Fall 12-22-2010

A Copula-Based Method for Analyzing Bivariate Binary Longitudinal Data

Seunghee Baek
University of Pennsylvania, seunghee@mail.med.upenn.edu

Recommended Citation

Baek, Seunghee, "A Copula-Based Method for Analyzing Bivariate Binary Longitudinal Data" (2010). *Publicly Accessible Penn Dissertations*. 1564.
http://repository.upenn.edu/edissertations/1564

A Copula-Based Method for Analyzing Bivariate Binary Longitudinal Data

Abstract

The work presented as part of this dissertation is primarily motivated by a randomized trial for HIV serodiscordant couples. Specifically, the Multisite HIV/STD Prevention Trial for African American Couples is a behavioral modification trial for African American, heterosexual, HIV discordant couples. In this trial, investigators developed and evaluated a couple-based behavioral intervention for reducing risky shared sexual behaviors and collected retrospective outcomes from both partners at baseline and at 3 follow-ups to evaluate the intervention efficacy. As the outcomes refer to the couples' shared sexual behavior, couples' responses are expected to be correlated, and modeling approaches should account for multiple sources of correlation: within-individual over time as well as within-couple both at the same measurement time and at different times. This dissertation details the novel application copulas to modeling dyadic, longitudinal binary data to estimate reliability and efficacy. Copulas have long been analytic tools for modeling multivariate outcomes in other settings. Particularly, we selected a mixture of max-infinitely divisible (max-id) copula because it has a number of attractive analytic features.

The dissertation is arranged as follows: Chapter 2 presents a copula-based approach in estimating the reliability of couple self-reported (baseline) outcomes, adjusting for key couple-level baseline covariates; Chapter 3 presents an extension of the max-id copula to model longitudinal (two measurement occasions), binary couples data; Chapter 4 further extends the copula-based model to accommodate more than two repeated measures in a different application examining two clinical depression measures. In this application, we are interested in estimating whether there are differential treatment effects on two different measures of depression, longitudinally.

The copula-based modeling approach presented in this dissertation provides a useful tool for investigating complex dependence structures among multivariate outcomes as well as examining covariate effects on the marginal distribution for each outcome. The application of existing statistical methodology to longitudinal, dyad-based trials is an important translational advancement. The methods presented here are easily applied to other studies that involve multivariate outcomes measured repeatedly.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Epidemiology & Biostatistics

First Advisor
Scarlett L. Bellamy

Second Advisor
Andrea B. Troxel

Keywords
bivariate binary outcomes, longitudinal data, copula, HIV study

Subject Categories
Biostatistics | Categorical Data Analysis | Longitudinal Data Analysis and Time Series | Multivariate Analysis | Statistical Models

A COPULA-BASED METHOD FOR ANALYZING BIVARIATE BINARY LONGITUDINAL DATA

Seunghee Baek

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2010

Supervisor of Dissertation	Co-supervisor
*Signature*____________________	*Signature*____________________
Scarlett L. Bellamy	Andrea B. Troxel
Associate Professor of Biostatistics	Associate Professor of Biostatistics

Graduate Group Chairperson

*Signature*____________________

Daniel F. Heitjan, Professor of Biostatistics

Dissertation Committee

Thomas R. Ten Have, Professor of Biostatistics

John B. Jemmott III, Professor of Psychiatry

Acknowledgments

I would like to thank all of people who have made this dissertation possible and what I am now. Writing PhD dissertation has been one of the most significant academic challenges. However, thanks to support and guidance of the following people, the past 4 years of my life have been built with the fulfillment of this dissertation work.

I am heartily thankful to my supervisors, Dr. Bellamy and Dr. Troxel, for their guidance, support, and understanding. I have felt fortunate to have such supervisors who always listen and give me advice. Especially, whenever I was frustrated, Dr. Bellamy has been always there to encourage me, help me overcome the difficulties and finally complete this dissertation work. I would also like to thank Dr. Troxel for her mentorship and commitment that made me confident and keep pursuing our research goal. This dissertation would not have been possible without their guidance and enthusiasm. In addition, I thank the other members of my committee, Dr. Ten Have and Dr. Jemmott, for their kind comments and help to improve the quality of this work.

My thanks also go out to Dr. Bae, who is my research assistantship supervisor at

Radiation Therapy Oncology Group (RTOG), American College of Radiology. I am grateful for her supervision and guidance, which provides a valuable life experience in my first career as a biostatistician. In addition, I thank Dr. Heitjan to make his efforts to support graduate students throughout the PhD program.

My graduate studies would not have been the same without my friend, Jing, who have shared everything and relied on each other during the past 4 years. I cannot imagine completing the entire process of PhD program without her. Also, I would like to show special thanks to Yimei, who was always willing to help and provide useful insight. In addition, many thanks go out to my friend, Yoon. She has been together like family and a mental supporter.

I would like to sincerely thank my parents and parents-in-law. Their support, quite patience and unwavering love enabled me to complete this dissertation. I greatly value their understanding and belief in me.

My tons of thanks go out to my husband, Goo, for his encouragement, support and love. Under his faith in me, I gained so much drive and an ability to tackle challenges head on. Completing this dissertation was not a challenge only for me, but for us and we could finally say, “We did it!”. I appreciate his devotion and endurance of such periods since I know no one can do that. I also want to thank my two little boys, who made the past 4 years happy and full of love.

Lastly, I thank God to make all of these possible, send such good people to me and enable me to overcome obstacles. I realize the Lord has been always with me to answer and give me the strength, and to him I dedicate this dissertation.

ABSTRACT

A COPULA-BASED METHOD FOR ANALYZING BIVARIATE BINARY LONGITUDINAL DATA

Seunghee Baek

Scarlett L. Bellamy

The work presented as part of this dissertation is primarily motivated by a randomized trial for HIV serodiscordant couples. Specifically, the Multisite HIV/STD Prevention Trial for African American Couples is a behavioral modification trial for African American, heterosexual, HIV discordant couples. In this trial, investigators developed and evaluated a couple-based behavioral intervention for reducing risky shared sexual behaviors and collected retrospective outcomes from both partners at baseline and at 3 follow-ups to evaluate the intervention efficacy. As the outcomes refer to the couples' shared sexual behavior, couples' responses are expected to be correlated, and modeling approaches should account for multiple sources of correlation: within-individual over time as well as within-couple both at the same measurement time and at different times. This dissertation details the novel application copulas to modeling dyadic, longitudinal binary data to estimate reliability and efficacy. Copulas have long been analytic tools for modeling multivariate outcomes in other settings. Particularly, we selected a mixture of max-infinitely divisible (max-id) copula because it has a number of attractive analytic features.

The dissertation is arranged as follows: Chapter 2 presents a copula-based approach in estimating the reliability of couple self-reported (baseline) outcomes, adjusting for key couple-level baseline covariates; Chapter 3 presents an extension of the max-id copula to model longitudinal (two measurement occasions), binary couples data; Chapter 4 further extends the copula-based model to accommodate more than two repeated measures in a different application examining two clinical depression measures. In this application, we are interested in estimating whether there are differential treatment effects on two different measures of depression, longitudinally.

The copula-based modeling approach presented in this dissertation provides a useful tool for investigating complex dependence structures among multivariate outcomes as well as examining covariate effects on the marginal distribution for each outcome. The application of existing statistical methodology to longitudinal, dyad-based trials is an important translational advancement. The methods presented here are easily applied to other studies that involve multivariate outcomes measured repeatedly.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The motivation for this dissertation work comes from a clinical trial for HIV serodiscordant couples. Specifically, the Multisite HIV/sexually transmitted disease (STD) Prevention Trial for African American Couples (AAC) is a behavioral modification trial for African American, heterosexual, HIV discordant couples, whose goal is to decrease risky sexual behaviors and increase health promoting behaviors among couples. The main purpose of the study is to examine the efficacy of a couple-focused HIV/STD risk reduction intervention versus an individual-focused health promotion intervention in reducing sexual risk behaviors and STD incidence. To accomplish the study's goals, reliable assessments of HIV sexual risk behaviors are critical. In this study, both partners are assessed independently for a number of shared sexual behaviors with their study partners. Thus, estimating the magnitude of the concordancy and discordancy among couple outcomes will allow us to measure the reliability of each partner's self-reported behavior. Potential issues with many self-report measures are their susceptibility to response bias, that is, a tendency for subjects to over/or under-report outcomes for a number of reasons. Therefore, developing such a tool for estimating reliability is an important issue in dealing with self-reported data. Copulas have been popular tools for modeling multivariate outcomes, since they allow us to construct the dependence structure separately from the marginal probabilities. Among several copulas, we select a max-id copula since it has desirable properties for modeling multivariate discrete data in that it has flexible positive dependence structures and has a closed form cumulative distribution function (cdf). Therefore, in Chapter 2, we will present a novel application using a mixture of max-infinitely divisible (max-id) bivariate copulas to quantify dependence measures by estimating dependency as a proxy for reliability among couples' outcomes in the context of a couple-based behavioral modification trial. Furthermore, the proposed approach provides an estimation procedure to explore individual-level factors that may be associated with the estimated dependency. We also conduct simulation studies to demonstrate that this approach performs well in terms of bias and coverage by estimating the covariate effect on the dependence measure as well as on the marginal probabilities. Another interesting feature of the motivating example is its complex dependence structure. Unlike other bivariate outcomes, the couples' outcomes of self-reported sexual behavior at the same time may be more correlated than the repeated measures within individual. Thus, ignoring the correlation between partner's outcomes at different time points may lead to invalid inferences regarding measures of effect. Therefore, the second aim of this dissertation is to develop statistical methods to estimate the longitudinal intervention effect as well as handle the complex correlation structure in a statistically feasible way. We propose a copula-based approach which is an extension of a multivariate logit model proposed in Nikoloulopoulos and Karlis (2008), and used in Chapter 2, which uses a mixture of max-infinitely divisible (max-id) bivariate copulas. This extension makes it applicable to bivariate longitudinal data by incorporating time as a regression parameter as well as constructing a complex dependence structure via the copula parameters. Through a number of simulations, we evaluate this approach by estimating the covariate effects on the marginal probabilities and of copula parameters in terms of bias and coverage. In Chapter 4, as an extension of Chapter 3, we develop a copula-based method to model bivariate longtidutinal binary outcomes to incorporate more than two repeated measures. In Chapter 3, we focus on evaluating the extension of max-id copula as a statistical tool for modeling longitudinal couples' outcomes, while in Chapter 4 we focus on generalizing the approach and the feasibility of modeling bivariate outcomes with additional repeated measures. Diverting from the HIV prevention trial, we illustrate the extended method using longitudinal data on depression among subjects treated in primary care practices using two outcomes, the diagnosis of major depressive disorder (MDD) and the Hamilton rating scale for depression (HAMD). Together, the work developed in this dissertation provides a novel application of the bivariate max-id copula approach to modeling dependence, specifically to evaluate reliability of paired responses. In addition, the approach is developed for more complicated longitudinal studies with potentially complex correlation structures, and provides a unified approach to handling clustered, correlated, repeated binary data.

Chapter 2

A copula approach for estimating the reliability of self-reported sexual behaviors among HIV serodiscordant couples

2.1 Introduction

Understanding relationships among multivariate outcomes is a fundamental problem in statistical science. In longitudinal and/or cluster-randomized trials, the dependency among outcomes may not be of primary interest, but it must be accounted for in order to make valid inference. In other settings where the dependency among outcomes is of primary interest, copulas have become an increasingly popular analysis tool.

The motivation for our work comes from a clinical trial for HIV serodiscordant couples. Specifically, the Multisite HIV/sexually transmitted disease (STD) Prervention Trial for African American Couples (AAC) is a behavioral modification trial for African American, heterosexual, HIV discordant couples, whose goal is to decrease risky sexual behaviors and increase health promoting behaviors among couples. The main purpose of the study is to examine the efficacy of a couple-focused HIV/STD risk reduction intervention versus an individual-focused health promotion intervention in reducing sexual risk behaviors and STD incidence. A comprehensive description of the study design and randomization process can be found in Bellamy et al. (2005) and the primary findings of the trial are presented in El-bassell et al. (2010).

To accomplish the study's goals, reliable assessments of HIV sexual risk behaviors are critical. In this study, both partners are assessed independently for a number of shared sexual behaviors with their study partners. Thus, estimating the magnitude of the concordancy and discordancy among couple outcomes will allow evaluation of the reliability of each individual partners self-reported behavior. We focus on estimating the agreement between partners of self-reported, shared sexual behaviors at baseline.

Copulas allow us to model the dependence structure of outcomes separately from the marginal probability in addition to constructing a joint multivariate distribution of all outcomes. A nice feature of the copula approach in settings like AAC, where each member of the couple dyad is assessed independently regarding shared behaviors (e.g., male participants provide data on condom use with their female study partner in the past 90 days and female participants provide data on condom use with their male study partner in the past 90 days) is that each pair of responses can be used to measure the reliability of these same self-reported, shared behaviors. Additionally, while we construct the dependence structure, we are able to estimate dependency, a measure of reliability in this context, adjusting for individual characteristics of interest. In AAC, we are interested in (1) measuring dependency of partner outcomes to provide a measure of reliability for a number of self-reported sexual behavioral measures and (2) adjusting these estimates of dependency for a number of potentially important couple-level characteristics. Thus, employing copulas is a feasible method for measuring reliability of self-reported data in our motivating example. Our primary outcome of interest is a correlated binary outcome, which is 'consistent condom use' at every sexual episode with study partner, for both male and female partners. Thus, we apply a multivariate logit model introduced in Nikoloulopoulos and Karlis (2008), which uses a mixture of max-infinitely divisible (max-id) bivariate copulas proposed in Joe and Hu (1996). The application of copulas is usually limited for modeling multivariate binary outcomes primarily because of theoretical and computational limitations since the probability mass function should be obtained using finite differences for discrete data (Nikoloulopoulos and Karlis (2008)). Therefore, in order to be able to use copula models for multivariate discrete data, we need to specify copulas with rather simple forms (Nikoloulopoulos and Karlis (2008)). There are some desired properties for a parametric family of multivariate copulas applicable to discrete data described in Joe and Hu (1996) and Nikoloulopoulos and Karlis (2008). The max-id copula approach is attractive since it allows flexible positive dependence structures and has closed form cumulative distribution function (cdf); no other copula family has both these properties Joe and Hu (1996). However, it allows only positive dependence between random variables. Unlike other copula families where dependence parameters have joint constraints among them (Joe and Hu (1996) and Joe (1997)), the max-id copula achieves dependency sufficiecy such that we can model the dependence parameter using the covariate information.

Therefore, flexibility in modeling dependency while adjusting for covariates will allow us to examine how a number of factors (e.g., sociodemographic or relationship characteristics) may be associated with the reliability of self-reported shared behaviors among couples.

