. This approach does not resort to factorization, and thus yields a symmetrical treatment of the outcomes. Using random effects, GLMMs incorporate subject-specific effects in the analysis, and embed a correlation structure between either clustered or longitudinal measurements for the same and/or different outcomes. Let $\mathbf{B}$ be the vector of random effects, either shared or correlated, and $f_{\mathbf{X},\mathbf{Y}|\mathbf{B}}(\mathbf{x},\mathbf{y}|\mathbf{b})$ the conditional joint density of $\mathbf{X}$ and $\mathbf{Y}$, given $\mathbf{B}$. Then the joint density of $\mathbf{X}$ and $\mathbf{Y}$ can be obtained as

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) \;=\; \int f_{\mathbf{X},\mathbf{Y}|\mathbf{B}}(\mathbf{x},\mathbf{y}|\mathbf{b}) f_{\mathbf{B}}(\mathbf{b}) d\mathbf{b}.$$

Typically, it is assumed that $f_{\mathbf{X},\mathbf{Y}|\mathbf{B}}(\mathbf{x},\mathbf{y}|\mathbf{b}) = f_{\mathbf{X}|\mathbf{B}}(\mathbf{x}|\mathbf{b}) f_{\mathbf{Y}|\mathbf{B}}(\mathbf{y}|\mathbf{b})$, i.e., $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent. In contrast, Gueorguieva and Agresti (2001) introduced a correlated probit model that accounts for conditional dependence between the discrete outcome $\mathbf{X}$ and the continuous outcome $\mathbf{Y}$, via the inclusion of correlated residual errors. However, the residual errors for their model are still assumed to be jointly normally distributed (Gueorguieva, 2013). Even though GLMMs overcome the drawbacks of the factorization models, they also have their shortcomings. For one, correlations may be restricted to lie

within artificially narrow ranges; for another, computational difficulties may become an issue in high-dimensional data settings.

Typically, normally distributed random effects and residual errors are assumed in GLMMs. One recent exception, where a non-normal distribution for the random effects is adopted, is studied by Lin *et al.* (2010). In their model, correlated random effects are incorporated to account for intra-cluster correlations between mixed bivariate outcomes X (binary) and Y (continuous), and the joint model for X and Y is specified by using Fitzmaurice and Laird's (1995) factorization approach, with a mixed logistic regression model for X. For interpretational ease, a bridge distribution is assumed for the random effect for X, so that the marginal regression model for X has a conveniently logistic form; a Gaussian copula (Song, 2007) is then used to construct its joint distribution with the corresponding normally distributed random effect for Y. However, Lin *et al.*'s (2010) approach still treats the two outcomes asymmetrically, and suffers from the same inadequacies of other factorization models.

A recent alternative strategy of indirectly specifying mixed-outcome joint models involves the use of copulas (Nelsen, 2006), as discussed by de Leon and Wu (2011), Dobra and Lenkoski (2011), Song *et al.* (2009), and Song (2007), to name a few. The approach embeds univariate marginal cumulative distribution functions (CDFs) $F_{Y_1}(\cdot), \cdots, F_{Y_P}(\cdot)$ into their corresponding P-dimensional CDF $F_{Y_1,\cdots,Y_p}(\cdot)$ via a copula $C(\cdot)$ as follows:

$$F_{Y_1,\cdots,Y_P}(y_1,\cdots,y_P) \;=\; C\left(F_{Y_1}(y_1),\cdots,F_{Y_P}(y_P)\right) \;=\; C\left(u_1,\cdots,u_P\right), \qquad (1.1)$$

where $u_1 = F_{Y_1}(y_1), \cdots, u_p = F_{Y_P}(y_P)$ are realizations of the probability integral transforms (PITs) $U_1 = F_{Y_1}(Y_1) \sim \mathit{uniform}[0,1], \cdots, U_P = F_{Y_P}(Y_P) \sim \mathit{uniform}[0,1]$, with $\mathit{uniform}[0,1]$ the uniform distribution over $[0,1]$. Sklar's Theorem (Nelsen, 2006) asserts that, assuming $Y_1, \cdots, Y_P$ are continuous random variables, the CDF $F_{Y_1,\cdots,Y_P}(\cdot)$ is uniquely specified via its margins and a copula that "glues" them together. In parametric contexts, the margins need not come from the same parametric family, allowing researchers great flexibility in modeling data with different types of outcomes. In statistical and practical respects, the

copula accounts for "dependence" between outcomes in a way that is separate from their marginal specifications. Copulas are particularly well-suited for constructing joint models for mixed outcomes $\mathbf{X}$ and $\mathbf{Y}$, where the relevant joint distribution is either not available or difficult to specify but marginal distributions for $\mathbf{X}$ and $\mathbf{Y}$ can be specified with confidence.

Adopting copula function in modeling mixed discrete and continuous outcomes is still a recent phenomenon. Trivedi and Zimmer (2007) provided a useful survey for researchers, including theoretical concepts and properties of copula models and generation of different copula families, as well as the attractive features of copula models in parametric contexts. Applications of copulas to discrete data are studied and discussed, for example, in Nikoloulopoulos and Karlis (2010,2009,2008) and Meester and MacKay (1994). As Genest and Nešlehovà (2007) showed, a number of complications arise from the direct application of copula models to discrete data (i.e., using discrete margins in copulas). One such complication concerns the failure of the copula to uniquely determine the distribution. Another more practical one involves the interpretability of the dependence parameters. Note that from a modeling perspective, the non-uniqueness of the copula is not an issue, as the parameters of the model are still identifiable and the copula still corresponds to a proper multivariate distribution. Common rank-based association measures like Kendall's tau and Spearman's rho, however, may now depend on the margins (see, e.g., Nešlehová, 2007; Mesfioui and Tajar, 2005), and the range of their possible values may be restricted — severely in some cases — rendering interpretations of such measures problematic. To get around these complications, de Leon and Wu (2011) recently developed a copula-based regression model for mixed bivariate discrete and continuous outcomes, where a t-latent variable is used to define the marginal distribution of the discrete outcome, and a normal distribution is assumed for the continuous outcome. A copula is employed to construct the joint distribution of the latent variable and the continuous outcome, from which the joint model for the mixed outcomes is indirectly specified. Because the latent variable is continuous, Sklar's Theorem still applies and the

resulting copula model is unique. To see how this is done, let a continuous latent vector $\mathbf{Y}^* = (Y_1^*, \cdots, Y_Q^*)^\top$ underly the vector $\mathbf{X} = (X_1, \cdots, X_Q)^\top$ of binary outcomes through the threshold model $X_j = \mathrm{I}\{Y_j^* > 0\}$, $j = 1, \cdots, Q$, where $\mathrm{I}\{\cdot\}$ is the indicator function. The joint CDF of $\mathbf{Y}^*$ and the vector $\mathbf{Y} = (Y_1, \cdots, Y_C)^\top$ of continuous outcomes is specified via a copula $C(\cdot)$ as

$$F_{\mathbf{Y}^*,\mathbf{Y}}(\mathbf{y}^*, \mathbf{y}) \;=\; C\left(F_{Y_1^*}(y_1^*), \cdots, F_{Y_Q^*}(y_Q^*), F_{Y_1}(y_1), \cdots, F_{Y_C}(y_C)\right), \tag{1.2}$$

where $F_{Y_1^*}(\cdot), \cdots, F_{Y_Q^*}(\cdot)$ are the marginal CDFs of $Y_1^*, \cdots, Y_Q^*$, respectively, and $F_{Y_1}(\cdot), \cdots, F_{Y_C}(\cdot)$ are the marginal CDFs of $Y_1, \cdots, Y_C$, respectively. The joint density of $\mathbf{X}$ and $\mathbf{Y}$ is then given by

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \;=\; \int_{A_1 \times \cdots \times A_Q} f_{\mathbf{Y}^*,\mathbf{Y}}(\mathbf{y}^*, \mathbf{y}) d\mathbf{y}^*, \tag{1.3}$$

where $f_{\mathbf{Y}^*,\mathbf{Y}}(\mathbf{y}^*, \mathbf{y}) = \partial^{Q+C} F_{\mathbf{Y}^*,\mathbf{Y}}(\mathbf{y}^*, \mathbf{y}) / \partial \mathbf{y}^* \partial \mathbf{y}$ is the joint density of $\mathbf{Y}^*$ and $\mathbf{Y}$, and the intervals A_j are either $(-\infty, 0]$ or $(0, +\infty)$ according as whether X_j is 0 or 1.

Another way to overcome the non-uniqueness problem when dealing with discrete outcomes is to use continuous extensions (or "continuous-ation") of discrete random variables. The method, introduced by Denuit and Lambert (2005), then forms the copula model for the "continued" discrete variables, thus ensuring the uniqueness of the copula-generated joint distribution. Madsen and Fang (2010) adopted the "continuous-ation" method, also known as "jittering", to jointly model correlated binary outcomes; see also the ensuing discussion in Sabo and Chaganty (2011) and Madsen and Fang (2011). Heinen and Rengifo (2007) and Shi and Valdez (2012) likewise employed jittering to jointly model multivariate count data; additional references include Heinen and Rengifo (2008) and Machado and Santos Silva (2005). Note that the jittering method can be used for any type of discrete variables (e.g., nominal or count data), which makes the method widely appealing in a variety of applications.

The choice of an appropriate copula with which to "couple" the margins depends on the suitability of the copula's dependence parameter for describing the data's dependence structure. Several copula families with different ways of modeling dependence are available in the

literature (Trivedi and Zimmer, 2007; Nelsen, 2006; Joe, 1997). For example, the elliptical family of copulas includes the Gaussian and Student's t-copulas while the Archimedean copula family comprises the so-called Clayton and Gumbel copulas, among others. In practice, the best fitting copula is the one that can best capture the data's dependence structure. Gaussian copulas are an important family which has been used in a variety of applications (e.g., Song, 2007) for several reasons. For one, the Gaussian copula is analytically tractable and closed under marginalization and conditionalization, the same way that the Gaussian (normal) distribution is. For another, it introduces a flexible model for the dependence since it allows both positive and negative associations for the outcomes. Specifically, the P-dimensional Gaussian copula is defined as

$$C(u_1, \cdots, u_P) \;=\; \Phi_P\left(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_P); \widetilde{\mathbf{R}}\right) \;=\; F_{Y_1,\cdots,Y_P}(y_1, \cdots, y_P), \quad (1.4)$$

where $u_1, \cdots, u_p$ and $U_1, \cdots, U_P$ are defined in (1.1), $\Phi^{-1}(\cdot)$ is the inverse of the standard normal CDF $\Phi(\cdot)$ (i.e., $\Phi^{-1}(\cdot)$ is the quantile function for the standard normal distribution), $\Phi_P(\cdot; \widetilde{\mathbf{R}})$ is the P-dimensional standard multivariate normal CDF (i.e., zero means and unit variances) with correlation matrix $\widetilde{\mathbf{R}}$. Note that $\widetilde{\mathbf{R}}$ is not the correlation matrix of the random variables $Y_1, \cdots, Y_P$; instead, $\widetilde{\mathbf{R}}$ is the correlation matrix of the so-called normal scores $\Phi^{-1}(U_1) = \Phi^{-1}(F_{Y_1}(Y_1)), \cdots, \Phi^{-1}(U_P) = \Phi^{-1}(F_{Y_P}(Y_P))$ and its elements $\widetilde{\rho}_{jj'} = corr\{\Phi^{-1}(F_{Y_j}(Y_j)), \Phi^{-1}(F_{Y_{j'}}(Y_{j'}))\}$ are called normal correlations. Note that $\widetilde{\mathbf{R}}$ is "margin-free", since $\widetilde{\rho}_{jj'}$ does not depend on the margins $F_{Y_j}(\cdot)$ and $F_{Y_{j'}}(\cdot)$. This follows from the fact that $\Phi^{-1}(F_{Y_j}(Y_j)) \sim normal(0, 1)$, for any continuous margin $F_{Y_j}(\cdot)$, for all $j = 1, \cdots, P$. To see this, we have

$$P(\Phi^{-1}(F_{Y_j}(Y_j)) \leq y) \;=\; P(F_{Y_j}(Y_j) \leq \Phi(y)) \;=\; \Phi(y),$$

for all real y, where the last equality followed from the fact that the PIT $F_{Y_j}(Y_j) \sim uniform[0, 1]$, for all $j = 1, \cdots, P$.

The flexibility and analytical tractability of Gaussian copulas make them a handy tool in many applications. In this thesis, we adopt the Gaussian copula for joint model specifications

in correlated mixed-outcome settings. Note however that the methodology we propose can be adapted to any copula family.

1.3 Overview of thesis

The objective of this thesis is the construction of joint models for correlated mixed discrete and continuous outcomes via copulas, in general, and Gaussian copulas, in particular. The resulting models can then be used for flexible joint analysis of mixed-outcome data. Methods for the ensuing estimation and inference for the copula models are also developed and their finite-sample performance is investigated empirically via simulations. Real-data applications in developmental toxicology, clinical trials, and applied econometrics are used all throughout the thesis to illustrate the methodology.

Chapter 2 considers the analysis of multiple correlated discrete and continuous outcomes that are observed on the same subjects over time in longitudinal settings. A Gaussian copula mixed model for the mixed outcomes that allows flexible, possibly non-Gaussian, distributions for the residual errors and random effects is constructed, for which the resulting marginal dependence structure between outcomes is obtained. Full likelihood estimation is discussed along with a variant based on pairwise likelihoods that provide necessary computational economies in the case of high-dimensional data. An application of the model to individual panel data on the wages, work hours, and union memberships of 545 men during the period 1980–1987 (Vella and Verbeek, 1998) is also included.

Chapter 3 concerns the analysis of clustered data from developmental toxicity studies with mixed responses, i.e., where each member of the cluster has binary and continuous outcomes. The same methodological approach outlined in Chapter 2 is adopted to construct a flexible joint model based on the Gaussian copula. The model is general enough to include earlier models developed by Gueorguieva and Agresti (2001), Faes *et al.* (2008), Faes (2013), and Masarotto and Varin (2012), among others. Estimation and inference based on the full

likelihood is discussed, and finite-sample properties of the resulting estimates are studied empirically via simulations. An application to quantitative risk assessment of ethylene glycol (EG) in mice is reported as well.

In Chapter 4, we adopt the "continuous-ation" approach in order to circumvent the complications, both theoretical and practical engendered by the direct application of copulas to discrete variables. The approach provides an alternative route to copula modeling of mixed outcomes without resorting to a latent variable formulation of discrete outcomes, as was used in Chapters 2 and 3. Likelihood estimation is implemented for the resulting joint model, and simulation results concerning the relative bias and efficiency of the resulting estimates are reported. The proposed methodology is illustrated by revisiting the burn injury data previously analyzed in de Leon and Wu (2011) and Song *et al.* (2009).

Chapter 5 introduces a methodology for calculating the sample size in clinical trials with multiple mixed binary and continuous co-primary endpoints modeled by the Gaussian copula. We adopt a latent variable description of the binary endpoints and makes use of tests on the latent means to test for difference in binary proportions. Empirical results on the sample size calculations are reported and compared with those recently obtained by Sozu *et al.* (2012). The PREMIER Study considered in Sozu *et al.* (2012) is used to illustrate the methodology.

Finally, Chapter 6 concludes the thesis with a brief summary and a discussion of promising areas for future research.

Chapter 2

Joint analysis of multiple mixed longitudinal outcomes via Gaussian copula mixed models

2.1 Introduction

In this chapter, we propose a copula-based approach for constructing joint models for mixed outcomes from longitudinal studies that account for associations between the outcomes (of the same or of different types) for the same subject at the same time point, and/or at different time points through the inclusion of subject-specific random effects. A latent-variable framework is adopted to avoid the direct use of discrete margins for copulas, thus sidestepping the complications discussed in Chapter 1. The models treat mixed outcomes symmetrically and do not resort to factorization and can be considered as a hybrid of GLMMs and the model introduced by de Leon and Wu (2011). The approach developed in this chapter can be used for studying the relationship between workers' union membership and their wages and work hours. The so-called union effect (Vella and Verbeek, 1998) has been shown to significantly impact wages of workers in unionized and non-unionized employment, with the former yielding a higher average wage than the latter. However, due to differences in union employment as well as other incremental individual effects, the total union effect has been shown to be highly variable across individuals. A GLMM, which incorporates individual-specific effects and accounts for individual-level heterogeneity, is a natural framework for analyzing such data. We used data on young male workers from the U. S. National Longitudinal Survey for the period 1980–1987 (Vella and Verbeek, 1998) to illustrate the application of our methodology, focusing our analysis on three mixed longitudinal outcomes, namely, hourly wage,

Wu *et al.* (2013), which appeared as a book chapter in the edited volume *Analysis of Mixed Data: Methods & Applications*, 2013, CRC/Chapman & Hall, is based on parts of this chapter.

annual worked hours, and union membership. Note that conventional GLMM methodology makes the assumption of conditional independence of the outcomes necessary in practice, in the absence of a flexible joint model for mixed variables. Our proposed approach relaxes this assumption and allows the construction of a joint model that can delineate the various associations in the data, including conditional dependence between outcomes, if needed.

Section 2.2 introduces a general approach based on copulas, in general, and the Gaussian copula, in particular, to constructing multivariate distributions for mixed discrete and continuous variables. We then adopt the approach to specify a joint model for longitudinal data comprising mixed discrete and continuous outcomes in Section 2.3. A brief discussion of how associations are incorporated in the model is presented in Section 2.4. Likelihood estimation is outlined in Section 2.5, along with a variant based on pairwise likelihood, a computationally more efficient alternative to full likelihood estimation. Section 2.6 is devoted to a case study involving data on the hourly wages, work hours, and union memberships of 545 male youths from the 1980–1987 National Longitudinal Survey (Vella and Verbeek, 1998). Finally, Section 2.7 concludes the chapter.

2.2 Gaussian copula joint models

Suppose we have discrete outcomes $X_1, \cdots, X_Q$, and continuous outcomes $Y_1, \cdots, Y_C$, where we suppressed the subject index i for convenience. Let X_j have $L_j + 1$ discrete values $x_j^{(0)} < \cdots < x_j^{(L_j)}$, $j = 1, \cdots, Q$. Underlying X_j is Y_j^*, a continuous latent variable whose relationship with X_j is defined by the following threshold model:

$$X_j = \begin{cases} x_j^{(L_j)} & \text{, if and only if } Y_j^* > \gamma_j^{(L_j)} \\ x_j^{(\ell_j)} & \text{, if and only if } \gamma_j^{(\ell_j)} < Y_j^* \leq \gamma_j^{(\ell_j+1)}, \quad \ell_j = 1, \cdots, L_j - 1, \\ x_j^{(0)} & \text{, if and only if } Y_j^* \leq \gamma_j^{(1)} \end{cases} \tag{2.1}$$

where $\gamma_j^{(1)}, \cdots, \gamma_j^{(L_j)}$, are unknown cutpoints or thresholds. Without loss of generality, we assume that $x_j^{(\ell_j)} = \ell_j$, $\ell_j = 0, \cdots, L_j$.

Given marginal CDFs $F_{Y_1^*}(\cdot), \cdots, F_{Y_Q^*}(\cdot)$, and $F_{Y_1}(\cdot), \cdots, F_{Y_C}(\cdot)$, for latent variables and continuous outcomes, respectively, we assume that the joint CDF of $\mathbf{Y}^* = (Y_1^*, \cdots, Y_Q^*)^\top$, and $\mathbf{Y} = (Y_1, \cdots, Y_C)^\top$, is determined by a Gaussian copula with correlation matrix $\widetilde{\mathbf{R}}$. With $P = C + Q$, the joint distribution of $\mathbf{X} = (X_1, \cdots, X_Q)^\top$ and $\mathbf{Y}$ is then

$$\begin{aligned} P\begin{pmatrix} X_1 = \ell_1, \cdots, X_Q = \ell_Q, \\ Y_1 \leq y_1, \cdots, Y_C \leq y_C \end{pmatrix} &= \sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} (-1)^{Q+\sum_{j=1}^{Q}\epsilon_j} \\ &\quad \times \Phi_P \begin{pmatrix} \Phi^{-1}(u_1^{(\ell_1+\epsilon_1)}), \cdots, \Phi^{-1}(u_Q^{(\ell_Q+\epsilon_Q)}), \\ \Phi^{-1}(v_1), \cdots, \Phi^{-1}(v_C) \end{pmatrix}; \widetilde{\mathbf{R}} \Bigg), \quad (2.2) \end{aligned}$$

where $u_j^{(\ell_j+\epsilon_j)} = F_{Y_j^*}(\gamma_j^{(\ell_j+\epsilon_j)})$ and $v_k = F_{Y_k}(y_k)$, for $j = 1, \cdots, Q$, and $k = 1, \cdots, C$, with

$$E(Y_j^*) = \mu_j^*(\mathbf{z}_{1j}, \boldsymbol{\alpha}_j) \quad \text{and} \quad E(Y_k) = \mu_k(\mathbf{z}_{2k}, \boldsymbol{\beta}_k), \quad (2.3)$$

for vectors $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_k$ of regression coefficients, outcome-specific covariate vectors $\mathbf{z}_{1j}$ and $\mathbf{z}_{2k}$, and link functions $\mu_j^*(\cdot)$ and $\mu_k(\cdot)$ specifying how the covariates are incorporated in the marginal means. For example, if $\mu_j^* = \mathbf{z}_{1j}^\top \boldsymbol{\alpha}_j$ and $\mu_k = \mathbf{z}_{2k}^\top \boldsymbol{\beta}_k$, then we have linear models for the means. For identifiability reasons, we assume that Y_j^* has unit scale (or variance parameter, if it is the scale parameter) and $\boldsymbol{\alpha}_j$ has no intercept term; if X_j is binary (i.e., $L_j = 1$), we may arbitrarily assume the single cutpoint to be zero, and $\boldsymbol{\alpha}_j$ will need to include an intercept term. Note that the margins $F_{Y_1^*}(\cdot), \cdots, F_{Y_Q^*}(\cdot)$, and $F_{Y_1}(\cdot), \cdots, F_{Y_C}(\cdot)$, can be any continuous CDFs, thus allowing researchers great flexibility in joint modeling of mixed outcomes. For example, a normal latent distribution for Y_j^* results in a marginal probit model for binary X_j.

The corresponding density of $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$\begin{aligned} f_{\mathbf{X},\mathbf{Y}}(\boldsymbol{\ell}, \mathbf{y}) &= \frac{\phi(z_{C-1|C})\phi(z_{C-2|C:C-1}) \cdots \phi(z_{1|C:2})}{\prod_{k=1}^{C-1}\sqrt{1-\widetilde{\rho}^2_{W_k W_C}} \prod_{k=1}^{C-2}\sqrt{1-\widetilde{\rho}^2_{W_k W_{C-1}|W_C}} \cdots \sqrt{1-\widetilde{\rho}^2_{W_1 W_2|W_C:W_3}}} \\ &\quad \times \sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} (-1)^{Q+\sum_{j=1}^{Q}\epsilon_j} \Phi_Q\left(s_{1|C:1}^{(\ell_1+\epsilon_1)}, \cdots, s_{Q|C:1}^{(\ell_Q+\epsilon_Q)}; \widetilde{\mathbf{R}}_{|W_C:W_1}\right) \end{aligned}$$

$$\times f_{Y_C}(y_C)\prod_{k=1}^{C-1}\frac{f_{Y_k}(y_k)}{\phi(z_k)}, \tag{2.4}$$

where $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_Q)^\top$ and $\mathbf{y} = (y_1, \cdots, y_C)^\top$, with $f_{Y_k}(\cdot)$ the marginal density of Y_k, $\phi(\cdot)$ the standard normal density, and

$$\begin{aligned} z_{C-k|C:C-k+1} &= \frac{z_{C-k|C:C-k+2} - \widetilde{\rho}_{W_{C-k}W_{C-k+1}|W_C:W_{C-k+2}} z_{C-k+1|C:C-k+2}}{\sqrt{1-\widetilde{\rho}^2_{W_{C-k}W_{C-k+1}|W_C:W_{C-k+2}}}}, \\ s^{(\ell_j+\epsilon_j)}_{j|C:k} &= \frac{s^{(\ell_j+\epsilon_j)}_{j|C:k+1} - \widetilde{\rho}_{W_j^*W_k|W_C:W_{k+1}} z_{k|C:k+1}}{\sqrt{1-\widetilde{\rho}^2_{W_j^*W_k|W_C:W_{k+1}}}}, \end{aligned}$$

where $W_j^* = \Phi^{-1}\{F_{Y_j^*}(Y_j^*)\}$ and $W_k = \Phi^{-1}\{F_{Y_k}(Y_k)\}$ are the normal scores (latent in the case of W_j^*), $s_j^{(\ell_j+\epsilon_j)} = \Phi^{-1}(u_j^{(\ell_j+\epsilon_j)})$, $z_k = \Phi^{-1}(v_k)$, $\widetilde{\rho}_{W_{C-k}W_{C-k+1}|W_C:W_{C-k+2}}$ is the partial correlation between W_{C-k} and W_{C-k+1}, after eliminating $W_{C-k+2}, \cdots, W_C$, $\widetilde{\rho}_{W_j^*W_{C-k}|W_C:W_{C-k+1}}$ is the partial correlation between W_j^* and W_{C-k}, after eliminating $W_{C-k+1}, \cdots, W_C$, and $\widetilde{\mathbf{R}}_{|W_C:W_{C-k+1}} = \widetilde{\mathbf{R}}_{W_1^*:W_{C-k}|W_C:W_{C-k+1}}$ is the partial correlation matrix for $W_1^*, \cdots, W_Q^*, W_1, \cdots, W_{C-k}$, after eliminating $W_{C-k+1}, \cdots, W_C$, for $k = 1, \cdots, C-1$. See de Leon *et al.* (2012) for more details; see also de Leon and Wu (2011) for the special case $Q = C = 1$.

The approach outlined above contrasts with Song *et al.*'s (2009) and Song's (2007) direct application of copulas to model discrete outcomes. Our approach adopts a latent variable description of discrete outcomes, and the resulting joint model arises from the joint distribution, constructed via a copula, of continuous outcomes and the latent variables. The joint model is thus indirectly specified through the copula-based joint distribution of continuous outcomes and the latent variables. In addition to its interpretational appeal in medical and health studies, using latent variables to describe discrete outcomes also makes statistical sense, since common regression models for discrete outcomes (e.g., logistic, probit models) have parallel formulations in terms of latent variables (see Teixeira-Pinto and Normand, 2009). In addition, it also leads to familiar association measures for capturing dependence between two discrete outcomes (i.e., polychoric correlation) or between mixed discrete and continuous outcomes (i.e., polyserial correlation). In contrast to Pearson's correlations as

measures of association between two discrete outcomes and between mixed discrete and continuous outcomes, polychoric and polyserial correlations are not constrained by marginal probabilities of the discrete outcomes, as they are just the usual pairwise correlations between continuous variables (latent or otherwise). Moreover, the number of polychoric and polyserial correlations remains the same for polychotomous discrete data, certainly not the case for odds ratios. Finally, note that our approach does not resort to factorization, resulting in a symmetrical treatment of outcomes while preserving the margins, an attractive feature in practice.

2.3 Case of longitudinal data

Similar to the previous section, let X_{ijt} and Y_{ikt} be the respective discrete and continuous outcomes for subject $i = 1, \cdots, N$, at time $t = 1, \cdots, T$, where we assume the same threshold model (2.1) for X_{ijt} and its corresponding latent variable Y^*_{ijt}, with the cutpoints not depending on time t for each $j = 1, \cdots, Q$. In what follows, we develop Gaussian copula mixed models for longitudinal mixed outcomes which allow for flexible non-normal residual errors and random effects, extending and generalizing previous models studied by de Leon and Wu (2011), Lin *et al.* (2010), Najita *et al.* (2009), and Gueorguieva and Agresti (2001), among others. For simplicity, we assume a shared subject-specific random effect $\mathbf{B}_i = (B_{i1}, B_{i2})^\top$ and construct a joint density for $\mathbf{X}_{it} = (X_{i1t}, \cdots, X_{iQt})^\top$, and $\mathbf{Y}_{it} = (Y_{i1t}, \cdots, Y_{iCt})^\top$ as

$$f_{\mathbf{X}_{it},\mathbf{Y}_{it}}(\boldsymbol{\ell}_{it}, \mathbf{y}_{it}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{\mathbf{X}_{it},\mathbf{Y}_{it}|\mathbf{B}_i}(\boldsymbol{\ell}_{it}, \mathbf{y}_{it}|\mathbf{b}_i) f_{\mathbf{B}_i}(\mathbf{b}_i) d\mathbf{b}_i, \tag{2.5}$$

for $\boldsymbol{\ell}_{it} = (\ell_{i1t}, \cdots, \ell_{iQt})^\top$, where

$$f_{\mathbf{X}_{it},\mathbf{Y}_{it}|\mathbf{B}_i}(\boldsymbol{\ell}_{it}, \mathbf{y}_{it}|\mathbf{b}_i) = \int_{\gamma_1^{\ell_{i1}}}^{\gamma_1^{\ell_{i1}+1}} \cdots \int_{\gamma_Q^{\ell_{iQ}}}^{\gamma_Q^{\ell_{iQ}+1}} f_{\mathbf{Y}^*_{it},\mathbf{Y}_{it}|\mathbf{B}_i}(\mathbf{y}^*_{it}, \mathbf{y}_{it}|\mathbf{b}_i) dy^*_{iQt} \cdots dy^*_{i1t}, \tag{2.6}$$

where $\mathbf{Y}^*_{it} = (Y^*_{i1}, \cdots, Y^*_{iQ})^\top$ and $\mathbf{y}^*_{it} = (y^*_{i1t}, \cdots, y^*_{iQt})^\top$. In (2.5), $f_{\mathbf{B}_i}(\cdot)$ is the density of $\mathbf{B}_i$, and $f_{\mathbf{Y}^*_{it},\mathbf{Y}_{it}|\mathbf{B}_i}(\cdot|\cdot)$ is the conditional density of $\mathbf{Y}^*_{it}$ and $\mathbf{Y}_{it}$, given $\mathbf{B}_i$. Note that (2.6) is of

the form (2.4), assuming a Gaussian copula is adopted to construct the conditional CDF of $\mathbf{Y}_{it}^*$ and $\mathbf{Y}_{it}$, as in

$$F_{\mathbf{Y}_{it}^*,\mathbf{Y}_{it}|\mathbf{B}_i}(\mathbf{y}_{it}^*,\mathbf{y}_{it}|\mathbf{b}_i) \;=\; \Phi_P\left(\begin{array}{c}\Phi^{-1}\{u_{i1t}(\mathbf{b}_i)\},\cdots,\Phi^{-1}\{u_{iQt}(\mathbf{b}_i)\},\\ \Phi^{-1}\{v_{i1t}(\mathbf{b}_i)\},\cdots,\Phi^{-1}\{v_{iCt}(\mathbf{b}_i)\}\end{array};\widetilde{\mathbf{R}}\right), \tag{2.7}$$

where $u_{ijt}(\mathbf{b}_i) = F_{Y_{ijt}^*|\mathbf{B}_i}(y_{ijt}^*|\mathbf{b}_i)$ and $v_{ikt}(\mathbf{b}_i) = F_{Y_{ikt}|\mathbf{B}_i}(y_{ikt}|\mathbf{b}_i)$, with $F_{Y_{ijt}^*|\mathbf{B}_i}(\cdot|\cdot)$ and $F_{Y_{ikt}|\mathbf{B}_i}(\cdot|\cdot)$ the respective (conditional) marginal CDFs of Y_{ijt}^* and Y_{ikt}, for $j = 1,\cdots,Q$, $k = 1,\cdots,C$, and $t = 1,\cdots,T$. Note that the margins $F_{Y_{ijt}^*|\mathbf{B}_i}(\cdot|\cdot)$ and $F_{Y_{ikt}|\mathbf{B}_i}(\cdot|\cdot)$ need not come from the same parametric family, allowing researchers great flexibility in modeling disparate outcomes. The (conditional) joint density $f_{\mathbf{X}_{i1},\mathbf{Y}_{i1},\cdots,\mathbf{X}_{iT},\mathbf{Y}_{iT}|\mathbf{B}_i}(\cdot|\cdot)$ of $\mathbf{X}_{i1},\mathbf{Y}_{i1},\cdots,\mathbf{X}_{iT},\mathbf{Y}_{iT}$, given $\mathbf{B}_i$, is constructed by assuming that $\{\mathbf{X}_{it},\mathbf{Y}_{it}\}$ and $\{\mathbf{X}_{it'},\mathbf{Y}_{it'}\}$ are conditionally independent, given $\mathbf{B}_i$, for all $t \neq t'$. We thus have

$$f_{\mathbf{X}_{i1},\mathbf{Y}_{i1},\cdots,\mathbf{X}_{iT},\mathbf{Y}_{iT}|\mathbf{B}_i}(\boldsymbol{\ell}_{i1},\mathbf{y}_{i1},\cdots,\boldsymbol{\ell}_{iT},\mathbf{y}_{iT}|\mathbf{b}_i) \;=\; \prod_{t=1}^{T} f_{\mathbf{X}_{it},\mathbf{Y}_{it}|\mathbf{B}_i}(\boldsymbol{\ell}_{it},\mathbf{y}_{it}|\mathbf{b}_i). \tag{2.8}$$

Suppose

$$Y_{ijt}^*|\mathbf{B}_i \sim \text{independent } F_{Y_{ijt}^*|\mathbf{B}_i}(\cdot|\cdot), \quad Y_{ikt}|\mathbf{B}_i \sim \text{independent } F_{Y_{ikt}|\mathbf{B}_i}(\cdot|\cdot),$$

for $i = 1,\cdots,N$, $t = 1,\cdots,T$. Assuming $E(B_{i1}) = E(B_{i2}) = 0$ and $var(B_{i1}) = var(B_{i2}) = 1$, with $var(Y_{ijt}^*|\mathbf{B}_i) = 1$ (assuming the scale parameter is the variance), for identifiability reasons, the conditional mean models, given the random effect $\mathbf{B}_i$, are defined by

$$\mu_{ijt}^*(\mathbf{B}_i) \;=\; E(Y_{ijt}^*|\mathbf{B}_i) \;=\; \mu_{ijt}^*(\mathbf{z}_{1ijt},\boldsymbol{\alpha}_j) + \lambda_{j1}^* B_{i1} + \lambda_{j2}^* B_{i2} t, \tag{2.9}$$

$$\mu_{ikt}(\mathbf{B}_i) \;=\; E(Y_{ikt}|\mathbf{B}_i) \;=\; \mu_{ikt}(\mathbf{z}_{2ikt},\boldsymbol{\beta}_k) + \lambda_{k1} B_{i1} + \lambda_{k2} B_{i2} t, \tag{2.10}$$

where $\mathbf{z}_{1ijt}$ and $\mathbf{z}_{2ikt}$ are known covariate vectors (assumed time-dependent), and $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_k$ are the corresponding unknown regression coefficients, with λ_{jh}^* and λ_{kh}, $h = 1, 2$, accounting for the difference in scales of Y_{ijt}^* and Y_{ikt}. The common marginal means (i.e., time-dependent) then follow as $E(Y_{ijt}^*) = \mu_{ijt}^*(\mathbf{z}_{1ijt},\boldsymbol{\alpha}_j) = \mathbf{z}_{1ijt}^\top\boldsymbol{\alpha}_j$, say, and $E(Y_{ikt}) = \mu_{ikt}(\mathbf{z}_{2ikt},\boldsymbol{\beta}_k) = \mathbf{z}_{2ikt}^\top\boldsymbol{\beta}_k$, say, for all $t = 1,\cdots,T$. Observe that B_{i1} corresponds to the random intercept while

B_{i2} to the random slope. As shown in Section 2.4, the inclusion of the random slope term is necessary to incorporate temporal effects in the marginal correlations between outcomes at fixed time points, since we assumed the correlation matrix $\widetilde{\mathbf{R}}$ to be time-independent.