In sum, this paper demonstrates how to model and estimate dependency (i.e., reliability) parameters from multivariate binary data using copulas, evaluates the performance of this approach through a number of simulation studies and applies the proposed method to our motivating example. Note that GEE Liang and Zeger (1986) is another commonly used method that provides estimates of correlation as well as covariates. Therefore, we conduct a simulation study to compare copula-based estimates to moment estimates of GEE for the correlation coefficients, which are common measure of dependency, and to examine how copula approach can estimate the covariate effect on the copula parameter. Alternating logistic regression (ALR) proposed in Carey et al. (1993) could be used to adjust for different levels of clustering in the pairwise odds ratio (Carey et al. (1993) and lipsitz et al. (1991)). However, one limitation of this approach is that it applies only when $n_i = n$ for all clusters, but the copula-based method presenting here has no restriction on that. Also, estimates of the covariate effect in the pairwise odds ratio are not directly comparable since the copula method adjusts for covariates in the copula parameter, not in the pairwise odds ratio.Based on the results from the simulation study, we apply what we believe is the most appropriate copula to the data from AAC.

The following sections give more details on the motivating clinical application, the copula-based approaches for bivariate binary data, parameter estimation based on the log-likelihood, and estimation of odds ratios and binary correlations.

2.2 Motivating clinical example

Our motivating example is a randomized controlled trial (RCT) of HIV serodiscordant, African American couples designed to assess the effect of a culturally tailored HIV/STD risk reduction intervention on sexually transmitted infections and risky sexual behaviors among couples. Heterosexual transmission of HIV is the dominant route of infection worldwide, indicating a critical need for heterosexual couple-focused interventions (Witte et al. (2007)). In addition, for couples who are serodiscordant (one partner is HIV positive and one partner is HIV negative) putting the seronegative partner at high risk of HIV, reliable assessments of HIV sexual risk behaviors are critical in informing the efficacy of behavioral modification interventions. Couple-based studies provide a unique opportunity to measure the reliability of self-reported shared sexual behaviors, as each partner is measured independently and one can measure the degree to which couple responses are consistent. Reliable measures of self-reported sexual behaviors in high-risk populations have direct and obvious implications for estimating intervention effects. In this trial, couples assessed their condom use and other shared sexual behaviors retrospectively. A strength of the study is that each shared sexual behavior of interest is reported independently by each study partner (males and females, separately); therefore concordance of responses for these shared couple behaviors can be readily evaluated and used as a measure of reliability.

Potential issues with many self-report measures are their susceptibility to response bias, that is, a tendency for subjects to over or under report outcomes for a number of reasons (Catania et al. (1990)). Additionally, there is no "gold standard" for quantifying the validity of sexual behaviors since these behaviors are largely unobtainable by more objective methods. Nevertheless, in couples based studies, examining the concordance of partner responses is a reasonable assessment of reliability for shared behaviors. Additionally, exploring the influence of individual factors on estimated concordance may also help explain sources of response biases, if they exist. The effects of demographics and the couples' relationship context on concordance of reported sexual behaviors were examined using a measure of agreement such as Kappa statistics, conditional probability and McNemar's Statistics in El-bassell et al. (2010).

A few studies have quantitatively explored individual characteristics associated with concordance of partner reporting of sexual behaviors (El-bassell et al. (2010), Ochs and Binik (1999) and Seal (1997)). For the few studies that do try to estimate the association of individual factors on discordant couple responses, often a single outcome is constructed for each measure of interest that is a simple indicator of whether or not both partners had identical responses, and a gender-stratified model is used to predict the constructed couple indicator of discordance as a function of gender-specific characteristics (El-bassell et al. (2010)). Since factors that are related to their responses should explain both concordance and discordance and not just one, ideally we would like to employ a statistical tool to quantify dependence while simultaneously adjusting for individual factors that may be associated with this dependence.

Accordingly, the purpose of this paper is to present a novel application to quantify dependence measures by estimating association as a proxy for reliability in the context of a couple-based behavioral modification trial. Also, the proposed method provides an estimation procedure to explore individual-level factors that may be associated with the estimated correlation. Specifically, we apply max-id copulas to the AAC project in order to estimate the dependency of couple responses and gain insight into the reliability of self-reported sexual risk behaviors among a sample of African American, serodiscordant heterosexual couples.

2.3 Statistical methods

2.3.1 A copula approach for binary data

A copula-based model involves the generation of a multivariate joint distribution for outcomes of interest given the marginal distributions of the correlated responses. In the case of the AAC couples data, we can model a bivariate joint distribution considering the correlated responses from male and female partners as bivariate outcomes. Specifically, each partner is asked to report on shared sexual behaviors with their study partner in the past 90 days. By construction, partner responses are expected to be correlated and a measure of the magnitude of this correlation can also serve as a measure of reliability of these self-reported, sexual behavior outcomes.

The definition of a copula $C(u_1, \ldots, u_m)$ is a multivariate distribution function defined over the unit cube linking uniformly distributed marginals $(u_1, \ldots, u_m)$ (Sklar (1959) and Nelsen (2006)). Let $F_j(Y_j)$ be the cumulative distribution function (cdf) of a univariate random variable Y$_j$ $(j = 1, \ldots, m)$. Then, $C(F_1(y_1), \ldots, F_m(y_m))$ is an m-variate distribution for $y = (y_1, \ldots, y_m)^T$ with marginal distributions $F_j(j = 1, \ldots, m)$. Sklar first showed that there exists an m-dimensional copula C such that for all y in the domain of H in Sklar (1959),

$$H(y_1, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m)). \tag{2.3.1}$$

If $F_1, \ldots, F_m$ are continuous, then the function C is unique; otherwise, there are many possible copulas as emphasized in Genest and Nešlehová (2007). However, all of these coincide on the closure of $Ran(H_1) \times \ldots \times Ran(H_m)$, where Ran(H) denotes the range of H. While it is relatively easy to derive a joint distribution in the continuous case, it is not so simple in the case of discrete data. The latter involves 2^m finite differences of H(y), thus, to compute the joint probability mass function, one needs to evaluate the copula repeatedly. Therefore, in order to be able to use copula models for multivariate discrete data, one needs to specify copulas with rather simple forms.

Joe and Hu (1996) proposed multivariate parametric families of copulas that are mixtures of max-id bivariate copulas, allowing flexible dependence structures, having closed form cdfs, and satisfying the closure property under marginalization. This meets three desired properties for a parametric family of multivariate copulas applicable to discrete data (Nikoloulopoulos and Karlis (2008)). One property this does not satisfy is allowing negative dependence.

Given that our primary outcomes of interest are binary responses collected from male and female partners within each couple, we will use mixtures of max-id copulas. The mixture of m-variate max-id copulas cdfs has the following form

$$C(\mathbf{u};\Theta) = \phi\left(\sum_{j<k} log\, C'_{jk}(e^{-p_j\phi^{-1}(u_j;\theta)}, e^{-p_k\phi^{-1}(u_k;\theta)};\theta_{jk}) + \sum_{j=1}^{m} v_j p_j \phi^{-1}(u_j;\theta);\theta\right) \tag{2.3.2}$$

where $C'_{jk}(\cdot;\theta,\theta_{jk})$ is a bivariate max-id copula, $\phi(\cdot;\theta)$ is a Laplace transform (LT), $\Theta = \{\theta, \theta_{jk} : j,k = 1,...,m, j < k\}$ denotes the vector of all dependence parameters of the copula, u_j is cdf of a univariate random variable and $p_j = (v_j + m - 1)^{-1}$ where v_j is arbitrary. Specifically, the (j,k) bivariate marginal copula is

$$\begin{aligned} C_{jk}(u_j, u_k;\theta,\theta_{jk}) = \phi(&-logC'_{jk}(e^{-p_j\phi^{-1}(u_j;\theta)}, e^{-p_k\phi^{-1}(u_k;\theta)};\theta_{jk}) \\ &+ (v_j + m - 2)p_j\phi^{-1}(u_j;\theta) + (v_k + m - 2)p_k\phi^{-1}(u_k;\theta);\theta). \end{aligned} \tag{2.3.3}$$

We can simplify Equation (2.3.3) by assuming $v_j + m - 2 = 0$, then Equation (2.3.2) would become max-id copula with m(m-1)/2+1 dependence parameters. In our bivariate model, we need only one copula parameter and therefore force θ_{jk} equal to θ for every pair. Some members of max-id bivariate copulas and LTs are presented in Table 2.1. Thus, a combination of each family (5) and corresponding LT (4) in Table 2.1 will result in 20 parametric families with flexible dependence structure.

In particular, this approach allows us to estimate the measure of association between two binary outcomes through the copula dependence parameter, θ, which represents the degree of association. Moreover, we can incorporate covariate information in estimating the copula parameter by using a log transformation; this will be explained in detail in Subsection 2.3.2. Thus, we can obtain an estimate of correlation, adjusting for covariates of interest. Note that the GEE method could be used to model bivariate binary data by regarding the correlation between two outcomes as a nuisance parameter (Liang and Zeger (1986)). Since GEE is a widely used method, we will proceed to compare the estimated dependence from the copula approach with the moment estimates of correlation coefficient from GEE.

2.3.2 Copula-based bivariate logit model

In this section, we will discuss how the copula-based method can be integrated into a logit model and how to introduce covariate information in the copula parameter, θ. For simplicity and to relate the notation to our couples data, we will describe the bivariate logit model where $j = 1, 2$ denotes female and male responses respectively. Note that these models can be easily extended to a multivariate logit model where $j > 2$. Consider Equation (2.3.2) where $\mathbf{y} = (y_1, y_2)$ denotes the bivariate binary response for a couple and F_j the cdf of the univariate Bernoulli distribution function with probability of success π_j,

$$F_j(y_j;\pi_j) = \begin{cases} 1-\pi_j & \text{if } \ y_j = 0 \\ \\ 1 & \text{if } \ y_j = 1 \end{cases} \qquad j = 1, 2.$$

The standard logistic regression model for the probability of success π_{ij} corresponding to the copula in Equation (2.3.2) is

$$logit(\pi_{ij}) = \beta_j^T x_{ij}, \ \ j = 1, 2$$

where β_j is the vector of marginal regression parameters and X_{ij} is a vector of covariates for the i^{th} couple with j^{th} partner (female or male). As mentioned previously, we can also model dependence structure and introduce a regression coefficient in the copula parameter θ_i by choosing the appropriate log transformation for a given family from Table 2.1. For example, if we assume a Frank copula with LTD, then we would use the following model

$$log(\theta_i) = b_i^T Z_i, \quad i = 1, ..., n \ ,$$

where b_i^T is a vector of regression coefficients in the dependence measure and Z_i is a vector of covariate for the i^{th} couple. θ_i in the dependence model above will be incorporated in Equation (2.3.2) and used for joint distribution modeling of bivariate outcomes.

2.3.3 Parameter estimation

When marginal models are discrete, a multivariate probability function is obtained by taking the Radon-Nikodym derivative for H(y) in Equation (2.3.2). Thus, for the binary case, the bivariate probability function is given by

$$P(Y_1 = y_1, Y_2 = y_2) = C(u_1, u_2) - C(u_1, v_2) - C(v_1, u_2) + C(v_1, v_2) \tag{2.3.4}$$

where $u_j = F_j(y_j)$ and $v_j = F_j(y_j - 1)$ (Song (2000)). It follows that the joint log-likelihood of the bivariate logit copula model with various choices of copula family and LT can be written as

$$L(\boldsymbol{\beta}, b) = log \sum_{i=1}^{n} [C(F_1(y_{i1}), F_2(y_{i2})) - C(F_1(y_{i1}), F_2(y_{i2} - 1))$$

$$- C(F_1(y_{i1} - 1), F_2(y_{i2})) + C(F_1(y_{i1} - 1), F_2(y_{i2} - 1)); X_{ij}, \boldsymbol{\beta_j}, b_i]$$

where C is max-id copula, F_1, F_2 are univariate marginal cdfs, $\boldsymbol{\beta} = (\beta_1, \beta_2)$ is a vector of regression coefficients in the marginal model and b_i is a vector of regression coefficients in copula parameter. In this study, we focus on the standard maximum likelihood (ML) method that maximizes the joint log-likelihood. By using the ML method, we will simultaneously obtain the estimates of both copula and marginal parameters.

2.3.4 Estimation of odds ratio and binary correlation

In this study, we present odds ratios and binary correlations as a measure of dependency. The copula parameter may be presented as a measure of dependency in the copula-based method. However, it is not directly comparable to other dependence measures even among the copula-based approach since they differ according to which copula families are used. Due to this limitation, many applications involving copula methods use Kendall's τ as a measure of association. Kendall's τ is appropriate for measuring the strength of dependence between continuous outcomes, but it is less appropriate as a measure of association when applied to discrete variables. In particular, it is no longer distribution-free and has a range narrower than $[-1, 1]$, and this has to be taken into account when assessing the strength of the dependence (Denuit and Lambert (2005)). The bounds on Kendalls τ are plotted for Bernoulli margins with success probabilities $p_1 = p_2 = p \in [0, 1]$ in Nikoloulopoulos and Karlis (2008). Given the marginal probabilities of p_1 and p_2, we can rewrite Kendalls τ using the copula-based joint distribution of success, $C(1, 1)$,

$$\tau(Y_1, Y_2) = 2[C(1, 1) - p_1 p_2]$$

Thus, even in the most favorable cases, Kendall's τ does not reach 1 or -1 and cannot be comparable to usual measures of dependence when outcomes are binary. On the other hand, the odds ratio, which is one of common measures of the association between pairs of responses, is not constrained by marginal probabilities in the same way that Kendall's τ is constrained. The odds ratio, φ, can take any value in $(-\infty, \infty)$ with $\varphi = 1$ corresponding to no association. Figure 2.1 shows the relationship between the odds ratio and correlation coefficient, for examples in which the marginal probability of response for both males and females is 0.1, 0.2, 0.3, or 0.5.

We can write the odds ratio as a function of the joint probability of failure for both outcomes and their marginal probabilities, p_1 and p_2. Denote the joint probability of failure, p_{00}, for both female and male partners as

$$p_{00} = Pr\,(Y_1 = 0, Y_2 = 0)$$

where Y_j (Y_1 =female response, Y_2 =male response) denotes binary bivariate outcomes from female and male partners considering our motivating example. By using a copula approach, we have the following form of the joint probability of failure derived from Equation 2.3.2:

$$\begin{aligned} p_{00} &= Pr\,(Y_1 = 0, Y_2 = 0) = C(u_1, u_2 : \Theta) \\ &= \phi(-logC'(e^{-p_1\phi^{-1}(u_1;\theta)}, e^{-p_2\phi^{-1}(u_2;\theta)}; \theta) \qquad (2.3.5) \\ &+ (v_1 + m - 2)w_1\phi^{-1}(u_1;\theta) + (v_2 + m - 2)w_2\phi^{-1}(u_2;\theta)) \end{aligned}$$

where $u_1 = F(Y_1 = 0) = 1 - \pi_1$, $u_2 = F(Y_2 = 0) = 1 - \pi_2$, $w_j = (v_j + m - 1)^{-1}$, m=2, v_j is arbitrary, and $j = 1, 2$. Let φ be the odds ratio between responses Y_1 and Y_2. The odds ratio for binary responses is defined as

$$\varphi = \frac{Pr(Y_1 = 1, Y_2 = 1)Pr(Y_1 = 0, Y_2 = 0)}{Pr(Y_1 = 1, Y_2 = 0)Pr(Y_1 = 0, Y_2 = 1)} = \frac{(p_{00} + p_1 + p_2 - 1)p_{00}}{(1 - p_1 - p_{00})(1 - p_2 - p_{00})}$$

where $p_1 = Pr(Y_1 = 1)$ and $p_2 = Pr(Y_2 = 1)$; this is equivalent to the ratio of the odds of concordant to discordant responses. p_{11} can be simply derived by using the equation $p_{11} = p_{00} + p_1 + p_2 - 1$, where p_{00} is derived from the joint probability estimated from Equation (2.3.5).