Note that in the case of binary X_{ijt}, the single cutpoint γ_j can be assumed to be 0, provided an intercept term is included in $\boldsymbol{\alpha}_j$. In addition, while the model (and inference) for X_{ijt} is done at the latent level, note that this corresponds to a GLMM for X_{ijt} with a specific link function that depends on the choice of latent distribution. For example, a logistic latent distribution for Y^*_{ijt} results in a logistic mixed model for X_{ijt}, while a normal latent distribution leads to a probit mixed model. With normal margins for Y^*_{ijt} and Y_{ikt}, our model specializes to a multivariate version of Najita *et al.*'s (2010) correlated bivariate probit model, albeit adapted to a longitudinal context.

The above approach deviates from the conventional one adopted for specifying GLMMs, where the residual error distribution drives that of the outcomes, as in Gueorguieva and Agresti's (2001) correlated probit model. Instead of specifying a Gaussian residual error distribution to construct the conditional joint density of the outcomes (given random effects), as is done in conventional GLMMs, our approach specifies the latter directly by specifying the (conditional) marginal response distributions and coupling them together using the Gaussian copula. The corresponding residual error distributions can then be obtained from the response distribution (given random effects) by transformation methods. The resulting GLMM is thus more flexible and more general than conventional GLMMs. Note as well that without the assumption of Gaussian errors for the model, assuming conditional independence of the outcomes becomes necessary.

Finally, we note that it is also possible to incorporate correlated outcome-specific random effects in our model. In addition, it is also possible to similarly use the Gaussian copula to build the density $f_{\mathbf{B}_i}(\cdot)$ of the random effect $\mathbf{B}_i$, as in Lin *et al.* (2010), instead of the usual normality assumption for $\mathbf{B}_i$, as in Gueorguieva and Agresti (2001). This affords flexibility

in specifying the marginal distributions of B_{i1} and B_{i2}; for example, for clustered data where correlated random intercepts are included in the model, a bridge distributed random intercept for the binary outcome along with a logistic distribution for the corresponding latent variable (i.e., a logistic mixed model for the binary outcome) yield a marginal logistic regression model for the binary outcome, which facilitates interpretation of results. This is discussed in detail in Chapter 3.

2.4 Associations

Copula models usually rely on rank-based association measures, such as Kendall's tau or Spearman's rho (Balakrishnan and Lai, 2009) to evaluate the strength of dependence between variables, as they are invariant to monotonic transformations (e.g., normal scores). However, unlike with continuous variables for which they provide margin-free measures of the level of dependence, this no longer holds in the discrete case (e.g., Genest and Nešlehová, 2007). Denuit and Lambert (2005) and Mesfioui and Tajar (2005) adopt a "continuousation" approach as a possible remedy; we adopt this approach in Chapter 4 to model mixed outcomes. Interpretability of the association measures is another issue since their range varies, and there is thus a need to re-scale them.

Directly applying copulas to model multivariate discrete data, as in Song *et al.* (2009), suffers from the same problem many multivariate discrete distributions do: correlations and dependence parameters are unnaturally constrained to ensure the propriety of the joint probabilities. However, merely using latent variables for the discrete variables and indirectly applying copula at the latent level do not always work. This is seen in the recent work of Li and Wong (2011), who adopt the multivariate Gumbel copula to specify a joint model for correlated binary data, via a threshold model as in (2.1); however, as shown by Nikoloulopoulos (2012), the dependence parameters of the Gumbel copula are constrained to ensure the positivity of the copula density. In contrast, our use of the Gaussian copula is advantageous

in that the matrix of normal correlations is not constrained; the only requirement is that the matrix be positive definite. Thus, using latent variable-based marginal models together with the Gaussian copula results in margin-free and unconstrained dependence measures (albeit measured at the latent level).

2.4.1 Conditional dependence

Consider first the conditional correlations, given the random effect $\mathbf{B}_i$, between outcomes for fixed time point t for the model in Section 2.3. Since the conditional correlations do not depend on the subject and time point, we suppress the indices i and t in what follows. The (conditional) correlation matrix $\widetilde{\mathbf{R}}$ contains three types of (conditional) correlations: correlation $\widetilde{\rho}_{W_j^* W_{j'}^*} = corr(W_j^*, W_{j'}^* | \mathbf{B}_i)$ between latent normal scores W_j^* and $W_{j'}^*$ based on latent variables Y_j^* and $Y_{j'}^*$, correlation $\widetilde{\rho}_{W_j^* W_k} = corr(W_j^*, W_k | \mathbf{B}_i)$ between a latent normal score W_j^* and normal score W_k based on continuous outcome Y_k, and correlation $\widetilde{\rho}_{W_k W_{k'}} = corr(W_k, W_{k'} | \mathbf{B}_i)$ between normal scores based on Y_k and $Y_{k'}$. Bodnar *et al.* (2010) show that any pair of variables (latent and otherwise) from $\mathbf{Y}^*$ and $\mathbf{Y}$, hence any pair from $\mathbf{X}$ and $\mathbf{Y}$, are (conditionally) independent if and only if the corresponding (conditional) correlation between their normal scores (latent and otherwise) is 0. The (conditional) correlations in $\widetilde{\mathbf{R}}$ are called (conditional) normal or dependence correlation coefficients (Bodnar *et al.*, 2010; Klaassen and Wellner, 1997), when variables are observable, as is the case with $\widetilde{\rho}_{W_k W_{k'}}$. Since $\mathbf{Y}^*$ is latent and unobservable and because $\widetilde{\rho}_{W_j^* W_{j'}^*}$ and $\widetilde{\rho}_{W_j^* W_k}$ are analogous to (conditional) polychoric and polyserial correlations $\rho_{Y_j^* Y_{j'}^*} = corr(Y_j^*, Y_{j'}^* | \mathbf{B}_i)$ and $\rho_{Y_j^* Y_k} = corr(Y_j^*, Y_k | \mathbf{B}_i)$, respectively, we refer to $\widetilde{\rho}_{W_j^* W_{j'}^*}$ as a (conditional) polychoric normal correlation coefficient, and to $\widetilde{\rho}_{W_j^* W_k}$ as a (conditional) polyserial normal correlation coefficient. From the nonlinearity of normal quantile transforms, it follows from Theorem 6.1 of Klaassen and Wellner (1997) that

$$\rho_{Y_j^* Y_{j'}^*} < |\widetilde{\rho}_{W_j^* W_{j'}^*}|, \quad \rho_{Y_j^* Y_k} < |\widetilde{\rho}_{W_j^* W_k}|, \quad \rho_{Y_k Y_{k'}} < |\widetilde{\rho}_{W_k W_{k'}}|. \tag{2.11}$$

Given $\widetilde{\rho}_{W_j^* W_{j'}^*}$, $\widetilde{\rho}_{W_j^* W_k}$, and $\widetilde{\rho}_{W_k W_{k'}}$, it is possible to obtain $\rho_{Y_j^* Y_{j'}^*}$, $\rho_{Y_j^* Y_k}$, and $\rho_{Y_k Y_{k'}}$, as

$$\rho_{Y_j^* Y_{j'}^*} = \psi_{jj'}^{**}(\widetilde{\rho}_{W_j^* W_{j'}^*}), \quad \rho_{Y_j^* Y_k} = \psi_{jk}^{*}(\widetilde{\rho}_{W_j^* W_k}), \quad \rho_{Y_k Y_{k'}} = \psi_{kk'}(\widetilde{\rho}_{W_k W_{k'}}), \tag{2.12}$$

for some functions $\psi_{jj'}^{**}(\cdot)$, $\psi_{jk}^{*}(\cdot)$, and $\psi_{kk'}(\cdot)$. Kugiumtzis and Bora-Senta (2010) use piecewise linear approximations based on truncated standard bivariate normal variables to obtain $\psi_{jj'}^{**}(\cdot)$, $\psi_{jk}^{*}(\cdot)$, and $\psi_{kk'}(\cdot)$. While easy to implement, it may occasionally yield a non-positive definite correlation matrix of $\rho_{Y_j^* Y_{j'}^*}$, $\rho_{Y_j^* Y_k}$, and $\rho_{Y_k Y_{k'}}$; this arises mainly due to such correlations needing to satisfy certain admissible ranges.

Because W_j^* and W_k are monotonic transformations of Y_j^* and Y_k, respectively, it follows that the (conditional) Kendall's tau measures are such that $\widetilde{\tau}_{W_j^* W_{j'}^*} = \tau(W_j^*, W_{j'}^* | \mathbf{B}_i) = \tau_{Y_j^* Y_{j'}^*} = \tau(Y_j^*, Y_{j'}^* | \mathbf{B}_i)$, $\widetilde{\tau}_{W_j^* W_k} = \tau(W_j^*, W_k | \mathbf{B}_i) = \tau_{Y_j^* Y_k} = \tau(Y_j^*, Y_k | \mathbf{B}_i)$, and $\widetilde{\tau}_{W_k W_{k'}} = \tau(W_k, W_{k'} | \mathbf{B}_i) = \tau_{Y_k Y_{k'}} = \tau(Y_k, Y_{k'} | \mathbf{B}_i)$. Assuming a Gaussian copula joint model as in (2.2) and (2.4) for $\mathbf{Y}^*$ and $\mathbf{Y}$, these measures are easy to calculate given $\widetilde{\mathbf{R}}$:

$$\widetilde{\tau}_{W_j^* W_{j'}^*} = \frac{2}{\pi} \sin^{-1}(\widetilde{\rho}_{W_j^* W_{j'}^*}), \;\; \widetilde{\tau}_{W_j^* W_k} = \frac{2}{\pi} \sin^{-1}(\widetilde{\rho}_{W_j^* W_k}), \;\; \widetilde{\tau}_{W_k W_{k'}} = \frac{2}{\pi} \sin^{-1}(\widetilde{\rho}_{W_k W_{k'}}); \tag{2.13}$$

they can also capture the full range of possible associations in $\mathbf{Y}^*$ and $\mathbf{Y}$, making them quite attractive in practice. A similar approach may be adapted to Spearman's correlation rho $\rho_S = 6 \sin^{-1}(\sin(\pi\tau/2)/2)/\pi$.

2.4.2 Marginal dependence

For the marginal correlations, we assume that the conditional correlations $\rho_{Y_j^* Y_{j'}^*}$ (i.e., the conditional polychoric correlation between X_{ijt} and $X_{ij't}$) and $\rho_{Y_j^* Y_k}$ (i.e., the conditional polyserial correlation between X_{ijt} and Y_{ikt}) have been obtained, say, via Kugiumtzis and Bora-Senta's (2010) piecewise-linear approximation, from $\widetilde{\rho}_{W_j^* W_{j'}^*}$ and $\widetilde{\rho}_{W_j^* W_k}$, the conditional polychoric and polyserial normal correlation coefficients, respectively. To simplify notations, we let

$$\mu_{ijt}^*(\mathbf{B}_i) \;=\; \mathbf{z}_{1ijt}^\top \boldsymbol{\alpha}_j + \lambda_{j1}^* B_{i1} + \lambda_{j2}^* B_{i2} t, \quad \mu_{ikt}(\mathbf{B}_i) \;=\; \mathbf{z}_{2ikt}^\top \boldsymbol{\beta}_k + \lambda_{k1} B_{i1} + \lambda_{k2} B_{i2} t.$$

Then the marginal means and variances are given by

$$E(Y^*_{ijt}) = \mathbf{z}^\top_{1ijt}\boldsymbol{\alpha}_j, \quad var(Y^*_{ijt}) = 1 + (\lambda^*_{j1})^2 + (\lambda^*_{j2}t)^2 + 2\lambda^*_{j1}\lambda^*_{j2}t\rho_{B_1B_2}, \tag{2.14}$$

$$E(Y_{ikt}) = \mathbf{z}^\top_{2ikt}\boldsymbol{\beta}_k, \quad var(Y_{ikt}) = \sigma^2_k + \lambda^2_{k1} + (\lambda_{k2}t)^2 + 2\lambda_{k1}\lambda_{k2}t\rho_{B_1B_2}, \tag{2.15}$$

where $\rho_{B_1B_2} = corr(B_{i1}, B_{i2})$ and $\sigma^2_k = var(Y_{ikt}|\mathbf{B}_i)$. If the Gaussian copula is used to construct the density $f_{\mathbf{B}_i}(\cdot)$, then $\rho_{B_1B_2}$ can be obtained by piecewise-linear approximation (Kugiumtzis and Bora-Senta, 2010) from the corresponding normal correlation coefficient $\widetilde{\rho}_{B_1B_2} = corr\{\Phi^{-1}(F_{B_{i1}}(B_{i1})), \Phi^{-1}(F_{B_{i2}}(B_{i2}))\}$. The marginal within-subject covariances at the same time point t or at the different time points t and t', with $t > t'$, are as follows:

$$\begin{aligned}
cov(Y^*_{ijt}, Y^*_{ijt'}) &= (\lambda^*_{j1})^2 + \lambda^*_{j1}\lambda^*_{j2}(2t - \triangle t)\rho_{B_1B_2} + (\lambda^*_{j2})^2 t(t - \triangle t), \\
cov(Y^*_{ijt}, Y^*_{ij't}) &= \rho_{Y^*_jY^*_{j'}} + \lambda^*_{j1}\lambda^*_{j'1} + (\lambda^*_{j1}\lambda^*_{j'2} + \lambda^*_{j2}\lambda^*_{j'1})t\rho_{B_1B_2} + \lambda^*_{j2}\lambda^*_{j'2}t^2, \\
cov(Y_{ikt}, Y_{ikt'}) &= \lambda^2_{k1} + \lambda_{k1}\lambda_{k2}(2t - \triangle t)\rho_{B_1B_2} + \lambda^2_{k2}t(t - \triangle t), \\
cov(Y_{ikt}, Y_{ik't}) &= \rho_{Y_kY_{k'}}\sigma_k\sigma_{k'} + \lambda_{k1}\lambda_{k'1} + (\lambda_{k1}\lambda_{k'2} + \lambda_{k2}\lambda_{k'1})t\rho_{B_1B_2} + \lambda_{k2}\lambda_{k'2}t^2, \\
cov(Y^*_{ijt}, Y^*_{ij't'}) &= \lambda^*_{j1}\lambda^*_{j'1} + \{\lambda^*_{j1}\lambda^*_{j'2}(t - \triangle t) + \lambda^*_{j'1}\lambda^*_{j2}t\}\rho_{B_1B_2} + \lambda^*_{j2}\lambda^*_{j'2}t(t - \triangle t), \\
cov(Y_{ikt}, Y_{ik't'}) &= \lambda_{k1}\lambda_{k'1} + \{\lambda_{k1}\lambda_{k'2}(t - \triangle t) + \lambda_{k'1}\lambda_{k2}t\}\rho_{B_1B_2} + \lambda_{k2}\lambda_{k'2}t(t - \triangle t), \\
cov(Y^*_{ijt}, Y_{ikt}) &= \rho_{Y^*_jY_k}\sigma_k + \lambda^*_{j1}\lambda_{k1} + (\lambda^*_{j1}\lambda_{k2} + \lambda_{k1}\lambda^*_{j2})t\rho_{B_1B_2} + \lambda^*_{j2}\lambda_{k2}t^2, \\
cov(Y^*_{ijt}, Y_{ikt'}) &= \lambda_{j1}\lambda_{k1} + \{\lambda^*_{j1}\lambda_{k2}(t - \triangle t) + \lambda^*_{j2}\lambda_{k1}t\}\rho_{B_1B_2} + \lambda^*_{j2}\lambda_{k2}t(t - \triangle t),
\end{aligned}$$

where $0 < \triangle t = t - t' < t$. Note that by including random slopes for time, we are able to account for temporal effects in the marginal correlations.

2.5 Likelihood estimation

Assume the Gaussian copula mixed model for longitudinal mixed outcomes in Section 2.3. Let $\boldsymbol{\Theta}$ be the vector containing all the parameters, including regression coefficients, correlation parameters, and scale parameters for the continuous margins, in the conditional joint density (2.6).

2.5.1 Full likelihood estimation

The conditional log-likelihood contribution of subject i at time t is given by

$$\begin{aligned}\ell_{it}(\boldsymbol{\Theta}|\mathbf{B}_i) &= \sum_{\boldsymbol{\ell}_{it}} \mathrm{I}\{\mathbf{x}_{it}=\boldsymbol{\ell}_{it}\}\log\left\{\sum_{\epsilon_1=0}^{1}\cdots\sum_{\epsilon_Q=0}^{1}(-1)^{Q+\sum_{j=1}^{Q}\epsilon_j}\Phi_Q\left(s_{1t|C:1}^{(\ell_{1t}+\epsilon_1)},\cdots,s_{Qt|C:1}^{(\ell_{Qt}+\epsilon_Q)};\widetilde{\mathbf{R}}_{|W_C:W_1}\right)\right\}\\ &+\sum_i\left(\sum_{k=1}^{C}\log\ f_{Y_{ikt}}(y_{ikt})+\sum_{k=1}^{C-1}\log\left\{\frac{\phi(z_{i,C-k,t|C:C-k+1})}{\phi(z_{ikt})}\right\}\right)-\frac{N}{2}\sum_{k=1}^{C-1}\log(1-\widetilde{\rho}^2_{W_kW_C})\\ &-\frac{N}{2}\sum_{k=1}^{C-2}\log(1-\widetilde{\rho}^2_{W_kW_{C-1}|W_C})-\cdots-\frac{N}{2}\log(1-\widetilde{\rho}^2_{W_1W_2|W_C:W_3}).\end{aligned}$$

Note that the (conditional) partial correlations in $\boldsymbol{\Theta}$ are a one-to-one re-parameterization of the conditional correlations in $\widetilde{\mathbf{R}}$. To illustrate this, let $C=2$ and $Q=1$, so that the conditional log-likelihood function in this case involves the conditional correlations $\widetilde{\rho}_{W_1^*W_1|W_2}, \widetilde{\rho}_{W_2^*W_1|W_2}$, and $\widetilde{\rho}_{W_1^*W_2^*|W_1}$. These can be represented in terms of the conditional correlations $\widetilde{\rho}_{W_1W_2}$, $\widetilde{\rho}_{W_1^*W_2}$, and $\widetilde{\rho}_{W_2^*W_2}$ as follows:

$$\begin{aligned}\widetilde{\rho}_{W_1^*W_1|W_2} &= \frac{\widetilde{\rho}_{W_1^*W_1}-\widetilde{\rho}_{W_1^*W_2}\widetilde{\rho}_{W_1W_2}}{\sqrt{(1-\widetilde{\rho}^2_{W_1^*W_2})(1-\widetilde{\rho}^2_{W_1W_2})}},\\ \widetilde{\rho}_{W_1^*W_2|W_1} &= \frac{\widetilde{\rho}_{W_1^*W_2}-\widetilde{\rho}_{W_1^*W_1}\widetilde{\rho}_{W_1W_2}}{\sqrt{(1-\widetilde{\rho}^2_{W_1^*W_1})(1-\widetilde{\rho}^2_{W_1W_2})}},\\ \widetilde{\rho}_{W_1W_2|W_1^*} &= \frac{\widetilde{\rho}_{W_1W_2}-\widetilde{\rho}_{W_1^*W_1}\widetilde{\rho}_{W_1^*W_2}}{\sqrt{(1-\widetilde{\rho}^2_{W_1^*W_1})(1-\widetilde{\rho}^2_{W_1^*W_2})}}.\end{aligned}$$

It is clear that $\widetilde{\mathbf{R}}$ can be recovered from $\boldsymbol{\Theta}$.

The marginal log-likelihood contribution of subject i is then

$$\ell_i(\boldsymbol{\Theta}) = \log\left\{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\prod_{t=1}^{T}\exp\{\ell_{it}(\boldsymbol{\Theta}|\mathbf{b}_i)\}f_{\mathbf{B}_i}(\mathbf{b}_i)d\mathbf{b}_i\right\}. \tag{2.16}$$

The full log-likelihood function is thus $\ell(\boldsymbol{\Theta})=\sum_{i=1}^{N}\ell_i(\boldsymbol{\Theta})$. Therefore, the maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\Theta}}$ is obtained by maximizing $\ell(\boldsymbol{\Theta})$ using an iterative technique such as a Newton-Raphson or quasi-Newton method. Under standard regularity conditions, the MLE $\widehat{\boldsymbol{\Theta}}$ is consistent and asymptotically multivariate normal with mean $\boldsymbol{\Theta}$ and covariance matrix given by the inverse of the Fisher information matrix $E\{-h(\boldsymbol{\Theta})\}=E\{s(\boldsymbol{\Theta})s^\top(\boldsymbol{\Theta})\}$, where

$s(\boldsymbol{\Theta}) = \partial\ell(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}$ is the score function and $h(\boldsymbol{\Theta}) = \partial^2\ell(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^\top$ is the Hessian matrix. Large-sample standard errors (SEs) for $\widehat{\boldsymbol{\Theta}}$ are calculated from diagonals of $\{s(\widehat{\boldsymbol{\Theta}})s^\top(\widehat{\boldsymbol{\Theta}})\}^{-1}$ or of $-h^{-1}(\widehat{\boldsymbol{\Theta}})$, provided either matrix is invertible. The MLE $\widehat{\boldsymbol{\Theta}}$ can be calculated easily using standard software, such as PROC NLMIXED in SAS and function `optim` in R. The former automatically includes SEs in its output which is convenient, while the latter outputs the inverse of the Hessian matrix, from whose diagonals the large-sample SEs can be obtained.

Although full likelihood estimation is straightforward, the multi-dimensional integrations involved in the marginal likelihood evaluations can take on prohibitive computational costs. As a consequence, full likelihood inference might be difficult for high-dimensional data settings (Faes *et al.*, 2008). A possible remedy based on pairwise likelihoods (Fieuws and Verbeke, 2006) which achieves important computational economies is discussed next.

2.5.2 Pairwise likelihood estimation

Instead of maximizing the full log-likelihood function based on (2.16), we fit all pairwise bivariate models, and estimate these bivariate models in a maximum likelihood framework Fieuws and Verbeke (2006). Under standard regularity conditions, the resulting estimates are still consistent and asymptotically multivariate normally distributed. Because some elements in $\boldsymbol{\Theta}$ appear in multiple pairwise bivariate models, these parameters will correspondingly have multiple estimates. The final single estimate for each parameter in $\boldsymbol{\Theta}$ is then obtained by averaging estimates obtained from the bivariate models. Fieuws and Verbeke (2006) showed that these averaged estimates are still asymptotically normally distributed with the correct parameter values as means, but the SEs do not directly follow from the combinations of the SEs of the individual results. Therefore, we need to have an additional step to calculate the sampling variability of the estimates based on pseudo-likelihood theory.

Fitting the pairwise bivariate models from the full joint model is tantamount to maximizing the pairwise log-likelihood functions separately. Suppose $q, q' = 1, \cdots, Q + C$, such that $q' < q$, and let $\boldsymbol{\Theta}_{qq'}$ denote the vector of all parameters in the bivariate joint model

corresponding to the specific outcome-pair $\{q, q'\}$. The pairwise log-likelihood function is then

$$\ell_{qq'}(\boldsymbol{\Theta}_{qq'}) \;=\; \sum_{i=1}^{N} \ell_{qq'i}(\boldsymbol{\Theta}_{qq'}). \tag{2.17}$$

where $\ell_{qq'i}(\boldsymbol{\Theta}_{qq'})$ is of the form (2.16) in which $\ell_{qq'it}(\boldsymbol{\Theta}_{qq'}|\mathbf{B}_i)$ is also of the form $\ell_{it}(\boldsymbol{\Theta}|\mathbf{B}_i)$ in Section 2.5.1, albeit for the bivariate case. Note that this follows from the closure property of the Gaussian copula under marginalization.

With a total of $P = (Q + C)(Q + C - 1)/2$ possible pairwise bivariate models, we can write (2.17) as $\sum_{i=1}^{N} \ell_{pi}(\boldsymbol{\Theta}_p)$, for $p = 1, \cdots, P$. Hence, if we let $\boldsymbol{\Theta}^{\dagger}$ be the stacked vector comprising all bivariate model-specific parameter vectors $\boldsymbol{\Theta}_p$, i.e., $\boldsymbol{\Theta}^{\dagger} = (\boldsymbol{\Theta}_1^{\top}, \cdots, \boldsymbol{\Theta}_P^{\top})^{\top}$, then the estimates for the elements of $\boldsymbol{\Theta}^{\dagger}$ are obtained by maximizing each of the P pairwise likelihoods $\ell_{qq'}(\boldsymbol{\Theta}_{qq'})$ separately to get $\widehat{\boldsymbol{\Theta}}_1, \cdots, \widehat{\boldsymbol{\Theta}}_P$. The final estimate $\widehat{\boldsymbol{\Theta}}_{\text{pair}}$, called the maximum pairwise likelihood estimate (MPLE), of the original parameter vector $\boldsymbol{\Theta}$ can be obtained by averaging all corresponding bivariate model-specific MLEs in $\widehat{\boldsymbol{\Theta}}^{\dagger}$. The linear combination of these estimates shares the same asymptotic properties with their constituting elements.

The large-sample SEs of the MPLEs of parameters in $\boldsymbol{\Theta}$ can be obtained through the large-sample SEs of elements of $\widehat{\boldsymbol{\Theta}}^{\dagger}$. The asymptotic covariance matrix of $\widehat{\boldsymbol{\Theta}}^{\dagger}$ is obtained using pseudo-likelihood theory (Renard *et al.*, 2004), which results in the familiar sandwich covariance matrix. Specifically, an asymptotic multivariate normal distribution for $\widehat{\boldsymbol{\Theta}}^{\dagger}$ can be derived (Fieuws and Verbeke, 2006), i.e.,

$$\sqrt{N}(\widehat{\boldsymbol{\Theta}}^{\dagger} - \boldsymbol{\Theta}^{\dagger}) \overset{a}{\sim} N_{P^{\dagger}}(\mathbf{0}, \boldsymbol{\Omega} = \mathbf{J^{-1}K J^{-1}}), \tag{2.18}$$

where $P^{\dagger}$ is the dimension of $\boldsymbol{\Theta}^{\dagger}$, $\mathbf{J}$ is a block-diagonal matrix with diagonal blocks $\mathbf{J}_{pp}$, and $\mathbf{K}$ is a symmetric matrix containing blocks $K_{pp'}$:

$$\mathbf{J}_{pp} \;=\; -\frac{1}{N}\sum_{i=1}^{N} E\left(\frac{\partial^{P^{\dagger}} \ell_{pi}(\boldsymbol{\Theta}_p)}{\partial \boldsymbol{\Theta}_p (\partial \boldsymbol{\Theta}_p)^{\top}}\right), \quad \mathbf{K}_{pp'} \;=\; \frac{1}{N}\sum_{i=1}^{N} E\left\{\frac{\partial \ell_{pi}(\boldsymbol{\Theta}_p)}{\partial \boldsymbol{\Theta}_p}\left(\frac{\partial \ell_{p'i}(\boldsymbol{\Theta}_{p'})}{\partial \boldsymbol{\Theta}_{p'}}\right)^{\top}\right\}.$$

Table 2.1: Summary statistics for data on union membership, hourly wage and annual hours from 1980-1987 U. S. National Longitudinal Survey.

Time	Union membership		log(Hourly wage)		log(Annual hours)	
	Number	Percent	Mean	SD	Mean	SD
1	137	0.251	1.394	0.558	3.258	0.185
2	136	0.25	1.513	0.531	3.291	0.16
3	140	0.257	1.572	0.497	3.304	0.144
4	134	0.246	1.619	0.481	3.33	0.118
5	137	0.251	1.69	0.524	3.343	0.1
6	122	0.224	1.739	0.523	3.347	0.099
7	115	0.211	1.8	0.515	3.352	0.106
8	143	0.262	1.867	0.467	3.361	0.1

Using the above approach, we thus are able to take the variability among the bivariate model-specific estimates into account. Therefore, the large-sample SEs can be obtained by simply dropping the expectations and replacing the unknown parameters by their MPLEs. It follows that $\widehat{\boldsymbol{\Theta}}_{\text{pair}} = \mathbf{A}\widehat{\boldsymbol{\Theta}}^{\dagger}$ has an asymptotic multivariate normal distribution with mean $\boldsymbol{\Theta}$ and covariance matrix $\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^{\top}$, where the matrix $\mathbf{A}$ contains the appropriate coefficients for averaging multiple estimates in $\widehat{\boldsymbol{\Theta}}^{\dagger}$ to obtain MPLEs of the corresponding parameters in $\boldsymbol{\Theta}$; large-sample SEs of $\widehat{\boldsymbol{\Theta}}_{\text{pair}}$ are obtained from the diagonals of $\mathbf{A}\widehat{\boldsymbol{\Omega}}_{\text{pair}}\mathbf{A}^{\top}$, where $\widehat{\boldsymbol{\Omega}}_{\text{pair}}$ is $\boldsymbol{\Omega}$ evaluated at $\widehat{\boldsymbol{\Theta}}_{\text{pair}}$. Computational and statistical performance (i.e. bias and efficiency) of this method has been shown to range from acceptably good to excellent (Fieuws and Verbeke, 2006).

2.6 Application to wages-hours-union memberships data

We consider annual data on $N = 545$ male youth workers from the Youth Sample of the U. S. National Longitudinal Survey for the period 1980-1987 (Vella and Verbeek, 1998). For our purposes, we focus our analysis in three mixed outcomes, namely, log-hourly wage Y_{i1t}, log-annual worked hours Y_{i2t}, which are both continuous, and union membership X_{it}, which

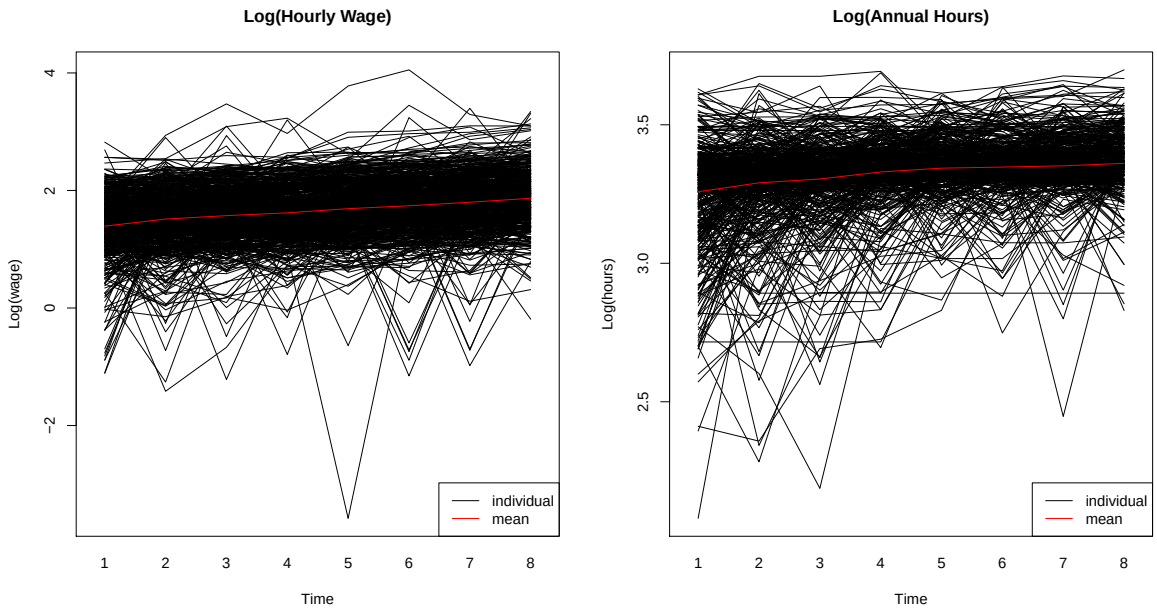

Figure 2.1: Spaghetti plots for log(Hourly wage) and log(Annual hours).

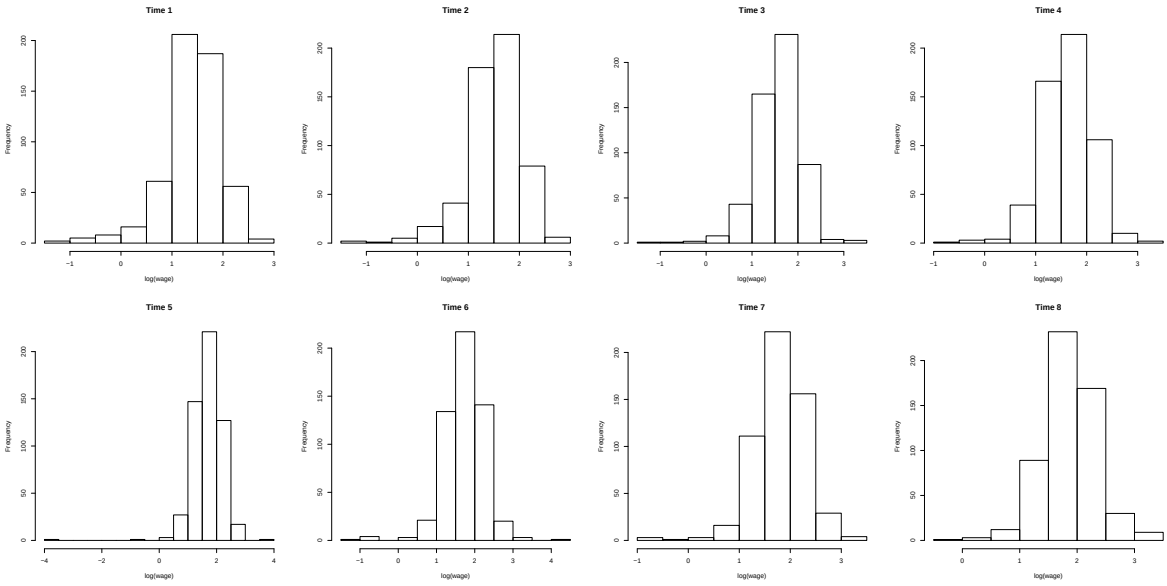

Figure 2.2: Histograms of log(Hourly wage) at time $t = 1$ to time $t = 8$.

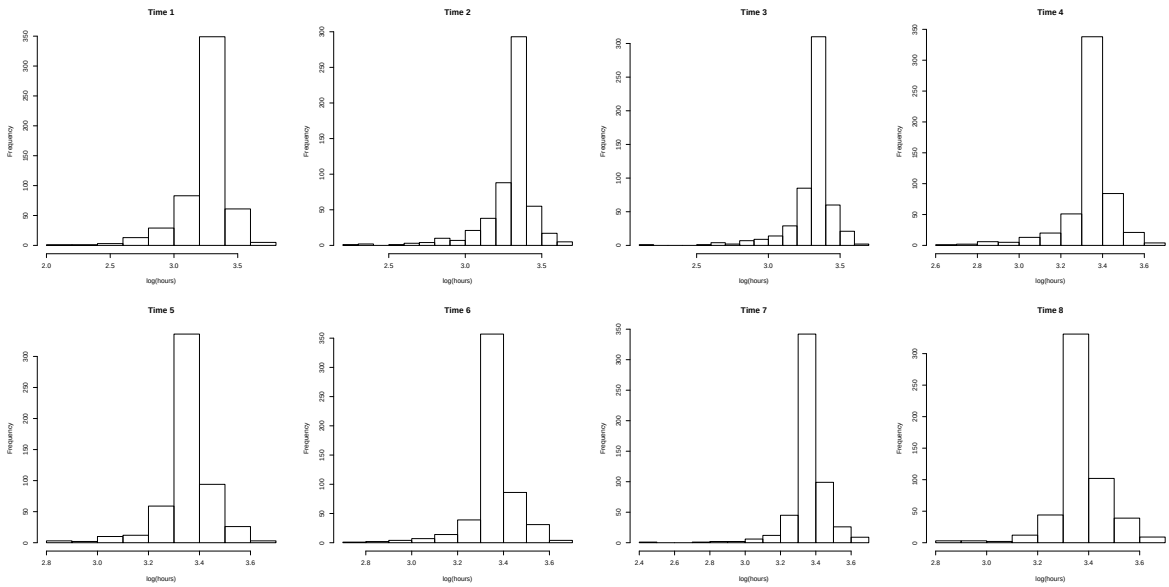

Figure 2.3: Histograms of log(Annual hours) at time $t = 1$ to time $t = 8$.

is binary indicator of whether a male youth worker holds a union membership ($X_{it} = 1$) or not ($X_{it} = 0$), for $t = 1, \cdots, 8$, $i = 1, \cdots, 545$, where $t = 1$ corresponds to year 1980, $t = 2$ to year 1981, and so on.

Summary statistics for log-hourly wage, log-annual worked hours, and union membership for the $N = 545$ male youth workers at the $T = 8$ time points are displayed in Table 2.1. From Table 2.1, we see the number of union memberships varied over the time. The means of log-hourly wage and log-annual worked hours steadily increased from over the entire 8-year period. From Figure 2.1, spaghetti plots of log-hourly wage and log-annual worked hours show slight a increase over the time.

In our analysis, we adopt the Gaussian copula mixed model in Section 2.3 for $Q = 1, C = 2$, and with $X_{it} = \mathrm{I}\{Y_{it}^* > 0\}$. We assume a within-subject shared random intercept B_{i1} and a within-subject random slope B_{i2}, such that B_{i1} and B_{i2} have a joint bivariate normal distribution with $E(B_{i1}) = E(B_{i2}) = 0$, $var(B_{i1}) = var(B_{i2}) = 1$, and $cov(B_{i1}, B_{i2}) = corr(B_{i1}, B_{i2}) = \rho_{B_1 B_2}$. Given the random effects B_{i1} and B_{i2}, we have the following mixed

Table 2.2: Full and pairwise likelihood estimates, their SEs, and p-values for wages-hours-union memberships data, using a logit-normal-normal Gaussian copula mixed model with shared random intercepts and slopes.

Effect	Parameter	Full ML			Pairwise ML		
		Est	SE	p-value	Est	SE	p-value
Union membership							
Intercept	α_1	–0.69	0.109	< 0.001	–0.527	0.007	< 0.001
Time	α_2	–0.512	0.06	< 0.001	–0.694	0.004	< 0.001
Scale factor (B_{i1})	λ_1^*	0.678	0.124	< 0.001	0.583	0.013	< 0.001
Scale factor (B_{i2})	λ_2^*	0.933	0.073	< 0.001	1.023	0.005	< 0.001
log(Hourly wage)							
Intercept	β_{11}	1.385	0.019	< 0.001	1.38	0.000	< 0.001
Time	β_{12}	0.062	0.007	< 0.001	0.064	0.000	< 0.001
Scale factor (B_{i1})	λ_{11}	0.379	0.014	< 0.001	0.386	0.001	< 0.001
Scale factor (B_{i2})	λ_{12}	–0.04	0.003	< 0.001	–0.041	0.000	< 0.001
Error variance	σ_1^2	0.359	0.005	< 0.001	0.296	0.000	< 0.001
log(Annual worked hours)							
Intercept	β_{21}	3.262	0.004	< 0.001	3.254	0.000	< 0.001
Time	β_{22}	0.014	0.001	< 0.001	0.015	0.000	< 0.001
Scale factor (B_{i1})	λ_{21}	0.031	0.003	< 0.001	–0.03	0.000	< 0.001
Scale factor (B_{i2})	λ_{22}	–0.001	0.001	0.238	–0.003	0.000	< 0.001
Error variance	σ_2^2	0.127	0.002	< 0.001	0.091	0.000	< 0.001
Associations							
Correlation between random effects	$\rho_{B_1B_2}$	0.346	0.038	< 0.001	–0.027	0.002	< 0.001
Correlations between normal scores	$\widetilde{\rho}_{Y^*Y_1}$	0.148	0.033	< 0.001	0.114	0.000	< 0.001
	$\widetilde{\rho}_{Y^*Y_2}$	–0.135	0.034	< 0.001	–0.076	0.001	< 0.001
	$\widetilde{\rho}_{Y_1Y_2}$	–0.233	0.02	< 0.001	–0.132	0.001	< 0.001
Correlations between outcomes	$\rho_{Y^*Y_1}$	0.138	—	—	0.105	—	—
	$\rho_{Y^*Y_2}$	–0.077	—	—	–0.034	—	—
	$\rho_{Y_1Y_2}$	–0.128	—	—	–0.056	—	—

Table 2.3: Estimated marginal biserial correlations between union membership indicator and log-hourly wage for time points $t = 1, \cdots, 8$, using full likelihood estimation.

$\widehat{corr(Y^*_{it}, Y_{i1t})}$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
$t=1$	0.154	0.049	0.041	0.033	0.025	0.017	0.009	0.002
$t=2$	—	0.14	0.032	0.021	0.011	0.000	–0.01	–0.02
$t=3$	—	—	0.123	0.014	0.002	–0.009	–0.02	–0.032
$t=4$	—	—	—	0.105	–0.003	–0.015	–0.027	–0.04
$t=5$	—	—	—	—	0.087	–0.019	–0.031	–0.044
$t=6$	—	—	—	—	—	0.07	–0.034	–0.048
$t=7$	—	—	—	—	—	—	0.053	–0.05
$t=8$	—	—	—	—	—	—	—	0.04

Table 2.4: Estimated marginal biserial correlations between union membership indicator and log-hourly wage for time points $t = 1, \cdots, 8$, using pairwise likelihood estimation.

$\widehat{corr(Y^*_{it}, Y_{i1t})}$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
$t=1$	0.096	0.024	0.016	0.009	0.002	–0.005	–0.011	–0.017
$t=2$	—	0.068	–0.005	–0.015	–0.024	–0.033	–0.041	–0.048
$t=3$	—	—	0.045	–0.027	–0.037	–0.046	–0.055	–0.063
$t=4$	—	—	—	0.026	–0.044	–0.054	–0.063	–0.071
$t=5$	—	—	—	—	0.008	–0.059	–0.068	–0.076
$t=6$	—	—	—	—	—	–0.007	–0.071	–0.082
$t=7$	—	—	—	—	—	—	–0.021	–0.082
$t=8$	—	—	—	—	—	—	—	–0.034

models for the conditional means:

$$E(Y_{it}^*|B_{i1}, B_{i2}) = \mu_{it}^*(B_{i1}, B_{i2}) = \alpha_1 + \alpha_2 t + \lambda_1^* B_{i1} + \lambda_2^* B_{i2} t, \tag{2.19}$$

$$E(Y_{i1t}|B_{i1}, B_{i2}) = \mu_{i1t}(B_{i1}, B_{i2}) = \beta_{11} + \beta_{12} t + \lambda_{11} B_{i1} + \lambda_{12} B_{i2} t, \tag{2.20}$$

$$E(Y_{i2t}|B_{i1}, B_{i2}) = \mu_{i2t}(B_{i1}, B_{i2}) = \beta_{21} + \beta_{22} t + \lambda_{21} B_{i1} + \lambda_{22} B_{i2} t, \tag{2.21}$$

for $t = 1, \cdots, 8$, $i = 1, \cdots, 545$. Given the shared random intercept B_{i1} and the random slope B_{i2}, we assume a logistic-latent (conditional) distribution for Y_{it}^* with unit scale. Note that this implies a logistic mixed regression model for X_{it}. The (conditional) distributions Y_{i1t} and Y_{i2t}, given the shared random effects, are assumed to be normal with variances σ_1^2 and σ_2^2, respectively. Note that this implies that the marginal distributions of Y_{i1t} and Y_{i2t} are also normal. Histograms in Figures 2.2 and 2.3 for Y_{i1t} and Y_{i2t} at the time points $t = 1, \cdots, 8$, appear to validate this assumption.

Both full and pairwise likelihood estimation methods outlined in Sections 2.5.1 and 2.5.2 were implemented for the data using SAS's PROC NLMIXED. The resulting estimates for the logit-normal-normal Gaussian copula mixed model are reported in Table 2.2. We can see that the both sets of estimates are very close, especially for the regression coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^\top$, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12})^\top$, and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22})^\top$. Positive fixed effects of time for log(hourly wage) and log(annual worked hours) indicate that a male youth worker's hourly wage and annual worked hours increase, on average, over time. A slightly negative fixed effect of time for union membership indicates that over time, the proportion of union membership (conditional on the random effects) tends to decrease slightly. In Table 2.2, we find slightly positive conditional correlations between union membership and log-hourly wage. This indicates that union members have slightly higher hourly wages than non-union members, after taking worker-level heterogeneity into account. The slightly negative correlations between union membership and log-annual worked hours and between log-hourly wage and log-annual worked hours suggest, given differences among workers, that union members tend to work fewer hours than non-union members, and workers with higher hourly wages tend

to work fewer hours annually than those with lower hourly wages. All this suggests that the 3 outcomes are conditionally dependent, and through our methodology, we are able to easily incorporate this in the analysis. Note that conventional GLMMs will necessitate assuming conditional independence of the outcomes.

Tables 2.3 and 2.4 show estimates of the marginal biserial correlations between union membership and log(hourly wage) based on full and pairwise MLEs, respectively. We see that the positive correlations decrease over time while the negative correlations increase over time.

2.7 Discussion

In this chapter, we developed a Gaussian copula-based joint model for multiple mixed discrete and continuous longitudinal outcomes. In the joint model, we introduced random effects to incorporate within-subject associations between the outcomes (of the same or different types) at the same time point, and/or at different time points; note that although we used shared random effects in our discussion and illustration, this is not necessary, as correlated random effects may likewise be adopted for our model (see Chapter 3 for details). The regression models for the outcomes are specified using GLMMs linking the outcomes' means to covariates, conditional on random effects. By using Gaussian copulas, we are able to separately specify the outcome-specific mixed models from the model for the outcomes' dependence structure. Therefore, our joint model is a flexible alternative to conventional approaches that typically assume the outcomes or some transformations of them are normally distributed in order to account for conditional dependence. Our Gaussian copula mixed model is able to incorporate conditional dependence regardless of the outcome type, which may be Gaussian or non-Gaussian (e.g., time-to-event, binary, ordinal).

By introducing latent variable descriptions of the discrete outcomes, our Gaussian copula mixed model is still unique and the non-identifiability problem is not an issue. Moreover,

such latent variable formulations are appealing in practice, both biologically and statistically, since regression models for discrete outcomes have equivalent parallel formulations in terms of regression models for latent variables. For example, a logistic or probit regression model for a binary response can be equivalently viewed in terms of a latent variable with a logistic or normal distribution. Additionally, latent variable models lead to familiar association measures for capturing dependence between two discrete outcomes (i.e., polychoric correlation) or between mixed discrete and continuous outcomes (i.e., polyserial correlation).

While our Gaussian copula mixed models can be readily implemented via standard statistics software and packages (e.g., PROC NLMIXED), the computations involved may become difficult to carry out in high dimensional settings. To circumvent this problem, we also adopted Fieuws and Verbeke's (2006) pairwise maximum likelihood method of estimation to simplify the multi-dimensional likelihood evaluations involved in the full likelihood approach as well as to reduce the incidence of non-convergence.

In this chapter, we exclusively used the Gaussian copula to develop the joint model, as in Song *et al.* (2009), because Gaussian copulas are mathematically and computationally tractable. Although they are quite popular in a variety of applications, they may not work in all cases. In this situation, other copula families could be considered as well. Note that our methodology is general enough to be easily and straightforwardly adapted to different copula families.

Chapter 3

Gaussian copula mixed models for clustered mixed outcomes in developmental toxicology

3.1 Introduction

The purpose of developmental toxicity studies is to determine toxicity levels of chemical and pharmaceutical substances and the aim is to regulate their properly usage dose levels on human beings. The studies are typically done on laboratory animals: monkeys, rabbits but mostly pregnant mice. The pregnant mice are exposed to toxicants, who have adverse effects such as low birth weight or/and congenital malformations, at increasing dose levels, then their offsprings are observed. Such observations often involve correlated mixed discrete and continuous outcomes measured on clustered fetuses. Consider, for example, the ethylene glycol (EG) data from a developmental toxicity study conducted by the National Toxicology Program (Price *et al.*, 1985): a total of 94 pregnant mice, called dams or clusters, are randomly exposed to EG at four different dose levels, 0, 0.75, 1.5, and 3g/kg/day, with dose 0g/kg/day as a control level, during the period of development of major fetal organs. The 1028 live fetuses from the 94 litters, with litter sizes ranging from 1 to 16, were examined for various defects, including fetal weight (continuous) and the presence/absence of fetal malformations (binary), both are sensitive indicators of toxicity.

In such studies, the basic sampling unit is typically a litter of fetuses, each yielding mixed outcomes, and the main interest is on the dose-response relationships. Risk assessment is traditionally undertaken for each outcome separately; however, simultaneously analyzing the outcomes and then carrying out a joint risk assessment may be more appropriate in practice,

Wu and de Leon (2013*a*), forthcoming in *Journal of Agricultural, Biological & Environmental Statistics*, is a slightly modified version of this chapter.

given that the outcomes are correlated. There is thus a need for flexible joint models that can meaningfully capture the dose-response relationships and the associations between outcomes on different and/or the same fetuses, including the outcomes' marginal and conditional characteristics. Ideally, such joint models should have (1) marginally interpretable dose-response models for the outcomes; in addition, a desirable joint model should (2) account for intra-litter (i.e., between fetuses from same litter) correlations, and (3) directly incorporate the intra-fetus association between mixed outcomes. Property (3) is necessary in quantitative risk assessment, where fetal risk (i.e., probability of fetal malformation and/or "low" fetal weight) needs to be estimated.

Joint model specifications for mixed discrete and continuous outcomes are discussed in Chapter 1. In this chapter, we adapt the Gaussian copula-based modeling strategy introduced in Chapter 2 to the analysis of clustered mixed data in developmental toxicology. The Gaussian copula mixed models developed in Chapter 2 extends and generalizes models previously proposed by Gueorguieva and Agresti (2001) and Lin *et al.* (2010). We introduce the Gaussian copula mixed effects model for clustered data in Section 3.2; the various associations in the data (between different/same outcomes) are provided as well. Note that there is some duplication in the discussion of the model; in addition, some notations are slightly modified for convenience. In Section 3.3, we discuss the likelihood representation of the model along with corresponding likelihood estimation. Empirical results for the simulation study of finite-sample properties of MLEs for the model are reported in Section 3.4. Section 3.5 illustrates the application of the model to the EG mice data. Finally, Section 3.6 concludes the chapter.

3.2 Gaussian copula mixed model

Recall the EG mice data, and consider the respective correlated binary and continuous outcomes X_{ih}, which indicates presence/absence of fetal malformations, and Y_{ih}, which rep-

resents fetal weight, for fetus $h = 1, \cdots, n_i$, in litter $i = 1, \cdots, N$. Let $X_{ih} = 1$, if fetus h in litter i is malformed, and $X_{ih} = 0$, otherwise. Similar to the development in Chapter 2, let Y_{ih}^* be an unobserved continuous latent variable underlying X_{ih}, such that $X_{ih} = \mathrm{I}\{Y_{ih}^* > 0\}$ (Gueorguieva and Agresti, 2001; Catalano and Ryan, 1992). Assuming a latent continuous structure for X_{ih} holds much intuitive appeal to toxicologists, as it provides a natural description of the biological process underlying fetal malformation (Ryan, 2002). This can be readily extended to a polychotomous ordinal fetal malformation outcome via so-called threshold models, as in (2.1) we introduced in Chapter 2.

To simplify the ensuing discussion, we assume litter-level correlated random effects B_i^* and B_i corresponding to Y_{ih}^* and Y_{ih}, respectively, and consider the underlying linear mixed models given by

$$Y_{ih}^* \;=\; \mathbf{z}_{1i}^\top \boldsymbol{\alpha} + \lambda^* B_i^* + \varepsilon_{ih}^*, \qquad Y_{ih} \;=\; \mathbf{z}_{2i}^\top \boldsymbol{\beta} + \lambda B_i + \varepsilon_{ih}, \tag{3.1}$$

where ε_{ih}^* and ε_{ih} are the error terms, $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ are known litter-level (possibly fetus-level) outcome-specific covariate vectors, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corresponding unknown regression coefficients, with λ^* and λ accounting for heterogeneity between B_i^* and B_i. Note that λ^* and λ are the respective standard deviations (SDs) of the re-scaled random effects $\bar{B}_i^* = \lambda^* B_i^*$ and $\bar{B}_i = \lambda B_i$. Instead of specifying the residual error distribution, usually Gaussian, to construct the conditional joint distribution of responses given random effects, as is usually done in GLMMs, our approach specifies the latter directly by specifying the outcome-specific conditional CDFs $F_{Y_{ih}^*|B_i^*}(\cdot|\cdot)$ and $F_{Y_{ih}|B_i}(\cdot|\cdot)$, and coupling them together using a Gaussian copula. In practice, statisticians oftentimes know very little about the joint behavior of outcomes but can specify their marginal behaviors reasonably well; copulas provide a flexible means of assembling a joint distribution in this case. Using the Gaussian copula, we have the conditional CDF of Y_{ih}^* and Y_{ih} as

$$F_{Y_{ih}^*, Y_{ih}|B_i^*, B_i}(y_{ih}^*, y_{ih}|b_i^*, b_i) \;=\; \Phi_2\left(\Phi^{-1}\{u_{ih}(b_i^*)\}, \Phi^{-1}\{v_{ih}(b_i)\}; \widetilde{\rho}\right), \tag{3.2}$$

where $u_{ih}(b_i^*) = F_{Y_{ih}^*|B_i^*}(y_{ih}^*|b_i^*)$ and $v_{ih}(b_i) = F_{Y_{ih}|B_i}(y_{ih}|b_i)$. In (3.2),

$$\widetilde{\rho} \;=\; corr(W_{ih}^*(B_i^*), W_{ih}(B_i)|B_i^*, B_i) \tag{3.3}$$

is the correlation between conditional normal scores $W_{ih}^*(B_i^*) = \Phi^{-1}\{U_{ih}(B_i^*)\}$ and $W_{ih}(B_i) = \Phi^{-1}\{V_{ih}(B_i)\}$, where $U_{ih}(B_i^*) = F_{Y_{ih}^*|B_i^*}(Y_{ih}|B_i^*)$ and $V_{ih}(B_i) = F_{Y_{ih}|B_i}(Y_{ih}|B_i)$ are the conditional PITs; note that $\widetilde{\rho}$ was referred to as a normal correlation in Chapter 2. The conditional CDF $F_{Y_{ih}^*,Y_{ih}|B_i^*,B_i}(\cdot|\cdot)$ is thus specified via its (conditional) margins (given B_i^* and B_i) and the Gaussian copula that glues them together. Note that the conditional margins $F_{Y_{ih}^*|B_i^*}(\cdot|\cdot)$ and $F_{Y_{ih}|B_i}(\cdot|\cdot)$ need not come from the same parametric family, allowing researchers much flexibility in modeling disparate outcomes. The use of Gaussian copula in (3.2) is also appealing, as it describes dependence in the same way that the multivariate normal distribution does, in addition to its analytical and computational tractability (de Leon and Wu, 2011; Song *et al.*, 2009). For identifiability reasons, we assume that $var(Y_{ih}^*|B_i^*) = 1$, for all i, j. If the variance depends on a scale parameter, we assume that Y_{ih}^*, given B_i^*, has unit scale instead. In either case, we assume that $var(Y_{ih}^*|B_i^*) = var(\varepsilon_{ih}^*|B_i^*) > 0$ is known.

Instead of the usual joint normality assumption for B_i^* and B_i (Faes *et al.*, 2008,2009; Gueorguieva and Agresti, 2001), it is possible to similarly use the Gaussian copula to build their joint density $f_{B_i^*,B_i}(\cdot)$ from

$$F_{B_i^*,B_i}(b_i^*, b_i) \;=\; \Phi_2\left(\Phi^{-1}\{F_{B_i^*}(b_i^*)\}, \Phi^{-1}\{F_{B_i}(b_i)\}; \widetilde{\eta}\right), \tag{3.4}$$

as in Lin *et al.* (2010), where $\widetilde{\eta} = corr\{\Phi^{-1}(F_{B_i^*}(B_i^*)), \Phi^{-1}(F_{B_i}(B_i))\}$, with $E(B_i^*) = E(B_i) = 0$ and $var(B_i^*) = var(B_i) = 1$. This affords flexibility in specifying the respective marginal CDFs $F_{B_i^*}(\cdot)$ and $F_{B_i}(\cdot)$ of B_i^* and B_i; for example, a bridge margin may be assumed for B_i^* along with a logistic mixed model for binary outcome X_{ih} (i.e., a logistic latent distribution for Y_{ih}^*) to facilitate interpretability of marginal effects. The approach outlined above also includes the case of a shared litter-level random effect, i.e., $B_i^* \equiv B_i$, for all i. Note that λ^* and λ in the shared random effects model account for the difference in measurement scales

between Y_{ih}^* and Y_{ih}. Unlike in the case of correlated random effects, λ^* can be negative and only λ is the SD of the random effect B_i is assumed positive. A more general random effects formulation with random slopes mentioned in the previous chapter, in addition to random intercepts, is detailed in Chapter 2 for longitudinal data.

The density of X_{ih} and Y_{ih} is then obtained as

$$f_{X_{ih},Y_{ih}}(x_{ih}, y_{ih}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X_{ih},Y_{ih}|B_i^*,B_i}(x_{ih}, y_{ih}|b_i^*, b_i) f_{B_i^*,B_i}(b_i^*, b_i) db_i^* db_i, \quad (3.5)$$

for $x_{ih} = 0, 1$, where

$$\begin{aligned} f_{X_{ih},Y_{ih}|B_i^*,B_i}(x_{ih}, y_{ih}|b_i^*, b_i) &= \left(1 - \int_0^{+\infty} f_{Y_{ih}^*,Y_{ih}|B_i^*,B_i}(y_{ih}^*, y_{ih}|b_i^*, b_i) dy_{ih}^*\right)^{1-x_{ih}} \\ &\quad \times \left(\int_0^{+\infty} f_{Y_{ih}^*,Y_{ih}|B_i^*,B_i}(y_{ih}^*, y_{ih}|b_i^*, b_i) dy_{ih}^*\right)^{x_{ih}}, \end{aligned} \quad (3.6)$$

where $f_{Y_{ih},Y_{ih}^*|B_i^*,B_i}(\cdot|\cdot)$ is the conditional density of Y_{ih}^* and Y_{ih}, given B_i^* and B_i. Note that model (3.4) above accommodates a direct intra-fetus association between the outcomes via the correlation $\widetilde{\rho}$, which accounts for conditional dependence between the outcomes; hence, the assumption of conditional independence of Y_{ih}^* and Y_{ih} (hence, of X_{ih} and Y_{ih}) is not necessary.

Note that since X_{ih} is a binary outcome, we assume the single cutpoint to be 0, so that an intercept term must be included in $\boldsymbol{\alpha}$. In addition, while the model (and inference) for X_{ih} is done at the latent level, this corresponds to a GLMM for X_{ih} with a specific link function that depends on the choice of latent distribution. For example, a logistic latent distribution for Y_{ih}^* results in a logistic mixed regression model for X_{ih}, while a normal latent distribution leads to a probit mixed model. A t-latent distribution for Y_{ih}^* generalizes the logistic and probit models for X_{ih} with the so-called robit model (de Leon and Wu, 2011). With normal margins for Y_{ih}^* and Y_{ih}, our model specializes to the correlated probit models of Najita *et al.* (2009) and Gueorguieva and Agresti (2001). Our model also extends the Gaussian copula marginal regression model of Masarotto and Varin (2012) with the inclusion of random effects.

3.2.1 Distribution of residual errors

Given $Y_{ih}^*|B_i^* \sim F_{Y_{ih}^*|B_i^*}(\cdot|\cdot)$ and $Y_{ih}|B_i \sim F_{Y_{ih}|B_i}(\cdot|\cdot)$, the corresponding residual error joint distribution can then be obtained directly by transformation methods. To see this, let $\varepsilon_{ih}^*|B_i^* \sim F_{\varepsilon_{ih}^*|B_i^*}(\cdot|\cdot)$ and $\varepsilon_{ih}|B_i \sim F_{\varepsilon_{ih}|B_i}(\cdot|\cdot)$, where

$$F_{\varepsilon_{ih}^*|B_i^*}(e_{ih}^*|b_i^*) = F_{Y_{ih}^*|B_i^*}(\mathbf{z}_{1i}^\top\boldsymbol{\alpha} + \lambda^* b_i^* + e_{ih}^*|b_i^*), \quad F_{\varepsilon_{ih}|B_i}(e_{ih}|b_i) = F_{Y_{ih}|B_i}(\mathbf{z}_{2i}^\top\boldsymbol{\beta} + \lambda b_i + e_{ih}|b_i).$$

It then follows that the (marginal) joint density $f_{\varepsilon_{ih}^*,\varepsilon_{ih}}(\cdot)$ of the residual errors is given by

$$\begin{aligned} f_{\varepsilon_{ih}^*,\varepsilon_{ih}}(e_{ih}^*, e_{ih}) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{f_{Y_{ih}^*,Y_{ih}|B_i^*,B_i}(y_{ih}^*, y_{ih}|b_i^*, b_i) f_{\varepsilon_{ih}^*|B_i^*}(e_{ih}^*|b_i^*) f_{\varepsilon_{ih}|B_i}(e_{ih}|b_i)}{f_{Y_{ih}^*|B_i^*}(y_{ih}^*|b_i^*) f_{Y_{ih}|B_i}(y_{ih}|b_i)} \\ &\quad \times f_{B_i^*,B_i}(b_i^*, b_i) db_i^* db_i, \end{aligned} \tag{3.7}$$

where $f_{Y_{ih}^*|B_i^*}(\cdot|\cdot)$ and $f_{Y_{ih}|B_i}(\cdot|\cdot)$ are the respective (conditional) marginal densities of Y_{ih}^* and Y_{ih} (given random effects B_i^* and B_i), and $f_{\varepsilon_{ih}^*|B_i^*}(\cdot|\cdot)$ and $f_{\varepsilon_{ih}|B_i}(\cdot|\cdot)$ are the respective (conditional) marginal densities of ε_{ih}^* and ε_{ih} (given random effects B_i^* and B_i). Note that (3.7) reduces to the usual joint normality assumption for ε_{ih}^* and ε_{ih} when Y_{ih}^* and Y_{ih} are assumed jointly normally distributed, given B_i^* and B_i. Hence, our model does not require $f_{\varepsilon_{ih}^*,\varepsilon_{ih}}(\cdot)$ to be normal, and is therefore more flexible than conventional GLMMs.

As a special case, consider $Y_{ih}^*|B_i^* \sim \mathit{logistic}(\mathbf{z}_{1i}^\top\boldsymbol{\alpha} + \lambda^* B_i^*, 1)$ (i.e., the logistic distribution with mean $\mathbf{z}_{1i}^\top\boldsymbol{\alpha} + \lambda^* B_i^*$ and unit scale) and $Y_{ih}|B_i \sim \mathit{normal}(\mathbf{z}_{2i}^\top\boldsymbol{\beta} + \lambda B_i, \sigma^2)$. Since both (conditional) margins belong to the location-scale family, it can be easily shown that $\varepsilon_{ih}^* \sim \mathit{logistic}(0, 1)$ and $\varepsilon_{ih} \sim \mathit{normal}(0, \sigma^2)$, with ε_{ih}^* and ε_{ih} independent of B_i^* and B_i. Figure 3.1 displays contour plots of $f_{\varepsilon_{ih}^*,\varepsilon_{ih}}(\cdot)$ in (3.7) in this case, with $\sigma^2 = var(Y_{ih}|B_i) = 1$. Observe that $f_{\varepsilon_{ih}^*,\varepsilon_{ih}}(\cdot)$ has heavier tails than the usual bivariate normal residual error distribution, and should be able to better accommodate extreme observations.

Our Gaussian copula mixed model is especially relevant in developmental toxicology, where data normally comprise clustered mixed outcomes. Because the marginal distributions are specified independently of the association, our model yields meaningful and interpretable outcome-specific dose-response models. In addition, the inclusion of random effects enables

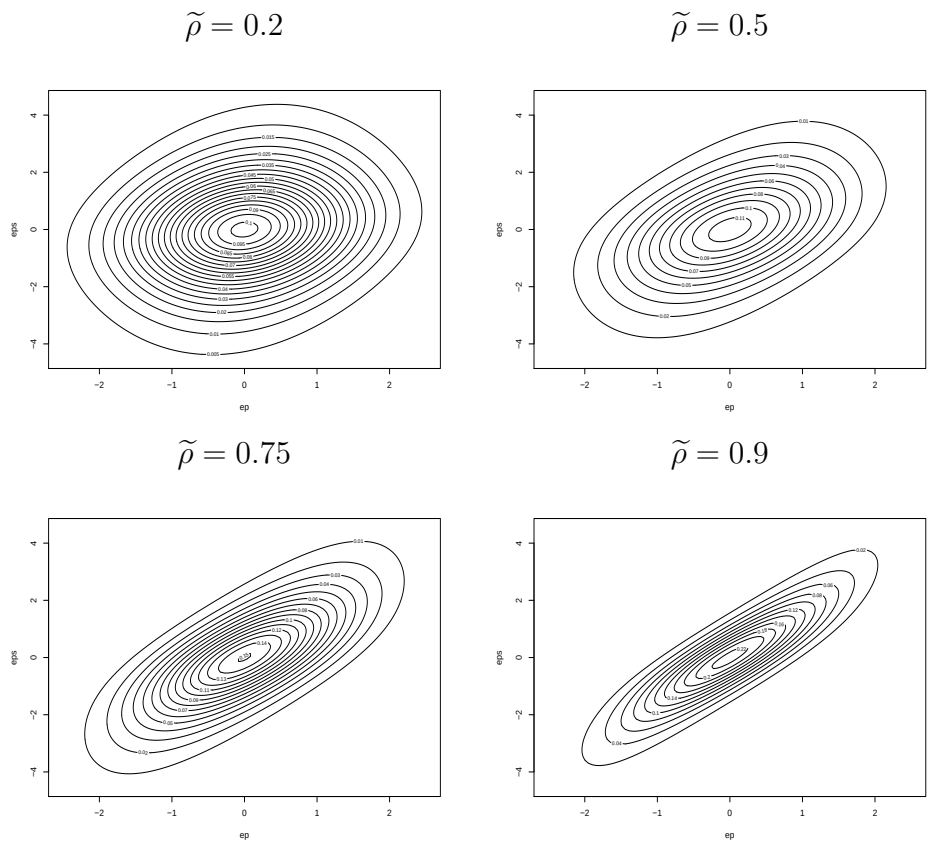

Figure 3.1: Contour plots for the joint density $f_{\varepsilon^*_{ih},\varepsilon_{ih}}(\cdot)$ of ε^*_{ih} and ε_{ih}, for $\widetilde{\rho} = 0.2, 0.5, 0.75, 0.9$, with $\varepsilon^*_{ih} \sim logistic(0,1)$ and $\varepsilon_{ih} \sim normal(0,1)$.

us to account for litter effects. The model also allows direct estimation of the intra-fetus association between the outcomes, a desirable feature in quantitative risk assessment. Hence, the model satisfies properties (1) to (3) outlined in Section 3.1. The flexibility in the specifications of both random effects distribution and residual errors distribution is also appealing in practice, where the conventional normality assumption may not be valid.