For the purpose of comparing a measure of dependence from the copula method to a moment estimate for correlation coefficient from GEE, we also estimate binary correlation using the formula of phi correlation in Streiner and Norman (1995). Binary correlation coefficient using above probabilities that will always lie in $[0, 1]$ is

$$Corr(\rho) = \frac{p_{11} - p_1 p_2}{\sqrt{p_1(1 - p_1)p_2(1 - p_2)}}.$$

2.4 Simulation studies

We conducted a simulation study to explore the performance of the copula approaches in estimating correlations and covariate effects on the copula parameter, and compared copula-based correlation estimates to moment estimates for correlation coefficients obtained from the GEE method. We use the results from the simulation study to determine which copula family might be the most appropriate for the AAC data. We summarize estimated odds ratios and correlations from 500 simulated samples with true correlation values of 0.05, 0.1, 0.25 and 0.5. As for the covariate effect on the copula parameter, data are simulated with true covariate coefficients of 0.01, 0.30 and 1.0. with 500 repetitions.

2.4.1 Data simulation method

We created correlated bivariate binary random variables by thresholding a normal distribution using the package 'bindata' in R, which applies algorithm presented in Qaqish (2003). We can set n samples from a multivariate normal distribution with mean and variance chosen in order to get the desired margin and common probabilities. We generated 500 simulation repetitions with 4 sets of correlation coefficients (0.05, 0.1, 0.25, and 0.5), and separately with 3 sets of covariate coefficients (0.01, 0.30 and 1.00) on the copula parameter. The sample size was equal to 1000 (500 pairs of correlated outcomes) in both settings. Parameter estimates and corresponding standard errors for the odds ratio and correlation coefficient from the copula method were estimated based on bootstrapped resampling (100 repetitions) within each simulated dataset. The simulated marginal probability was set as 0.25 for both treatment groups and couple measures. The marginal probability of 0.25 is a crude estimate of the marginal probability for the primary outcome of interest for both treatment and control groups at baseline in the AAC sample.

2.4.2 Bias and efficiency in estimating correlations

For each of the four simulation scenarios, we estimate odds ratios and correlation coefficients using the copula approach with a combination of 5 choices of max-id copula families and 4 choices of corresponding Laplace transformations (LT) (Table 2.2). Here we present detailed results only for Gumbel copula with Laplace transformation D (Gumbel D) and Frank copula with Laplace tranformation A (Frank A) as the other families provided similar results. To examine the performance of each estimator, we present bias, 95% coverage probability and mean squared error (MSE). For all underlying true correlations, Frank A performs the best or as well as the best compared to both Gumbel D and the GEE method, providing a good 95% coverage probability (93.6-95.6%), small bias (0.0005-0.0027), and small MSE (0.0019-0.0024). Gumbel D provides good estimates for modest and strong correlation ($\rho = 0.25, 0.5$), but performs worse when the true correlation is small ($\rho = 0.05$). For weak correlation ($\rho = 0.1$), Frank A provides a close estimate to the real value with corresponding bias 0.001, while estimates from Gumbel D provide highest 95% coverage probability. For modest ($\rho = 0.25$) and strong ($\rho = 0.5$) correlations, all methods provide similar results while the copula-based methods perform slightly better than GEE in terms of bias. In sum, the copula approach with Frank A performs the best or as well as the best in estimating correlations with respect to bias. Gumbel D does slightly better than Frank A when ρ is 0.5 in terms of coverage. The standard errors from all methods are similar.

2.4.3 Bias and efficiency in estimating the covariate effect on the dependence parameter

In order to examine the performance of the copula-based methods in estimating the covariate effect on the dependence parameter, we also perform a simulation study. We create one binary variable for covariate, and adjust for it on the dependence parameter. Thus, as described in 2.3.2, we have two regression coefficients, b_0 and b_1, on the dependence parameter, where b_1 represents a regression coefficient for the covariate. We set the value of b_0 as 0.262 which corresponds to copula parameter $\theta = 1.3$, and represents moderate level of correlation. For each of simulation, we set the covariate coefficient, b_1, as 0.01, 0.30 and 1.00, respectively, where covariate effect ranges from small to large with corresponding p-value from large to small. We also use Gumbel D and Frank A. We present bias and 95% coverage probability to show their performance (Table 2.3). We could fit the dependence model using different levels of regression coefficients and different values of b_0 (not presented here), but the results appear consistent.

Both methods provide good estimators of both regression coefficients on the dependence parameter. Gumbel D performs slightly better than Frank A with respect to bias, while Frank A does better than Gumbel D with respect to coverage. As expected, mean p-values for b_1 decrease when true value of b_1 gets bigger. Significant p-value (< 0.05) represents a significant difference in dependency among two groups we adjust for on the dependence parameter. We do not present the performance of copula-based approach in estimating marginal probability, but it appears to provide unbiased estimates (bias=0.000-0.002).

2.5 Application

In this section, we analyze the clinical trial of HIV serodiscordant African American couples designed to assess the effect on HIV/STD Prevention trial and apply our method to this data. As noted, we focus more on the dependency parameters than the marginal parameters, so that the estimated dependency can serve as a measure of reliability of self-reported sexual behaviors. The following subsections will describe the study design and data, characteristics of study patients and study outcomes. In the last subsection, the results of the analysis by applying copula approach are shown.

2.5.1 Study design and data

We use baseline data from HIV/STD Prevention trial, a two-arm, couple-based randomized controlled intervention trial of HIV serodiscordant African American couples from four cities in the US (Atlanta, GA; Los Angeles, CA; New York, NY; and Philadelphia, PA). The study was designed to test the efficacy of a couple-focused HIV/STD risk reduction intervention vs. an individual-focused health promotion intervention in reducing sexual risk behaviors and STD incidence (see NIMHa (2008) and NIMHb (2008)).

The study includes 535 couples (1070 individuals) recruited from HIV care clinics, HIV testing and counseling sites, primary care clinics, AIDS services organizations, substance abuse treatment programs, churches and HIV/AIDS ministries, HIV/AIDS providers and community-based coalitions and advocacy organizations. Participants met specific study criteria.

Data were obtained from three sources. First, participants completed a 90-minute Audio Computer-Assisted Survey Interview (ACASI), which assessed sociodemographic and relationship characteristics, sexual behaviors and condom use, and psychosocial mediators that had sound psychometric properties and had previously been implemented with adult African American populations. Although both participating male and female partners completed the same ACASI assessments, the sexual behavior items were written to be appropriate for each gender. Subsequently, a trained African American interviewer administered validated and reliable assessments on sexual and physical abuse and a brief index assessing study participants' commitment to the African American community.

2.5.2 Characteristics of study participants

Study partners were asked to indicate their age (in years), education, income, type of health insurance, and incarceration history. HIV status at baseline was determined via biological testing in order to confirm that couples were HIV serodiscordant. Study participants were also asked questions that addressed relationship characteristics including length of relationship with their study partner, whether or not participants were married to their study partner (yes/no), and sexual dysfunction items (yes/no).

To illustrate previously described copula methods in this data, we created categorized couple variables with 3 levels for the following items: high school graduate, income (over $850/month), insurance, incarceration history indicating whether each characteristic was observed in neither, one or both partners within each couple. HIV status refers to whether the female partner was the HIV positive partner. These 9 items were considered covariates of interest in measuring the the dependence parameter. Because our primary interest was to estimate the dependence parameter, the only covariate used in estimating the marginal probability of the outcome of interest was randomized treatment assignment.

2.5.3 Primary outcome

Participants provided data on the use of male and female condoms during sex they had engaged in with study partners (vaginal, anal and oral intercourse) over the past 90 days. Proportion of condom use was calculated first. For the purpose of illustrating the present copula methodology and because it is a common primary endpoint in HIV/STD risk modification trials, we constructed an outcome, 'consistent condom use', that equals one when condom use was reported at every sexual episode with study partner and zero otherwise, for both male and female partners.

2.5.4 Results: Correlations according to different level of covariates

Our main purpose in this study is to determine how the associations between male and female consistent condom use responses vary across the different sub-populations and how the self-reported outcomes are reliable. We fit the model that incorporates the covariates of interest in the copula parameter as described in Subsection 2.5.2. Our preliminary analysis using `PROC CORR` showed estimated correlation of reported consistent condom was 0.34. From the previous simulation study with modest correlation, both Gumbel D and Frank A performed similarly. We arbitrarily fit our model using the Frank A copula since we expect which copula family we choose does not affect the results based on simulations in the previous section.

Table 2.4 summarizes the results of the estimated regression coefficients of covariates on the dependence parameter, θ. Wald test is done at 0.05 significance level. The result shows a statistically significant difference in the copula dependence parameter between couples where both have insurance and those where neither or either has insurance. Couples where both have insurance are likely to have more correlated outcomes than those where neither or either has insurance. In addition, education, income and duration of relationship have relatively smaller p-values indicating there may be differences in the correlation of male and female partner responses between subgroups of those covariates. Also, couples where both have high income ($> \$850$) or high school diploma are likely to have more correlated outcomes than those where neither or either has high income or high school diploma, but these were not statistically significantly different. Other covariates such as age, incarceration history, HIV status (female or not), married to study partner, and sexual dysfunction had similar correlations among subgroups.

Table 2.5 summarizes the odds ratios and correlation across different levels of the covariates estimated from its corresponding dependence parameter. The average odds ratio and correlation between female and male partners are 5.59 and 0.34, respectively. As expected, as sociodemographic indices such as education, income, insurance status and no incarceration history increases, the correlation between couples response increases. The correlation where both have insurance is 0.401, while the correlation where neither have insurance is 0.099, demonstrating the impact of these covariates on correlation.

2.6 Conclusions and discussion

In this work, we applied a bivariate copula-based logit model to data from the HIV serodiscordant couples study and estimated the correlation of the couples' responses according to the different covariates. We have illustrated a novel application of estimating dependency, while adjusting for covariates, as an estimator of reliability of self-reported shared sexual behavior from a couple-based study. Prior to the application, we conducted a simulation study to examine how copula-based models perform relative to GEE, and to determine which copula family works well in a number of settings with varying levels of underlying correlations and covariate on the copula parameter. We also compared the results of the copula-based method with those of GEE. In prior work with copulas, many have used the Kendall's τ or copula parameters as the measure of concordance. Since neither the Kendall's τ nor copula parameter is directly comparable to moment estimates for the correlation coefficient from GEE, we introduced a copula-based estimator of binary correlation and odds ratio to compare the moment estimates for the correlation coefficient from GEE. Based on the results from the simulation study, we found that most of the models with copula families perform well and provide similar results for moderate and strong correlation. Frank A performed well for the weak correlation, however, Gumbel D did not do well when true ρ is 0.05. In terms of small bias, both copula models performed better than GEE when true correlations are 0.25 and 0.5. For correlation 0.1, Frank A worked the best. In terms of 95% coverage rate, the results from all methods were similar for all levels of correlation except for Gumbel D when true ρ is 0.05. Both methods with Gumbel D and Frank A also performs well in estimating the regression coefficient on the dependence parameter. Gumbel D performs better in terms of bias, while Frank A does better in terms of coverage.

Finally, we fitted copula-based models to our data and focused on estimation of the correlation between responses of consistent condom use from couples adjusting for couple-level covariate information. We selected 9 different couple-based covariates. The findings show that there is a statistically significant difference in the correlation between couples where both have insurance and those where neither or either has insurance. Couples where both have insurance have more correlated outcomes than those where neither or either has insurance. Among couples where females and males have high school diplomas compared to couples where neither or either has a diploma, the responses are more highly correlated. Couples where both have high income ($> \$850$) are likely to have more correlated outcomes than those where neither or either has high income. Interestingly, in terms of relationship duration, the responses from couples with more than 5 year relationship seem to be less correlated. Whether the female partner is HIV infected or not does not affect the correlation. These findings suggest that we need to pay attention to those couples with the covariates such as no insurance, low income and low education, indicating low correlation, to improve the reliability of self-reports.

To summarize, this work provide a good measure of reliability of self-reported sexual behaviors among HIV serodiscordant couples by estimating correlation using a copula-based method. This work also introduces systematic research on the influence of the factors on the responses of self-reported sexual behaviors. Thus, we can see the magnitude of matching responses based on the estimated correlation adjusting for important covariates of interest, which can tell us the reliability of paired or couple-based self-reported data.

This approach has the advantages of constructing separate models for the marginal probabilities and the dependence parameters, which is more efficient. In addition, this method is fully specified allowing joint and conditional probabilities to be derived easily, and is straightforward to apply using a standard and direct maximum likelihood inference procedure. Also, it allows us to model the dependence parameter with covariate information of interest without computational difficulty. Therefore, it leads to a better understanding of couple-level issues related to self-reported sexual behaviors.

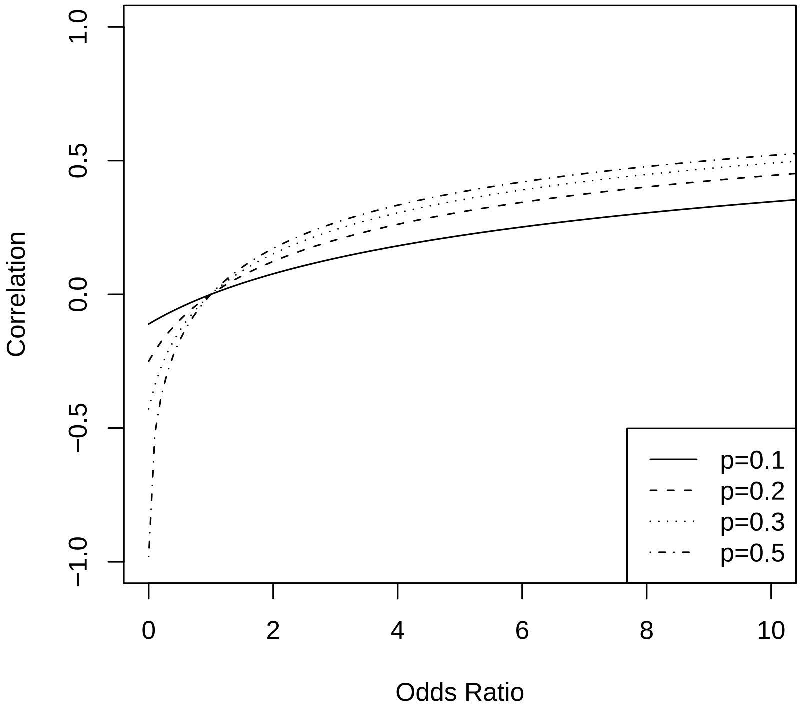


Figure 2.1: Relationship between the odds ratio and correlation coefficient.