3.2.2 Marginal associations between outcomes

Using latent variable Y^*_{ih} to describe binary outcome X_{ih} implies that correlations between X_{ih} and Y_{ih} (for same fetus) and between X_{ih} and $X_{ih'}$ (for different fetuses from the same litter), are measured respectively, by the biserial correlation $corr(Y^*_{ih}, Y_{ih})$ between X_{ih} and

Y_{ih}, and the tetrachoric correlation $corr(Y^*_{ih}, Y^*_{ih'})$ between X_{ih} and $X_{ih'}$. This is standard practice in psychometrics, for example, where latent variable models are commonplace. In contrast to Pearson's correlations $corr(X_{ih}, Y_{ih})$ and $corr(X_{ih}, X_{ih'})$ as measures of association between X_{ih} and Y_{ih} and between X_{ih} and $X_{ih'}$, biserial and tetrachoric correlations are not constrained by marginal probabilities of X_{ih} (and of $X_{ih'}$), as they are nothing but the usual pairwise correlations between continuous variables (latent or otherwise). Moreover, the number of tetrachoric and biserial correlations remains the same for polychotomous data, certainly not the case for odds ratios.

For the Gaussian copula mixed model with correlated random effects in Section 3.2.1, the marginal inter-fetus and intra-fetus correlations are found to be $corr(Y_{ih}, Y_{ih'}) = \lambda^2/(\sigma^2+\lambda^2)$, $corr(Y^*_{ih}, Y^*_{ih'}) = (\lambda^*)^2/\{var(Y^*_{ih}|B_i) + (\lambda^*)^2\}$,

$$corr(Y^*_{ih}, Y_{ih}) = corr(Y^*_{ih}, Y_{ih'}) \left(1 + \frac{\rho\sigma}{\eta\lambda^*\lambda}\sqrt{var(Y^*_{ih}|B^*_i)}\right), \tag{3.8}$$

$$corr(Y^*_{ih}, Y_{ih'}) = \frac{\eta\lambda^*\lambda}{\sqrt{\sigma^2+\lambda^2}\sqrt{var(Y^*_{ih}|B^*_i) + (\lambda^*)^2}}, \tag{3.9}$$

where $\rho = corr(Y^*_{ih}, Y_{ih}|B^*_i, B_i)$ and $\eta = corr(B^*_i, B_i)$, and where we again assume $var(Y^*_{ih}|B_i) > 0$ is known and $var(Y_{ih}|B_i) = \sigma^2 > 0$. Observe that $|corr(Y^*_{ih}, Y_{ih})| \geq |corr(Y^*_{ih}, Y_{ih'})|$, confirming that the association between X_{ih} and Y_{ih} from the same fetus j is stronger than that between X_{ih} and $Y_{ih'}$ from different fetuses j and j' within cluster i. The case with shared random effects yields a slightly more restrictive association structure for the mixed outcomes from different fetuses, thus motivating consideration of correlated random effects. We illustrate this in Section 3.5.

Note that dependence between outcomes is not modeled directly in our approach; instead, correlations are incorporated for the normal scores, which are transformations of the original variables (latent in the case of Y^*_{ih}). Similarly as shown in Chapter 2, $\widetilde{\rho} = corr(W^*_{ih}(B^*_i), W_{ih}(B_i)|B^*_i, B_i)$ in (3.3) is used as proxy for ρ; it can be shown that $\rho \leq |\widetilde{\rho}|$ (Klaassen and Wellner, 1997), which can then be used to bound marginal correlations (3.8) and (3.9). Alternatively, a piecewise-linear approximation may be used to recover

ρ from $\widetilde{\rho}$ (Kugiumtzis and Bora-Senta, 2010). The same can be done to recover η from $\widetilde{\eta} = corr\{\Phi^{-1}(F_{B_i^*}(B_i^*)), \Phi^{-1}(F_{B_i}(B_i)))\}$ in (3.4).

We can also use nonparametric rank-based measures like Kendall's tau to gauge associations, since they are invariant to monotonic transformations. For the Gaussian copula mixed model with correlated random effects, we get the conditional Kendall's tau as $\tau(Y_{ih}^*, Y_{ih}|B_i^*, B_i) = \tau(W_{ih}^*(B_i^*), W_{ih}(B_i)|B_i^*, B_i) = 2\arcsin(\widetilde{\rho})/\pi$. This is the approach adopted recently by Parzen *et al.* (2011).

3.2.3 Conditional assessment of outcomes

One important feature of this joint model is that we can obtain the conditional distribution of one outcome given the other. It is thus possible to obtain conditional interpretations for one response variable given the other. For the EG mice data, we obtain the conditional probability of malformation as a function of dose, given fetal weight, or vice versa, and the conditional mean fetal weight as a function of dose, given fetal malformation, as follows:

$$\begin{aligned} P(X_{ih} = x_{ih}|Y_{ih} = y_{ih}) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{f_{X_{ih},Y_{ih}|B_i^*,B_i}(x_{ih}, y_{ih}|b_i^*, b_i)}{f_{Y_{ih}|B_i}(y_{ih}|b_i)} f_{B_i^*,B_i}(b_i^*, b_i) \\ &\quad \times db_i^* db_i, \qquad (3.10) \\ E(Y_{ih}|X_{ih} = x_{ih}) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y \frac{f_{X_{ih},Y_{ih}|B_i^*,B_i}(x_{ih}, y_{ih}|b_i^*, b_i)}{P(X_{ih} = x_{ih}|B_i^* = b_i^*)} f_{B_i^*,B_i}(b_i^*, b_i) \\ &\quad \times dy db_i^* db_i. \qquad (3.11) \end{aligned}$$

Plots of (3.10) and (3.11) can then be used to assess the conditional behavior of one outcome given the other, which are shown in Section 3.5.2. From (2.4) (see also de Leon and Wu, 2011), we get a simple expression for $f_{X_{ih},Y_{ih}|B_i^*,B_i}(x_{ih}, y_{ih}|b_i^*, b_i)$ in (3.10) and (3.11) as

$$\begin{aligned} f_{X_{ih},Y_{ih}|B_i^*,B_i}(x_{ih}, y_{ih}|b_i^*, b_i) &= \left\{1 - \Phi\left(\frac{\frac{\widetilde{\rho}}{\sigma}\{y_{ih} - \mu_i(b_i)\} - \Phi^{-1}\{u_{ih}(b_i^*)\}}{\sqrt{1-\widetilde{\rho}^2}}\right)\right\}^{1-x_{ih}} \\ &\quad \times \left\{\Phi\left(\frac{\frac{\widetilde{\rho}}{\sigma}\{y_{ih} - \mu_i(b_i)\} - \Phi^{-1}\{u_{ih}(b_i^*)\}}{\sqrt{1-\widetilde{\rho}^2}}\right)\right\}^{x_{ih}} \\ &\quad \times f_{Y_{ih}|B_i}(y_{ih}|b_i), \qquad (3.12) \end{aligned}$$

so that with $Y_{ih}|B_i \sim normal(\mu_i(B_i), \sigma^2)$, the conditional mean (3.11) simplifies further as

$$\begin{aligned} E(Y_{ih}|X_{ih} = x_{ih}) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left\{ \left(\mu_i^*(b_i^*) - \frac{\widetilde{\rho}\sigma}{u_{ih}(b_i^*)} \phi[\Phi^{-1}\{u_{ih}(b_i^*)\}] \right) \right\}^{1-x_{ih}} \\ &\quad \times \left\{ \left(\mu_i^*(b_i^*) + \frac{\widetilde{\rho}\sigma}{1-u_{ih}(b_i^*)} \phi[\Phi^{-1}\{u_{ih}(b_i^*)\}] \right) \right\}^{x_{ih}} \\ &\quad \times f_{B_i^*,B_i}(b_i^*, b_i) db_i^* db_i. \end{aligned} \tag{3.13}$$

In Section 3.5, we use Monte Carlo integration to evaluate (3.10) and (3.11) for the EG data.

3.3 Likelihood estimation

Let $\{x_{ih}, y_{ih}, \mathbf{z}_{1i}, \mathbf{z}_{2i}\}$ denote the observed data, for $h = 1, \cdots, n_i$, and $i = 1, \cdots, N$. Putting $\boldsymbol{\Theta}$ as the parameter vector containing the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the correlation $\widetilde{\rho}$, and respective parameters $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$ of (conditional) margins $F_{Y_{ih}^*|B_i^*}(\cdot|\cdot)$ and $F_{Y_{ih}|B_i}(\cdot|\cdot)$, the likelihood contribution of litter i is given by

$$L_i(\boldsymbol{\Theta}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \prod_{j=1}^{n_i} f_{X_{ih},Y_{ih}|B_i^*,B_i}(x_{ih}, y_{ih}|b_i^*, b_i) f_{B_i^*,B_i}(b_i^*, b_i) db_i^* db_i, \tag{3.14}$$

where $f_{X_{ih},Y_{ih}|B_i^*,B_i}(\cdot|\cdot)$ is given in (3.12), so that the likelihood function becomes $L(\boldsymbol{\Theta}) = \prod_{i=1}^N L_i(\boldsymbol{\Theta})$. The log-likelihood function is then $\ell(\boldsymbol{\Theta}) = \log L(\boldsymbol{\Theta}) = \sum_{i=1}^N \log L_i(\boldsymbol{\Theta})$. Note that the evaluation of $L_i(\boldsymbol{\Theta})$ in 3.14 is obtained numerically using standard software, such as PROC NLMIXED in SAS, with Gaussian quadrature and importance sampling as possible numerical integration techniques. The same large-sample approximations discussed in Chapter 2, based on asymptotic normality of the MLE $\widehat{\boldsymbol{\Theta}}$, apply in this case. For example, large-sample SEs for MLEs in $\widehat{\boldsymbol{\Theta}}$ are calculated from diagonals of $\{s(\widehat{\boldsymbol{\Theta}})s^\top(\widehat{\boldsymbol{\Theta}})\}^{-1}$ or $-h^{-1}(\widehat{\boldsymbol{\Theta}})$, provided either matrix is invertible.

3.4 Simulation study

In this section, we report simulation results for finite-sample bias and efficiency of MLEs for the robit-normal Gaussian copula mixed model with shared random effects, a particular

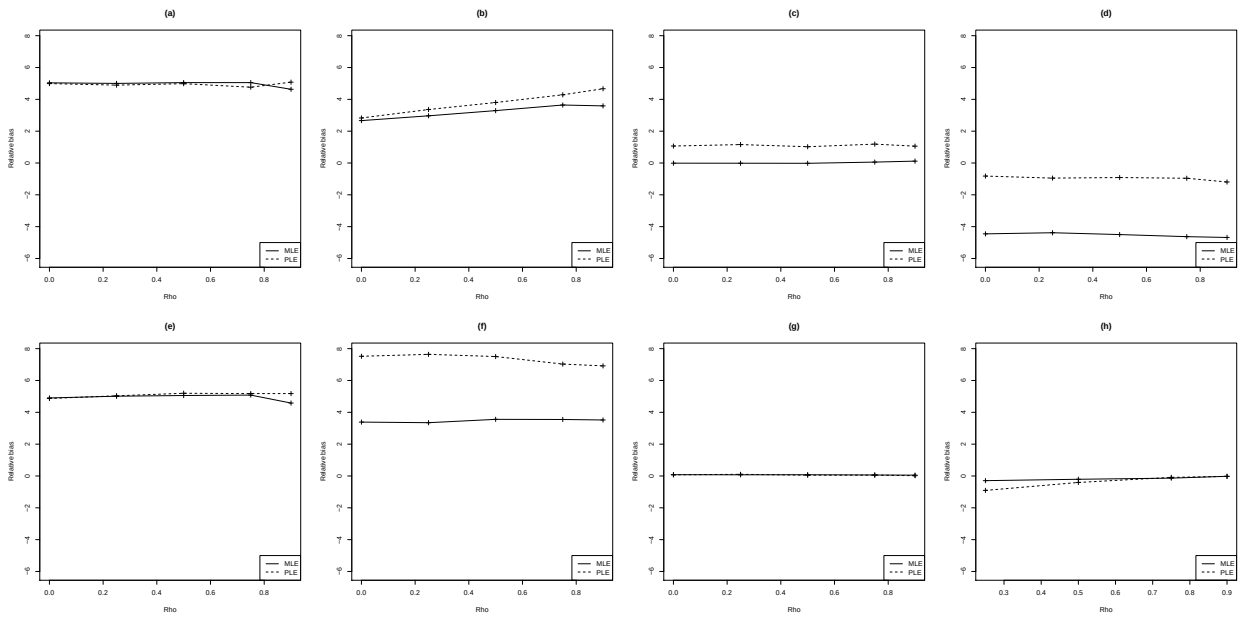

Figure 3.2: Relative bias "100 × (mean of estimates–true value)/true value" of MLEs and PLEs of (a) α_1, (b) α_2, (c) λ^*, (d) β_1, (e) β_2, (f) λ, (g) σ, and (h) $\widetilde{\rho}$ for robit-normal mixed model with shared litter-level normal random effects.

version of our model with a normal distribution for Y_{ih} and a t-distribution for Y^*_{ih}, the latent variable underlying the binary outcome $X_{ih} = \mathrm{I}\{Y^*_{ih} > 0\}$, $h = 1, \cdots, n_i$, $i = 1, \cdots, N$. Specifically, given a random effect $B_i \sim normal(0, 1)$ shared by measurements in litter i, we have $Y^*_{ih}|B_i \sim t_{\nu^*}(\mu^*_i(B_i), 1)$ (i.e., the t-distribution with mean $\mu^*_i(B_i)$, ν^* degrees of freedom, and unit scale) and $Y_{ih}|B_i \sim normal(\mu_i(B_i), \sigma^2)$, where $\mu^*_i(B_i) = \alpha_1 + \alpha_2 z_{1i} + \lambda^* B_i$ and $\mu_i(B_i) = \beta_1 + \beta_2 z_{2i} + \lambda B_i$, with outcome-specific litter-level covariates z_{1i} and z_{2i}, and with λ and λ^* accounting for the difference in measurement scales between Y^*_{ih} and Y_{ih}. As shown in Section 3.2, λ^* in the above model can be negative and only λ, the SD of the random effect B_i, is assumed positive.

We refer to the above model as the robit-normal Gaussian copula mixed model, since it yields a robit mixed model for X_{ih}, a robust alternative to the logistic and probit mixed models. As shown by Liu (2004), the resulting latent mixed model is equivalent to a mixed effects logistic regression model for X_{ih}, for $\nu^* \approx 7$; with large ν^*, it approximates a mixed

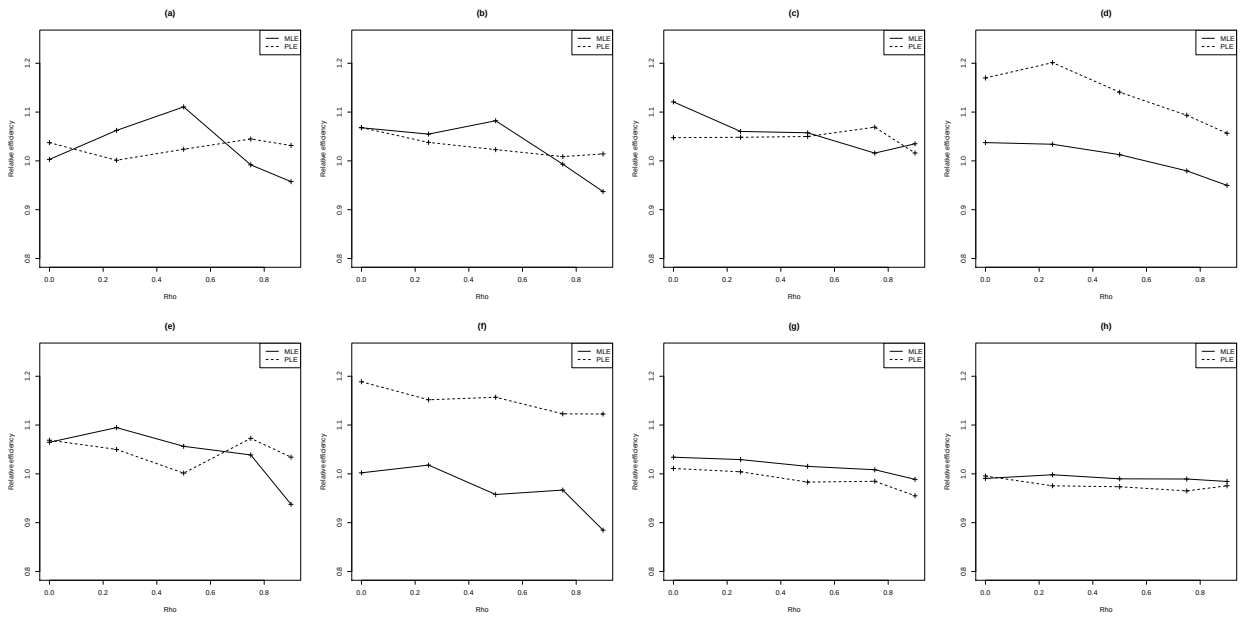

Figure 3.3: Relative efficiency "mean of SEs/empirical SD" of MLEs and PLEs of (a) α_1, (b) α_2, (c) λ^*, (d) β_1, (e) β_2, (f) λ, (g) σ, and (h) $\widetilde{\rho}$ for robit-normal mixed model with shared litter-level normal random effects.

effects probit regression model. The model generalizes the robit-normal model of de Leon and Wu (2011) to clustered data with mixed binary and continuous outcomes, with the inclusion of random effects.

For the simulations, a total of R repeated samples were generated with $N = 200$ clusters, each with varying sizes from $n_i = 1, \cdots, 10$, generated randomly via a truncated binomial distribution. Cluster-specific covariates z_{1i} and z_{2i} were generated using a uniform distribution over $(-1, 1)$. The following parameter configuration was considered: $\alpha_1 = \beta_1 = \lambda^* = \lambda = \sigma = 1$, $\alpha_2 = \beta_2 = 2$, and $\nu^* = 5$ with varying correlation $\widetilde{\rho} \in \{0, 0.25, 0.5, 0.75, 0.9\}$; the relative bias and relative efficiency of estimates were then obtained. To avoid problems associated with estimating the degrees of freedom ν^*, we employed the method of profile likelihood (Song *et al.*, 2007). The method entails maximizing the profile log-likelihood $\ell_{\nu^*}(\boldsymbol{\Theta}_{(-\nu^*)}) = \ell(\boldsymbol{\Theta}_{(-\nu^*)}; \nu^*)$ at fixed grid points $\nu^* \in (2, M]$, for some suitably large constant $M > 2$, where $\boldsymbol{\Theta}_{(-\nu^*)}$ is $\boldsymbol{\Theta}$ with ν^* removed. The profile likelihood

estimate (PLE) $\widehat{\boldsymbol{\Theta}}_{(-\nu^*)}$ corresponds to $\widehat{\nu^*}$ at which $\ell_{\nu^*}(\boldsymbol{\Theta}_{(-\nu^*)})$ is maximum on $(2, M]$. This method is easy to implement using dense grid points on $(2, M]$, and obviates the convergence issues associated with estimating ν^*. Asymptotic properties analogous to those for MLEs can be similarly established. We used PROC NLMIXED in SAS in the simulations, where we fixed $R = 1000$ for full maximum likelihood estimation and $R = 500$ for profile likelihood estimation, with $M = 8$; the smaller number of repeats for profile likelihood estimation is due to the method's extra computational requirements.

Figure 3.2 plots relative biases of MLEs and PLEs as functions of $\widetilde{\rho} = 0.25, 0.5, 0.75, 0.9$; note that the relative bias of the MLE and PLE of $\widetilde{\rho}$ is undefined for $\widetilde{\rho} = 0$. Figure 3.3 plots the corresponding relative efficiencies of MLEs and PLEs. From Figure 3.2, we can see that both MLEs and PLEs are relatively unbiased, with all plots close to 0. Figure 3.3 displays relative efficiencies that are generally close to 1, indicating that both MLEs and PLEs have SEs that reflect the estimates' true sampling variabilities. Results suggest that estimates behave well in that they generally have small or negligible bias; their true sampling variabilities can likewise be captured reasonably well by their SEs obtained by conventional methods.

3.5 Developmental toxicity of ethylene glycol

In this section, we adopt the methodology outlined in Sections 3.2 and 3.3 to analyze the EG data described in Section 3.1. The primary interest lies in the dose-response relationships for the outcomes fetal weight Y_{ih} and fetal malformation X_{ih} (equal to 1 if malformation is present, and 0, otherwise) for $h = 1, \cdots, n_i \leq 16$, and $i = 1, \cdots, N = 94$. We first adopt a logit-normal Gaussian copula mixed model, where $z_i = 0, 0.75, 1.5, 3$g/kg/day, is the dose level. For ease of interpretation and to allow the marginal model for X_{ih} to have a logistic structure, we assume a bridge distribution for $\bar{B}_i^*$ as follows:

$$f_{\bar{B}_i^*}(\bar{b}_i^*) = \frac{\sin(\varphi\pi)}{2\pi\{\cosh(\varphi\bar{b}_i^*) + \cos(\varphi\pi)\}}, \quad -\infty < \bar{b}_i^* < +\infty, \tag{3.15}$$

where $\varphi \in (0,1)$ is the attenuation parameter. A discussion of the bridge distribution and its nice marginalization properties is given in Lin *et al.* (2010). Note that with $\bar{B}_i^* \sim bridge(\varphi)$ and a logistic mixed model for X_{ih} (i.e., a logistic latent distribution for Y_{ih}^*, given B_i^*), the marginal regression model for X_{ih} has a logistic structure, i.e.,

$$\text{logit}\{P(X_{ih}=1)\} \;=\; \log\left\{\frac{P(X_{ih}=1)}{P(X_{ih}=0)}\right\} \;=\; \varphi(\alpha_1+\alpha_2 z_i).$$

As in Lin *et al.* (2010), with $B_i \sim normal(0,1)$, a joint density for $\bar{B}_i^*$ and $\bar{B}_i$ is constructed from (3.4).

The above model (Model 1) is fit using PROC NLMIXED in SAS and the results are displayed in Table 3.2. The positive estimated dose effect for fetal malformation indicates that fetal malformation rate increases with increasing dose, while the negative estimated dose effect on fetal weight suggests that mean fetal weight decreases with increasing dose. The dose coefficients for fetal malformation and weight are both statistically significant.

LOESS-smoothed residual plots of the model in Figure 3.4 exhibit slight quadratic trends for dose for both outcomes. Results of a re-analysis of the model with an additional term for quadratic dose effect (Model 2) are also displayed in Table 3.2. The estimated quadratic dose coefficients for fetal malformation and fetal weight are both significant, and the LOESS-smoothed residual plots displayed in Figure 3.4 seem to bear this out. The absence of any apparent trend suggests that including the quadratic dose effect has improved the fit of the

Table 3.1: Summary statistics for fetal malformation and fetal weight outcomes for EG mice data.

Dose (g/kg)	Number of dams	Number of live fetuses	Malformations		Weight (g)	
			Number	Percent	Mean	SD
0	25	297	1	0.3	0.972	0.098
0.75	24	276	26	9.42	0.877	0.104
1.5	22	229	89	38.86	0.764	0.107
3	23	226	129	57.08	0.704	0.124

Table 3.2: Estimates, their SEs and t-values for EG data using the logit-normal Gaussian copula mixed model with correlated random effects (REs), with only linear dose effects (Model 1) and with both linear and quadratic dose effects (Model 2), and using the robit-t Gaussian copula mixed model with correlated REs, with only linear dose effects (Model 3) and with both linear and quadratic dose effects (Model 4).

		Model 1			Model 2			Model 3			Model 4		
		Est	SE	t	Est	SE	t	Est	SE	t	Est	SE	t
Malformation													
Random intercept	λ^*	2.010	0.206	7.72	1.644	0.231	7.12	1.112	0.177	6.28	0.995	0.502	1.98
Intercept	α_1	–4.265	0.399	–10.94	–5.85	0.695	–8.42	–2.898	0.406	–7.14	–4.918	3.751	–1.31
Dose	α_2	1.834	0.205	8.97	5.014	0.908	5.52	1.147	0.176	6.57	4.261	3.671	1.16
Dose^2	α_3	—	—	—	–1.004	0.245	–4.10	—	—	—	–0.843	0.821	–1.03
Degrees of freedom	ν^*	—	—	—	—	—	—	5.815	3.079	1.89	3.275	5.448	0.60
Fetal weight													
Random intercept	λ	0.094	0.008	11.65	0.081	0.006	13.62	0.101	0.007	14.06	0.075	0.004	18.06
Intercept	β_1	0.949	0.013	71.35	0.979	0.015	66.07	0.956	0.010	97.81	0.983	0.012	80.97
Dose	β_2	–0.090	0.008	–12.02	–0.167	0.026	–6.42	–0.082	0.006	–13.80	–0.164	0.024	–6.81
Dose^2	β_3	—	—	—	0.028	0.008	3.45	—	—	—	0.025	0.008	3.24
Degrees of freedom	ν	—	—	—	—	—	—	5.536	0.953	5.81	5.439	0.927	5.87
SD	σ	0.075	0.002	43.13	0.075	0.002	43.09	0.059	0.002	25.29	0.059	0.002	25.34
Associations													
Normal correlation for REs	$\widetilde{\eta}$	–0.722	0.057	–12.68	–0.588	0.077	–7.64	–0.810	0.033	–24.68	–0.521	0.101	–5.17
Correlation for REs	η	–0.693	—	—	–0.559	—	—	—	—	—	—	—	—
Conditional biserial normal correlation	$\widetilde{\rho}$	–0.208	0.055	–3.80	–0.214	0.055	–3.88	–0.221	0.055	–4.00	–0.228	0.055	–4.12
Conditional biserial correlation	ρ	–0.174	—	—	–0.177	—	—	–0.194	—	—	–0.191	—	—

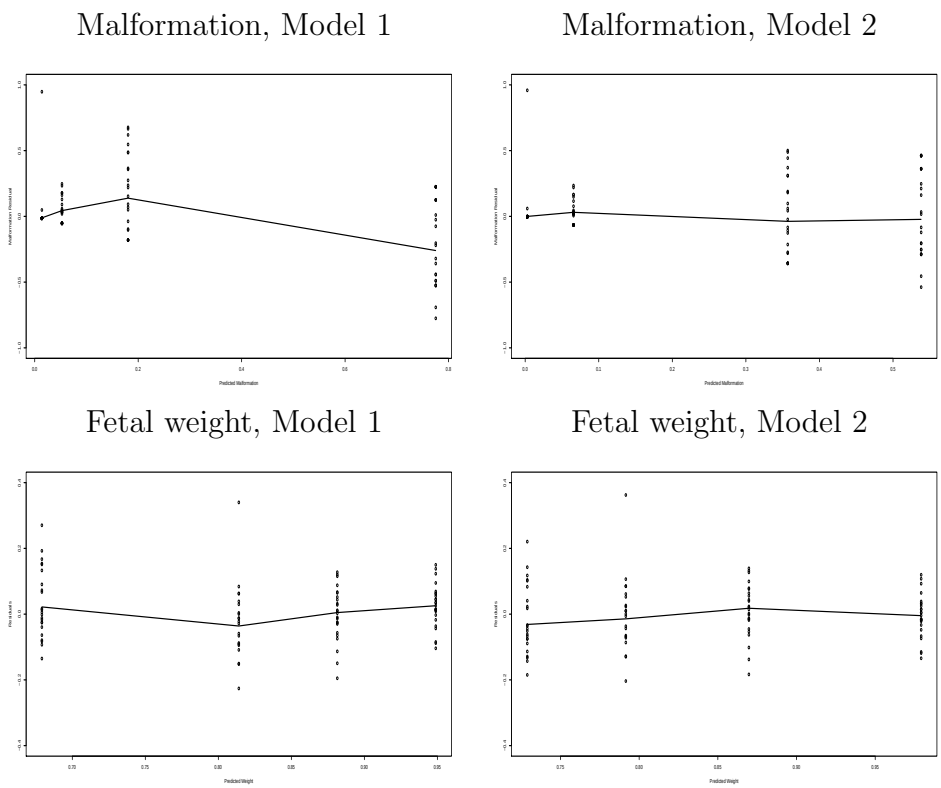

Figure 3.4: LOESS-smoothed marginal residual plots for EG mice data using the logit-normal Gaussian copula mixed model with correlated random effects.

model. An estimate of φ can be obtained from $\widehat{\varphi} = (3\widehat{\lambda}^*/\pi^2 + 1)^{-1/2}$ as 0.788 for Model 1 and 0.816 for Model 2. These estimates are somewhat greater than that reported in Lin *et al.* (2010), indicating moderate attenuation of dose effects due to cluster heterogeneity.

We next consider a joint model with $Y^*_{ih}|B^*_i \sim t_{\nu^*}(\mu^*_i(B^*_i) = \alpha_1 + \alpha_2 z_i + \lambda^* B^*_i, 1)$ and $Y_{ih}|B_i \sim t_\nu(\mu_i(B_i) = \beta_1 + \beta_2 z_i + \lambda B_i, \sigma^2)$; we call this a robit-$t$ Gaussian copula mixed model with correlated random effects. By properties of the t-distribution, note that the logit-normal mixed models (Models 1 and 2) considered earlier are special cases of this more robust model (i.e., $\nu^* \approx 7$ implies $t_{\nu^*} \approx$ *logistic* and $\nu \to \infty$ means $t_\nu \approx$ *normal*). Assuming correlated outcome-specific random effects $B^*_i \sim$ normal(0,1) and $B_i \sim$ normal(0,1) (i.e., B^*_i and B_i are jointly standard bivariate normally distributed), we fit this model first with only a linear dose effect (Model 3), and next with both linear and quadratic dose effects (Model 4), based

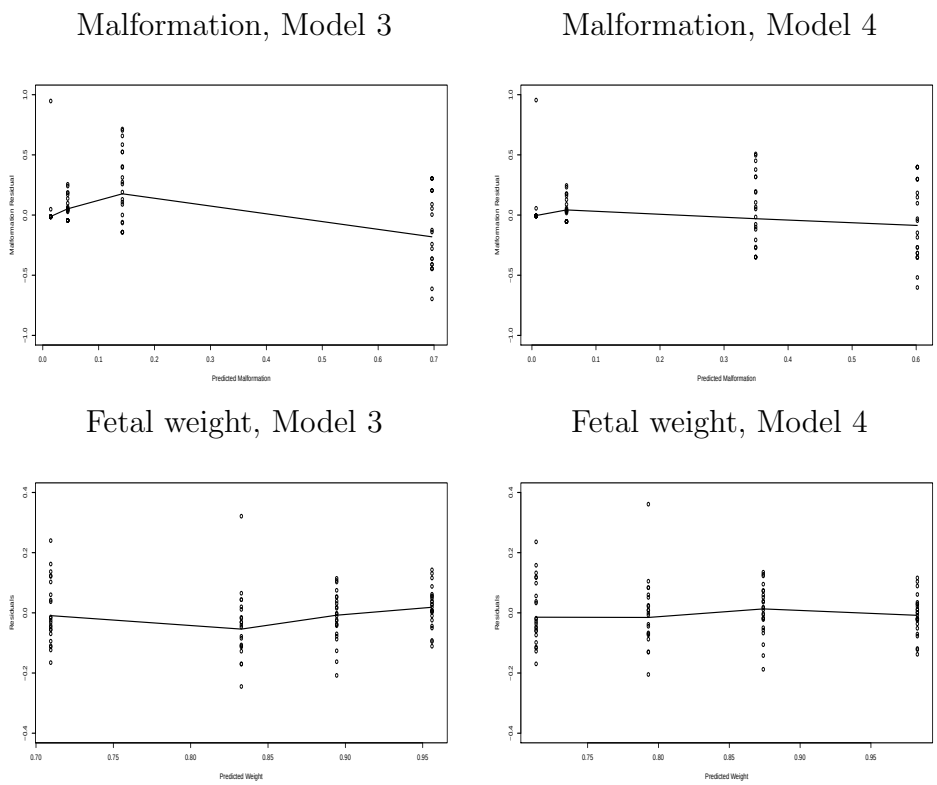

Figure 3.5: LOESS-smoothed marginal residual plots for EG mice data using the robit-t Gaussian copula mixed models with correlated random effects.

on the slight curvilinear pattern in the LOESS-smoothed residual plots for Model 3 in Figure 3.5. Results for Models 3 and 4 are also displayed in Table 3.2. They were obtained using the profile likelihood method implemented via PROC NLMIXED, with importance sampling as the numerical integration technique.

Using piecewise linear approximations (Kugiumtzis and Bora-Senta, 2010), the respective estimates for Models 1 to 4 of the (conditional) correlation $\widetilde{\rho}$ between normal scores (based on Y_{ih}^* and Y_{ih}) yield the corresponding estimates in Table 3.2 of the (conditional) biserial correlation ρ between fetal malformation and weight. They suggest that the two outcomes are not conditionally independent given the random effects, and are, in fact, slightly conditionally negatively associated. Estimates of $\widetilde{\eta}$, the correlation between normal scores based on random effects B_i^* and B_i, are also shown. Again using piecewise linear approximations, they give

estimates of the correlation η between the random effects for Models 1 and 2, as shown in Table 3.2. Note that since Models 3 and 4 have Gaussian random effects, we have $\widetilde{\eta} = \eta$, and no approximation is needed.