Table 2.1: Max-id bivariate copulas and Laplace transforms (LTs)

Family	$C'(u_j, u_k; \theta)$	LTs: $\phi(t;\theta)$	$\theta \in$	Log transformation
Gumbel(LTA)	$e^{-(\tilde{u_j}^{\theta}+\tilde{u_k}^{\theta})^{1/\theta}}$	$e^{-t^{1/\theta}}$	$[1,\infty)$	$log(\theta-1)$
Kimeldorf(LTB)	$(u_j^{-\theta}+u_k^{-\theta}-1)^{-1/\theta}$	$(1+t)^{-1/\theta}$	$(0,\infty)$	$log\theta$
Joe(LTC)	$1-(\bar{u_j}^{\theta}+\bar{u_k}^{\theta}-\bar{u_j}^{\theta}\bar{u_k}^{\theta})^{1/\theta}$	$1-(1-e^{-t})^{1/\theta}$	$[1,\infty)$	$log(\theta-1)$
Frank(LTD)	$-\frac{1}{\theta}log\left\{1+\frac{(e^{-\theta u_j}-1)(e^{-\theta u_k}-1)}{e^{-\theta}-1}\right\}$	$-\frac{log(1-(1-e^{-\theta})e^{-t})}{\theta}$	$(0,\infty)$	$log\theta$
Galambos	$u_j u_k e^{(\tilde{u_j}^{-\theta}+\tilde{u_k}^{-\theta})^{-1/\theta}}$		$[0,\infty)$	$log\theta$

Note that $\bar{u_j} = 1 - u_j$ and $\tilde{u_j} = -logu_j$ where $u_j = F_j(y_j)$

Table 2.2: The average estimates and standard errors of odds ratios and correlation coefficient for simulated data using copula approach and GEE

True			Odds Ratio		Correlation				
Corr.	Method	Group	Est.	S.E.	Est.	S.E	MSE	Bias	Coverage
ρ=0.05	Gumbel D	Trt.	1.664	0.144	0.100	0.016	0.0027	0.0498	0.188
		Ctrl.	1.664	0.144	0.099	0.016	0.0026	0.0495	0.182
	Frank A	Trt.	1.325	0.323	0.050	0.046	0.0023	-0.0005	0.950
		Ctrl.	1.323	0.324	0.049	0.046	0.0023	-0.0009	0.944
	GEE	N.A.	-	-	0.047	0.046	0.0023	-0.0026	0.944
		N.A.	-	-					
ρ=0.1	Gumbel D	Trt.	1.803	0.289	0.115	0.028	0.0010	0.0147	0.980
		Ctrl.	1.803	0.289	0.115	0.028	0.0010	0.0150	0.980
	Frank A	Trt.	1.684	0.406	0.098	0.047	0.0020	-0.0017	0.950
		Ctrl.	1.686	0.405	0.099	0.047	0.0020	-0.0014	0.956
	GEE		-	-	0.096	0.047	0.0020	-0.0037	0.958
ρ=0.25	Gumbel D	Trt.	3.440	0.833	0.248	0.048	0.0024	-0.0024	0.934
		Ctrl.	3.445	0.836	0.247	0.048	0.0024	-0.0027	0.930
	Frank A	Trt.	3.440	0.825	0.248	0.048	0.0024	-0.0023	0.936
		Ctrl.	3.445	0.827	0.247	0.048	0.0024	-0.0027	0.944
	GEE		-	-	0.245	0.048	0.0024	-0.0048	0.942
ρ=0.5	Gumbel D	Trt.	12.165	3.368	0.498	0.045	0.0019	-0.0023	0.954
		Ctrl.	12.114	3.344	0.498	0.045	0.0019	-0.0022	0.954
	Frank A	Trt.	12.157	3.383	0.498	0.045	0.0019	-0.0024	0.950
		Ctrl.	12.115	3.361	0.498	0.045	0.0019	-0.0022	0.950
	GEE		-	-	0.496	0.045	0.0020	-0.0043	0.956

Table 2.3: The average estimates and standard errors of covariate coefficients on the dependence parameter using copula approach

True	Copula		$b0$				$b1$			
$b = (b0, b1)$	family	Est.	S.E.	Bias	Coverage	Est.	S.E.	Bias	Coverage	p-value*
(0.262, 0.010)	Gumbel D	0.266	0.082	0.004	94.0	0.010	0.117	0.000	94.4	0.506
	Frank A	0.265	0.089	0.008	94.6	0.011	0.128	0.001	94.6	0.536
(0.262, 0.300)	Gumbel D	0.266	0.082	0.003	94.0	0.310	0.137	0.010	94.4	0.102
	Frank A	0.266	0.089	0.002	94.2	0.314	0.154	0.014	93.6	0.128
(0.262, 1.000)	Gumbel D	0.264	0.082	0.001	93.6	0.979	0.203	0.021	91.2	0.001
	Frank A	0.263	0.089	0.001	94.4	1.043	0.229	0.043	94.0	0.000

* p-value obtained from the Wald test

Table 2.4: The estimates of dependence parameters adjusting for each of covariate using max-id copula with Frank A

	Covariate effect		
covariates	Est.	S.E.	p-value
Age	-0.002	0.010	0.874
Education	0.225	0.133	0.092
Income	0.128	0.102	0.206
Insurance	0.176	0.081	0.029*
Incarceration	0.091	0.100	0.358
HIV	0.007	0.134	0.959
Duration	-0.171	0.135	0.205
Married	-0.021	0.136	0.876
Sexual dysfunction	-0.087	0.131	0.505

* statistically significant at 0.05 level

Table 2.5: The corresponding odds ratios and correlation of each covariate level based on estimated dependence parameter using max-id copula with Frank A

	Odds ratio				Correlation			
	Treatment group		Control group		Treatment group		Control group	
covariates	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
Age	5.566	1.689	5.572	1.723	0.341	0.056	0.341	0.056
Education								
0	3.515	1.334	3.525	1.342	0.247	0.073	0.247	0.074
1	8.107	3.286	8.142	3.276	0.416	0.067	0.416	0.066
Income								
0	4.120	1.484	4.117	1.455	0.279	0.064	0.280	0.065
1	6.667	2.439	6.657	2.434	0.377	0.062	0.378	0.062
2	10.202	8.625	10.174	8.651	0.458	0.103	0.460	0.102
Insurance								
0	1.709	1.587	1.708	1.491	0.099	0.378	0.100	0.360
1	3.958	1.173	3.941	1.167	0.269	0.059	0.271	0.059
2	7.645	2.762	7.588	2.732	0.401	0.060	0.404	0.060
Incarceration History								
0	3.902	2.677	3.911	2.702	0.269	0.108	0.269	0.109
1	5.545	1.544	5.558	1.533	0.341	0.050	0.342	0.051
2	7.629	4.037	7.644	4.049	0.404	0.084	0.405	0.085
HIV								
0	5.383	2.395	5.363	2.371	0.333	0.076	0.335	0.076
1	5.537	2.001	5.516	1.953	0.339	0.065	0.341	0.065
Duration(>5yrs)								
0	7.654	3.182	7.599	3.156	0.402	0.071	0.404	0.072
1	4.046	1.539	4.029	1.520	0.274	0.070	0.276	0.070
Married								
0	5.636	1.792	5.614	1.774	0.343	0.059	0.344	0.060
1	5.185	2.638	5.167	2.581	0.326	0.086	0.327	0.086
Sexual dysfunction								
0	6.297	2.333	6.275	2.294	0.365	0.065	0.366	0.066
1	4.573	2.007	4.562	1.971	0.300	0.077	0.301	0.077

Chapter 3

Extension of Max-id Copula to Longitudinal Binary Couples Data

3.1 Introduction

By definition, in a longitudinal study, multiple outcomes are measured repeatedly for individuals or groups over time. In our motivating example, investigators developed and evaluated a couple-based behavioral intervention for reducing shared sexual risk behavior and collected retrospective self-report outcomes from both partners at baseline and at 3 follow-ups. See El-Bassell et al (2010) for more details of the trial. As the outcomes refer to the couples' shared sexual behavior, couples' responses are expected to be correlated. Thus, modeling couple responses should account for multiple sources of correlation: within-individual over time as well as within-couple both at the same measurement time and at different times. In this paper, we present a copula-based approach for modeling multivariate longitudinal binary outcomes to estimate intervention effects for male and female partners as well as to account for multiple sources of correlation in practical settings where dyads are measured repeatedly over time.

There are several approaches to modeling multivariate longitudinal binary outcomes. One is the generalized estimating equations (GEE) approach using a logit link function proposed by Liang and Zeger (1986). The GEE approach is easy to implement and gives efficient estimates of regression coefficients, although estimates of the association among the binary outcomes can be inefficient [Carey et al, 1993]. When the association model is of primary interest, the second-order generalized estimating equation (GEE2) approach developed by Liang, Zeger, and Qaqish (1992) gives more efficient estimates of association parameters. However, GEE2 becomes computationally intense when the cluster size is large. Moreover, covariate effects on marginal probabilities are biased if the association model is mis-specified.

Random effects approaches proposed by Ten Have, Kunselman and Tran (1999) are often applied in analyses of longitudinal data with nested levels of clustering. However, this approach would be challenging to accommodate several sources of correlation, especially for binary outcomes, as in our motivating example. Also, it involves complex assumptions regarding the distributions of the random effects, and can be more challenging to implement. Additionally, this approach offers different covariate effects on the marginal probabilities, as these effects are interpreted conditional on random effects. Also, the estimates of these effects could be biased if the random effects structure is mis-specified.

Another general analytic approach for modeling correlated outcomes is to employ copulas. There exist several copula-based approaches. Song, Li, and Yuan (2009) proposed joint regression analysis of correlated data using Gaussian copulas, also referred to as vector generalized linear models (VGLM). Escarela, Luis Carlos, and Russel (2009) proposed a copula-based Markov chain model for the analysis of binary longitudinal data using a probit link. Both of these copula-based methods use three correlated outcomes (a single outcome measured at baseline and two follow-up times) as an approach for modelling longitudinal data using Gaussian copula. Lambert and Vandenhende (2002) also proposed a new model for multivariate non-normal longitudinal data, where three responses are measured repeatedly and assumed to follow different parametric marginal distributions. The normal copula was used to relate those responses because the dependence structure can easily be specified through the variance-covariance matrix, but it can be computationally intense when applied to more than three correlated outcomes.

We extend the multivariate max-id copula logit model proposed by Nikoloulopoulos and Karlos (2008), which was originally applied to multivariate outcomes. We analyze longitudinal bivariate binary outcomes by incorporating possibly different sets of covariate structure including time indicator for each marginal model and using a max-id copula to accommodate the correlations for all pairs of outcomes. Specifically, to assess the longitudinal effect, a time indicator was added to the covariate structure in the marginal model. Thus, a single response at different time points has a different covariate structure since time indicator changes depending on time point. The application of copulas is usually limited in modeling multivariate discrete outcomes primarily because of theoretical and computational restrictions; when the multivariate case is considered, the form of dependence is usually quite limited (Nikoloulopoulos and Karlis 2009). For example, Archimedian copulas only allow for simple correlation which is similar to an exchangeable structure and their range of dependence becomes narrower as the dimension increases. However, the max-id approach is attractive in that more flexible dependence structures can be accommodated for each possible pair of outcomes and the cumulative distribution functions are available in closed form, which can increase computational feasibility for discrete data. In the max-id approach, the dependence structure is modeled via the copula parameter, which permits separate estimates of the dependency with respect to clusters or covariates of interest. It allows direct and intuitive interpretation of the correlation structure through several sets of copula parameters describing various components (e.g., across time, within-cluster, etc). Components are intuitive and interpretable as sources of variation. Finally, the max-id approach uses its likelihood for estimation, which is useful for conducting statistical inference and model selection.

The paper is organized as follows: we introduce our motivating example in Section 2; in Section 3, we present the max-id copula-based approach for bivariate, longitudinal binary outcomes; Section 4 evaluates the performance of the max-id approach through a number of simulations; and in Section 5, we apply the max-id approach to our motivating example, the Multisite HIV/STD Prevention Trial for African American Couples (AAC)(NCT00644163).

3.2 Statistical methods

3.2.1 Copula-based model for bivariate longitudinal data

Copula models involve the generation of a multivariate, joint distribution for outcomes of interest, given the marginal distributions of the correlated responses. Let $F_j(y_j)$ be the cumulative distribution function (cdf) of a univariate random variable $(j = 1, \ldots, m)$. Sklar (1957) first showed that there exists an m-dimensional copula, C, such that for all y in the domain of H,

$$H(y_1, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m)). \tag{3.2.1}$$

Joe and Hu (1996) proposed multivariate parametric families of copulas that are mixtures of max-id bivariate copulas, which have flexible dependence structures, closed form cdfs, and satisfy the closure property under marginalization, all desired properties for modeling binary data using a parametric family of multivariate copulas (see Nikoloulopoulos and Karlis 2009). The mixture of m-variate max-id copulas cdfs has the following form

$$C(\mathbf{u};\Theta) = \phi\left(\sum_{j<k} \log\, C'_{jk}(e^{-p_j\phi^{-1}(u_j;\theta)}, e^{-p_k\phi^{-1}(u_k;\theta)}; \theta_{jk}) + \sum_{j=1}^{m} v_j p_j \phi^{-1}(u_j;\theta); \theta\right) \tag{3.2.2}$$

where $C^{'}_{jk}(\cdot;\theta,\theta_{jk})$ is a bivariate max-id copula, $\phi(\cdot;\theta)$ is a Laplace transform (LT), $\Theta = \{\theta, \theta_{jk} : j, k = 1, ..., m, j < k\}$ denotes the vector of all dependence parameters of the copula, u_j is cdf of a univariate random variable and $p_j = (v_j + m - 1)^{-1}$ where

v_j is arbitrary. Specifically, the (j,k) bivariate marginal copula is

$$C_{jk}(u_j,u_k;\theta,\theta_{jk}) = \phi(-\log C'_{jk}(e^{-p_j\phi^{-1}(u_j;\theta)}, e^{-p_k\phi^{-1}(u_k;\theta)};\theta_{jk}) \tag{3.2.3}$$

$$+ (v_j+m-2)p_j\phi^{-1}(u_j;\theta) + (v_k+m-2)p_k\phi^{-1}(u_k;\theta);\theta). \tag{3.2.4}$$

Without loss of generality, we assume $v_j + m - 1 = 0$ and obtain simple max-id copulas with $\frac{m\times(m-1)}{2} + 1$ dependence parameters. Note that the several choices of copula families with 4 possible LTs to construct a max-id copula will result in a rich class of 20 different parametric copula families (Table 2.1).

In the context of our motivating example, we consider 4 correlated outcomes: $(y_{1m}, y_{1f}, y_{2m}, y_{2f})$ where the first subscript $(1,2)$ denotes the baseline and follow-up measurement, respectively and the second subscript (m/f) denotes male/female partner, respectively. We model the multivariate joint distribution considering these 4 correlated responses using the max-id copula described above.

3.2.2 Copula-based model with a logit link

In this section, we demonstrate how the copula-based method can be integrated into a logit model with a set of covariates in the univariate marginal probability, π_j, as described in Nikouloulopoulos and Karlis (2008). Consider Equation (3.2.2) where $y = (y_1,\ldots,y_m)$ denotes the multivariate binary responses and F_j the cdf of the univariate Bernoulli distribution function with probability of success π_j $(j = 1,\ldots,m)$,

$$F_j(y_j;\pi_j) = \begin{cases} 1-\pi_j & \text{if } \ y_j = 0 \\ 1 & \text{if } \ y_j = 1 \end{cases} \qquad j = 1,2.$$

The standard logistic regression model for the probability of success π_{ij} corresponding to the copula in Equation (3.2.2) is

$$logit(\pi_{ij}) = \beta_j^T X_{ij}, \;\; j = 1, 2$$

where β_j is the vector of marginal regression parameters including time-varying covariates and X_{ij} is a vector of covariates for the j^{th} partner's outcome for the i^{th} couple. Incorporating a time variable allows us to model longitudinal data, which differs from the multivariate modeling in Nikoloulopoulos and Karlis (2008). This requires construction of a separate covariate structure for the outcomes at each time point by modeling the probabilities as a function of time.

3.2.3 Estimation of marginal model parameters

When marginal models are discrete, a multivariate probability function is obtained by taking the Radon-Nikodym derivative for $H(y)$ in Equation (3.2.1). Let $c = (c_1, \ldots, c_m)$ be vertices where y_j is discrete and each c_j is equal to either y_j or $y_j - 1$, $j = 1, \ldots, m$. Then, the multivariate joint probability function h is given by the copula representation

$$h(y_1, y_2, \ldots, y_m) = \sum sgn(c) C(F_1(c_1), \ldots, F_m(c_m)) \tag{3.2.5}$$

where the sum is taken over all vertices c, and $sgn(c)$ is given by

$$sgn(c) = \begin{cases} 1, & \text{if } \; c_j = y_j - 1 \text{ for an even number of j's} \\ \\ -1, & \text{if } \; c_j = y_j - 1 \text{ for an odd number of j's.} \end{cases}$$

See more details in Song (2000). Using this formulation, we can construct the joint log-likelihood of the multivariate logit copula model for a variety of copulas paired with a variety of LTs, where C is a max-id copula, F_j are j univariate marginal cdfs and β_j is a vector of regression coefficients in each marginal model.

We focus on standard maximum likelihood (ML). To obtain possible starting values, the j univariate log-likelihoods (Equation (3.2.6)) are maximized independently in order to obtain j separate $\hat{\beta}_j$, e.g.,

$$L_j(\beta_j) = \sum_{i=1}^{n} \log(h_j(y_{ij};\beta_j))\ j = 1,\dots,m \tag{3.2.6}$$

where $h_1,\dots,h_m$ are the univariate probability functions. Next, the joint log-likelihood (Equation (3.2.7)) is maximized over the copula and marginal regression parameters, simultaneously, using optimization technique with the Nelder-Mead method.

$$L(\beta,\Theta) = \sum_{i=1}^{n} \log(h(y_{i1},\dots,y_{im};\beta,\Theta)),\ j = 1,\dots,m \tag{3.2.7}$$

3.2.4 Estimation of binary correlation

We can write the pairwise correlation as a function of the joint probability of failure for both outcomes, y_j and y_k $(j,k = 1,\dots,m)$ using the bivariate max-id copula and the marginal probabilities of success, p_1 and p_2. Denote the joint probability of failure, p_{00}, as

$$p_{00} = P(y_j = 0; y_k = 0)$$

where y_j and y_k denote any two correlated outcomes when $j,k = 1,\dots,m$. Using a bivariate max-id copula, the joint probability of failure derived from Equation (3.2.3) takes the form

$$p(y_j = 0, y_k = 0) = C_{jk}(1 - \pi_j, 1 - \pi_k; \beta_j, \beta_k, \theta_{jk})$$

where β_j and β_k are regression coefficients in the marginal model and θ_{jk} is a subset of corresponding copula parameters for both outcomes, y_j and y_k. The binary correlation for binary responses is defined as

$$\phi = \frac{p_{11} - p_1 p_2}{\sqrt{p_1(1 - p_1)p_2(1 - p_2)}}$$

where $p_{11} = p(y_j = 1, y_k = 1) = p_{00} + p_1 + p_2 - 1$, $p_1 = p(y_j = 1)$ and $p_2 = p(y_k = 1)$.

3.3 A simulation study

3.3.1 Simulation set-up

We conducted a simulation study to evaluate the performance of the max-id copula in estimating marginal probabilities, copula parameters (Θ) and correlation coefficients. We generated 200 samples of 4 correlated outcomes using inputs that were similar to our motivating example. To generate correlated binary outcomes for a given correlation matrix, we used the R package, 'mvtBinaryEP', which applies an algorithm developed by Emrich and Piedmonte (1991).

Specifically, the primary focus of the motivating example was to estimate intervention efficacy in reducing risky behavior during follow-up, in a randomized trial; therefore we simulated marginal probabilities such that these probabilities were equal in the two interventions at baseline and different at follow-up. We assume the marginal probabilities of the outcomes for intervention (p_{1t}) and control (p_{1c}) at baseline are both 0.2 and the probabilities for both groups (p_{2t} and p_{2c} for intervention and control, respectively) at follow-up time are 0.4 and 0.6 (0.3 and 0.5 in some settings), respectively, which covers a range of plausible scenarios. We arbitrarily choose a Gumbel copula with the Laplace transformation D (Gumbel D). Based on simple correlation estimates of the primary outcome of interest from AAC, we assume there is a background correlation for all pairs (θ) and 3 possible additional pair-wise correlations: within-individual correlation over time (θ_{12}), within-couple correlation at baseline ($\theta_{1.}$) and within-couple correlation at follow-up ($\theta_{2.}$). For simplicity, we assume there are no more additional sources of correlation (e.g., within-couple correlation at different time points). Additionally, we assume that the additional within-individual correlation (θ_{12}) is equal for males and females. See Figure 1 for a graphical description of these associations where y_{1m}, y_{1f}, y_{2m}, y_{2f} represent the outcomes from male and female partners, respectively, at baseline and follow-up. We chose to estimate three sets of models depending on what copula parameters are estimated. The first and second model use the copula parameters of θ, θ_{12}, $\theta_{1.}$ and $\theta_{2.}$. However, the first model assumes that the within-couple correlation at each time point is the same ($\theta_{1.}$=$\theta_{2.}$), while the second model estimates these two parameters ($\theta_{1.}$ and $\theta_{2.}$) separately. The third model only estimates the copula parameter of θ. We consider 5 simulation scenarios (see Table 1) : (1) **strong** background association (θ) with **weak** additional within-individual correlation over time (θ_{12}) and **weak** additional within-couple correlation ($\theta_{1.}$, $\theta_{2.}$); (2) **weak** background association (θ) with **strong** additional within-individual correlation over time (θ_{12}) and **strong** additional within-couple correlation ($\theta_{1.}$, $\theta_{2.}$); (3) **moderate** background association (θ) with **strong** additional within-individual correlation over time (θ_{12}) and **strong** additional within-couple correlation ($\theta_{1.}$, $\theta_{2.}$); (4) **weak** background association (θ) with **weak** additional within-individual correlation over time (θ_{12}) and **weak** additional within-couple correlation ($\theta_{1.}$, $\theta_{2.}$) and (5) **strong** background association (θ) with no additional correlation for any pair. The third model can be applied in setting 9 only. For each setting within each model, 1,000 simulations were performed in order to evaluate the bias of the marginal probabilities (p's) and copula parameters (θ, θ_{12}, $\theta_{1.}$ and $\theta_{2.}$). In addition to estimating the copula parameters, we estimate all possible pair-wise correlation coefficients (ρ_i where $i = 1, \ldots, 6$ for 4 correlated outcomes) and compare them to the true values. We also present 95% coverage probabilities for the estimates of the copula parameters. We perform both Wald and likelihood ratio tests (LRT) to evaluate whether each copula parameter is significantly different from their respective null values (e.g., 0 for θ and 1 for θ_{12}, $\theta_{1.}$ and $\theta_{2.}$).

3.3.2 Simulation results

We explore the mean bias of the marginal probabilities, correlation coefficients and copula parameters. Table 2 displays the mean bias of the marginal probabilities of the outcomes at each time for each group (p_{1t}, p_{2t}, p_{1c}, p_{2c}). The mean bias ranges between 0.000 and 0.005, and is smaller either when both the background and additional associations are low (settings 7 and 8) or when the model has only one copula parameter (θ) to be estimated (setting 9). However, the differences in mean bias from different settings were very small ranging from 0.000 to 0.005. Table 3 displays the mean bias of the correlation coefficient for all possible pairs (ρ_1 and ρ_6 : within-individual correlation over time for males and females, respectively; ρ_2 and ρ_5 : within-couple correlation at baseline and follow-up, respectively; and ρ_3 and ρ_4: within-couple correlation at different times). The mean bias is smaller when there is no or low additional association (settings 1-2, 7-8 and 9). Within-couple correlation at different times seems to be overestimated, while within-couple correlation at the same time seems to be underestimated. However, the mean biases are very small in general ranging from 0.000 to 0.019. Table 4 demonstrates the results for copula parameters. Even though the mean bias of the correlation coefficients, which are derived from estimated copula parameters, are very small, we found some biased estimates for the copula parameters, especially . When additional association is weak (settings 7 and 8), we found less bias for copula parameters (θ_{12}, $\theta_{1.}$ and $\theta_{2.}$). Table 5 summarizes 95% coverage probabilities (CP) for the copula parameters. The 95% CP for the copula parameter (θ) seems close to its nominal value, and those for the other copula parameters (θ_{12}, $\theta_{1.}$ and $\theta_{2.}$) are fairly close to 0.95 in settings 1, 2, 7 and 8. However, when there is strong additional correlation with weak background correlation (settings 3 and 4), the 95% CP for the other copula parameters are low. Table 6 displays the test results for the copula parameters using Wald and likelihood ratio test (LRT). We expect a large p-value when there is weak additional or low background association, to help determine whether we need those parameters in the model. The Wald test seems consistently more conservative than LRT in all simulation settings. In settings 3 and 5, all three copula parameters are significant in the model. However, in settings 4 and

6, $\theta_{1.}$ becomes a non-significant copula parameter when we add one more parameter $\theta_{2.}$ in the model. Also, when we add $\theta_{2.}$ in the model, the likelihood does not change significantly. In analyzing data from the motivating AAC study, we use LRT for model selection and obtain p-values for θ using LRT by comparing the model with only θ to one with an independence correlation structure. In summary, the max-id logit copula model performs well in estimating marginal probabilities in all simulation settings. The estimates for copula parameters are biased in some settings where we set the copula parameters for additional pair-wise correlation as strong. However, the estimates of the pair-wise correlation coefficients are robust and consistent with the true values.

3.4 A motivating example

Our motivating example is a randomized controlled trial (RCT) of HIV serodiscordant, African American couples designed to assess the effect of a culturally tailored HIV/STD prevention intervention on sexually transmitted infections and risky sexual behaviors among couples. In this trial, couples assess their condom use and other shared sexual behaviors retrospectively at baseline, immediately following the eight week intervention (IPT), and at 6 and 12 months following the conclusion of the intervention. One strength of the couple-based design is that each outcome of shared sexual behaviors is reported independently by each partner, so that each can be modeled independently (male and female partners, separately) in a multivariate setting. In this context, there exist two sources of correlation: within-couple correlations from measuring each partner of each couple regarding their shared sexual behaviors and within-individual correlations induced from the repeated measures over time. In order to estimate the intervention effect for each partner while accounting for these multiple sources of correlation, we apply the max-id copulas to data from the HIV/STD Prevention Trial for African American Couples (AAC). As described in the simulation study, we choose 4 outcomes: male and female responses from baseline and one follow-up, respectively. The efficacy of HIV/STD prevention intervention was compared to a general health promotion intervention in terms of reducing sexual risk behaviors. We separately fit the model to two different sets of outcomes, one using baseline and IPT and the other using baseline and the 12 month follow-up. We fit a copula model with 3 copula parameters (θ, θ_{12} and $\theta_{1.}$), assuming that background, additional within-couple at baseline and the follow-up time and additional within-individual over time correlations exist. We further assume that the copula parameters (θ_{12}) for within-individual correlation are the same for males and females, and that copula parameters for within-couple correlation ($\theta_{1.}$, $\theta_{2.}$) are the same at both baseline and at follow-up time. We could fit more complex models with combinations of all possible copula parameters (e.g., estimating total 7 copula parameters is possible in this setting), however, we found that adding more copula parameters for these additional association does not improve the model fit (p-value from LRT: 0.389 (baseline vs IPT), 0.643 (baseline vs 12 mo f/u)). We added treatment, time and a treatment x time interaction terms as regression parameters in each univariate model. The estimates of the univariate regression parameters and 3 copula parameters are presented in Tables 7 and 8. We fit a model with 4 copula parameters by adding $\theta_{2.}$, but the model fit of this model is not statistically different from the model with 3 copula parameters (p-value from LRT$=$ 0.8 and 0.123 for Tables 7 and 8, respectively). All parameters are statistically significant except 'treatment', which represents no treatment effect at baseline between the two interventions as expected. The estimate for each copula parameter is also statistically significant from both the Wald and LR tests. LR tests are performed by setting corresponding copula parameters (θ_{12} and $\theta_{1.}$) equal to 1, which is the null value of the copula parameter for bivariate Gumbel copula. Table 9 displays the estimates and standard errors of the correlation for 6 possible pairs, which are derived from a bivariate max-id copula using the estimated 3 copula parameters. Couple-level correlations are consistent at around 0.4-0.48 regardless of time point; all other correlations are lower especially for the correlation between the outcome at baseline and the outcome at 12 month follow-up time.

There is strong evidence that the HIV/STD prevention intervention is more effective at reducing risky sexual behavior than the general health promotion intervention for both females and males, at both IPT and the 12 month follow-up. The size of the treatment effect seems larger at IPT, and decreases at the 12 month follow-up even though it remains statistically significant. The estimated correlation between couples is largest at baseline and decreases slightly over time. Within-individual correlation over time for both males and females seems moderate (0.317-0.353). Interestingly, within-couple correlations from different time points also appear to be moderate (0.22-0.29).

3.5 Discussion

In this work, we proposed an extension of max-id copula to couple-based longitudinal binary data. Note that this approach can be extended to binary longitudinal data or multivariate data with different kinds of outcomes. This modeling approach was selected based on the analysis goals of a recently concluded RCT to promote less risky sexual behaviors in an HIV-affected sample of couples. Also, our primary interest was in modeling correlation (refer to Figure 1). The copula approach proposed here allows direct and intuitive interpretation of the correlation, with decomposition of correlation into various components using several sets of copula parameters (e.g., across time, within-cluster, etc). An appealing feature of the model is the ability to allow the components of the copula parameter to depend on covariates of interest.