Estimates of marginal correlations are easily calculated from Section 3.2.2. For example, the intra-litter tetrachoric correlation between fetal malformation for different fetuses is 0.014 for Model 1 while the corresponding intra-litter correlation between fetal weights is 0.614; the intra-litter biserial correlation between fetal malformation and weight for the same fetus is -0.163, suggesting a negative association between fetal malformation and fetal weight. This is stronger than -0.056, that for different fetuses. The results reported here are consistent with earlier analyses reported in, among others, Gueorguieva and Agresti (2001), Fitzmaurice and Laird (1995), and Catalano and Ryan (1992). Our estimates are especially comparable to those reported in Lin *et al.* (2010), which similarly use a bridge margin for the fetal malformation random effect.

We also calculated the AIC (Akaike's information criterion) and BIC (Bayesian information criterion) values (not reported), along with the likelihood ratio statistic (LR $= -2 \times$ log-likelihood), for all four models. Models 2 and 4 yielded smaller LR, AIC and BIC values than Models 1 and 3, respectively; this indicates that models incorporating a quadratic term for dose provided better fits than those with only a linear dose effect. In addition, the robit-t mixed models yielded smaller LR, AIC, and BIC than the logit-normal mixed models, which implies that the former fit the EG data better than do the latter.

3.5.1 Conditional plots

For Models 1 to 4, we plot in Figure 3.6 the estimated conditional malformation probability (see (3.10) in Section 3.2.3) as a function of dose level $z_i = 0, 0.75, 1.5, 3$g/kg, at each fixed fetal weight $y_{ih} = 0.25, 0.5, 0.75, 1$g, and the conditional mean fetal weight (see (3.11) in Section 3.2.3) as a function of dose, given malformation indicator $x_{ih} = 0$ or $x_{ih} = 1$. For comparisons, we include in Figures 3.6a and 3.6b and Figures 3.6d and 3.6e, respectively, the

marginal probability of malformation and marginal mean fetal weight, as functions of dose. Figures 3.6c and 3.6f plot the estimated conditional malformation probability as a function of fetal weight at each fixed dose level $z_i = 0, 0.75, 1.5, 3$g/kg. Similar plots for models with the additional quadratic terms are shown in Figure 3.7. It is clear that the malformation rate decreases with increasing weight, confirming the negative association between the two outcomes.

3.5.2 Quantitative risk assessment

As mentioned previously, our model possesses property (3), which directly accounts for the intra-litter association between fetal malformation and fetal weight. This property is very useful in quantitative risk assessment, where such association is needed. Let the benchmark dose BMD_q be the dose corresponding to a specified level $q = 1, 5, 10\%$, say, of increased response above background risk (see definition below). Given a dose-response model, BMD_q is used for determining acceptable low-risk exposure levels for humans, along with 95% lower confidence limit LED_q (known as the lower effective dose), which accounts for estimation variability.

In estimating the overall risk to a fetus of exposure to EG at dose z, let $P(z)$ be the probability that a fetus is malformed or has "low" birth weight, where "low" in this case is specified as fetal weight smaller than 2 SDs below the average fetal weight at control dose 0g (i.e., $0.972 - 2 \times 0.098 = 0.776$g, see Geys *et al.*, 2001). That is, we obtain $P(z)$ for fetus i in litter j from (3.5) as

$$P(z) = 1 - \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Phi_2 \left(\Phi^{-1}\{u_{ih}^*(b_i^*)\}, \frac{\mu_i(b_i) - 0.776}{\sigma}; -\widetilde{\rho} \right) f_{B_i^*, B_i}(b_i^*, b_i) db_i^* db_i. \quad (3.16)$$

In determining BMD_q, the focus is on either the additional risk $r_a(z) = P(z) - P(0)$ or the extra risk $r_e(z) = r_a(z)/\{1 - P(0)\}$, where $P(0)$ is the so-called background risk (i.e., the risk at dose $z = 0$). The benchmark dose BMD_q is then defined as the estimated dose such that $r_e(\widehat{\text{BMD}_q}) = q$ (or alternatively, $r_a(\widehat{\text{BMD}_q}) = q$), where $\widehat{r_e(z)}$ is the plug-in estimate of

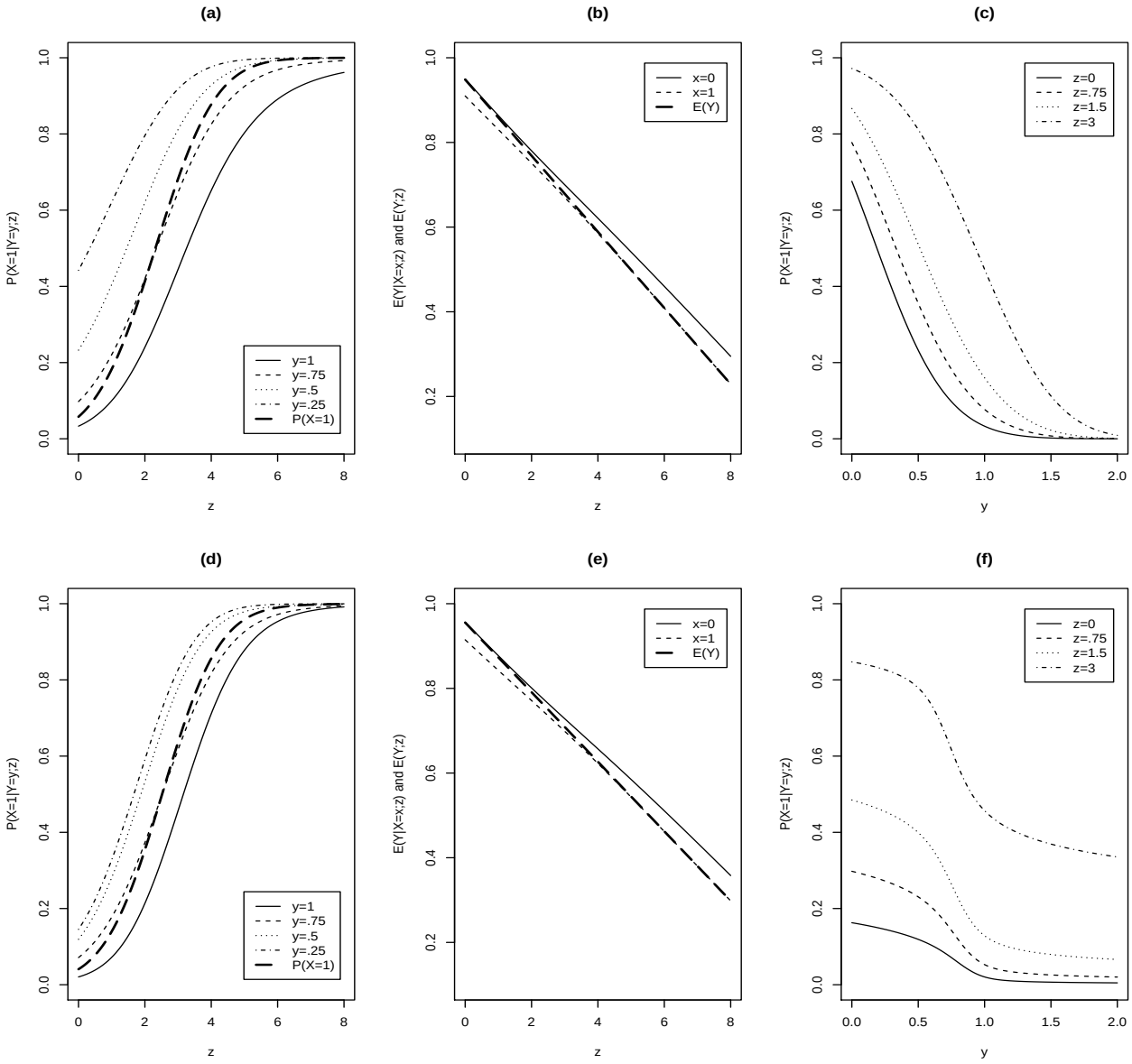

Figure 3.6: Plots of (a and d) $P(X_{ih} = 1|Y_{ih} = y)$ as function of dose z; (b and f) $E(Y_{ih}|X_{ih} = x)$ as function of dose z; and (c and g) $P(X_{ih} = 1|Y_{ih} = y)$ as function of y, for fixed dose z; $E(Y_{ih})$ and $P(X_{ih} = 1)$, as functions of dose z, are shown for comparison. Top panel obtained using Model 1, bottom panel using Model 3.

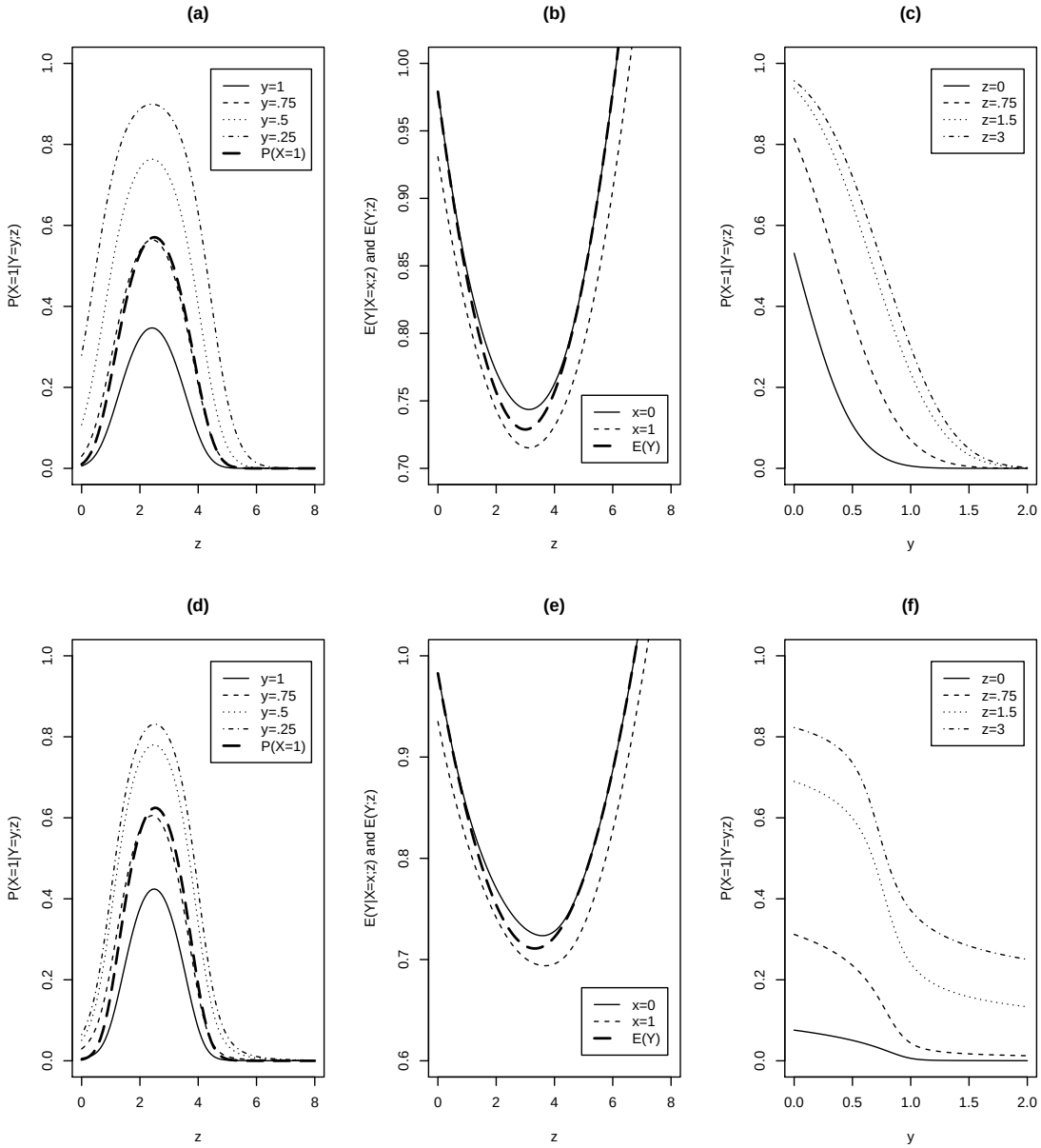

Figure 3.7: Plots of (a and d) $P(X_{ih} = 1|Y_{ih} = y)$ as function of dose z; (b and f) $E(Y_{ih}|X_{ih} = x)$ as function of dose z; and (c and g) $P(X_{ih} = 1|Y_{ih} = y)$ as function of y, for fixed dose z; $E(Y_{ih})$ and $P(X_{ih} = 1)$, as functions of dose z, are shown for comparison. Top panel obtained using Model 2, bottom panel using Model 4.

Table 3.3: Estimates of overall risk probability $P(z)$, additional risk $r_a(z)$, and extra risk $r_e(z)$, for dose $z = 0, 0.75, 1.5, 3$g/kg, based on Models 1 to 4.

z	Model 1			Model 2			Model 3			Model 4		
	$P(z)$	$r_a(z)$	$r_e(z)$	$P(z)$	$r_a(z)$	$r_e(z)$	$P(z)$	$r_a(z)$	$r_e(z)$	$P(z)$	$r_a(z)$	$r_e(z)$
0	0.099	0	0	0.037	0	0	0.099	0	0	0.034	0	0
0.75	0.237	0.138	0.153	0.243	0.206	0.214	0.222	0.122	0.136	0.229	0.129	0.144
1.5	0.421	0.322	0.357	0.518	0.480	0.499	0.409	0.309	0.343	0.599	0.499	0.554
3	0.674	0.575	0.638	0.653	0.616	0.640	0.805	0.705	0.783	0.832	0.732	0.813

$r_e(z)$, for some dose z. In addition, the lower effective dose LED_q is defined as the estimated dose such that $q = \widehat{r_e(\text{LED}_q)} + 1.645 \times \text{SE}\{\widehat{r_e(\text{LED}_q)}\}$. The standard error $\text{SE}\{\widehat{r_e(z)}\}$, for some dose z, is usually obtained by delta method (Regan and Catalano, 1999) via the plug-in estimate $\widehat{P(z)}$; however, because (3.16) is a complicated function of $\boldsymbol{\Theta}$, we instead employ the jackknife method (Lipsitz *et al.*, 1994) to obtain the estimated variance of $\widehat{r_e(z)}$.

Results of the quantitative risk assessment for EG mice data using Models 1 to 4 are displayed in Tables 3.2 and 3.3. We can see that the benchmark dose corresponding to a 10% extra risk over the background ranged from a minimum of $\text{BMD}_{0.1} = 0.413$g/kg for Model 4 to a maximum of $\text{BMD}_{0.1} = 0.589$g/kg for Model 3. Corresponding results for a 10% additional risk are quite similar since the estimated background risk $\widehat{P(0)}$ is small for all models. Note that these are somewhat higher than those reported in Regan and Catalano (1999); this is not surprising in light of our estimates of $P(z)$, $z = 0, 0.75, 1.5, 3$g/kg, shown in Table 3.3, which are generally smaller than those reported in Regan and Catalano (1999), perhaps because unlike in Regan and Catalano (1999), we did not consider embryolethality in the calculation of $P(z)$.

3.6 Discussion

In this chapter, we adapted the Gaussian copula mixed model first introduced in Chapter 2 to the analysis clustered mixed binary and continuous outcomes, with particular application in developmental toxicology. The resulting joint model possesses three ideal properties required in the analysis of data from developmental toxicity studies. First, the model yields interpretable outcome-specific dose-response models for the mixed outcomes; this is accomplished by using the Gaussian copula to couple together the models for the mixed outcomes and account for dependence between them. The copula approach enables the specification of the outcome-specific models independently of the dependence structure, and permits non-normal residual errors. Second, the model accounts for cluster effects by the introduction of cluster-

Table 3.4: Benchmark dose BMD_q and lower effective dose LED_q for $100q = 1, 5, 10\%$, based on extra risk, estimated using Models 1 to 4. Note that $\mathrm{SE} = \mathrm{SE}\{r_e\widehat{(\mathrm{LED}_q)}\}$.

$100q\%$	Model 1			Model 2			Model 3			Model 4		
	BMD_q	SE	LED_q	BMD_q	SE	LED_q	BMD_q	SE	LED_q	BMD_q	SE	LED_q
1	0.065	0.002	0.043	0.068	0.003	0.040	0.074	0	0.010	0	0	0.001
5	0.291	0.002	0.271	0.266	0.003	0.247	0.329	0.015	0.174	0.184	0.001	0.170
10	0.529	0.003	0.512	0.443	0.003	0.427	0.589	0.017	0.453	0.413	0.002	0.402

level random effects whose distribution (possibly non-normal) may be flexibly constructed via the Gaussian copula. Last, the model directly incorporates within-subject associations between the mixed outcomes, a necessary ingredient in quantitative risk assessment. The model adopts a latent variable description of the binary outcome, a quite appealing formulation of such binary outcomes as fetal malformation in developmental toxicology. As discussed in Chapter 2, this also makes statistical sense, as common binary regression models (e.g., logistic, probit) have parallel formulations in terms of latent variables (Teixeira-Pinto and Normand, 2009); in addition, it also leads to familiar association measures for capturing dependence between two binary outcomes (i.e., tetrachoric correlation) or between mixed binary and continuous outcomes (i.e., biserial correlation).

As an illustration of our methodology, we applied the robit-normal mixed models with shared random effects, the logit-normal Gaussian copula mixed models with non-normal correlated random effects, and the robit-t Gaussian copula mixed models with normal correlated random effects to the EG mice data. The models we considered extend and generalize previous ones adopted in Lin *et al.* (2010), Gueorguieva and Agresti (2001), and Regan and Catalano (1999), among others. Although estimates are not directly comparable with those from previous analyses, estimates for dose effects for fetal weight and malformation rate can be qualitatively compared with those from Gueorguieva and Agresti (2001) and Lin *et al.* (2010), with the former being a special case of our model, and the latter being akin to ours in its use of a Gaussian copula to construct the random effects distribution, with a marginally bridge-distributed random effect for fetal malformation. While all three models produced similar results, ours is, on the one hand, a more general formulation of the correlated probit model of Gueorguieva and Agresti (2001) that allows for non-normal residual errors and non-normal random effects; Lin *et al.* (2010), on the other hand, is a factorization model and as such, suffers from the inadequacies of factorization models (e.g., non-invariance to direction of factorization, asymmetrical treatment of outcomes).

One drawback is our exclusive use of the Gaussian copula. However, as pointed out in Chapter 2, our methodology is sufficiently general to accommodate other copulas (e.g., Clayton, t). The optimal choice of the copula can then be investigated using likelihood-based model selection criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC).

Finally, our methodology can be easily implemented using standard statistical software/packages; in particular, we used PROC NLMIXED in SAS in our simulations and analysis.

Chapter 4

Gaussian copula joint models for mixed outcomes with "continued" binary variables

4.1 Introduction

Chapters 2 and 3 developed a copula-based approach to joint analysis of mixed discrete and continuous data that utilized a latent variable framework for describing the discrete variables. The resulting Gaussian copula joint model is akin to and can be viewed as a generalization of the Gaussian distribution-based CGCM (see Chapter 1) to applications involving non-Gaussian latent variables (e.g., t-latent variables) and non-Gaussian continuous outcomes (e.g., gamma distributed time-to-event outcomes). The latter, which yields probit models for the discrete data, has been in the literature for some time and has been used in several applications in psychometrics (Muthèn, 1984) and econometrics (Hausman and Wise, 1978), among others.

Although a latent variable description of discrete variables is practically appealing in many applications, as in the wages-hours-union memberships example in Chapter 2 and the developmental toxicity study considered in Chapter 3, such an approach does not work in all cases. Count data, for example, present a situation for which a latent variable framework may not be appropriate. This is also the case for nominally scaled categorical outcomes, such as gender and hair colour, for example. It is thus imperative that an alternative approach be developed that does not necessitate the use of latent variables for the discrete data.

The present chapter aims to develop a parallel methodology to the one based on latent variables developed previously in Chapters 2 and 3. The approach we adopt in this chapter makes use of the "continuous-ation" or "jittering" method first proposed by Machado and

Santos Silva (2005) and later formalized by Denuit and Lambert (2005). The method entails transforming the discrete data into continuous, hence the name, by the addition of independently generated continuous random variables, called "jitters". It has been recently applied in the analysis of correlated count data in economics (Heinen and Rengifo, 2007; Heinen and Rengifo, 2008), of spatially dependent discrete data in ecology (Madsen, 2009), and of longitudinal insurance claim counts in actuarial studies (Shi and Valdez, 2012).

A major motivation for the jittering method in copula modeling is the failure of Sklar's Theorem, mainly due to the non-uniformity of the PIT, to guarantee the uniqueness of the copula-generated distribution when discrete margins are directly used for the copula. This "unidentifiability" issue was recently addressed by Genest and Nešlehovà (2007), who showed that the distribution is unique only in the product of the ranges of the individual discrete margins, and multiple copula representations may exist for the same distribution; in fact, there may be infinitely many of them for discrete variables with only a few mass points, as in the case of binary data.

While an important theoretical issue, the non-uniqueness of the copula representation when discrete margins are used may be of minor consequence in practical applications. This is because the resulting copula-generated model is still a valid and proper distribution, which perhaps explains the large number of applications of copula models to discrete data in the literature. See, for example, Li and Wong (2011), Nikoloulopoulos and Karlis (2010,2009,2008), Trégouët *et al.* (1999), and Meester and MacKay (1994), among others. However, as Genest and Nešlehovà (2007) documented in their paper, the non-uniqueness issue has serious consequences on the use of copulas as far as dependence modeling is concerned, which is the main rationale for copula models in the first place. For one, the copula fails to completely characterize dependence among the discrete variables, unlike in the continuous case (e.g., see Examples 5 and 6 of Genest and Nešlehovà, 2007). For another, dependence measures such as Kendall's tau and Spearman's rho are no longer margin-free (Nešlehovà, 2007; Mes-

fioui and Tajar, 2005). This latter drawback is particularly detrimental in practice, as the margin-free nature of copula-based dependence models — which allows for the specification of marginal models independently of the dependence structure — is the cornerstone of copula modeling.

We briefly review in Section 4.2 the jittering method and adapt the Gaussian copula joint model in Chapter 2 to mixed binary and continuous data, where the binary variables have been "continued" or jittered. The various associations between the outcomes arising from the copula model are then briefly discussed in Section 4.3. Section 4.4 focuses on likelihood estimation for our model while Section 4.5 reports simulation results concerning the empirical performance of MLEs for the model. In Section 4.6, data on burn injuries, previously considered by de Leon and Wu (2011), are used to illustrate our methodology. Finally, Section 4.7 concludes the chapter.

4.2 Gaussian copula joint model with jittered binary outcomes

Assume X is a discrete random variable that takes on non-negative integers as mass points. Denote the density of X by $p_x = P(X = x) > 0$, where x is a mass point, so that its CDF is $F_X(x) = \sum_{x' \leq x} p_{x'} \geq 0$. Following Denuit and Lambert (2005), define

$$Y^* \;=\; X + (V^* - 1), \tag{4.1}$$

where V^* is a continuous random variable, independent of X, and with a known strictly increasing CDF $F_{V^*}(\cdot)$ sharing no parameters with the distribution of X. We say that X is "continued" by V^*, and Y^* is a "continued" discrete random variable.

Let $F_{Y^*}(\cdot)$ and $F_{V^*}(\cdot)$ be the CDFs of Y^* and V^*, respectively, and $[\cdot]$ be the greatest integer function (i.e., $[w]$ is equal to the greatest integer less than or equal to $w \in \mathbb{R}$). Then $\forall y^* \in \mathbb{R}$, we get the CDF $F_{Y^*}(\cdot)$ of Y^* as

$$F_{Y^*}(y^*) \;=\; F_X([y^*]) + F_{V^*}(y^* - [y^*])p_{[y^*+1]}. \tag{4.2}$$

The corresponding density of Y^* is then $f_{Y^*}(y^*) = f_{V^*}(y^* - [y^*])p_{[y^*+1]}$, where $f_{Y^*}(\cdot)$ and $f_{V^*}(\cdot)$ are the corresponding densities of Y^* and V^*, respectively. One natural and simple choice of V^* that satisfies all constraints on $F_{V^*}(\cdot)$ is a uniform random variable on $(0, 1)$. We assume this is the case in what follows.

We now adapt the above development to the case of binary data. Consider a binary outcome X_i for subject $i = 1, \cdots, N$, where $X_i = 1$ indicates "success" and 0, "failure". Let $p_i = P(X_i = 1) = E(X_i)$ be the "success" probability for subject i. Then, the "continued" binary random variable Y_i^* is then given by (4.1). For $V_i^* \sim \textit{uniform}(0, 1)$, we get the CDF $F_{Y_i^*}(\cdot)$ and density $f_{Y_i^*}(\cdot)$ of Y^* from (4.2) as

$$F_{Y_i^*}(y_i^*) = \begin{cases} 0 & \text{, if } y_i^* < -1 \\ (1-p_i)(y_i^*+1) & \text{, if } -1 \leq y_i^* < 0 \\ 1 + p_i(y_i^* - 1) & \text{, if } 0 \leq y_i^* < 1 \\ 1 & \text{, if } y_i^* \geq 1 \end{cases}, \tag{4.3}$$

$$f_{Y_i^*}(y_i^*) = \begin{cases} 1 - p_i & \text{, if } -1 < y_i^* < 0 \\ p_i & \text{, if } 0 \leq y_i^* < 1 \\ 0 & \text{, elsewhere} \end{cases}. \tag{4.4}$$

We can now construct a joint model for bivariate mixed data comprising a binary outcome X_i and a continuous outcome Y_i, where Y_i^* is treated as a proxy for X_i. By using (4.3) as a margin for the Gaussian copula, we can then model the joint CDF of the "continued" binary outcome Y_i^* and the continuous outcome $Y_i \sim F_{Y_i}(\cdot)$ via the Gaussian copula as follows

$$F_{Y_i^*, Y_i}(y_i^*, y) = \Phi_2\left(\Phi^{-1}(u_i), \Phi^{-1}(v_i); \widetilde{\rho}\right), \tag{4.5}$$

where $\widetilde{\rho}$ is the correlation between the normal scores $\Phi^{-1}(U_i) = \Phi^{-1}(F_{Y_i^*}(Y_i^*))$ and $\Phi^{-1}(V_i) = \Phi^{-1}(F_{Y_i}(Y_i))$. The corresponding joint density $f_{Y_i^*, Y_i}(\cdot)$ is then given by

$$f_{Y_i^*, Y_i}(y_i^*, y_i) = \frac{\phi_2\left(\Phi^{-1}(u_i), \Phi^{-1}(v_i); \widetilde{\rho}\right)}{\phi(\Phi^{-1}(u_i))\phi(\Phi^{-1}(v_i))} f_{Y_i^*}(y_i^*) f_{Y_i}(y_i), \tag{4.6}$$

where $f_{Y_i}(\cdot)$ is the density of Y_i. With outcome specific covariate vectors $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$, the outcome-specific regression models $E(X_i) = p_i = \mu_{1i}(\mathbf{z}_{1i}, \boldsymbol{\alpha})$ and $E(Y_i) = \mu_{2i}(\mathbf{z}_{2i}, \boldsymbol{\beta})$ can then be embedded in the above joint model, where $\mu_{1i}(\cdot)$ and $\mu_{2i}(\cdot)$ are link functions specifying the relationships between the outcomes and the corresponding covariates. For example, $\mu_{1i}(\cdot)$ can be the logit or probit link while $\mu_{2i}(\cdot)$ can be the identity link for a Gaussian outcome or the log link for a gamma-distributed outcome.

The Gaussian copula joint model in (4.5) and (4.6) for Y_i^* and Y_i can be used to obtain a joint model for X_i and Y_i, since X_i can be recovered from Y_i^* via $X_i = [Y_i^* + 1]$. This joint model is an alternative to that in (2.2) and (2.4) which is not based on a latent variable description of X_i. Note that since Y_i^* is continuous, the use of the Gaussian copula in (4.5) and (4.6) does not engender the problems discussed in Section 4.1 arising from the use of discrete margins in copula models. For example, the Gaussian copula representation in (4.5) is still unique by Sklar's Theorem. In addition, the dependence between Y_i^* and Y_i (hence, between X_i and Y_i) is completely characterized by (4.5) via $\widetilde{\rho}$, which is still margin-free (see (1.5) in Chapter 1). Finally, we note that while the Gaussian copula joint model in (4.5) and (4.6) was developed for the bivariate case, it is straightforward to extend it to the multivariate setting with multiple binary outcomes X_{ij}, $j = 1, \cdots, Q$, and multiple continuous outcomes Y_{ik}, $k = 1, \cdots, C$.

4.3 Associations

The major justification for adopting copula models in practice is that they allow the specification of the dependence among the outcomes independently of the outcomes' marginal models. However, the margin-dependent dependence model arising from the use of discrete margins in copulas makes copula models unattractive in applications involving discrete data. Fortunately, Denuit and Lambert (2005) proved that the jittering method is able to preserve the concordance order — a nonparametric measure of dependence — between a pair of

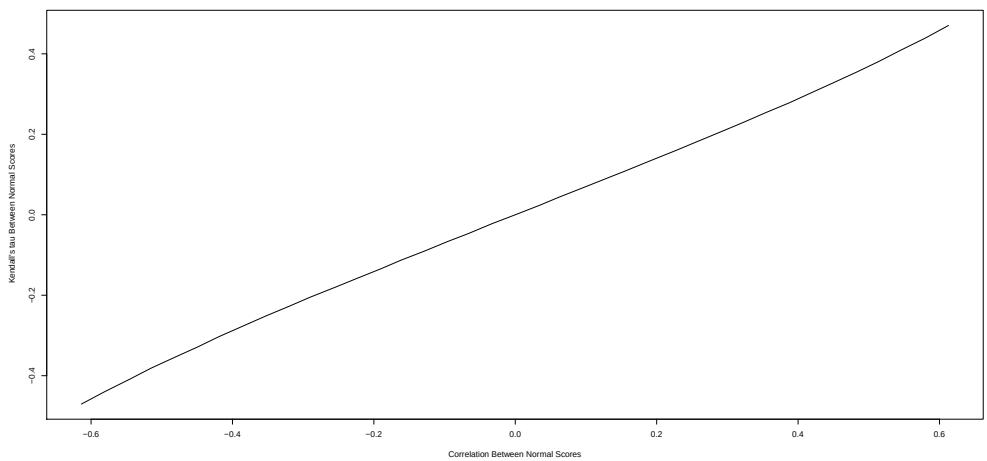

Figure 4.1: Relationship between normal correlation $\widetilde{\rho} = corr\{\Phi^{-1}(F_{Y_i^*}(Y_i^*)), \Phi^{-1}(F_{Y_i}(Y_i))\}$ vs. the corresponding Kendall's tau $\tau\{\Phi^{-1}(F_{Y_i^*}(Y_i^*)), \Phi^{-1}(F_{Y_i}(Y_i))\} = \tau(Y_i^*, Y_i)$.

discrete variables and the corresponding pair of their "continued" versions. Specifically, if $\tau(X_{ij}, X_{ij'})$ is the Kendall's tau between the jth and j'th binary outcomes for subject i, then $\tau(X_{ij}, X_{ij'}) = \tau(Y_{ij}^*, Y_{ij'}^*) = \tau\{\Phi^{-1}(F_{Y_{ij}^*}(Y_{ij}^*)), \Phi^{-1}(F_{Y_{ij'}^*}(Y_{ij'}^*))\}$, where Y_{ij}^* and $Y_{ij'}^*$ are the respective "continued" versions of X_{ij} and $X_{ij'}$. It is straightforward to prove that this also holds for mixed binary and continuous outcomes. For example, in the bivariate case considered in Section 4.2, we have $\tau(X_i, Y_i) = \tau(Y_i^*, Y_i) = \tau\{\Phi^{-1}(F_{Y_i^*}(Y_i^*)), \Phi^{-1}(F_{Y_i}(Y_i))\}$. Note that the equality of the Kendall's tau between the "continued" binary response and the continuous outcome, and that between their normal scores follows from the invariance of Kendall's tau under monotonic transformations.

In the case of the Gaussian copula joint model in (4.5) and (4.6), the relationship between Kendall's tau and the normal correlation $\widetilde{\rho}$ is well-known and is given by

$$\tau\{\Phi^{-1}(F_{Y_i^*}(Y_i^*)), \Phi^{-1}(F_{Y_i}(Y_i))\} \;=\; \tau(Y_i^*, Y_i) \;=\; \frac{2}{\pi}\arcsin(\widetilde{\rho}). \tag{4.7}$$

Figure 4.1 shows the relationship (4.7), which is monotonically increasing indicating that one or the other can be used to gauge the dependence between Y_i^* and Y_i.

The point-biserial correlation $corr(X_i, Y_i)$ between X_i and Y_i is obtained using (4.1) and

the assumed independence of X_i and the jitter V_i^*. This gives

$$corr(X_i, Y_i) \;=\; corr(Y_i^*, Y_i)\sqrt{1 + \frac{var(V_i^*)}{var(X_i)}} \;=\; \rho\sqrt{1 + \frac{1}{12p_i(1-p_i)}}, \tag{4.8}$$

where the last equality follows from the fact that $var(V_i^*) = 1/12$ (i.e., the variance of a *uniform*$(0,1)$ random variable) and $var(X_i) = p_i(1-p_i)$ (since $X_i \sim$ *Bernoulli*(p_i)). Note that $\rho = corr(Y_i^*, Y_i)$ can be calculated from $\widetilde{\rho}$ using piece-wise linear approximations (Kugiumtzis and Bora-Senta, 2010). Because ρ is margin-free and since $var(X_i)$ can be obtained independently of the marginal model for X_i, $corr(X_i, Y_i)$ is also margin-free. In addition, observe that $|corr(X_i, Y_i)| > |\rho|$, i.e., the correlation between the observed outcomes is attenuated by the "continuous-ation" of X_i.

4.4 Likelihood estimation

Suppose $\{x_i, y_i, \mathbf{z}_{1i}, z_{2i}\}$ denotes the observed data, for subject $i = 1, \cdots, N$, with outcome-specific covariate vectors $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$. Given randomly generated realizations $\{v_1^*, \cdots, v_N^*\}$ of the jitters $V_1, \cdots, V_N \overset{iid}{\sim}$ *uniform*$(0,1)$, we get the "continued" binary data $y_1^*, \cdots, y_N^*$. Let $\boldsymbol{\Theta}$ be the parameter vector containing the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the normal correlation $\widetilde{\rho}$, and the parameters for the continuous margins; note that $\boldsymbol{\Theta}$ is the same parameter vector for both the observed data and the pseudo-data $\{y_i^*, y_i\}$, $i = 1, \cdots, N$. The log-likelihood function based on (4.6) for the pseudo-data, is then given by

$$\ell_{\text{pseudo}}(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \left\{\log \phi_2(w_i^*, w_i; \widetilde{\rho}) + \log f_{Y_i^*}(y_i^*) + \log f_{Y_i}(y_i) - \log \phi(w_i^*) - \log \phi(w_i)\right\}, \tag{4.9}$$

where $w_i^* = \Phi^{-1}(F_{Y_i^*}(y_i^*))$ and $w_i = \Phi^{-1}(F_{Y_i}(y_i))$. Nikoloulopoulos (2013) calls log-likelihood function (4.9) a surrogate log-likelihood function for the observed outcome data. Although the MLE $\widehat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$ can be obtained from (4.9) by direct maximization, the noise introduced by the jittering may result in biased estimates. To minimize, if not eliminate, this bias, Heinen and Rengifo (2007,2008) proposed to generate M sets of pseudo-data $\{y_{im}^*, y_{im}\}$ by generating M sets of jitters $\{v_{1m}^*, \cdots, v_{Nm}^*\}$, $m = 1, \cdots, M$. For each set of pseudo-data, the

corresponding MLE $\widehat{\boldsymbol{\Theta}}_m$ is then calculated, and the final MLE $\widehat{\boldsymbol{\Theta}}$ is obtained by averaging the M estimates from the M pseudo-data sets. The large-sample SEs of the MLEs in $\widehat{\boldsymbol{\Theta}}$ are obtained by averaging the large-sample SEs for $\widehat{\boldsymbol{\Theta}}_m$ obtained from the inverse Hessian matrix.

Another approach, discussed by Madsen (2009) and Madsen and Fang (2010) (see also Madsen and Fang, 2011, and Sabo and Changanty, 2011, for further discussions), utilizes the log-likelihood $\ell_{\text{observed}}(\boldsymbol{\Theta})$ for the observed data $\{x_i, y_i\}$, $i = 1, \cdots, N$, given by

$$\ell_{\text{observed}}(\boldsymbol{\Theta}) \;=\; \log\left\{E_{V_1^*,\cdots,V_N^*}\left[\exp\{\ell_{\text{pseudo}}(\boldsymbol{\Theta})\}\right]\right\}, \tag{4.10}$$

where the expectation is taken with respect to the jitters $V_1, \cdots, V_N \overset{iid}{\sim} \textit{uniform}(0,1)$. In practice, (4.10) is approximated by the simulated log-likelihood, obtained by averaging over M pseudo-data sets obtained from M randomly generated sets of jitters, i.e.,

$$\ell_{\text{observed}}(\boldsymbol{\Theta}) \;\approx\; \log\left\{\frac{1}{M}\sum_{m=1}^{M}\exp\{\ell_{\text{pseudo},m}(\boldsymbol{\Theta})\}\right\}, \tag{4.11}$$

where $\ell_{\text{pseudo},m}(\boldsymbol{\Theta})$ is the likelihood function based on the mth pseudo-data set from the mth set of jitters. The MLE $\widehat{\boldsymbol{\Theta}}$ is obtained by maximizing $\ell_{\text{observed}}(\boldsymbol{\Theta})$ directly via some numerical optimization routine (e.g., function `optim` in R), with the large-sample SEs obtained conventionally from the inverse Hessian matrix.

For our Gaussian copula joint model (4.6), we adopt Heinen and Rengifo's (2007,2008) surrogate likelihood approach. We report the empirical performance of the resulting estimates in the next section.

4.5 Simulation study

For the simulations, we consider the Gaussian copula joint model in Section 4.2 with a Gaussian continuous outcome Y_i with $var(Y_i) = \sigma^2$, for all i, and a logit link for the binary outcome X_i, for $i = 1, \cdots, N$. Specifically, we assumed $E(X_i) = p_i = \{1 + \exp(\alpha_1 + \alpha_2 z_{1i})\}^{-1}$ and $E(Y_i) = \beta_1 + \beta_2 z_{2i}$, where the covariates z_{1i} and z_{2i} were randomly generated from

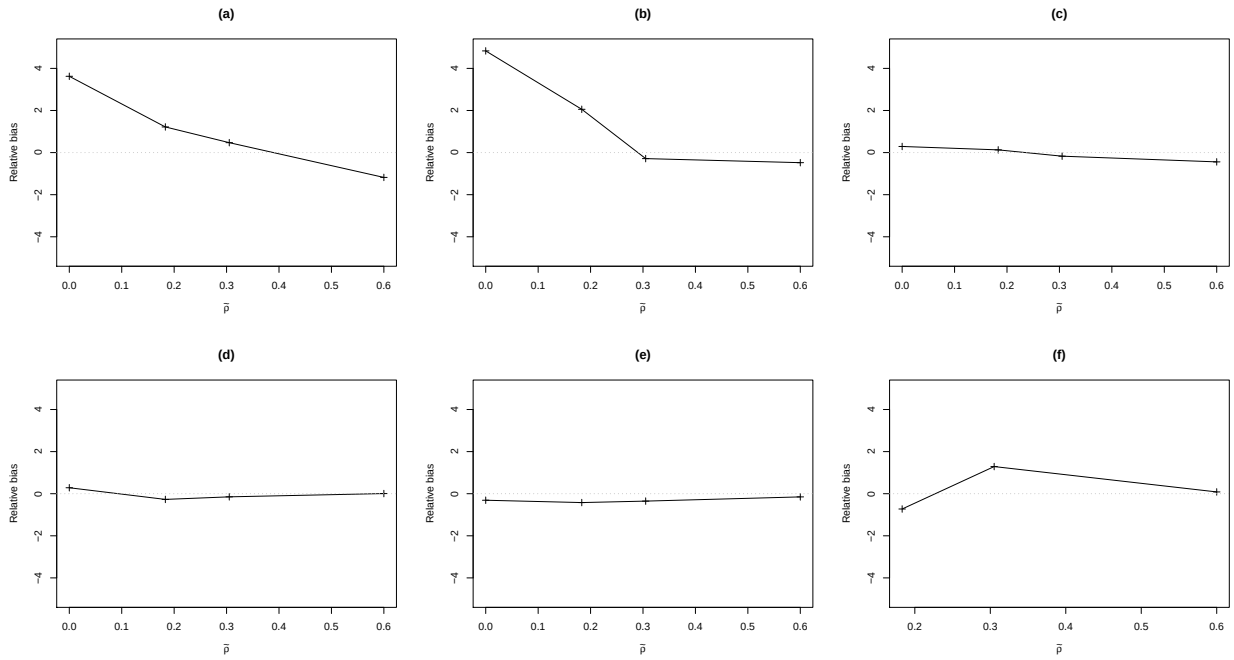

Figure 4.2: Plots (a), (b), (c), (d), (e), and (f) represent the relative bias of MLEs of α_1, α_2, β_1, β_2, σ, and $\widetilde{\rho}$, respectively, based on $R = 500$ repeats and $M = 1000$ jitters for a Gaussian copula joint model with logit link for the binary outcome.

uniform$(0, 1)$. We generated a total of $R = 500$ repeated samples, each of size $N = 200$. For each repeat, we randomly generated $M = 1000$ jitters, so that we generated $M = 1000$ pseudo-data sets. For each pseudo-data set, we used Heinen and Rengifo's (2007,2008) surrogate likelihood approach to obtain the MLEs. The $M = 1000$ estimates from the $M = 1000$ pseudo-data sets are then averaged to obtain the MLE for each repeat. The relative bias and relative efficiency are calculated based on the MLEs from the $R = 500$ repeats.

The following parameter configurations were used in the simulations: $\alpha_1 = \beta_1 = \sigma = 1$, $\alpha_2 = -1$, $\beta_2 = 2$, and with normal correlation $\widetilde{\rho} = 0, 0.183, 0.305, 0.6$. These are the observed Monte Carlo normal correlations from the pseudo-data (Madsen and Fang, 2010), not the actual correlations between the observed or the pseudo-data. Figure 4.2 presents the relative bias of MLEs for $\widetilde{\rho} = 0, 0.183, 0.305, 0.6$. Note that the relative bias is undefined when $\widetilde{\rho} = 0$; hence, we only considered $\widetilde{\rho} = 0.183, 0.305, 0.6$, in Figure 4.2f. Note that all plots show the

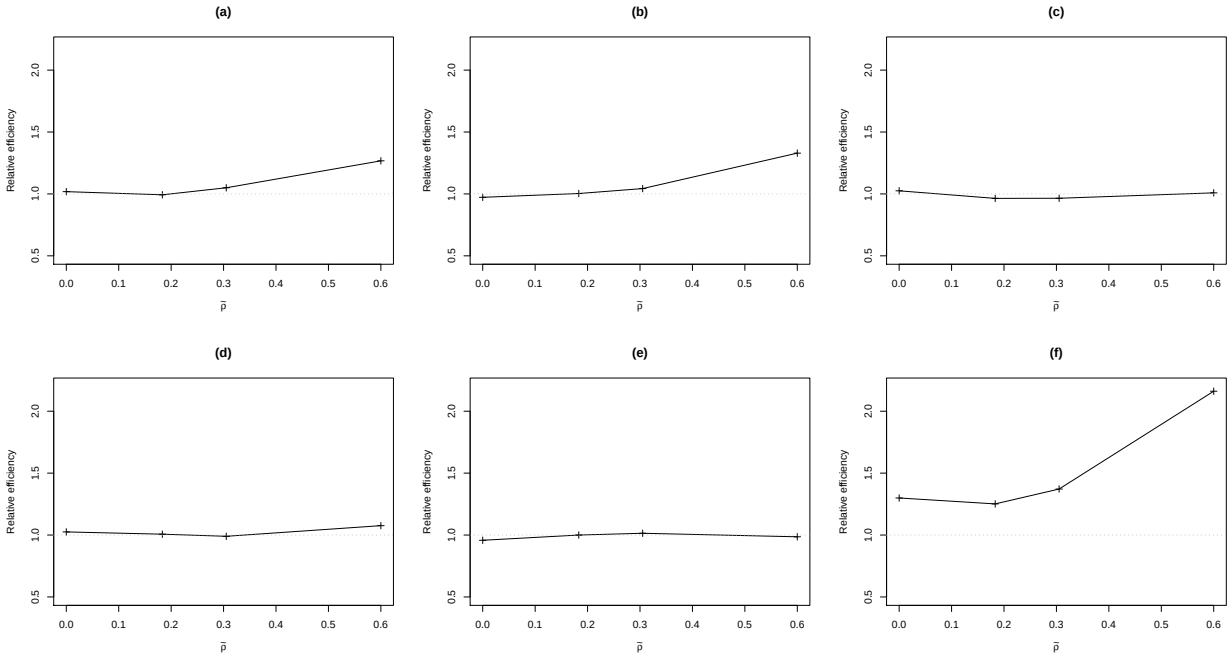

Figure 4.3: Plots (a), (b), (c), (d), (e), and (f) represent the relative efficiency of MLEs of α_1, α_2, β_1, β_2, σ, and $\widetilde{\rho}$, respectively, based on $R = 500$ repeats and $M = 1000$ jitters for a Gaussian copula joint model with logit link for the binary outcome.

relative biases to be relatively small, all below 5% in magnitudes, which indicate that the surrogate likelihood approach yields reasonably unbiased estimates of the parameters.

Figure 4.3 displays the relative efficiency of MLEs for $\widetilde{\rho} = 0, 0.183, 0.305, 0.6$. From the plots, we can see that the efficiencies of the MLEs of the regression coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^\top$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ are all relatively close to unity, indicating that our method of precision estimation for the regression parameters yields SEs which reflect the true sampling variability. There is, however, clear inefficiency in the MLE of $\widetilde{\rho}$ with increasing $\widetilde{\rho}$.

4.6 Application to burn injury data

In this section, we illustrate the methodology outlined in Sections 4.2, 4.3, and 4.4 on a real-data set, namely, the burn injury data set previously analyzed by de Leon and Wu (2011).

Table 4.1: MLEs (based on $M = 1000$ jitters), their SEs, and t-values for the Gaussian copula joint model (Model 1) with marginal models $\text{logit}(p_i) = \text{logit}\{E(X_i)\} = \alpha_1 + \alpha_2 \times \text{age}$ and $Y_i \sim normal(\mu_i = \beta_1 + \beta_2 \times \text{age}, \sigma^2)$, $i = 1, \cdots, 981$. For comparison, we also included Song *et al.*'s (2009) result based on a Gaussian copula joint model with $X_i \sim Bernoulli(p_i)$ used directly as a discrete margin and $Y_i \sim normal(\mu_i, \sigma^2)$, and de Leon and Wu's (2011) result based on the robit-normal Gaussian copula joint model in Chapter 2 with a t-latent distribution (hence, a robit regression model for X_i) and $Y_i \sim normal(\mu_i, \sigma^2)$.

Parameter	Model 1			Song *et al.* (2009)			de Leon and Wu (2011)		
	Est	SE	t	Est	SE	t	Est	SE	t
Disposition									
α_1	–3.684	0.239	–15.42	–3.346	0.201	–16.64	–3.910	0.383	–10.22
α_2	0.050	0.005	10.72	0.043	0.004	10.86	0.055	0.006	8.79
$log(Burn\ area + 1)$									
β_1	6.708	0.067	100.41	6.693	0.069	96.95	6.672	0.067	99.58
β_2	0.004	0.002	3.18	0.004	0.002	2.26	0.004	0.002	2.35
σ	1.257	0.028	44.41	1.255	0.028	44.37	1.205	0.027	44.79
Associations									
$\widetilde{\rho}$	0.290	0.029	10.11	0.863	0.018	46.97	0.820	0.026	31.41
$corr(Y_i^*, Y_i)$	0.265	—	—	—	—	—	0.613	—	—
$corr(X_i, Y_i)$	0.338	—	—	0.796	—	—	—	—	—
τ	0.187	0.019	—	0.663	0.023	—	0.612	0.029	—

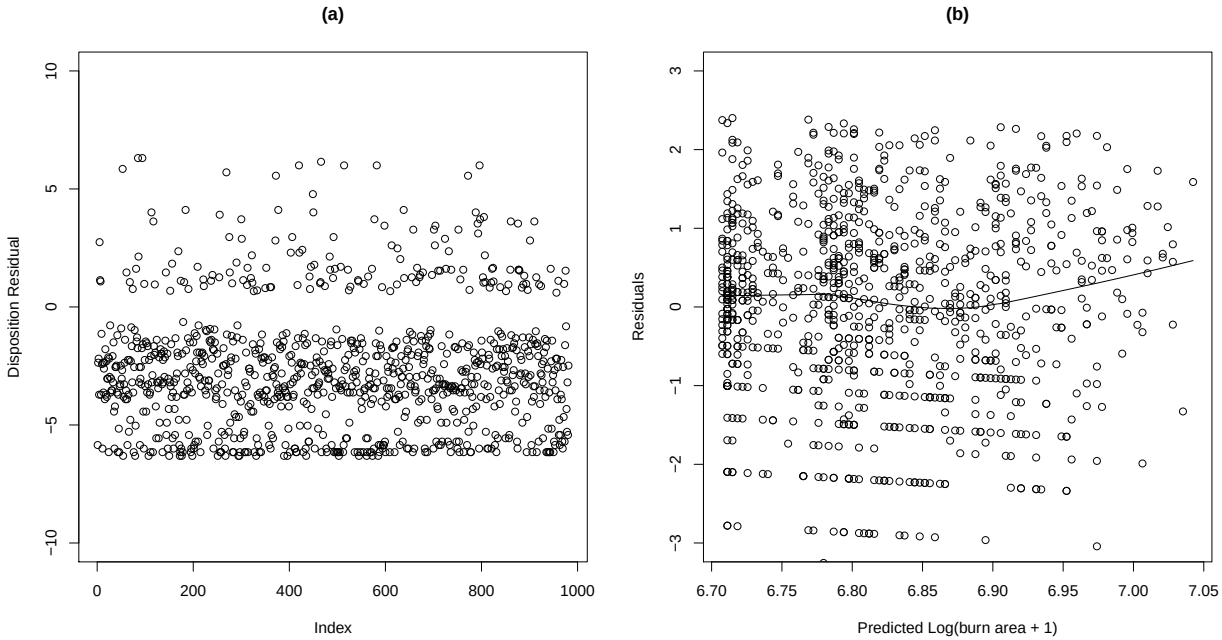

Figure 4.4: Index plot (a) and LOESS-smoothed plot (b) of marginal residuals against predicted $Y_i = \log(\text{burn area} + 1)$ for burn injury data, using a Gaussian copula joint model with marginal models $\text{logit}(p_i) = \text{logit}\{E(X_i)\} = \alpha_1 + \alpha_2 \times \text{age}$ and $Y_i \sim normal(\mu_i = \beta_1 + \beta_2 \times \text{age}, \sigma^2)$ (Model 1).

The burn injury data involved $N = 981$ patients with different ages suffering from burn injuries. Our goal is to jointly model two outcomes, disposition X_i (a binary outcome with 1 for death, and 0 for survival) and severity of burn injury $Y_i = \log(\text{burn area}+1)$ (a continuous outcome representing the total burn area), as a function of the patient's age z_i. We adopt the Gaussian copula joint model with marginal models $\text{logit}(p_i) = \text{logit}\{E(X_i)\} = \alpha_1 + \alpha_2 z_i$ and $Y_i \sim normal(\mu_i = \beta_1 + \beta_2 z_i, \sigma^2)$ (Model 1); we also considered the same model but with an additional quadratic effect of age (Model 2).

The estimates for Model 1 are presented in Table 4.1. The additional regression coefficients for the quadratic effect of age turned out to be not significant for Model 2; hence, we do not report the results for Model 2. Table 4.1 also displays the estimates for Song *et al.*'s (2009) Gaussian copula joint model, where $X_i \sim Bernoulli(p_i)$ is used directly as a discrete

margin and $Y_i \sim normal(\mu_i, \sigma^2)$, and de Leon and Wu's (2011) Gaussian copula joint model with a t-latent distribution (i.e., a robit model for X_i) and $Y_i \sim normal(\mu_i, \sigma^2)$. Note that our marginal models are marginally meaningful in that the regression coefficients have the same interpretation as when the outcomes are analyzed marginally.

We see that $\widehat{\alpha}_2 = 0.05$ with SE $= 0.005$ in Model 1 for the death disposition outcome, indicating that the estimated death rate increases with increasing age. Similarly, $\widehat{\beta}_2 = 0.004$ with SE $= 0.002$ for the burn area outcome, implying that the estimated mean total burn area increases with increasing age as well. Comparing the estimates from the three models in Table 4.1, we find that all the age effect coefficients for death disposition and total burn area are relatively close in magnitude with the same signs.

The associations in Model 1 are smaller than those in Song *et al.*'s (2009) direct approach and in de Leon and Wu's (2011) robit-normal Gaussian copula joint model. The estimate of $\widetilde{\rho}$, the correlation between normal scores, is 0.29 with SE 0.029. Using the piecewise linear approximation method (Kugiumtzis and Bora-Senta, 2010), the biserial correlation is estimated as 0.265. The estimate of the point-biserial correlation $corr(X_i, Y_i)$ between patient's death disposition and total burn area is calculated to be 0.338 by using (4.8). These associations suggest that the two outcomes are not independent, but have a slight positive association. Note that the same estimate from Song *et al.* (2009) differs quite a bit, which is closer to the estimated biserial correlation in de Leon and Wu (2011).

The index plot in Figure 4.4a and the corresponding LOESS-smoothed residual plot in Figure 4.4b for predicted log(burn area + 1) based on Model 1 show no apparent patterns and trends. The absence of any apparent trends suggests that our model fits the burn injury data quite well.

One important feature of our Gaussian copula joint model is that we can obtain and assess the conditional distribution of one outcome given the other. For instance, we can obtain the conditional probability of a patient's death at fixed ages, given the patient's total

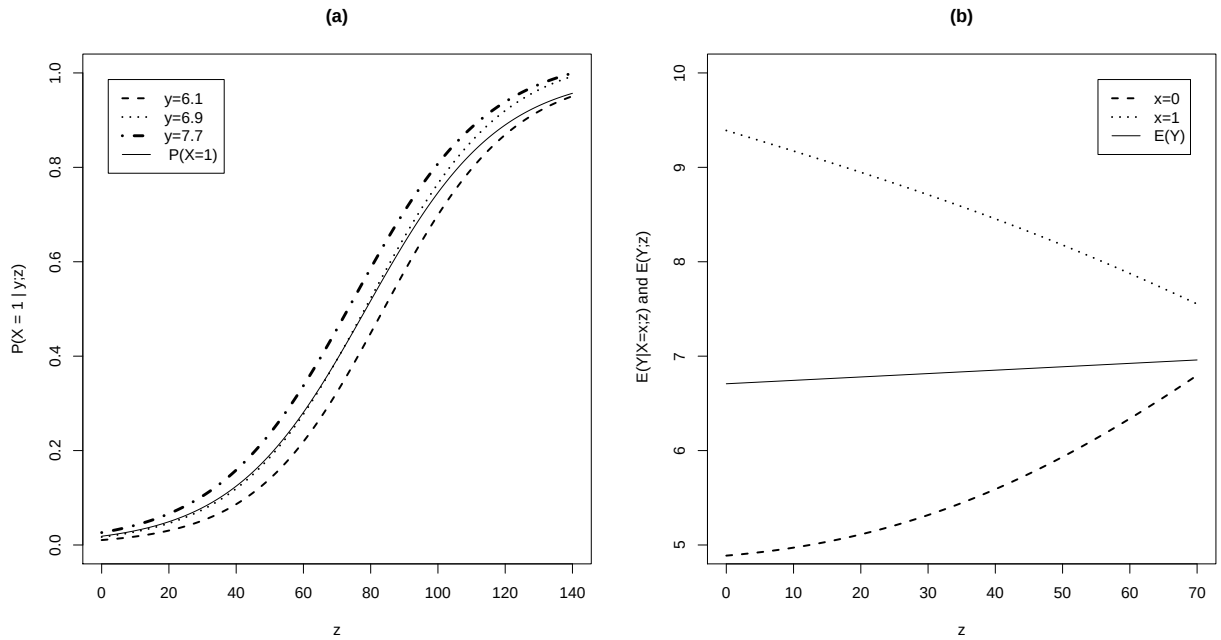

Figure 4.5: Plots of (a) the conditional probability of death as a function of age z at fixed $y = \log(\text{burn area}+1) = 6.1, 6.9, 7.7$, and (b) the conditional mean of $y = \log(\text{burn area}+1)$, given the disposition of death $x = 1$, as a function of age z.

burn area. We have

$$P(X_i = x_i | Y_i = y_i) \quad = \quad \int_{A_{x_i}} \frac{\phi_2\left(\Phi^{-1}(u_i), \Phi^{-1}(v_i); \tilde{\rho}\right)}{\phi(\Phi^{-1}(u_i))\phi(\Phi^{-1}(v_i))} f_{Y_i^*}(y_i^*) dy_i^*, \tag{4.12}$$

where $A_{x_i} = \{y_i^* \in (-1, 1) : [y_i^* + 1] = x_i\}$, for $x_i = 0, 1$. Similarly, we have conditional mean total burn area at fixed ages, given patient's disposition, as

$$E(Y_i | X_i = x_i) \quad = \quad \int_{-\infty}^{+\infty} y_i \left\{ \frac{1}{P(X_i = x_i)} \int_{A_{x_i}} f_{Y_i^*, Y_i}(y_i^*, y_i) dy_i^* \right\} dy_i. \tag{4.13}$$

We plot the estimated conditional probability (4.12) of death as a function of age, at each fixed $y_i = \log(\text{burn area} + 1) = 6.1, 6.9, 7.7$, in Figure 4.5a; the estimated conditional mean (4.13) of $Y_i = \log(\text{burn area}+1)$ is also plotted as a function of age, given disposition indicator $x_i = 0$ or $x_i = 1$, in Figure 4.5b. From Figure 4.5a, we observe that the estimated conditional probability of disposition of death increases with age, for given total burn area. We also find that a patient with smaller total burn area is less likely to die than another with a larger

total burn area. Figure 4.5b shows that, given the death disposition indicator, the total burn area, on average, decreases with age. We also included the marginal probability of death and the marginal mean total burn area, as functions of age, in Figure 4.5 for comparison.

4.7 Discussion

In this chapter, we developed a Gaussian copula joint model for mixed discrete and continuous outcomes based on continuous extension of the discrete responses. By employing continuous extensions of integer-valued outcomes to avoid the direct use of discrete margins in the Gaussian copula, our joint model uniquely determines the joint distribution; therefore, non-identifiability is not a concern. The association between mixed discrete and continuous variables is measured by the biserial correlation obtained through piece-wise linear approximation (Kugiumtzis and Bora-Senta, 2010) from the normal correlation.

The simulation results for finite-sample relative bias and efficiency of MLEs are also reported. The relative bias of the estimates suggests that they behave well, exhibiting generally small or negligible bias. The MLEs' SEs, obtained by conventional methods, appear to capture the estimates' true sampling variability, with relative efficiencies all close to unity. The unique exception is the MLE for the normal correlation; this same observation was reported recently by Nikoloulopoulos (2013).

We use the burn injury data, analyzed earlier by de Leon and Wu (2011) and Song *et al.* (2009), to illustrate our proposed methodology. In contrast to Song *et al.*'s (2009) approach, our model employs the jittering method to overcome the non-identifiability issue arising from the direct use of discrete margins in copulas; compared with de Leon and Wu's (2011) approach, our model does not rely on a latent variable description of the binary outcome. The estimates of the regression coefficients from our joint model are qualitatively comparable to the results from Song *et al.* (2009) and de Leon and Wu (2011), with the exception of the estimated correlations. This is to be expected since different dependence models are used in

the three approaches.

Finally, we assumed the data are independent in our development of the model. Extending our methodology to allow for clustering, as in the EG data in Chapter 3, or for longitudinally correlated observations, as in the wages-hours-union memberships data in Chapter 2, should be a worthwhile project to pursue in the future.

Chapter 5

Sample size determination in clinical trials with mixed co-primary endpoints via Gaussian copulas

5.1 Introduction

A clinical trial is typically designed to characterize the efficacy (and safety) of drugs or treatments based on a so-called primary endpoint, which is considered the most clinically relevant and measurable outcome, among many, in a clinical trial. However, it is often difficult in practice to rely on a single primary endpoint, since it may fail to provide a comprehensive summary of a drug's efficacy. The recent PREMIER Study (Breedveld *et al.*, 2006) on early aggressive rheumatoid arthritis, for example, utilized two such correlated outcomes, referred to as co-primary endpoints, to compare the efficacy of two treatments, namely, combination therapy (adalimumab plus methotrexate) and monotherapy (methotrexate alone). The same is true of chronic kidney disease, where two co-primary endpoints were used for assessing kidney disease symptoms in the VALOR trial (Rudnick *et al.*, 2008). Sozu *et al.* (2012) provided other examples of diseases (e.g., irritable bowel syndrome and Alzheimer's disease), where at least two co-primary endpoints are adopted to evaluate the symptomatic improvement or non-improvement due to treatments.

Statistical significance in a clinical trial with co-primary endpoints are based on the co-primary endpoints all being statistically significant; there is thus no need to adjust the Type I error rate when designing such a clinical trial. That is, the hypothesis test corresponding to each co-primary endpoint is based on the same significance level, say 5%, and control of the maximum Type I error rate is not a concern. What is of concern, however, is the

Wu and de Leon (2013*b*), forthcoming in *Biometrical Journal*, is partially based on this chapter.

Type II error rate (or equivalently, the power), which increases with the number of co-primary endpoints. Hence, when determining the sample size needed for a clinical trial with co-primary endpoints, the major consideration is the overall power of the tests for the co-primary endpoints. Controlling the tests' overall power requires that the analysis properly accounts for the associations among the co-primary endpoints. A model for the joint distribution of the endpoints is thus necessary.

In this chapter, we outline a general methodology for calculating the sample size for a confirmatory and superiority clinical trial on two treatments' efficacy, as measured by multiple co-primary endpoints comprising a mixture of binary and continuous outcomes. Our methodology is based on the Gaussian copula joint model, as introduced in Section 2.3, which affords great flexibility when constructing a joint distribution for the mixed endpoints in that non-Gaussian endpoints (e.g., time-to-event outcomes) are easily accommodated. The Gaussian copula joint model includes the CGCM (see Chapter 1 for a discussion), which is the model Sozu *et al.* (2012) adopted, as a special case. While the use of latent variables to describe the binary endpoints may have biological basis in practice, it is also appealing statistically since tests on proportions for the binary endpoints can be equivalently formulated in terms of the means of the corresponding latent variables. As we show later in the chapter, the latter approach leads to higher overall power, thus yielding smaller sample sizes than those calculated via Sozu *et al.*'s (2012) methodology.

The Gaussian copula joint model is adapted to a clinical trial setting with mixed binary and continuous co-primary endpoints in Section 5.2. Section 5.3 proposes an approach to sample size determination based on latent means-tests for the binary co-primary endpoints. In Section 5.4, we provide empirical comparisons between our methodology and that of Sozu *et al.* (2012) in terms of sample sizes and overall power. Section 5.5 discusses a numerical illustration of our approach using the PREMIER Study (Breedveld *et al.*, 2006) on aggressive rheumatoid arthritis. Finally, we conclude the chapter in Section 5.6

5.2 Gaussian copula joint model

Suppose we have binary and continuous co-primary endpoints $\mathbf{X}_{gi} = (X_{gi1}, \cdots, X_{giQ})^\top$ and $\mathbf{Y}_{gi} = (Y_{gi1}, \cdots, Y_{giC})^\top$, respectively, for subject $i = 1, \cdots, n_g$, in group $g = T_0$ (control), T_1 (treatment). We assume that a latent vector $\mathbf{Y}^*_{gi} = (Y^*_{gi1}, \cdots, Y^*_{giQ})^\top$ underlies $\mathbf{X}_{gi}$, such that $X_{gij} = \mathrm{I}\{Y^*_{gij} > 0\}$, $j = 1, \cdots, Q$. Following the development of the Gaussian copula joint model in Section 2.3, we assume that the joint CDF $F_{\mathbf{Y}^*_{gi}, \mathbf{Y}_{gi}}(\cdot)$ of $\mathbf{Y}^*_{gi}$ and $\mathbf{Y}_{gi}$ is determined by a Gaussian copula with correlation matrix $\widetilde{\mathbf{R}}$, given marginal CDFs $F_{Y^*_{gi1}}(\cdot), \cdots, F_{Y^*_{giQ}}(\cdot)$ for the latent variables $Y^*_{gi1}, \cdots, Y^*_{giQ}$, and $F_{Y_{gi1}}(\cdot), \cdots, F_{Y_{giC}}(\cdot)$ for the continuous co-primary endpoints. With $P = C + Q$, we have

$$F_{\mathbf{Y}^*_{gi}, \mathbf{Y}_{gi}}(\mathbf{y}^*_{gi}, \mathbf{y}_{gi}) = \Phi_P \left(\begin{array}{c} \Phi^{-1}(u_{gi1}), \cdots, \Phi^{-1}(u_{giQ}), \\ \Phi^{-1}(v_{gi1}), \cdots, \Phi^{-1}(v_{giC}) \end{array} ; \widetilde{\mathbf{R}}_g \right), \tag{5.1}$$

and the corresponding joint density is

$$\begin{aligned} f_{\mathbf{Y}^*_{gi}, \mathbf{Y}_{gi}}(\mathbf{y}^*_{gi}, \mathbf{y}_{gi}) &= \phi_P \left(\begin{array}{c} \Phi^{-1}(u_{gi1}), \cdots, \Phi^{-1}(u_{giQ}), \\ \Phi^{-1}(v_{gi1}), \cdots, \Phi^{-1}(v_{giC}) \end{array} ; \widetilde{\mathbf{R}}_g \right) \left\{ \prod_{j=1}^{Q} \phi(\Phi^{-1}(u_{gij})) \right\}^{-1} \\ &\quad \times \left\{ \prod_{k=1}^{C} \phi(\Phi^{-1}(v_{gik})) \right\}^{-1} \prod_{j=1}^{Q} f_{Y^*_{gij}}(y^*_{gij}) \prod_{k=1}^{C} f_{Y_{gik}}(y_{gik}), \end{aligned} \tag{5.2}$$

where $u_{gij} = F_{Y^*_{gij}}(y^*_{gij})$, $v_{gik} = F_{Y_{gik}}(y_{gik})$, $f_{Y^*_{gij}}(\cdot)$ and $f_{Y_{gik}}(\cdot)$ are the respective marginal densities of Y^*_{gij} and Y_{gik}, and $\phi_P(\cdot; \widetilde{\mathbf{R}}_g)$ is the P-dimensional standard multivariate normal density with correlation matrix $\widetilde{\mathbf{R}}_g$, for $j = 1, \cdots, Q$, $k = 1, \cdots, C$, $i = 1, \cdots, n_g$. The corresponding joint density $f_{\mathbf{X}_{gi}, \mathbf{Y}_{gi}}(\cdot)$ of $\mathbf{X}_{gi}$ and $\mathbf{Y}_{gi}$ is obtained as

$$f_{\mathbf{X}_{gi}, \mathbf{Y}_{gi}}(\mathbf{x}_{gi}, \mathbf{y}_{gi}) = \int_{A_{giQ} \times \cdots \times A_{gi1}} f_{\mathbf{Y}^*_{gi}, \mathbf{Y}_{gi}}(\mathbf{y}^*_{gi}, \mathbf{y}_{gi}) d\mathbf{y}^*_{gi}, \tag{5.3}$$

where interval A_{gij} is either $(-\infty, 0]$ or $(0, +\infty)$ according as whether x_{gij} is 0 or 1. Note that (5.3) is given by (3.12), which is (2.4) with $L_1 = \cdots = L_Q = 1$. In (5.1) and (5.2), the matrix $\widetilde{\mathbf{R}}_g$ of normal correlations is partitioned as

$$\widetilde{\mathbf{R}}_g = \begin{pmatrix} \widetilde{\mathbf{R}}_{g11} & \widetilde{\mathbf{R}}_{g12} \\ & \widetilde{\mathbf{R}}_{g22} \end{pmatrix}, \tag{5.4}$$

where $\widetilde{\mathbf{R}}_{g11}$ is the $Q \times Q$ matrix of normal correlations between the normal scores $\Phi^{-1}(U_{gij})$ and $\Phi^{-1}(U_{gij'})$, $\widetilde{\mathbf{R}}_{g12}$ is the $Q \times C$ matrix of normal correlations between the normal scores $\Phi^{-1}(U_{gij})$ and $\Phi^{-1}(V_{gik})$, and $\widetilde{\mathbf{R}}_{g22}$ is the $C \times C$ matrix of normal correlations between the normal scores $\Phi^{-1}(V_{gik})$ and $\Phi^{-1}(V_{gik'})$, with $U_{gij} = F_{Y^*_{gij}}(Y^*_{gij})$ and $V_{gik} = F_{Y_{gik}}(Y_{gik})$ as the PITS. As in Chapter 2, the tetrachoric correlations $\rho_{Y^*_{gj}Y^*_{gj'}} = corr(Y^*_{gij}, Y^*_{gij'})$ between X_{gij} and $X_{gij'}$, the biserial correlations $\rho_{Y^*_{gj}Y_{gk}} = corr(Y^*_{gij}, Y_{gik})$ between X_{gij} and Y_{gik}, and the Pearson's correlations $\rho_{Y_{gk}Y_{gk'}} = corr(Y_{gik}, Y_{gik'})$ between Y_{gik} and $Y_{gik'}$, can all be obtained from $\widetilde{\mathbf{R}}_g$ via Kugiumtzis and Bora-Senta's (2010) piece-wise linear approximation method. Note that these correlations are nothing but the usual Pearson's pairwise correlations and as such, do not restrict the correlation parameter space beyond that the corresponding correlation matrix be positive definite. Note also that the associations between two binary endpoints and between a binary endpoint and a continuous endpoint are measured by correlations involving the latent variables corresponding to the binary endpoints. Although Pearson's correlations $corr(X_{gij}, X_{gij'})$ and $corr(X_{gij}, Y_{gik})$ can be calculated, note that their values are restricted by the binary endpoints' marginal probabilities (Sozu *et al.*, 2012). Sozu *et al.* (2012) used Pearson's correlations in their methodology to measure the associations among the mixed endpoints, which is disadvantageous in practice.