Figure 3.1: Correlation structure among outcomes presented by copula parameters

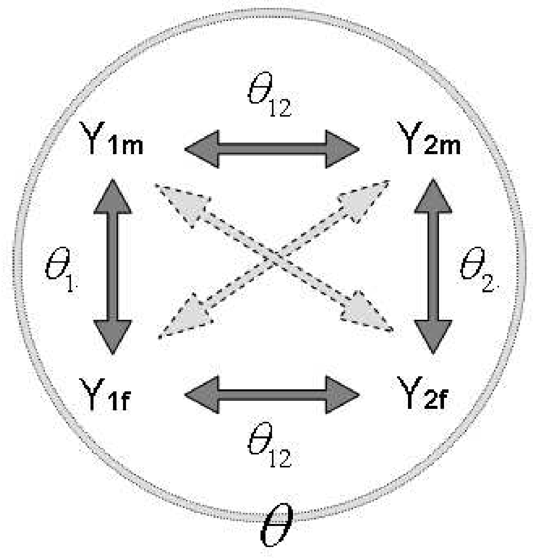


Table 3.1: Simulation set-up: copula parameter choices in different settings

Setting	Copula parameters	Background association	Additional association
1	$(\theta, \theta_{12}, \theta_{1.})$	High	Low
2	$(\theta, \theta_{12}, \theta_{1.}, \theta_{2.})$	High	Low
3	$(\theta, \theta_{12}, \theta_{1.})$	Low	High
4	$(\theta, \theta_{12}, \theta_{1.}, \theta_{2.})$	Low	High
5	$(\theta, \theta_{12}, \theta_{1.})$	Moderate	High
6	$(\theta, \theta_{12}, \theta_{1.}, \theta_{2.})$	Moderate	High
7	$(\theta, \theta_{12}, \theta_{1.})$	Low	Low
8	$(\theta, \theta_{12}, \theta_{1.}, \theta_{2.})$	Low	Low
9	$(\theta, \theta_{12}, \theta_{1.})$	High	

Table 3.2: Mean bias for estimators of the parameters of the marginal probabilities.

		Mean Bias			
		Treatment Group		Control Group	
Setting	True p	Baseline	F/U	Baseline	F/U
1	$(0.2, 0.6, 0.2, 0.6)$	0.0013	0.0038	0.0027	0.0009
2	$(0.2, 0.6, 0.2, 0.6)$	0.0019	0.0051	0.0024	0.0019
3	$(0.2, 0.5, 0.2, 0.3)$	0.0034	0.0010	0.0035	0.0017
4	$(0.2, 0.5, 0.2, 0.3)$	0.0030	0.0021	0.0026	0.0026
5	$(0.2, 0.5, 0.2, 0.3)$	0.0030	0.0027	0.0045	0.0021
6	$(0.2, 0.5, 0.2, 0.3)$	0.0027	0.0041	0.0037	0.0023
7	$(0.2, 0.6, 0.2, 0.4)$	0.0020	0.0004	0.0003	0.0016
8	$(0.2, 0.6, 0.2, 0.4)$	0.0004	0.0016	0.0003	0.0001
9	$(0.2, 0.6, 0.2, 0.4)$	0.0004	0.0013	0.0012	0.0032

F/U = follow-up

Table 3.3: Mean correlation structures and bias of the correlation coefficient estimators

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_4 & \rho_5 \\ \rho_2 & \rho_4 & 1 & \rho_6 \\ \rho_3 & \rho_5 & \rho_6 & 1 \end{pmatrix}$$

	True ρ						
Setting	$(\rho_1,\ \rho_2,\ \rho_3,\ \rho_4,\ \rho_5,\ \rho_6)$	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6
1	$(0.40, 0.40, 0.30, 0.30, 0.48, 0.40)$	0.002	0.002	0.009	0.009	0.003	0.002
2	$(0.40, 0.40, 0.30, 0.30, 0.50, 0.40)$	0.008	0.016	0.008	0.007	0.003	0.004
3	$(0.35, 0.35, 0.10, 0.10, 0.35, 0.35)$	0.014	0.008	0.019	0.019	0.004	0.017
4	$(0.35, 0.35, 0.10, 0.10, 0.35, 0.35)$	0.012	0.022	0.017	0.017	0.003	0.011
5	$(0.40, 0.40, 0.20, 0.20, 0.43, 0.40)$	0.010	0.008	0.014	0.013	0.008	0.010
6	$(0.40, 0.40, 0.20, 0.20, 0.40, 0.40)$	0.008	0.025	0.010	0.010	0.004	0.008
7	$(0.20, 0.25, 0.10, 0.10, 0.26, 0.20)$	0.005	0.003	0.005	0.005	0.002	0.005
8	$(0.20, 0.25, 0.10, 0.10, 0.25, 0.20)$	0.003	0.004	0.005	0.005	0.002	0.003
9	$(0.28, 0.28, 0.28, 0.28, 0.36, 0.28)$	0.001	0.001	0.001	0.002	0.000	0.002

Table 3.4: Mean difference $(\hat{\theta} - \theta)$ for the copula parameters

	True $\theta's$				
Setting	$(\theta, \theta_{12}, \theta_{1.}, \theta_{2.})$	θ	θ_{12}	$\theta_{1.}$	$\theta_{2.}$
1	$(3.500, 1.290, 1.161, -)$	0.368	0.025	-0.024	-
2	$(3.500, 1.290, 1.161, 1.219)$	0.002	0.104	-0.032	0.002
3	$(1.015, 1.399, 1, 336, -)$	0.243	-0.042	-0.027	-
4	$(1.015, 1.399, 1, 336, 1.331)$	0.219	-0.030	-0.021	-0.006
5	$(2.077, 1.378, 1.317, -)$	0.259	-0.027	-0.029	-
6	$(2.077, 1.378, 1.317, 1.253)$	0.217	0.002	-0.031	0.007
7	$(1.015, 1.139, 1.182, -)$	0.096	-0.007	-0.005	-
8	$(1.015, 1.139, 1.182, 1.169)$	0.101	-0.002	-0.005	0.007
9	$(3.249, -, -, -)$	0.113			

Table 3.5: 95% Coverage probabilities for the copula parameters

	Copula parameters			
Setting	θ	θ_{12}	$\theta_{1.}$	$\theta_{2.}$
1	94.4	94.6	91.2	
2	94.6	94/4	87.3	93.6
3	93.0	84.3	90.4	
4	93.2	87.5	85.9	93.3
5	94.2	89.0	89.8	
6	94.6	91.2	86.7	92.3
7	95.4	92.6	95.2	
8	94.0	92.2	91.3	94.9
9	94.8			

Table 3.6: Test results for the copula parameters by Wald test and Likelihood Ratio (LR) Tests

		θ		θ_{12}		$\theta_{1.}$		$\theta_{2.}$	
Setting	Copula parameters	Wald	LRT	Wald	LRT	Wald	LRT	Wald	LRT
1	$(\theta, \theta_{12}, \theta_{1.}, -)$	0.000	0.000	0.088	0.058	0.259	0.195	-	-
2	$(\theta, \theta_{12}, \theta_{1.}, \theta_{2.})$	0.000	0.000	0.098	0.054	0.352	0.272	0.265	0.246
3	$(\theta, \theta_{12}, \theta_{1.}, -)$	0.124	0.000	0.002	0.000	0.010	0.003	-	-
4	$(\theta, \theta_{12}, \theta_{1.}, \theta_{2.})$	0.131	0.000	0.003	0.000	0.072	0.026	0.043	0.029
5	$(\theta, \theta_{12}, \theta_{1.}, -)$	0.007	0.000	0.007	0.002	0.029	0.013	-	-
6	$(\theta, \theta_{12}, \theta_{1.}, \theta_{2.})$	0.010	0.000	0.009	0.002	0.120	0.061	0.123	0.101
7	$(\theta, \theta_{12}, \theta_{1.}, -)$	0.116	0.000	0.176	0.155	0.105	0.079	-	-
8	$(\theta, \theta_{12}, \theta_{1.}, \theta_{2.})$	0.134	0.000	0.180	0.156	0.201	0.152	0.266	0.244
9	$(\theta, -, -, -)$	0.000	0.000						

Table 3.7: Parameter estimation for the marginal model using copula model (outcomes at baseline vs IPT)

				p-value	p-value
	Covariates	Est.	S.E.	(Wald)	(LRT)
Female	Intercept	-0.448	0.134	0.001	
	Trt	0.063	0.190	0.739	
	Time	0.727	0.147	0.000	
	Trt×Time	1.081	0.232	0.000	
Male	Intercept	-0.312	0.133	0.019	
	Trt	-0.016	0.188	0.932	
	Time	0.650	0.146	0.000	
	Trt×Time	1.497	0.250	0.000	
Copula	θ_{12}	1.153	0.057	0.007	0.002
parameters	$\theta_{1.}$	1.254	0.077	0.001	0.000
	θ	3.036	0.388	0.000	-

Table 3.8: Parameter estimation for the marginal model using copula model (outcomes at baseline vs 12mos f/u)

				p-value	p-value
	Covariates	Est.	S.E.	(Wald)	(LRT)
Female	Intercept	-0.416	0.135	0.002	
	Trt	0.057	0.196	0.771	
	Time	0.493	0.152	0.001	
	Trt×Time	0.626	0.227	0.006	
Male	Intercept	-0.342	0.134	0.011	
	Trt	0.053	0.195	0.786	
	Time	0.388	0.151	0.010	
	Trt×Time	0.563	0.223	0.012	
Copula	θ_{12}	1.156	0.051	0.002	0.000
parameters	$\theta_{1.}$	1.323	0.073	0.000	0.000
	θ	2.340	0.338	0.000	-

Table 3.9: Estimated correlation coefficient among outcomes using copula model

	Correlations			
Outcomes	Baseline vs IPT		Baseline vs 12mo F/U	
(Pairs)	Est.	S.E.	Est.	S.E.
(y_{1f}, y_{1m})	0.479	0.039	0.447	0.035
(y_{2f}, y_{2m})	0.411	0.038	0.427	0.035
(y_{1f}, y_{2f})	0.353	0.042	0.317	0.039
(y_{1m}, y_{2m})	0.337	0.043	0.325	0.039
(y_{1f}, y_{2m})	0.267	0.044	0.223	0.036
(y_{1m}, y_{2f})	0.289	0.047	0.225	0.037

Chapter 4

A Copula-based Model for Longitudinal Data with Bivariate Binary Outcomes, with Application to Depression Data

4.1 Introduction

In longitudinal studies where multiple outcomes are measured repeatedly over time which may be correlated, it is important to investigate the marginal effect of covariates on each outcome as well as to accommodate the serial dependence of a single outcome within each subject and the dependence between responses at the same time and at different times across subjects. This paper presents a copula-based model for longitudinal bivariate binary data that estimates marginal covariate effects on outcomes of interest while taking multiple sources of dependence into account. The suggested model is illustrated by an examination of factors that are potentially associated with two different measures of depression.

Generalized linear models with random effects are commonly used to model repeated multivariate data. Using different choices of models for the marginal and dependence structures, several approaches have been proposed for longitudinal bivariate binary data. Ten Have and Morabia (1999) extended the original Dale (1986) model to accommodate the time component in analyzing longitudinal bivariate binary outcomes. Ribaudo et al. (2002) considered a generalized linear random coefficient model (Longford 1995) for repeated multivariate binary data as a hierarchical logistic regression model. The marginalized transition random effects models (MTM) were proposed to model multivariate longitudinal binary data by Ilk and Daniels (2007). The marginalized random effect model (MREM) was extended to model multivariate longitudinal binary data using a new covariance matrix with a Kronecker decomposition in Lee et al. (2009). However, common drawbacks of these methods include their reliance on full likelihood approaches with many nuisance parameters leading to diminished power, and the conceptual difficulty of modeling higher order associations in a flexible and interpretable manner (Bahadur 1961).

Generalized estimating equations (GEE) are an alternative to a full likelihood approach and are another commonly used method for analyzing longitudinal data. For example, Lipsitz et al. (2009) recently proposed a modified GEE for modeling multivariate longitudinal binary outcomes. Such marginal methods do not yield fully efficient estimates, nor consistent estimates when data are missing at random since the model is not fully specified. Thus, even using the modified GEE can still produce biased estimates of marginal regression parameters when outcomes are missing at random. Additionally, the GEE approach requires the specification of the association parameters among the different outcomes.

Using a non-parametric approach, Agresti (1997) proposed multivariate Rasch models for multivariate longitudinal binary data. Those models unfortunately do not allow continuous covariates, but are restricted to categorical predictors, which may be restrictive in many longitudinal settings.

Our proposed approach is an extension of the max-id copula approach of Nikolou lopoulos and Karlis (2008), enabling modeling of bivariate longitudinal data with more than two repeated measures. We consider a conditional joint distribution of the current observation of bivariate outcomes, given the joint distribution of the previous observation of the pair. This provides a dependence structure similar to a first-order Markov type structure when considering bivariate outcomes at each time as one unit. We construct the conditional joint distribution using a max-id copula approach so that we can separately model the univariate marginal probability of each outcome as well as accommodate the dependence between outcomes and over time. Unlike other likelihood-based approaches, this model is algebraically simple; we have a closed-form cumulative density function of joint probabilities as a function of the model parameters using the copula. Thus, maximum likelihood estimation is feasible and the number of terms in the likelihood is moderate with increased numbers of repeated measures. Treating missing data in this approach has not been explored in this context and is beyond the scope of the current article so we assume MAR and use multiple imputation for missing outcomes in the analysis presented here.

We illustrate the proposed approach using longitudinal depression data from an observational study. The two main study outcomes were different instruments used to assess depression in a general primary-care population. One instrument is a diagnostic tool that results in a binary indicator of major depressive disorder (MDD). The second is a screening measure of depression, the Hamilton Rating Score - Depression (HAMD), that results in a continuous score; an established cutoff value for this scale is indicative of high risk for a diagnosis of depression. Assuming that MDD and HAMD are correlated, a bivariate model is not only necessary for the estimation of the correlation between two outcomes, but should provide a better fit to the data than two separate univariate models. Both outcomes were collected at baseline and at two follow-up visits. Using the proposed model, we can accommodate the dependence between two outcomes at each measurement occasion and the serial dependence as well as capture the marginal effect on each outcome. In this paper, we primarily focus on estimating the marginal regression parameters for each outcome; for this goal the proposed approach is quite flexible in that the effects of covariates on the bivariate depression outcomes need not be identical. The goal of this analysis is to assess changes in longitudinal trend while identifying possible predictors of each outcome separately. This paper is organized as follows. In Section 4.2, we present a brief introduction for constructing the joint probability model for binary bivariate outcomes and a description of the extended model including max-id copula, the marginal regression model and the parameter estimation approach. In Section 3, we evaluate the performance of this approach using several sets of simulations. In Section 4.4, we illustrate this approach by analyzing the depression data from the motivating example. Concluding remarks appear in Section 4.5.

4.2 Statistical methods

4.2.1 Modified conditional model

Consider r correlated outcomes for any subject i, $y_{ij}^{(r)}$,$_{r=1,2}$ with t measurement times, $j = 1, \ldots, t$. We drop the index i to simplify notation. Let $f(y_1^{(1)}, y_1^{(2)}, y_2^{(1)}, y_2^{(2)},$ $\ldots, y_t^{(1)}, y_t^{(2)})$ denote the joint probability of all responses. The joint probability can be factorized as follows:

$$\begin{aligned} & f(y_1^{(1)}, y_1^{(2)}, y_2^{(1)}, y_2^{(2)}, \ldots, y_t^{(1)}, y_t^{(2)}) \\ & = f(y_1^{(1)}, y_1^{(2)}) f(y_2^{(1)}, y_2^{(2)} | y_1^{(1)}, y_1^{(2)}) f(y_3^{(1)}, y_3^{(2)} | y_1^{(1)}, y_1^{(2)}, y_2^{(1)}, y_2^{(2)}) \qquad (4.2.1) \\ & \ldots f(y_t^{(1)}, y_t^{(2)} | y_1^{(1)}, y_1^{(2)}), y_2^{(1)}, y_2^{(2)}, \ldots, y_{t-1}^{(1)}, y_{t-1}^{(2)}) \end{aligned}$$

For all j, we assume the conditional distribution of joint probability of the two correlated responses at the j^{th} observation $(y_j^{(1)}, y_j^{(2)})$ is dependent only on the $(j-1)^{th}$ bivariate response, $(y_{j-1}^{(1)}, y_{j-1}^{(2)})$. That is, for all j, the conditional distribution of $(y_j^{(1)}, y_j^{(2)})$ given $(y_1^{(1)}, y_1^{(2)}, \ldots, y_{(j-1)}^{(1)}, y_{j-1}^{(2)})$ is identical to the conditional distribution of $(y_j^{(1)}, y_j^{(2)})$ given $(y_{j-1}^{(1)}, y_{j-1}^{(2)})$ alone. Therefore, (1) can be re-expressed as

$$\begin{aligned} & f(y_1^{(1)}, y_1^{(2)}, y_2^{(1)}, y_2^{(2)}, \ldots, y_t^{(1)}, y_t^{(2)}) \\ & = f(y_1^{(1)}, y_1^{(2)}) f(y_2^{(1)}, y_2^{(2)} | y_1^{(1)}, y_1^{(2)}) f(y_3^{(1)}, y_3^{(2)} | y_2^{(1)}, y_2^{(2)}) \ldots f(y_t^{(1)}, y_t^{(2)} | y_{t-1}^{(1)}, y_{t-1}^{(2)}) \\ & = f(y_1^{(1)}, y_1^{(2)}) \prod_{j=2}^{t} \left[\frac{f(y_j^{(1)}, y_j^{(2)}, y_{j-1}^{(1)}, y_{j-1}^{(2)})}{f(y_{j-1}^{(1)}, y_{j-1}^{(2)})} \right] \end{aligned} \tag{4.2.2}$$

4.2.2 Max-id copula

In this subsection, we introduce a copula-based approach to obtain the joint probability function in Equation (4.2.2). By definition, a copula is a multivariate joint distribution defined on the m-dimensional unit cube $[0, 1]^m$ such that every marginal distribution is uniform on the interval $[0, 1]$. Let $F_j(Y_j)$ be the cumulative distribution function (cdf) of a univariate random variable y_j $(j = 1, \ldots, m)$; then there exists an m-dimensional copula, C, such that for all y in the domain of H,

$$H(y_1, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m)). \tag{4.2.3}$$

Equation (4.2.2) can now be re-expressed using copulas. At each time point j, for a given subject, we observe two related responses, and the corresponding likelihood contribution for a given subject can be written using the copula density function c and a set of copula parameters, Θ, as follows:

$$\begin{aligned}
&f(y_1^{(1)}, y_1^{(2)}, y_2^{(1)}, y_2^{(2)}, \ldots, y_t^{(1)}, y_t^{(2)}) \\
&= f(y_1^{(1)}, y_1^{(2)}) \prod_{j=2}^{t} \left[\frac{f(y_j^{(1)}, y_j^{(2)}, y_{j-1}^{(1)}, y_{j-1}^{(2)})}{f(y_{j-1}^{(1)}, y_{j-1}^{(2)})} \right] \\
&= f(y_1^{(1)}) f(y_1^{(2)}) c(F_1(y_1^{(1)}, F_2(y_1^{(2)}) | \Theta) \\
&\cdot \prod_{j=2}^{t} \left[\frac{f(y_j^{(1)}) f(y_j^{(2)}) f(y_{j-1}^{(1)}) f(y_{j-1}^{(2)}) c(F(y_j^{(1)}), F(y_j^{(2)}), F(y_{j-1}^{(1)}), F(y_{j-1}^{(2)}) | \Theta)}{f(y_{j-1}^{(1)}) f(y_{j-1}^{(2)}) c(F(y_{j-1}^{(1)}), F(y_{j-1}^{(2)}) | \Theta)} \right]
\end{aligned} \tag{4.2.4}$$

For discrete responses, a multivariate probability function is obtained by taking the Radon-Nikodym derivative for H(y) in Equation (4.2.3)(see Song 2000). Then, the multivariate joint probability of $(y_j^{(1)}, y_j^{(2)}, y_{j-1}^{(1)},\ y_{j-1}^{(2)})$ given by the copula representation is

$$\begin{aligned}
&c(F_1(y_j^{(1)}), F_2(y_j^{(2)}), F_1(y_{j-1}^{(1)}), F_2(y_{j-1}^{(2)}) | \Theta) \\
&= \sum_{k_1=1}^{2} \sum_{k_2=1}^{2} \sum_{k_3=1}^{2} \sum_{k_4=1}^{2} (-1)^{k_1+k_2+K_3+k_4} \cdot C[F_1(u_{j,k_1}^{(1)}), F_2(u_{j,k_2}^{(2)}), F_1(u_{j-1,k_3}^{(1)}), F_1(u_{j-1,k_4}^{(2)}) | \Theta]
\end{aligned}$$

where each $u_j^{(r)}$ is equal to $y_j^{(r)}$ or $y_j^{(r)} - 1$ for all j, r $(j = 2, \ldots, t, r = 1, 2, k_1 = 1, 2, k_2 = 1, 2, k_3 = 1, 2, k_4 = 1, 2)$.

Using Equation (4.2.4), we can construct the joint likelihood for the bivariate longitudinal outcomes of interest. Among several copula families, we chose the mixtures of max-id bivariate copulas proposed by Joe and Hu (1996) and applied to multivariate binary outcomes with a logit link proposed by Nikoloulopoulos and Karlis (2008). The mixtures of max-id copulas have flexible dependence structures, closed form cdfs, and satisfy the closure property under marginalization, which are desired properties for modeling binary data using a parametric family of multivariate copulas (Nikoloulopoulos and Karlis 2008). However, this approach does not allow negative correlation.

Specifically, the cdf of the mixture of m-variate max-id copulas has the following form

$$C(\mathbf{u};\Theta) = \phi\left(\sum_{j<k} log\, C'_{jk}(e^{-p_j\phi^{-1}(u_j;\theta)}, e^{-p_k\phi^{-1}(u_k;\theta)}; \theta_{jk}) + \sum_{j=1}^{m} v_j p_j \phi^{-1}(u_j;\theta);\theta\right) \tag{4.2.5}$$

where $C^{'}_{jk}(\cdot;\theta,\theta_{jk})$ is a bivariate max-id copula, $\phi(\cdot;\theta)$ is a Laplace transform (LT), $\Theta = \{\theta,\theta_{jk} : j,k = 1,...,m, j < k\}$ denotes the vector of all dependence parameters of the copula, u_j is cdf of a univariate random variable and $p_j = (v_j + m - 1)^{-1}$ where v_j is arbitrary. For details regarding the (j,k) bivariate marginal copula, $C^{'}_{jk}(\cdot;\theta,\theta_{jk})$, see Joe (1996). θ of LT represents a common level of dependence among all outcomes, whereas θ_{jk} describes the pairwise dependence among any pairs of outcomes, which is additional to a common level of dependence described by θ. Without loss of generality, we assume $v_j + m - 1 = 0$ and obtain simple max-id copulas with $\frac{(m\times(m-1))}{2} + 1$ dependence parameters for 4 correlated outcomes. We incorporate this formulation of max-id copulas to model bivariate longitudinal responses with adjacent time points, $y^{(1)}_{j+1}, y^{(2)}_{j+1}, y^{(1)}_{j}, y^{(2)}_{j}$ into Equation (4.2.4) using $\frac{4\times3}{2}+1$ possible dependence parameters. Thus, to model bivariate longtitudinal data measured with t times, we obtain $[(t - 1) \times (\frac{(4\times3)}{2} + 1) + (t - 2)]$ dependence parameters in total. Note that the number of dependence parameters can flexibly be reduced depending on how we parameterize the dependence model. For example, we can simplify the model by letting the additional dependence parameter between two adjacent single responses such as $y^{(1)}_{j+1}$ and $y^{(1)}_{j}$ or $y^{(2)}_{j+1}$ and $y^{(2)}_{j}$ be the same for each time lag and the additional dependence parameter between two different responses such as $y_j^{(1)}$ and $y_j^{(2)}$ be the same at each time j. Note that the several choices of copula families with 4 possible Laplace Transforms (LT) to construct a max-id copula will result in 20 different parametric copula families (Table 2.1).

4.2.3 Copula-based logit model

In this section, we demonstrate how to construct the marginal model for binary responses integrated into the copula-based approach described in Subsection 4.2.2. Considering two correlated responses $y_{ij}^{(r)}{}_{,r=1,2}$ with t measurement times $(j = 1, \ldots, t)$ as demonstrated in Subsection 2.1 and 2.2, we have a total of $(2 \times t)$ responses assuming that those have two separate cdfs, F_1 and F_2, for each response. Here we note one attractive feature of the copula approach in being able to estimate covariate effects independently for each outcome from the marginal distributions as well as construct the dependence structure separately from the margin. We assume that the two separate cdfs for each response, F_1 and F_2, are the cdfs of the univariate Bernoulli distribution functions with probability of success $(r = 1, 2)$.

$$F_r(y_j^{(r)}; \pi_r) = \begin{cases} 1 - \pi_r & \text{if } \ y_j^{(r)} = 0 \\ \\ 1 & \text{if } \ y_j^{(r)} = 1 \end{cases} \qquad j = 1, \ldots, t \ , r = 1, 2 \qquad (4.2.6)$$

The standard logistic regression model for the probability of success $\pi_r(r = 1, 2)$ is integrated using each cdf, F_1 and F_2 in Equation (4.2.4),

$$logit(\pi_r) = \beta_r^T X_{ijr}, \ \ i = 1, \ldots, n, \ j = 1, \ldots, t, \ r = 1, 2 \qquad (4.2.7)$$

where β_r is the vector of marginal regression parameters including coefficients for the effect of time and X_{ijr} is a vector of covariates for the r^{th} response at j^{th} time for the subject i. Adding a time variable in X_{ijr} allows us to estimate the longitudinal study effect.

4.2.4 Parameter estimation

We focus on the two-step inference functions of margins (IFM) method (Joe 1997) to reduce computational effort. In the first step, the r univariate log-likelihoods (Equation (4.2.8) are maximized independently in order to obtain r separate $\hat{\beta}_r$,

$$L_r(\beta_r) = \sum_{i=1}^{n} log f_r(y_{ijr}; \beta_r), \quad r = 1, 2 \tag{4.2.8}$$

where f_1 and f_2 are the univariate probabilities for each response. In the second step, the joint log-likelihood (Equation (4.2.9)) incorporating the formulation in Equation (4.2.4) is maximized over the set of the copula parameters (Θ) with $\hat{\beta}_r$ from the first step:

$$\begin{aligned} L(\beta, \Theta) &= \sum_{i=1}^{n} log f(y_1^{(1)}, y_1^{(2)}, \ldots, y_j^{(1)}, y_j^{(2)}; \hat{\beta}, \Theta) \\ &= \sum_{i=1}^{n} log[c(F_1(y_1^{(1)}), F_2(y_2^{(2)})|\Theta) \prod_{j=2}^{t} \frac{c(F_1(y_j^{(1)}), F_2(y_j^{(2)}), F_1(y_{j-1}^{(1)}), F_2(y_{j-1}^{(2)})|\Theta)}{c(F_1(y_{j-1}^{(1)}), F_2(y_{j-1}^{(2)})|\Theta)}; \hat{\beta}]. \end{aligned} \tag{4.2.9}$$

4.3 Simulation study

4.3.1 Simulation setting

We conducted a simulation study to evaluate the performance of the extended max-id copula approach in estimating marginal probabilities. We generated 100 samples of 8 outcomes for each setting (bivariate responses at 4 different time points). We considered 6 settings with 1000 repetitions of each (Figure 4.1). In settings 1 and 2, we assumed that bivariate outcomes measured at the same time and adjacent times have the same level of correlation for these 4 outcomes, mimicking an exchangeable correlation structure. Non-adjacent outcomes were assumed to be uncorrelated. The magnitude of correlation differs in setting 1 and 2. In settings 3 and 4, we set the correlation between 8 outcomes to be correlated with the exchangeable correlation structure and assumed the magnitude of correlation differs in setting 3 and 4. Setting 5 and 6 are similar to setting 1 and 2 in terms of that bivariate outcomes at the same time and adjacent time are set to be correlated among 4 outcomes, however, but the magnitude of correlation between two outcomes at the same time differs from those of the other pairs of correlation; there is a moderate level of correlation for bivariate outcomes at the same time and a weaker level of correlation for the other pairs.

In settings 1, 2, 3 and 4, among 23 possible copula parameters $((4-1)\times(\frac{(4\times3)}{2}+1)+(4-2)=23)$, we reduce them to 1 copula parameter, θ, to incorporate simple dependence structure by estimating only minimal level of correlations among 4 outcomes (bivariate outcomes measured at the same time and adjacent times). Settings 5 and 6 use two copula parameters, θ and θ_{13}, to estimate more complicated dependence structure, where θ describes a common level of correlation among 4 outcomes and θ_{13} is used to estimate an additional correlation among bivariate outcomes at each time point. Correlated binary data for a given correlation matrix were generated using the R package 'mvtBinaryEP,' which applies an algorithm developed in Emrich and Piedmonte (1991). In each setting, we generated marginal probabilities for each outcome where the true probabilities are different depending on the time and covariate (Table 4.1). In this simulation study, we add only one covariate and 3 indicator variables for time in each marginal probability. Table 4.2 displays the estimates and p -values for the copula parameters in each setting, where bigger estimates of the copula parameters represent a stronger dependence.

4.3.2 Simulation results

We explored the mean bias of the marginal probabilities. Table 4.1 displays the mean bias of the marginal probabilities of the outcomes at each time for each covariate group. The mean bias ranges between 0.000 and 0.016. The biases are slightly bigger in settings 3 and 4 where we generated the outcomes based on the exchangeable correlation structure among 8 outcomes, which is different from our scheme of modeling dependence. However, even when the dependence model, where only adjacent and bivariate outcomes are correlated, is different from the true dependence structure, the biases were small. In general, when we generated the outcomes with stronger association, the biases were bigger. Settings 5 and 6 have one more copula parameter to estimate more complicated correlation structures and provide smaller bias than settings 1 and 2. Table 4.2 describes the estimates of the copula parameters with corresponding p-values. Deviation of the estimates of the copula parameter from 1 indicates stronger association, while the estimates of the parameter close to 1 means no minimal or additional association for corresponding pairs (for θ or θ_{13} respectively). The Wald test was performed to show whether the estimates of the copula parameters are equal to 1, which is a null value. The LR test was done to describe whether there is a significant difference in likelihood after adding the copula parameters with the null value, 1. For example, in setting 1, the p-value for θ is 0.013 and 0.005 from both tests showing that the estimate of θ is not the null value. In setting 2, bigger estimates and smaller p-value of θ indicate stronger association.