Denote by μ^*_{gj} and μ_{gk} the means of Y^*_{gij} and Y_{gik}, respectively, and let $var(Y_{gik}) = \sigma^2_{gk} > 0$. Note that we assumed $var(Y^*_{gij}) = 1$, for identifiability reasons. We likewise assumed the cutpoints for the threshold models relating X_{gij} to Y^*_{gij}, to be all equal to 0 for identifiability reasons.

The Gaussian copula joint model in (5.1) and (5.2) is a flexible means of accounting for associations between the mixed co-primary endpoints in that its margins $F_{Y^*_{gij}}(\cdot)$, $j = 1, \cdots, Q$, and $F_{Y_{gik}}(\cdot)$, $k = 1, \cdots, C$, need not come from the same parametric family. In fact, any (absolutely) continuous CDF can be used as margins; hence, non-Gaussian endpoints (e.g., a gamma distributed time-to-event endpoint) can be easily accommodated

in the model. In addition, the Gaussian distribution-based CGCM, which is the default joint model for mixed binary and continuous variables, is a special case of the Gaussian copula joint model with Gaussian continuous variables and Gaussian latent variables. Our methodology is therefore a generalization of the CGCM-based methodology developed by Sozu *et al.* (2012).

5.3 Sample size calculation via latent-means tests

We are interested in testing the hypothesis on the difference $\delta_j^\dagger = \pi_{T_1 j} - \pi_{T_0 j}$ in the proportions and that on the difference $\delta_k = \mu_{T_1 k} - \mu_{T_0 k}$ in the means, assuming homogeneity of the two groups (i.e., $\sigma^2_{T_1 k} = \sigma^2_{T_0 k} = \sigma^2_k$, for all k, and $\widetilde{\mathbf{R}}_{T_1} = \widetilde{\mathbf{R}}_{T_0}$), where $\pi_{gj} = P(X_{gij} = 1) = P(Y^*_{gij} > 0)$. Note that a positive difference indicates a treatment benefit of interest. We can then assert the superiority of the test treatment over the control in terms of all $P = Q + C$ co-primary endpoints if and only if $\delta_j^\dagger > 0$ and $\delta_k > 0$, for all $j = 1, \cdots, Q$, $k = 1, \cdots, C$. The hypotheses of interest are then

$$H_0 : \begin{cases} \exists j \text{ such that } \delta_j^\dagger \leq 0 \\ \text{or} \\ \exists k \text{ such that } \delta_k \leq 0 \end{cases} \quad \text{and} \quad H_1 : \begin{cases} \delta_j^\dagger > 0 \ \forall j \\ \text{and} \\ \delta_k > 0 \ \forall k \end{cases} . \tag{5.5}$$

In testing the above hypotheses, the null hypothesis H_0 is rejected if and only if all the null hypotheses associated with each of the P co-primary endpoints are rejected at significance level α.

Since the latent means $\mu^*_{g1} = \Phi^{-1}(\pi_{g1}), \cdots, \mu^*_{gQ} = \Phi^{-1}(\pi_{gQ})$ are all estimable (e.g., by maximum likelihood), and noting the following equivalence:

$$\left.\begin{array}{c} H_0 : \exists j \text{ such that } \delta_j^\dagger \leq 0 \\ H_1 : \delta_j^\dagger > 0 \ \forall j \end{array}\right\} \text{ if and only if } \left\{\begin{array}{c} H_0^* : \exists j \text{ such that } \delta_j^* \leq 0 \\ H_1^* : \delta_j^* > 0 \ \forall j \end{array}\right. , \tag{5.6}$$

where $\delta_j^* = \mu^*_{T_1 j} - \mu^*_{T_0 j}$, we propose instead the following statistics for testing the difference

in latent means for each binary endpoint:

$$Z_j^* \;=\; \frac{\widehat{\mu}_{T_1 j}^* - \widehat{\mu}_{T_0 j}^*}{\sqrt{avar(\widehat{\mu}_{T_1 j}^*) + avar(\widehat{\mu}_{T_0 j}^*)}} \;=\; \frac{\widehat{\mu}_{T_1 j}^* - \widehat{\mu}_{T_0 j}^*}{\sqrt{n_{T_1}^{-1} + n_{T_0}^{-1}}} \;=\; \frac{\widehat{\mu}_{T_1 j}^* - \widehat{\mu}_{T_0 j}^*}{\sqrt{\frac{1+\kappa}{\kappa n_{T_1}}}}, \tag{5.7}$$

where $\kappa = n_{T_0}/n_{T_1}$, $\widehat{\mu}_{gj}^*$ is the MLE of μ_{gj}^* (Poon and Lee, 1987), and the large-sample variance $avar(\widehat{\mu}_{gj}^*) = n_g^{-1}$ is obtained, by standard asymptotic results on MLEs, from the inverse of the Fisher information matrix

$$n_g E\left\{-\frac{\partial^P}{\partial \boldsymbol{\Theta}_g \partial \boldsymbol{\Theta}_g^\top} \log f_{\mathbf{Y}_{gi}^*, \mathbf{Y}_{gi}}(\mathbf{y}_{gi}^*, \mathbf{y}_{gi})\right\}, \tag{5.8}$$

with $\boldsymbol{\Theta}_g$ containing $\boldsymbol{\mu}_g^* = (\mu_{g1}^*, \cdots, \mu_{gQ}^*)^\top$, $\boldsymbol{\mu}_g = (\mu_{g1}, \cdots, \mu_{gC})^\top$, and the unknown correlations in $\widetilde{\mathbf{R}}$. Test statistics (5.7) are a simpler alternative to Sozu *et al.*'s (2012) Pearson-type test statistic for the difference in the proportions, which requires the calculation of a SE for the estimated difference. Note that the MLE of $\boldsymbol{\mu}_g^*$ has no closed form and is obtained numerically.

For testing the hypothesis on the difference in the means of the continuous endpoints for the treatment and control groups, we follow Sozu *et al.* (2012) by using the same Z-statistics

$$Z_k \;=\; \frac{\widehat{\mu}_{T_1 k} - \widehat{\mu}_{T_0 k}}{\sqrt{var(\widehat{\mu}_{T_1 k}) + var(\widehat{\mu}_{T_0 k})}} \;=\; \frac{\overline{Y}_{T_1 k} - \overline{Y}_{T_0 k}}{\sigma_k \sqrt{n_{T_1}^{-1} + n_{T_0}^{-1}}} \;=\; \frac{\overline{Y}_{T_1 k} - \overline{Y}_{T_0 k}}{\sigma_k \sqrt{\frac{1+\kappa}{\kappa n_{T_1}}}}. \tag{5.9}$$

Note that $\overline{Y}_{T_1 k}$ and $\overline{Y}_{T_0 k}$ are the MLEs of $\mu_{T_1 k}$ and $\mu_{T_0 k}$, respectively.

The overall power function based on combining the test statistics Z_j^* in (5.7) with the test statistics Z_j in (5.9) is then

$$1 - \beta \;=\; P\left(\bigcap_{j=1}^{Q}\{Z_j^* > z_\alpha\} \bigcap_{k=1}^{C}\{Z_k > z_\alpha\} \,\middle|\, \boldsymbol{\delta}\right) \;\simeq\; P\left(\bigcap_{h=1}^{P}\{Z_h^\dagger > z_h^\dagger\} \,\middle|\, \boldsymbol{\delta}\right), \tag{5.10}$$

for $\boldsymbol{\delta} = (\delta_1^*, \cdots, \delta_Q^*, \delta_1, \cdots, \delta_C)^\top \neq \mathbf{0}$,

$$Z_h^\dagger \;=\; \begin{cases} Z_j^* - \delta_j^* \sqrt{\frac{\kappa n_{T_1}}{1+\kappa}} \;=\; \frac{\widehat{\mu}_{T_1 j}^* - \widehat{\mu}_{T_0 j}^* - \delta_j^*}{\sqrt{\frac{1+\kappa}{\kappa n_{T_1}}}} & , j = 1, \cdots, Q \\ Z_k - \frac{\delta_k}{\sigma_k}\sqrt{\frac{\kappa n_{T_1}}{1+\kappa}} \;=\; \frac{\overline{Y}_{T_1 k} - \overline{Y}_{T_0 k} - \delta_k}{\sigma_k \sqrt{\frac{1+\kappa}{\kappa n_{T_1}}}} & , k = 1, \cdots, C \end{cases}, \tag{5.11}$$

$$z_h^\dagger \;=\; \begin{cases} z_\alpha - \delta_j^* \sqrt{\frac{\kappa n_{T_1}}{1+\kappa}} & , j = 1, \cdots, Q \\ z_\alpha - \frac{\delta_k}{\sigma_k}\sqrt{\frac{\kappa n_{T_1}}{1+\kappa}} & , k = 1, \cdots, C \end{cases}, \tag{5.12}$$

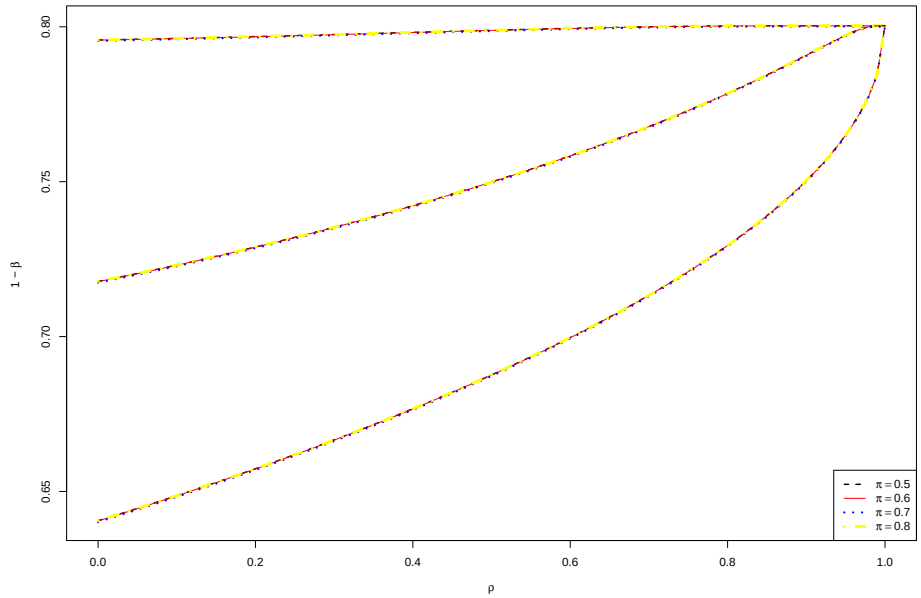

Figure 5.1: Overall power as a function of ρ for $Q = C = 1$, with the same set-up considered in Sozu *et al.* (2012).

where z_α is the $(1-\alpha)100$th standard normal percentile. The asymptotic multivariate normality of the respective MLEs $\widehat{\boldsymbol{\Theta}}_{T_0}$ and $\widehat{\boldsymbol{\Theta}}_{T_1}$ of $\boldsymbol{\Theta}_{T_0}$ and $\boldsymbol{\Theta}_{T_1}$ implies that $(Z_1^\dagger, \cdots, Z_P^\dagger)^\top$ has an approximate P-dimensional standard multivariate normal distribution with correlation matrix $\mathbf{R}$, the matrix of tetrachoric correlations between the binary endpoints, biserial correlations between the binary and continuous endpoints, and Pearson's correlations between the continuous outcomes. That is, we have

$$1 - \beta \;\simeq\; \Phi_P(-z_1^\dagger, \cdots, -z_P^\dagger; \mathbf{R}). \tag{5.13}$$

Note that our approach is completely based on $\widehat{\boldsymbol{\Theta}}_{T_0}$ and $\widehat{\boldsymbol{\Theta}}_{T_1}$, rendering our approach, which makes direct use of latent means and correlations, as a simpler and more coherent alternative to that adopted in Sozu *et al.* (2012). This is because our treatment of mixed endpoints is no different from that of continuous endpoints considered in Sozu *et al.* (2011), thus yielding a unified and more streamlined methodology.

Note as well that the overall power function in (5.13) does not depend on the marginal means (latent or otherwise). This is evident from Figure 5.1, a plot of the overall power as a function of the biserial correlation ρ for the case of one binary endpoint and one continuous endpoint (i.e., $Q = C = 1$), where we let $\pi_{T_1} - \pi_{T_0} = 0.05$, $\pi_{T_0} = 0.5, 0.6, 0.7$, and 0.8, at significance level $\alpha = 0.025$ and marginal power of 80% for each endpoint. This contrasts with that in Sozu *et al.* (2012), which depends on π_{T_0k} and π_{T_1k}, through the Pearson's correlations $corr(Z_1^\dagger, Z_2^\dagger) = corr(Z^*, Z)$. Finally, as we demonstrate in the next section, our approach is more powerful resulting in smaller sample sizes.

5.4 Empirical comparisons

For convenience, we consider the case of one binary and one continuous co-primary endpoints (i.e., $Q = C = 1$). The corresponding joint density (5.3) simplifies as

$$f_{X_{gi},Y_{gi}}(x_{gi}, y_{gi}) = \left(\int_{-\infty}^{0} f_{Y_{gi}^*,Y_{gi}}(y_{gi}^*, y_{gi})dy_{gi}^*\right)^{1-x_{gi}} \left(\int_{0}^{+\infty} f_{Y_{gi}^*,Y_{gi}}(y_{gi}^*, y_{gi})dy_{gi}^*\right)^{x_{gi}}. \quad (5.14)$$

In our empirical comparisons, we let $\widetilde{\rho} = 0, 0.31, 0.52$, and 0.82. Using piece-wise linear approximations (Kugiumtzis and Bora-Senta, 2010), we obtained the biserial correlation $\rho = \rho_{Y_g^*Y_g} = corr(Y_{gi}^*, Y_{gi}) = 0, 0.3, 0.5$, and 0.8, the same values used in Sozu *et al.* (2012) for their bivariate CGCM.

In the following calculations, we relax the conventional Gaussian assumption, as in Sozu *et al.* (2012), by considering $Y_{gi}^* \sim normal(\mu_g^*, 1)$ (i.e., a Gaussian latent variable for X_{gi}) and $Y_{gi} \sim logistic(\mu_g, \sqrt{3}/\pi)$ (i.e., a logistic distributed continuous endpoint). We use the normal-logistic Gaussian copula joint model for the joint distribution of Y_{gi}^* and Y_{gi}, with normal correlation $\widetilde{\rho}$. We use significance level $\alpha = 0.025$ and overall power $1 - \beta \approx 0.8$ in our sample size calculations.

In Tables 5.1 and 5.2, given (π_{T_1}, π_{T_0}), we obtained $\mu_{T_1}^* = \Phi^{-1}(\pi_{T_1})$ and $\mu_{T_0}^* = \Phi^{-1}(\pi_{T_0})$, so that we get $\delta^* = \mu_{T_1}^* - \mu_{T_0}^*$. For fixed $z_2^\dagger/z_1^\dagger$ in Table 5.1, the standardized effect size

Table 5.1: Sample sizes $n = n_{T_0} = n_{T_1}$ for overall power $1 - \beta \approx 80\%$, $\alpha = 0.025$, $Q = C = 1$, and with fixed $z_2^\dagger / z_1^\dagger$ and $\delta^* = \mu^*_{T_1} - \mu^*_{T_0} = \Phi^{-1}(\pi_{T_1}) - \Phi^{-1}(\pi_{T_0})$. Marginal sample sizes E_1 and E_2 for the continuous and binary endpoints, respectively, were calculated with individual power of at least 80%.

$\frac{z_2^\dagger}{z_1^\dagger}$	$\bar{\delta}$	δ^*	π_{T_1}	π_{T_0}	$n = n_{T_0} = n_{T_1}$				E_2	E_1
					$\rho = 0.0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.8$		
1	0.126	0.126	0.55	0.50	1306	1273	1240	1160	995	995
	0.132	0.132	0.65	0.60	1184	1154	1124	1052	902	902
	0.150	0.150	0.75	0.70	916	893	869	814	697	697
	0.195	0.195	0.85	0.80	544	530	516	483	414	414
	0.253	0.253	0.60	0.50	322	314	305	286	245	245
	0.271	0.271	0.70	0.60	281	274	267	250	214	214
	0.317	0.317	0.80	0.70	205	200	195	183	156	156
	0.440	0.440	0.90	0.80	107	104	102	95	82	82
1.5	0.144	0.126	0.55	0.50	1156	1130	1105	1046	752	995
	0.152	0.132	0.65	0.60	1048	1025	1002	948	682	902
	0.173	0.150	0.75	0.70	811	792	774	733	527	697
	0.224	0.195	0.85	0.80	481	471	460	436	313	414
	0.291	0.253	0.60	0.50	285	278	272	258	185	245
	0.312	0.271	0.70	0.60	249	243	238	225	162	214
	0.365	0.317	0.80	0.70	182	178	174	164	118	156
	0.506	0.440	0.90	0.80	95	93	91	86	62	82
3	0.201	0.126	0.55	0.50	1006	1002	998	995	388	995
	0.211	0.132	0.65	0.60	912	908	905	902	352	902
	0.240	0.150	0.75	0.70	705	702	700	679	272	697
	0.312	0.195	0.85	0.80	419	417	416	414	162	414
	0.406	0.253	0.60	0.50	248	247	246	245	96	245
	0.434	0.271	0.70	0.60	217	216	215	214	84	214
	0.508	0.317	0.80	0.70	158	158	157	157	61	156
	0.704	0.440	0.90	0.80	83	82	82	82	32	82

$\bar{\delta} = \delta/\sigma$ for the continuous endpoint is then given by

$$\bar{\delta} \;=\; \frac{\delta^*}{(z_{0.975} + z_{0.8})}\left\{z_{0.975} + \left(\frac{z_2^\dagger}{z_1^\dagger}\right) z_{0.8}\right\} \;=\; \frac{\delta^*}{2.8}\left\{1.96 + \left(\frac{z_2^\dagger}{z_1^\dagger}\right) 0.84\right\}. \tag{5.15}$$

To derive (5.15), we first fixed the sample size $n^* = \kappa^{-1}(1+\kappa)(z_{0.975} + z_{0.8})^2/(\delta^*)^2$ for the binary endpoint at 80% power, and use it along with $z_2^\dagger/z_1^\dagger$ to get (5.15). Given α, β, κ, and (π_{T_1}, π_{T_0}) (to get $(\mu^*_{T_1}, \mu^*_{T_0})$, and hence, δ^*), we obtain $\bar{\delta}$ from (5.15), so that we get $z_1^\dagger$ and $z_2^\dagger$. For fixed normal correlation $\widetilde{\rho}$, we can obtain the biserial correlation ρ using piece-wise linear approximations, then evaluate the overall power (5.13) iteratively to get the sample size n that yields an approximate power of 80%. In our empirical comparisons, we fixed $\kappa = 1$, i.e., $n_{T_1} = n_{T_0} = n$.

Our results displayed in Table 5.1, with $z_2^\dagger/z_1^\dagger = 1, 1.5, 3$, clearly show the superiority of our approach over that proposed by Sozu *et al.* (2012) in that our sample sizes (i.e., last 6 columns of Table 5.1) are much lower than those obtained by Sozu *et al.* (2012) while maintaining an approximate overall power of 80%. In some cases, the sample sizes in Sozu *et al.* (2012) are more than twice as large as those in Table 5.1; for example, we obtained a sample size of $n = 82$, which is much smaller than the sample size of $n = 200$ obtained by Sozu *et al.* (2012), for $z_2^\dagger/z_1^\dagger = 3$, $\delta^* = 0.44$ (which corresponds to $\pi_{T_1} - \pi_{T_0} = 0.9 - 0.8 = 0.1$), and $\rho = 0.8$. Note that the values of $\bar{\delta}$, the standardized effect size for the continuous endpoint, shown in Table 5.1 are not the same as those in Sozu *et al.* (2012) because δ^* is not the difference between proportions as in Sozu *et al.* (2012), it being the difference between latent means. For comparison, we also considered fixing $\bar{\delta}$ at values used by Sozu *et al.* (2012) and along with the given δ^* values, determined $z_2^\dagger/z_1^\dagger$ from

$$\frac{z_2^\dagger}{z_1^\dagger} \;=\; \frac{\bar{\delta}(z_{0.975} + z_{0.8}) - z_{0.975}\delta^*}{z_{0.8}\delta^*} \;=\; \frac{2.8\bar{\delta} - 1.96\delta^*}{0.86\delta^*}. \tag{5.16}$$

Table 5.2 shows the sample size $n = n_{T_1} = n_{T_0}$ for fixed δ^* and $\bar{\delta}$, at 80% approximate overall power and significance level $\alpha = 0.025$. Note that the marginal sample size E_1 is very close to what Sozu *et al.* (2012) reported, since we used the same $\bar{\delta}$ values. Although

closer for this case than for those considered in Table 5.1, our sample sizes are again much smaller than those in Sozu *et al.* (2012). The results in Tables 5.1 and 5.2 suggest that our approach is more powerful than Sozu *et al.*'s (2012) and attains the required overall power at much smaller sample sizes. Table 5.3 gives the overall power using our normal-logistic Gaussian copula joint model for the same sample sizes tabulated in Table 1 of Sozu *et al.* (2012), obtained with their method at 80% approximate overall power and significance level $\alpha = 0.025$. In Table 5.3, we find that the overall power using our method is at least 94% in all cases, a big increase from the 80% reported in Sozu *et al.* (2012).

Finally, we checked the accuracy of the large-sample normal approximation we used in calculating the approximate overall power for the normal-logistic Gaussian copula joint model. To do this, we used Monte Carlo approximation to calculate the exact power based on the sample sizes in Table 5.1, with fixed $z_2^\dagger/z_1^\dagger$. With $n = n_{T_0} = n_{T_1}$, overall power $1-\beta \approx 80\%$, and $\alpha = 0.025$, we generated $R = 5000$ samples from normal-logistic Gaussian copula joint model with a normal correlation $\widetilde{\rho}$ that corresponds to the value of the biserial correlation ρ in Tables 5.1 and 5.2. Based on the sample sizes in Table 5.1, we calculated $z_h^\dagger$ using (5.12). We then counted the number of samples out of R for which $min(z_1^\dagger, z_2^\dagger) > z_\alpha = 1.96$. The exact overall power is then $1-\beta = \{\text{number of samples such that } min(z_1^\dagger, z_2^\dagger) > 1.96\} \div R$. Results in Table 5.4 show that the exact overall powers generally relatively closely approximate the large-sample approximate power 0.8 that we used to calculate the sample sizes in our approach. Note that the sample sizes obtained using the large-sample normal approximation are conservative (with approximate powers slightly less than the exact values), which is important when one calculates sample sizes in clinical trials.

5.5 Application to PREMIER Study

In this section, we re-calculate the sample sizes for the PREMIER Study, considered in Sozu *et al.* (2012), using our method. The PREMIER Study is a multicenter, randomized,

Table 5.2: Sample sizes $n = n_{T_0} = n_{T_1}$ for overall power $1 - \beta \approx 80\%$, $\alpha = 0.025$, $Q = C = 1$, and with fixed $\delta^* = \mu^*_{T_1} - \mu^*_{T_0} = \Phi^{-1}(\pi_{T_1}) - \Phi^{-1}(\pi_{T_0})$ and $\bar{\delta}$. Marginal sample sizes E_1 and E_2 for the continuous and binary endpoints, respectively, were calculated with individual power of at least 80%.

$\frac{z_2^\dagger}{z_1^\dagger}$	$\bar{\delta}$	δ^*	π_{T_1}	π_{T_0}	$n = n_{T_0} = n_{T_1}$				E_2	E_1
					$\rho = 0.0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.8$		
0.321	0.100	0.126	0.55	0.50	1723	1692	1661	1600	1570	995
0.270	0.103	0.132	0.65	0.60	1607	1579	1553	1502	1480	902
0.155	0.112	0.150	0.75	0.70	1330	1311	1293	1261	1252	697
–0.072	0.132	0.195	0.85	0.80	926	918	912	903	901	414
0.314	0.201	0.253	0.60	0.50	426	418	411	396	389	245
0.252	0.210	0.271	0.70	0.60	385	379	373	361	356	214
0.095	0.231	0.317	0.80	0.70	310	305	302	296	296	156
–0.188	0.281	0.440	0.90	0.80	202	201	200	199	199	82
0.718	0.115	0.126	0.55	0.50	1437	1403	1368	1287	1187	995
0.673	0.119	0.132	0.65	0.60	1326	1295	1263	1190	1109	902
0.532	0.129	0.150	0.75	0.70	1088	1064	1041	987	944	697
0.252	0.151	0.195	0.85	0.80	745	733	721	698	687	414
0.72	0.232	0.253	0.60	0.50	354	345	337	317	292	245
0.644	0.242	0.271	0.70	0.60	318	311	304	286	269	214
0.462	0.266	0.317	0.80	0.70	252	247	242	230	222	156
0.126	0.323	0.440	0.90	0.80	159	157	155	152	151	82
1.909	0.160	0.126	0.55	0.50	1083	1064	1046	1010	614	995
1.832	0.165	0.132	0.65	0.60	992	974	956	920	577	902
1.641	0.179	0.150	0.75	0.70	790	774	758	722	490	697
1.276	0.211	0.195	0.85	0.80	505	493	480	451	353	414
1.900	0.322	0.253	0.60	0.50	267	262	258	249	152	245
1.796	0.336	0.271	0.70	0.60	237	232	228	219	140	214
1.554	0.370	0.317	0.80	0.70	180	176	172	164	115	156
1.075	0.450	0.440	0.90	0.80	105	102	99	93	78	82

Table 5.3: Sample sizes $n = n_{T_0} = n_{T_1}$ in Sozu *et al.* (2012) (obtained with approximate overall power of 80% using their approach) and corresponding power values $1-\beta$ obtained using our approach, for $\alpha = 0.025$, $Q = C = 1$, and with fixed $\bar{\delta}$ (as determined by fixing $z_2^\dagger/z_1^\dagger$ in Table 5.1) and $\delta^* = \mu^*_{T_1} - \mu^*_{T_0} = \Phi^{-1}(\pi_{T_1}) - \Phi^{-1}(\pi_{T_0})$. The rows are grouped according to the fixed values $z_2^\dagger/z_1^\dagger = 1, 1.5$, and 3, in Table 5.1; note, however, that these are not the actual values of $z_2^\dagger/z_1^\dagger$ for the given $\bar{\delta}, \delta^*$, and n.

$\bar{\delta}$	δ^*	π_{T_1}	π_{T_0}	$\rho = 0.0$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.8$	
				n	$1-\beta$	n	$1-\beta$	n	$1-\beta$	n	$1-\beta$
0.126	0.126	0.55	0.50	2055	0.962	2016	0.959	1980	0.958	1904	0.957
0.132	0.132	0.65	0.60	1931	0.968	1895	0.966	1862	0.964	1793	0.964
0.150	0.150	0.75	0.70	1643	0.981	1614	0.979	1588	0.978	1534	0.978
0.195	0.195	0.85	0.80	1189	0.995	1171	0.994	1154	0.994	1122	0.993
0.253	0.253	0.60	0.50	509	0.963	499	0.961	490	0.959	472	0.958
0.271	0.271	0.70	0.60	468	0.971	459	0.969	451	0.968	435	0.967
0.317	0.317	0.80	0.70	385	0.985	379	0.984	373	0.983	361	0.983
0.440	0.440	0.90	0.80	262	0.998	258	0.998	254	0.997	248	0.997
0.144	0.126	0.55	0.50	1819	0.958	1789	0.956	1760	0.955	1702	0.953
0.152	0.132	0.65	0.60	1710	0.965	1682	0.963	1656	0.961	1603	0.959
0.173	0.150	0.75	0.70	1454	0.978	1432	0.977	1411	0.976	1370	0.974
0.224	0.195	0.85	0.80	1053	0.993	1038	0.993	1025	0.992	1000	0.991
0.291	0.253	0.60	0.50	451	0.960	443	0.958	436	0.956	422	0.954
0.312	0.271	0.70	0.60	414	0.968	407	0.966	401	0.965	389	0.963
0.365	0.317	0.80	0.70	341	0.983	336	0.982	331	0.981	322	0.979
0.506	0.440	0.90	0.80	232	0.997	229	0.997	226	0.996	221	0.996
0.201	0.126	0.55	0.50	1583	0.942	1578	0.942	1574	0.941	1568	0.940
0.211	0.132	0.65	0.60	1487	0.949	1483	0.949	1479	0.948	1474	0.948
0.240	0.150	0.75	0.70	1265	0.965	1262	0.965	1259	0.964	1254	0.964
0.312	0.195	0.85	0.80	916	0.986	914	0.986	912	0.986	909	0.986
0.406	0.253	0.60	0.50	392	0.943	391	0.943	390	0.943	388	0.942
0.434	0.271	0.70	0.60	360	0.953	359	0.953	358	0.952	357	0.952
0.508	0.317	0.80	0.70	297	0.972	296	0.971	295	0.971	294	0.970
0.704	0.440	0.90	0.80	202	0.993	201	0.993	201	0.993	200	0.993

Table 5.4: Sample sizes $n = n_{T_0} = n_{T_1}$ in Table 5.1 (obtained with approximate or asymptotic overall power of 80% using the normal-logistic Gaussian copula joint model) and corresponding power values $1-\beta$ obtained using Monte Carlo approximation, for $\alpha = 0.025$ and with fixed $\bar{\delta}$ (as determined by fixing $z_2^\dagger/z_1^\dagger$ in Table 5.1) and $\delta^* = \mu^*_{T_1} - \mu^*_{T_0} = \Phi^{-1}(\pi_{T_1}) - \Phi^{-1}(\pi_{T_0})$. The rows are grouped according to the fixed values $z_2^\dagger/z_1^\dagger = 1, 1.5$, and 3, in Table 5.1.

$\bar{\delta}$	δ^*	π_{T_1}	π_{T_0}	$\rho = 0.0$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.8$	
				n	$1-\beta$	n	$1-\beta$	n	$1-\beta$	n	$1-\beta$
0.126	0.126	0.55	0.50	1306	0.819	1273	0.806	1240	0.811	1160	0.805
0.132	0.132	0.65	0.60	1184	0.820	1154	0.810	1124	0.803	1052	0.804
0.150	0.150	0.75	0.70	916	0.824	893	0.806	869	0.806	814	0.805
0.195	0.195	0.85	0.80	544	0.822	530	0.806	516	0.807	483	0.807
0.253	0.253	0.60	0.50	322	0.805	314	0.814	305	0.809	286	0.809
0.271	0.271	0.70	0.60	281	0.803	274	0.809	267	0.807	250	0.802
0.317	0.317	0.80	0.70	205	0.802	200	0.804	195	0.802	183	0.808
0.440	0.440	0.90	0.80	107	0.801	104	0.806	102	0.806	95	0.804
0.144	0.126	0.55	0.50	1156	0.804	1130	0.801	1105	0.807	1046	0.806
0.152	0.132	0.65	0.60	1048	0.802	1025	0.805	1002	0.800	948	0.805
0.173	0.150	0.75	0.70	811	0.802	792	0.802	774	0.802	733	0.807
0.224	0.195	0.85	0.80	481	0.812	471	0.812	460	0.817	436	0.816
0.291	0.253	0.60	0.50	285	0.800	278	0.801	272	0.808	258	0.810
0.312	0.271	0.70	0.60	249	0.804	243	0.810	238	0.806	225	0.802
0.365	0.317	0.80	0.70	182	0.806	178	0.800	174	0.800	165	0.804
0.506	0.440	0.90	0.80	95	0.802	93	0.808	91	0.807	86	0.803
0.201	0.126	0.55	0.50	1006	0.800	1002	0.803	998	0.800	995	0.803
0.211	0.132	0.65	0.60	912	0.808	908	0.805	905	0.802	902	0.803
0.240	0.150	0.75	0.70	705	0.802	702	0.801	700	0.808	697	0.804
0.312	0.195	0.85	0.80	419	0.805	417	0.811	416	0.803	414	0.809
0.406	0.253	0.60	0.50	248	0.806	247	0.804	246	0.813	245	0.801
0.434	0.271	0.70	0.60	217	0.810	216	0.806	215	0.811	214	0.807
0.508	0.317	0.80	0.70	158	0.804	158	0.801	157	0.801	157	0.802
0.704	0.440	0.90	0.80	83	0.807	82	0.803	82	0.810	82	0.805

double-blind, active comparator-controlled, Phase III clinical trial on patients with early aggressive rheumatoid arthritis. The study is based on two co-primary endpoints, namely, the percentage of patients who achieved a 50% improvement (i.e., ACR50=1) according to the criteria of the American College of Rheumatology (ACR), and the mean change from baseline in the modified Total Sharp Score (mTSS). Following Sozu *et al.* (2012), we are interested in determining the sample size of a future study comparing the efficacy of the two treatments combination therapy (adalimumab plus methotrexate) and monotherapy (methotrexate alone), using the two co-primary endpoints ACR50 response (binary) and mTSS (continuous).

For the Gaussian copula joint model, we assume a Gaussian latent variable for ACR50 and we treat mTSS as a Gaussian continuous outcome. This implies a bivariate CGCM for the joint distribution of ACR50 and mTSS, which is the same model used in Sozu *et al.* (2012). However, although we adopted the same model as in Sozu *et al.* (2012), our sample size methodology relies on latent-means tests. Note that the $z_2^\dagger/z_1^\dagger$ values, calculated using (5.16), are not the same as those in Table 2 of Sozu *et al.* (2012), since $\delta^* = \Phi^{-1}(\pi_{T_1}) - \Phi^{-1}(\pi_{T_0}) \neq \pi_{T_1} - \pi_{T_0}$. Our calculations show that our sample sizes, shown in Table 5.5, are smaller than those obtained by Sozu *et al.* (2012). Note as well the smaller marginal sample size $\mathrm{E}_1 = 146$, which is about 40% smaller than the sample size 231 reported in Sozu *et al.* (2012). We also computed the overall power using our method based on the sample sizes in Table 2 of Sozu *et al.* (2012), obtained using their method at 80% approximate overall power and significance level $\alpha = 0.025$. The overall powers in Table 5.6 show superiority of our method, albeit the values are only between 1% to 5% higher than 80%.

5.6 Discussion

This chapter proposed an alternative method, employing the Gaussian copula joint model introduced in Chapter 2 to construct a joint model for the correlated binary and continuous

Table 5.5: Sample sizes $n = n_{T_0} = n_{T_1}$ for endpoints mTSS (continuous) and ACR50 (binary), for overall power $1 - \beta \approx 80\%$ and $\alpha = 0.025$, with fixed $\bar{\delta}$ and δ^*. Sample sizes E_1 and E_2 for mTSS and ACR50, respectively, were calculated with individual power of at least 80%.

$\frac{z_2^\dagger}{z_1^\dagger}$	mTSS		ACR50			$n = n_{T_0} = n_{T_1}$				E_2	E_1
	δ	σ	π_{T_1}	π_{T_0}	δ^*	$\rho = 0.0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.8$		
0.022	4.4	19	0.59	0.46	0.328	305	302	299	294	293	146
–0.095	4.4	20	0.59	0.46	0.328	333	330	328	325	325	146
–0.202	4.4	21	0.59	0.46	0.328	363	361	360	358	358	146
–0.298	4.4	22	0.59	0.46	0.328	396	395	394	393	393	146

Table 5.6: Sample sizes $n = n_{T_0} = n_{T_1}$ in Sozu *et al.* (2012) (obtained with approximate overall power of 80% using their approach) for endpoints mTSS (continuous) and ACR50 (binary), and corresponding power values $1 - \beta$ obtained using our approach, for $\alpha = 0.025$, with fixed $\bar{\delta}$ and δ^*.

mTSS		ACR50			$\rho = 0.0$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.8$	
δ	σ	π_{T_1}	π_{T_0}	δ^*	n	$1-\beta$	n	$1-\beta$	n	$1-\beta$	n	$1-\beta$
4.4	19	0.59	0.46	0.328	346	0.853	340	0.849	334	0.845	323	0.836
4.4	20	0.59	0.46	0.328	369	0.843	363	0.838	358	0.835	347	0.826
4.4	21	0.59	0.46	0.328	394	0.833	389	0.830	384	0.826	374	0.817
4.4	22	0.59	0.46	0.328	422	0.826	417	0.822	413	0.819	404	0.811

endpoints, and using latent means-tests for the binary endpoints, to calculate the sample size for clinical trials with mixed binary and continuous co-primary endpoints in clinical trials Sozu *et al.* (2012). The Gaussian copula joint model is especially appropriate in this setting, where non-Gaussian endpoints are commonplace.

By introducing a latent variable framework to describe binary endpoints in our joint model, we are able to carry out tests on the difference of proportions for the binary endpoints in terms of the difference in the latent means for the corresponding latent variables. As we demonstrated empirically, our approach yields generally higher overall power than the method of Sozu *et al.* (2012), thus yielding comparatively smaller sample sizes. Hence, our method can attain the required overall power at smaller sample sizes than what Sozu *et al.*'s (2012) method requires. Our approach based on means tests on the latent structure of the binary endpoints results in a simpler and more streamlined methodology — all based on MLEs and which is no different from the methodology for the case when all the endpoints are continuous — and yields better performance for the tests in terms of power. We conjecture the reason for this superiority lies in the asymptotic approximations used in evaluating the overall power.

Chapter 6

Conclusion

6.1 Summary

The main aim of this thesis is to develop joint models for correlated mixed discrete and continuous outcomes via copulas, in general, and Gaussian copulas, in particular. Toward this goal, introduced the Gaussian copula joint model in Chapter 2 which allows the joint modeling of mixed discrete and continuous, possibly non-Gaussian, variables with the specification of the outcomes' marginal models whose dependence is independently modeled by the Gaussian copula. We then generalized the joint model to the longitudinal setting, which necessitates the modeling of associations between the outcomes (of the same or of different types) for the same subject at the same time point, and/or at different time points. This is accomplished with the incorporation of random effects (i.e., random intercepts and random slopes) in the individual (generalized) linear mixed models (LMMs/GLMMs) for the outcomes. Our approach extends the conventional LMM and GLMM methodology to non-Gaussian data which may exhibit conditional dependence. An attractive feature of the Gaussian copula mixed model is its use of copulas to separately model dependencies between outcomes, thereby preserving the outcomes' distinct marginal properties. They thus offer a flexible alternative to conventional approaches that generally rely on the assumption that outcomes, or some suitable transformations of them, follow a Gaussian distribution, or if this is not possible, that the outcomes are conditionally independent. In parametric contexts, the margins need not come from the same parametric family, allowing researchers great flexibility in modeling data of different types. The model adopts a latent variable description of discrete variables, which is an appealing means of explaining the conceptual basis underlying the discrete data. In addition, it also leads to familiar association measures for capturing

dependence between two discrete outcomes (i.e., polychoric correlation) or between mixed discrete and continuous outcomes (i.e., polyserial correlation). To overcome computational difficulties in a problems involving multiple outcomes, we introduce a pairwise maximum likelihood method (Fieuws and Verbeke, 2006). The usefulness of our methodology was illustrated through a real-data application on hourly wages, annual worked hours, and union memberships of male youth workers in the U. S.

In Chapter 3, we adapted the Gaussian copula mixed model in Chapter 2 to the joint analysis of clustered mixed binary and continuous outcomes in developmental toxicology. The resulting joint model possesses three ideal properties required in the analysis of data from developmental toxicity studies. First, the model yields interpretable outcome-specific dose-response models for the mixed outcomes; this is accomplished by using modeling the dependence between the outcomes via the Gaussian copula, which allowed the independent specification of the marginal models for the outcomes. Second, the model accounts for cluster effects by the introduction of cluster-level random effects whose distribution (possibly non-normal) may also be flexibly constructed via the Gaussian copula. Last, the model directly incorporates within-subject association between the mixed outcomes, a necessary ingredient in quantitative risk assessment. The model adopts a latent variable description of the binary outcome, a quite appealing formulation of such binary outcomes as fetal malformation in developmental toxicology. This also makes statistical sense, as common binary regression models (e.g., logistic, probit) have parallel formulations in terms of latent variables (Teixeira-Pinto and Normand, 2009).

The proposed methodology was applied to the EG mice data using several versions of the Gaussian copula mixed model, among which are that with a robit mixed model for the binary outcome and a Gaussian continuous outcome, and another with a logit mixed model for the binary outcome and a t-distributed continuous outcome. These models extend and generalize, for example, Lin *et al.*'s (2010), Gueorguieva and Agresti's (2001), and Regan

and Catalano's (1999) approaches. In particular, our model is a more general formulation of the correlated probit model of Gueorguieva and Agresti (2001) that allows for non-normal residual errors and non-normal random effects. In addition, our approach is not based on factorization, like Lin *et al.*'s (2010), and as such, does not suffer from the inadequacies of factorization models (e.g., non-invariance to direction of factorization, asymmetrical treatment of outcomes). Our methodology can be easily implemented using standard statistical software/packages; in particular, we used PROC NLMIXED in SAS in our simulations and analysis.

In Chapter 4, we adapted the Gaussian copula joint model in Chapter 2 to the case of mixed continuous and discrete outcomes, where the latter are transformed into continuous variables via "continuous-ation" (Denuit and Lambert, 2005). This avoids assuming a latent structure underlying the discrete data, like we did in Chapters 2 and 3, which may not work in applications involving count and nominal categorical data. Heinen and Rengifo's (2007,2008) surrogate likelihood approach to estimation was adopted for the model, and simulation results on the finite-sample relative bias and efficiency of resulting estimates indicate the adequacy of the method. Data on burn injuries, which were analyzed earlier by de Leon and Wu (2011) and Song *et al.* (2009), were revisited and used to illustrate the applicability of our approach.

Finally, Chapter 5 discussed a methodology for sample size calculation when designing confirmatory and superiority clinical trials with multiple mixed binary and continuous co-primary endpoints. The approach, based on the Gaussian copula joint model in Chapter 2, provides an alternative to Sozu *et al.*'s (2012) recent work that adopted the Gaussian distribution-based CGCM, which is a special case of our model. In addition, we introduced latent means-tests for the binary outcomes as alternatives to the Pearson-type statistics used by Sozu *et al.* (2012) to test for difference in proportions. As we demonstrated empirically, our method has higher overall power than the method of Sozu *et al.* (2012), and can attain

the required overall power at sample sizes smaller than what their method requires.

6.2 Future research

The approaches we outlined in the thesis make exclusive use of the Gaussian copula to describe the mixed-outcome joint model, as in Song *et al.* (2009). Even though Gaussian copulas are mathematically and computationally tractable and have been used in a variety of applications, they may not work in all cases. Other copula families could be considered as well, and the choice of the copula could then be investigated via various model selection criteria.

Computationally more efficient approaches to likelihood evaluation and optimization needed in likelihood estimation need to be developed. In Chapter 2, a pairwise likelihood approach was used to alleviate the computational demands of full likelihood analysis, while in Chapter 4, a surrogate likelihood approach was implemented for the Gaussian copula joint model with "continued" binary data. While it appears that these methods worked in our settings, a more complete investigation of their properties needs to be undertaken. Recent work by Nikoloulopoulos (2013) seems to suggest the adequacy of the methods; however, the bias in the scale and correlation parameters that we observed will have to be addressed at some future time.

Finally, there is a need to device theoretically sound methods of dealing with missing value problems in mixed data via the Gaussian copula joint model. Simply performing separate analyses on the discrete and continuous outcomes does not work because different sets of observations may be used in each analysis, and interpreting the results then becomes difficult. Little and Schluchter (1985), along with Belin *et al.* (1999), Fitzmaurice and Laird (1997), and Schafer (1997), addressed this issue in the case of GLOMs, where the missing data are assumed to be "missing at random" (MAR) (Little and Rubin, 1987). It may be possible to extend their methods to the Gaussian copula joint models, including the Gaussian

copula mixed model. An approach that can deal with non-monotone and non-MAR missing data in this context should be considered.

Appendix A

SAS codes for Gaussian copula mixed models

This appendix includes two sets of sample SAS codes: the first one is for maximum likelihood estimation of the logit-normal-normal Gaussian copula mixed model with shared random intercepts and slopes in Chapter 2, and the second is for obtaining the MLEs of the robit-t Gaussian copula mixed model with correlated random effects, with only linear dose effects in Chapter 3.

A.1 Logit-normal-normal Gaussian copula mixed model

```
proc nlmixed /* logit-normal-normal Gaussian copula mixed model */
data=dataw method=ISAMP qpoints=200 tech=NRRIDG NOAD;
pi=constant('pi');
mut = a0+a1*time+lamIs*bI+lamSs*bS*time;
p=exp(mut)/(1+exp(mut));
mun1 = b10+b11*time+lamI1*bI+lamS1*bS*time;
mun2 = b20+b21*time+lamI2*bI+lamS2*bS*time;
/* Case I */
if (union=0) then do;
qs = probit(1-p);
q1 = probit(cdf('normal', lwage, mun1, sdn1));
q2 = probit(cdf('normal', lhours, mun2, sdn2));
rho1s12 = (rho1s1-rho1s2*rho12)/sqrt(1-rho1s2*rho1s2)/sqrt(1-rho12*rho12);
q1s2 = (qs-rho1s2*q2)/sqrt(1-rho1s2*rho1s2);
q12 = (q1-rho12*q2)/sqrt(1-rho12*rho12);
```

```
q1s12 = (q1s2-rho1s12*q12)/sqrt(1-rho1s12*rho1s12);
pdfnor1 =1/(sdn1*sqrt(2*pi))*exp(-(lwage-mun1)*(lwage-mun1)/(2*sdn1*sdn1));
pdfnor2 =1/(sdn2*sqrt(2*pi))*exp(-(lhours-mun2)*(lhours-mun2)/(2*sdn2*sdn2));
const1 = 0.5*log(1-rho12*rho12);
const2 = log(pdf('normal', q12));
const3 = log(pdf('normal', q1));
Phi1 = cdf('normal', q1s12);
llik1 = log(pdfnor1) + log(pdfnor2) - const1 + const2 - const3 + log(Phi1);
llik = log(pdfnor1) + log(pdfnor2) - const1 + const2 - const3 + log(Phi1);
end;
/* Case II */
if (union=1) then do;
qs = probit(1-p);
q1 = probit(cdf('normal', lwage, mun1, sdn1));
q2 = probit(cdf('normal', lhours, mun2, sdn2));
rho1s12 = (rho1s1-rho1s2*rho12)/sqrt(1-rho1s2*rho1s2)/sqrt(1-rho12*rho12);
q1s2 = (qs-rho1s2*q2)/sqrt(1-rho1s2*rho1s2);
q12 = (q1-rho12*q2)/sqrt(1-rho12*rho12);
q1s12 = (q1s2-rho1s12*q12)/sqrt(1-rho1s12*rho1s12);
pdfnor1 =1/(sdn1*sqrt(2*pi))*exp(-(lwage-mun1)*(lwage-mun1)/(2*sdn1*sdn1));
pdfnor2 =1/(sdn2*sqrt(2*pi))*exp(-(lhours-mun2)*(lhours-mun2)/(2*sdn2*sdn2));
const1 = 0.5*log(1-rho12*rho12);
const2 = log(pdf('normal', q12));
const3 = log(pdf('normal', q1));
Phi1 = cdf('normal', q1s12);
llik2 = log(pdfnor1) + log(pdfnor2) - const1 + const2 - const3 + log(1-Phi1);
```

```
llik =llik2;
end;
model zzz   general(llik);
random bI bS   normal([0, 0], [1, rhob, 1]) subject=nr;
estimate'alpha0' a0;
estimate'alpha1' a1;
estimate'lamIs' lamIs;
estimate'lamSs' lamSs;
estimate'beta10' b10;
estimate'beta11' b11;
estimate'lamI1' lamI1;
estimate'lamS1' lamS1;
estimate'sdn1' sdn1;
estimate'beta20' b20;
estimate'beta21' b21;
estimate'lamI2' lamI2;
estimate'lamS2' lamS2;
estimate'sdn2' sdn2;
estimate'rhob' rhob;
estimate'rho1s1' rho1s1;
estimate'rho1s2' rho1s2;
estimate'rho12' rho12;
ods output ParameterEstimates=pars;
ods output AdditionalEstimates=margest;
run;
```

A.2 Robit-t Gaussian copula mixed model

```
proc nlmixed /* robit-t Gaussian copula mixed model */
data=EG method=ISAMP qpoints=200 tech=NRRIDG NOAD;
pi=constant('pi');
mut = a1 + a2*dose + bD;
mun = b1 + b2*dose + bC;
p1=min(max(cdf('T',(-mut), dfint),1.0E-50),0.9999999999);
q1=probit(p1);
p2=min(max(cdf('T',(weight-mun)/sdn, dfn),1.0E-50),0.9999999999);
q2=probit(p2);
pdfT = gamma((dfn+1)/2)*(1+(weight-mun)**2/(sdn*sdn*dfn))**(-(dfn+1)/2)
/(sdn*sqrt(pi*dfn))/(gamma(dfn/2));
cdf2 = probnorm((q1-rho*q2)/sqrt(1-rho*rho));
llik=malformation*(log(1.00001- cdf2)+log(pdfT))
+(1-malformation)*(log(cdf2)+log(pdfT));
sdDv=sdD*sdD;
sdCv=sdC*sdC;
sdDC=sdD*sdC*rhob;
model zzz   general(llik);
random bD bC   normal([0, 0], [sdDv, sdDC, sdCv]) subject=litter_id;
estimate'alpha1' a1;
estimate'alpha2' a2;
estimate'sdD' sdD;
estimate'dfint' dfint;
estimate'beta1' b1;
estimate'beta2' b2;
```

```
estimate'sdn' sdn;

estimate'sdC' sdC;

estimate'dfn' dfn;

estimate'rhob' rhob;

estimate'rho' rho;

ods output ParameterEstimates=pars;

ods output AdditionalEstimates=margest;
```

Appendix B

Derivation of conditional mean (3.13)

As shown in (3.11) in Chapter 3, the conditional mean fetal weight as a function of dose, given fetal malformation, is

$$E(Y_{ih}|X_{ih}=x_{ih}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y\frac{f_{X_{ih},Y_{ih}|B_i^*,B_i}(x_{ih},y_{ih}|b_i^*,b_i)}{P(X_{ih}=x_{ih}|B_i^*=b_i^*)} f_{B_i^*,B_i}(b_i^*,b_i) \times dy db_i^* db_i. \tag{B.1}$$

With $Y_{ih}|B_i \sim normal(\mu_i(B_i),\sigma^2)$, (3.12) simplifies as

$$\begin{aligned} f_{X_{ih},Y_{ih}|B_i^*,B_i}(x_{ih},y_{ih}|b_i^*,b_i) &= \left\{1-\Phi\left(\frac{\frac{\widetilde{\rho}}{\sigma}\{y_{ih}-\mu_i(b_i)\}-\Phi^{-1}\{u_{ih}(b_i^*)\}}{\sqrt{1-\widetilde{\rho}^2}}\right)\right\}^{1-x_{ih}} \\ &\times\left\{\Phi\left(\frac{\frac{\widetilde{\rho}}{\sigma}\{y_{ih}-\mu_i(b_i)\}-\Phi^{-1}\{u_{ih}(b_i^*)\}}{\sqrt{1-\widetilde{\rho}^2}}\right)\right\}^{x_{ih}} \\ &\times\frac{1}{\sigma}\phi\left(\frac{y_{ih}-\mu_i(b_i)}{\sigma}\right). \end{aligned} \tag{B.2}$$

Based on entries (10,010.8) and (10,011.3) in Owen (1980), the conditional mean $E(Y_{ih}|X_{ih}=0,b_i^*,b_i)$ of Y_{ih}, given random effects b_i^* and b_i, and with $X_{ih}=0$, can be expressed as

$$\begin{aligned} E(Y_{ih}|X_{ih}=0,b_i^*,b_i) &= \frac{1}{u_{ih}(b_i^*)}\int_{-\infty}^{+\infty}\{\sigma z+\mu_i(b_i)\}\phi(z)\Phi(a+bz)dz \\ &= \frac{1}{u_{ih}(b_i^*)}\left\{\sigma\frac{b}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)+\mu_i(b_i)\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)\right\} \\ &= \frac{1}{u_{ih}(b_i^*)}\left[-\sigma\widetilde{\rho}\phi\{\Phi^{-1}(u_{ih}(b_i^*))\}+\mu_i(b_i)\Phi(\Phi^{-1}(u_{ih}))\right] \\ &= \mu_i(b_i)-\frac{\widetilde{\rho}\sigma}{u_{ih}(b_i^*)}\phi\{\Phi^{-1}(u_{ih}(b_i^*))\}, \end{aligned} \tag{B.3}$$

where $z=(y-\mu_i(b_i))/\sigma$, $a=\Phi^{-1}(u_{ih}(b_i^*))/\sqrt{1-\widetilde{\rho}^2}$, and $b=-\widetilde{\rho}/\sqrt{1-\widetilde{\rho}^2}$. Similarly, we obtain the conditional mean $E(Y_{ih}|X_{ih}=1,b_i^*,b_i)$ of Y_{ih}, given random effects b_i^* and b_i, and with $X_{ih}=1$, as

$$E(Y_{ih}|X_{ih}=1,b_i^*,b_i) = \frac{1}{1-u_{ih}(b_i^*)}\int_{-\infty}^{+\infty}\{\sigma z+\mu_i(b_i)\}\phi(z)\Phi(a+bz)dz$$

$$
\begin{aligned}
&= \frac{1}{1-u_{ih}(b_i^*)}\left\{\sigma\frac{b}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)+\mu_i(b_i)\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)\right\} \\
&= \frac{1}{1-u_{ih}(b_i^*)}\left[\sigma\widetilde{\rho}\phi\{-\Phi^{-1}(u_{ih}(b_i^*))\}+\mu_i(b_i)\Phi(-\Phi^{-1}(u_{ih}))\right] \\
&= \mu_i(b_i)+\frac{\widetilde{\rho}\sigma}{1-u_{ih}(b_i^*)}\phi\{\Phi^{-1}(u_{ih}(b_i^*))\}. \qquad \text{(B.4)}
\end{aligned}
$$

Combining (B.3) and (B.4), we thus get (3.13).

Bibliography

Balakrishnan, N. and C-D. Lai (2009). *Continuous Bivariate Distributions*. 2nd ed.. Springer.

Belin, T. R., M-Y. Hu, A. S. Young and O. Grusky (1999). Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Statistics in Medicine* **18**, 3123–3135.

Bodnar, O., T. Bodnar and A. K. Gupta (2010). Estimation and inference for dependence in multivariate data. *Journal of Multivariate Analysis* **101**, 869–881.

Breedveld, F. C., M. H. Weisman, A. F. Kavanaugh, S. B. Cohen, K. Pavelka, R. van Vollenhoven, J. Sharp, J. L. Perez and G. T. Spencer-Green (2006). The PREMIER Study: a multicenter, randomized, double-blind clinical trial of combination therapy with adalimumab plus methotrexate versus methotrexate alone or adalimumab alone in patients with early, aggressive rheumatoid arthritis who had not had previous methotrexate treatment. *Arthritis & Rheumatism* **51**, 26–37.

Catalano, P. J. (1997). Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine* **16**, 883–900.

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

de Leon, A. R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters* **75**, 49–57.

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

de Leon, A. R. and K. C. Carrière (2007). General mixed-data model: extension of general location and grouped continuous models. *Canadian Journal of Statistics* **35**, 533–548.

de Leon, A. R. and K. Carrière Chough (2010). Mixed-outcome data. In: *Encyclopedia of Biopharmaceutical Statistics* (S. C. Chow, Ed.). 3rd ed.. pp. 817–822. Chapman & Hall.

de Leon, A. R. and K. Carrière Chough (2013). *Analysis of Mixed Data: Methods & Applications*. Chapman & Hall/CRC.

de Leon, A. R., B. Wu and N. Withanage (2012). Gaussian copula-generated distributions for correlated discrete and continuous variables. Under review.

Denuit, M. and P. Lambert (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis* **93**, 40–57.

Dobra, A. and A. Lenkoski (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics* **5**, 969–993.

Faes, C. (2013). Hierarchical modeling of endpoints of different types with generalized linear mixed models. In: *Analysis of Mixed Data: Methods & Applications* (A. R. de Leon and K. Carrière Chough, Eds.). pp. 125–138. Chapman & Hall/CRC.

Faes, C., Geys H. and P. Catalano (2009). Joint models for continuous and discrete longitudinal data. In: *Longitudinal Data Analysis* (G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs, Eds.). pp. 327–348. Chapman & Hall.

Faes, C., M. Aerts, G. Molenberghs, H. Geys, G. Teuns and L. Bijnens (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Statistics in Medicine* **27**, 4408–4427.

Fedorov, V., Y. Wu and R. Zhang (2012). Optimal dose-finding designs with correlated continuous and discrete responses. *Statistics in Medicine* **31**, 217–234.

Fieuws, S. and G. Verbeke (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* **62**, 424–431.

Fitzmaurice, G. M. and N. M. Laird (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* **90**, 845–852.

Fitzmaurice, G. M. and N. M. Laird (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics* **53**, 110–122.

Genest, C. and J. Nešlehovà (2007). A primer on copula for count data. *ASTIN Bulletin* **37**, 475–515.

George, E. O., D. Armstrong, P. J. Catalano and D. K. Srivastava (2007). Regression models for analyzing clustered binary and continuous outcomes under an assumption of exchangeability. *Journal of Statistical Planning & Inference* **137**, 3462–3474.

Gueorguieva, R. (2013). Random effects models for joint analysis of repeatedly measured discrete and continuous outcomes. In: *Analysis of Mixed Data: Methods & Applications* (A. R. de Leon and K. Carrière Chough, Eds.). pp. 109–123. Chapman & Hall/CRC.

Gueorguieva, R. V. and A. Agresti (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association* **96**, 1102–1112.

Hausman, J. and D. Wise (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* **45**, 319–339.

Heinen, A. and E. Rengifo (2007). Multivariate autoregressive modeling of time series count data using copulas. *Journal of Emprical Finance* **14(4)**, 564–583.

Heinen, A. and E. Rengifo (2008). Multivariate reduced rank regression in nonGaussian contexts, using copulas. *Computational Statistics & Data Analysis* **52**, 2931–2944.

Hirakawa, A. (2012). An adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase I oncology trials. *Statistics in Medicine* **31**, 516–532.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall.

Kang, J. and Y. Yang (2013). Joint modeling of mixed count and continuous longitudinal data. In: *Analysis of Mixed Data: Methods & Applications* (A. R. de Leon and K. Carrière Chough, Eds.). pp. 63–79. Chapman & Hall/CRC.

Klaassen, C. A. J. and J. A. Wellner (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli* **3**, 55–77.

Kugiumtzis, D. and E. Bora-Senta (2010). Normal correlation coefficient of non-normal variables using piecewise linear approximation. *Computational Statistics* **25**, 645–662.

Li, J. and W. K. Wong (2011). Two-dimensional toxic dose and multivariate logistic regression, with application to decompression sickness. *Biostatistics* **12**, 143–155.

Lin, L., D. Bandyopadhyay, S. R. Lipsitz and D. Sinha (2010). Association models for clustered data with binary and continuous responses. *Biometrics* **66**, 287–293.

Lipsitz, S. R., K. B. G. Dear and L. Zhao (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics* **50**, 842–846.

Little, R. J. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley.

Little, R. J. and M. D. Schluchter (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 496–512.

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* (A. Gelman and X-L. Meng, Eds.). pp. 227–238. Wiley.

Machado, J. and J. Santos Silva (2005). Quantiles for counts. *Journal of the American Statistical Association* **100(472)**, 1226–1237.

Madsen, L. (2009). Maximum likelihood estimation of regression parameters with spatially dependent discrete data. *Journal of Agricultural, Biological & Environmental Statistics* **14**, 375–391.

Madsen, L. and Y. Fang (2010). Joint regression analysis for discrete longitudinal data. *Biometrics* **67**, 1171–1176.

Madsen, L. and Y. Fang (2011). Response to the Letter to the Editor by Sabo, R. T. and Chaganty, N. R.. *Biometrics* **67**, 1670–1671.

Masarotto, G. and C. Varin (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics* **6**, 1517–1549.

McCulloch, C. (2008). Joint modeling of mixed outcome types using latent variables. *Statistical Methods in Medical Research* **17**, 53–73.

Meester, S. and J. MacKay (1994). A parametric model for cluster correlated categorical data. *Biometrics* **50**, 954–963.

Mesfioui, M. and A. Tajar (2005). On the properties of some nonparametric concordance measures in the discrete case. *Journal of Nonparametric Statistics* **17**, 541–554.

Muthèn, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**, 115–132.

Najita, J. S., Y. Li and P. J. Catalano (2009). A novel application of a bivariate regression model for binary and continuous outcomes to studies of fetal toxicity. *Journal of the Royal Statistical Society–C* **58**, 555–573.

Nelsen, R. B. (2006). *An Introduction to Copulas*. 2nd ed.. Springer.

Nešlehovà, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis* **95**, 544–567.

Nikoloulopoulos, A. K. (2012). Comment on "Two-dimensional toxic dose and multivariate logistic regression, with application to decompression sickness," by J. Li and W. K. Wong. *Biostatistics* **13**, 1–3.

Nikoloulopoulos, A. K. (2013). On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood. Under review.

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

Nikoloulopoulos, A. K. and D. Karlis (2009). Finite normal mixture copulas for multivariate discrete data modeling. *Journal of Statistical Planning & Inference* **139**, 3878–3890.

Nikoloulopoulos, A. K. and D. Karlis (2010). Modeling multivariate count data using copulas. *Communications in Statistics–Simulations & Computation* **39**, 172–187.

Olkin, I. and R. F. Tate (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **32**, 448–465. (with correction in **36**, 343–344).

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics–Simulations & Computation* **B9**, 389–419.

Parzen, M., S. Ghosh, S. Lipsitz, D. Sinha, G. M. Fitzmaurice, B. K. Mallick and J. G. Ibrahim (2011). A generalized linear mixed model for longitudinal binary data with marginal logit link function. *Annals of Applied Statistics* **5**, 449–467.

Poon, W.-Y. and S.-Y. Lee (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika* **52**, 409–430.

Price, C. J., C. A. Kimmel, R. W. Tyl and M. C. Marr (1985). The developmental toxicity of ethylene glycol in rats and mice. *Toxicology and Applied Pharmacology* **81**, 113–127.

Regan, M. M. and P. J. Catalano (1999). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics* **55**, 760–768.

Renard, D., G. Molenberghs and H. Geys (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis* **44**, 649–667.

Rudnick, M. R., C. Davidson, W. Laskey, J. L. Stafford, P. F. Sherwin and VALOR Trial Investigators (2008). Nephrotoxicity of iodixanol versus ioversol in patients with chronic kidney disease: the Visipaque Angiography/Interventions with Laboratory Outcomes in Renal Insufficiency (VALOR) Trial. *American Heart Journal* **156**, 776–782.

Ryan, L. M. (2002). Issues in modeling clustered data. In: *Topics in Modelling of Clustered Data* (M. Aerts, H. Geys, G. Molenberghs and L. Ryan, Eds.). pp. 37–45. Chapman & Hall/CRC.

Sabo, R. T. and N. R. Chaganty (2011). Letter to the Editor on "Joint regression analysis for discrete longitudinal data," by Madsen, L. and Fang, Y.. *Biometrics* **67**, 1669–1670.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC.

Shi, P. and E. A. Valdez (2012). Longitudinal modeling of insurance claim counts using jitters. *Scandinavian Actuarial Journal*. To appear.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer.

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

Song, P. X.-K., P. Zhang and A. Qu (2007). Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. *Statistica Sinica* **17**, 929–943.

Sozu, T., T. Sugimoto and T. Hamasaki (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *Journal of Biopharmaceutical Statistics* **21**, 650–668.

Sozu, T., T. Sugimoto and T. Hamasaki (2012). Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometrical Journal* **54**, 716–729.

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

Trégouët, D.-A., P. Ducimetière, V. Bocquet, S. Visvikis, F. Soubrier and L. Tiret (1999). A parametric copula model for analysis of familial binary data. *American Journal of Human Genetics* **64**, 886–893.

Trivedi, P. and D. Zimmer (2007). Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics* **1**, 1–111.

Vella, F. and M. Verbeek (1998). Whose wages do unions raise? a dynamic model of unionism and wage rate determination for young men. *Journal of Applied Econometrics* **13**, 163–183.

Wu, B., A. R. de Leon and N. Withanage (2013). Joint analysis of mixed discrete and continuous outcomes via copulas. In: *Analysis of Mixed Data: Methods & Applications* (A. R. de Leon and K. Carrière Chough, Eds.). pp. 139–156. Chapman & Hall/CRC.

Wu, B. and A. R. de Leon (2013*a*). Gaussian copula mixed models for clustered mixed outcomes, with application in developmental toxicology. *Journal of Agricultural, Biological & Environmental Statistics*. To appear.

Wu, B. and A. R. de Leon (2013*b*). Letter to the Editor re: "Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables," by Sozu, T., Sugimoto, T., & Hamasaki, T.. *Biometrical Journal*. To appear.

Zimmer, D. and P. Trivedi (2006). Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. *Journal of Business & Economic Statistics* **24**, 63–76.