4.4 Application

As an illustration, we applied our method to an observational study for depression in patients in primary care. Primary care providers (PCPs), including physicians (MDs), nurse practitioners (NPs), and physician assistants (PAs) were recruited from Clinical Care Associates (CCA) of the University of Pennsylvania Health System. There are 37 primary care practices and over 80 providers in CCA. Consenting PCPs completed brief entry and exit questionnaires about their attitudes, communication style and care delivery with regard to depression. Among several outcomes for depression diagnosis, we selected the two likely correlated primary depression outcomes (MDD and HAMD). These two outcomes were collected at baseline and at two follow-up visits during the study. For our purpose, MDD represents a binary outcome indicating a diagnosis of major depression and HAMD score was dichotomized using a standard cut-off score characterizing moderate to severe depression. To handle the missing data, we used a multiple imputation procedure (Rubin 1987) to predict missing values using PROC MI in SAS with The expectation-maximization (EM) algorithm (Little and Rubin 1987) and produced 100 multiple imputed data sets. Analysis was carried out on each imputed data set and combined to produce one overall analysis using PROC MIANALYZE in SAS.

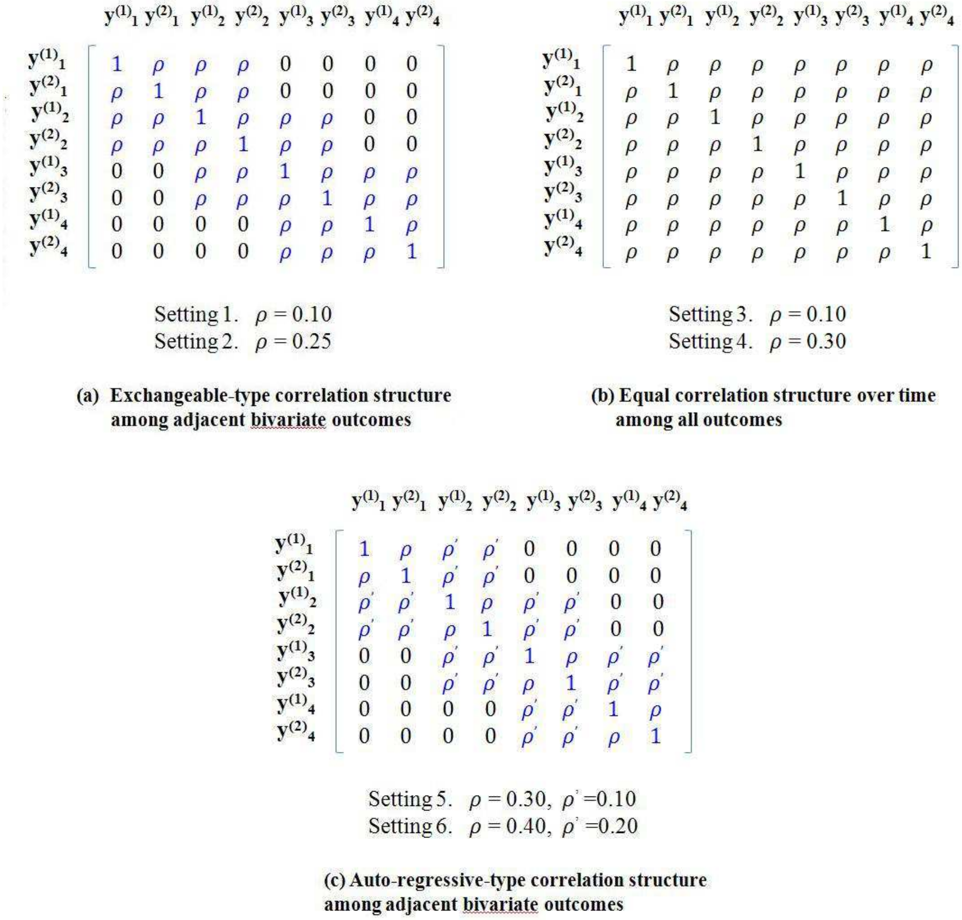
Using the proposed approach, we can model the marginal probability of MDD and HAMD using separate sets of covariate as predictors as well as incorporating the association between the two outcomes and serial dependence. As described in Section 2, we assume a logit model for each marginal probability using standard logistic regression. As possible covariates, we chose baseline neuroticism score (continuous), previous history of MDD (binary), baseline SCID score, insurance status and time. We used standard backward elimination to find the best subset of predictors for the two marginal models, while simultaneously incorporating dependence. In the marginal model of MDD, neuroticism has strong interaction with follow-up time (p-value$= 0.017$). Therefore, the probability of MDD is similar across follow-up times for patients who have low baseline neuroticism scores, however, this probability significantly decreases over time among patients with moderate or high baseline neuroticism scores (Table 4.4). In the marginal model of HAMD, the final subset of predictors is different from those predicting MDD. Although baseline neuroticism is a significant predictor in the model for HAMD, there was no significant interaction between baseline neuroticism and follow-up times as was observed in the marginal model for MDD. Thus, the probability of a HAMD score above the cutoff decreased over time regardless of baseline neuroticism (Table 4.4).

To incorporate correlations among outcomes, we used three copula parameters, $\theta_{.}$, θ_{12}, θ, in the model. θ describes a common level of dependence among all bivariate outcomes with time j and $j+1$ ($j = 1, 2$), whereas $\theta_{.}$ and θ_{12} describe an additional pairwise dependence among bivariate outcomes at j^{th} time and among single outcomes at j and $(j+1)^{th}$ measurement time ($j = 1, 2$). We found a moderate correlation among all outcomes ($\theta = 1.253$, p-value=< 0.000) and additional strong association between the two responses, MDD diagnosis and elevated HAMD score, at the same measurement times ($\theta_{.} = 1.444$, p-value=< 0.000). Additional serial correlation within each outcome described by θ_{12} was not significantly strong ($\theta_{12} = 1.039$, p-value$= 0.303$).

4.5 Conclusions

In this paper, we have presented an extended max-id copula approach to modeling bivariate longitudinal binary data. The proposed approach has an attractive feature in terms of separately estimating the marginal probability of each outcome as well as constructing flexible dependence structure. We assumed the current outcomes only depend on the previous outcomes, which mimics a first-order Markov type correlation structure for bivariate outcomes. Thus, the outcomes at non-adjacent times are not included in estimating their correlations in this approach. The model could be extended to incorporate the dependence among outcomes at non-adjacent times, but will be computationally tedious. The results of simulations have shown that this approach provides unbiased estimates of the marginal probability even when the correlation model is misspecified. We obtained the estimates of copula parameters with p-values, which provide some information of the magnitude of dependence. However, we find it hard to transform the values to a commonly used one such as a correlation coefficient.

Figure 4.1: Simulation set-up based on the correlation structure among outcomes



$$\begin{array}{c|cccccccc} & y^{(1)}_1 & y^{(2)}_1 & y^{(1)}_2 & y^{(2)}_2 & y^{(1)}_3 & y^{(2)}_3 & y^{(1)}_4 & y^{(2)}_4 \\ \hline y^{(1)}_1 & 1 & \rho & \rho & \rho & 0 & 0 & 0 & 0 \\ y^{(2)}_1 & \rho & 1 & \rho & \rho & 0 & 0 & 0 & 0 \\ y^{(1)}_2 & \rho & \rho & 1 & \rho & \rho & \rho & 0 & 0 \\ y^{(2)}_2 & \rho & \rho & \rho & 1 & \rho & \rho & 0 & 0 \\ y^{(1)}_3 & 0 & 0 & \rho & \rho & 1 & \rho & \rho & \rho \\ y^{(2)}_3 & 0 & 0 & \rho & \rho & \rho & 1 & \rho & \rho \\ y^{(1)}_4 & 0 & 0 & 0 & 0 & \rho & \rho & 1 & \rho \\ y^{(2)}_4 & 0 & 0 & 0 & 0 & \rho & \rho & \rho & 1 \end{array}$$

Setting 1. $\rho = 0.10$
Setting 2. $\rho = 0.25$

(a) Exchangeable-type correlation structure among adjacent bivariate outcomes

$$\begin{array}{c|cccccccc} & y^{(1)}_1 & y^{(2)}_1 & y^{(1)}_2 & y^{(2)}_2 & y^{(1)}_3 & y^{(2)}_3 & y^{(1)}_4 & y^{(2)}_4 \\ \hline y^{(1)}_1 & 1 & \rho & \rho & \rho & \rho & \rho & \rho & \rho \\ y^{(2)}_1 & \rho & 1 & \rho & \rho & \rho & \rho & \rho & \rho \\ y^{(1)}_2 & \rho & \rho & 1 & \rho & \rho & \rho & \rho & \rho \\ y^{(2)}_2 & \rho & \rho & \rho & 1 & \rho & \rho & \rho & \rho \\ y^{(1)}_3 & \rho & \rho & \rho & \rho & 1 & \rho & \rho & \rho \\ y^{(2)}_3 & \rho & \rho & \rho & \rho & \rho & 1 & \rho & \rho \\ y^{(1)}_4 & \rho & \rho & \rho & \rho & \rho & \rho & 1 & \rho \\ y^{(2)}_4 & \rho & \rho & \rho & \rho & \rho & \rho & \rho & 1 \end{array}$$

Setting 3. $\rho = 0.10$
Setting 4. $\rho = 0.30$

(b) Equal correlation structure over time among all outcomes

$$\begin{array}{c|cccccccc} & y^{(1)}_1 & y^{(2)}_1 & y^{(1)}_2 & y^{(2)}_2 & y^{(1)}_3 & y^{(2)}_3 & y^{(1)}_4 & y^{(2)}_4 \\ \hline y^{(1)}_1 & 1 & \rho & \rho' & \rho' & 0 & 0 & 0 & 0 \\ y^{(2)}_1 & \rho & 1 & \rho' & \rho' & 0 & 0 & 0 & 0 \\ y^{(1)}_2 & \rho' & \rho' & 1 & \rho & \rho' & \rho' & 0 & 0 \\ y^{(2)}_2 & \rho' & \rho' & \rho & 1 & \rho' & \rho' & 0 & 0 \\ y^{(1)}_3 & 0 & 0 & \rho' & \rho' & 1 & \rho & \rho' & \rho' \\ y^{(2)}_3 & 0 & 0 & \rho' & \rho' & \rho & 1 & \rho' & \rho' \\ y^{(1)}_4 & 0 & 0 & 0 & 0 & \rho' & \rho' & 1 & \rho \\ y^{(2)}_4 & 0 & 0 & 0 & 0 & \rho' & \rho' & \rho & 1 \end{array}$$

Setting 5. $\rho = 0.30$, $\rho' = 0.10$
Setting 6. $\rho = 0.40$, $\rho' = 0.20$

(c) Auto-regressive-type correlation structure among adjacent bivariate outcomes

Table 4.1: Mean bias of the marginal probabilities for each outcome

			Mean Bias			
			Outcome 1			
Settings	true prob.	Covariate	Time1	Time2	Time3	Time4
1	(0.3,0.4,0.4,0.4)	0	0.000	0.001	0.004	0.002
	(0.3,0.7,0.7,0.7)	1	0.005	0.001	0.002	0.004
2	(0.3,0.4,0.4,0.4)	0	0.007	0.009	0.009	0.009
	(0.3,0.7,0.7,0.7)	1	0.012	0.006	0.007	0.000
3	(0.3,0.4,0.4,0.4)	0	0.000	0.001	0.005	0.003
	(0.3,0.7,0.7,0.7)	1	0.005	0.001	0.001	0.003
4	(0.3,0.4,0.4,0.4)	0	0.010	0.010	0.014	0.011
	(0.3,0.7,0.7,0.7)	1	0.015	0.009	0.009	0.003
5	(0.3,0.4,0.4,0.4)	0	0.001	0.001	0.005	0.003
	(0.3,0.7,0.7,0.7)	1	0.002	0.000	0.002	0.003
6	(0.3,0.4,0.4,0.4)	0	0.005	0.005	0.007	0.006
	(0.3,0.7,0.7,0.7)	1	0.005	0.003	0.004	0.000
			Outcome 2			
1	(0.3,0.4,0.4,0.4)	0	0.001	0.002	0.003	0.001
	(0.3,0.7,0.7,0.7)	1	0.006	0.003	0.001	0.001
2	(0.3,0.4,0.4,0.4)	0	0.009	0.010	0.009	0.008
	(0.3,0.7,0.7,0.7)	1	0.012	0.002	0.007	0.004
3	(0.3,0.4,0.4,0.4)	0	0.001	0.002	0.003	0.001
	(0.3,0.7,0.7,0.7)	1	0.006	0.003	0.001	0.000
4	(0.3,0.4,0.4,0.4)	0	0.012	0.011	0.011	0.009
	(0.3,0.7,0.7,0.7)	1	0.016	0.006	0.009	0.005
5	(0.3,0.4,0.4,0.4)	0	0.004	0.003	0.003	0.003
	(0.3,0.7,0.7,0.7)	1	0.003	0.003	0.001	0.000
6	(0.3,0.4,0.4,0.4)	0	0.006	0.006	0.006	0.006
	(0.3,0.7,0.7,0.7)	1	0.006	0.000	0.003	0.003

Table 4.2: Estimates and p-values for the copula parameters in each setting

				p-value	p-value
Settings	Copula parameters	Est.	S.E.	(Wald)	(LRT)
1	θ	1.148	0.048	0.013	0.005
2	θ	1.574	0.096	0.000	0.009
3	θ	1.149	0.048	0.016	0.010
4	θ	1.781	0.118	0.000	< 0.000
5	θ_{13}	1.544	0.126	0.000	< 0.000
	θ	1.134	0.057	0.076	0.004
6	θ_{13}	1.566	0.127	0.000	< 0.000
	θ	1.420	0.089	0.000	< 0.000

Table 4.3: Parameter estimation for the marginal model using extended max-id copula and corresponding predictive probabilities of MDD diagnosis and high HAMD score (number of multiple imputation:100)

Outcomes	Covariates	Est.	S.E.	p-value (Wald)	p-value (LRT)
MDD	Intercept	-3.970	0.728	< 0.001	
	Neuroticism	0.036	0.007	< 0.001	
	Time1	1.522	0.999	0.128	
	Time2	1.589	1.105	0.151	
	Neuro × Time1	-0.023	0.009	0.017	
	Neuro × Time2	-0.025	0.010	0.017	
HAMD	Intercept	-2.311	0.529	< 0.001	
	Neuroticism	0.018	0.005	0.000	
	Time1	-0.383	0.208	0.066	
	Time2	-0.952	0.247	0.000	
Copula	$\theta_{\cdot}$	1.444	0.137	< 0.001	< 0.001
parameters	θ_{12}	1.039	0.049	< 0.001	0.303
	θ	1.253	0.086	< 0.001	< 0.001

Neuroticism	Time	Prob. of MDD	Prob. of HAMD
50	Baseline	0.102	0.196
	Time 1	0.142	0.143
	Time 2	0.138	0.086
100	Baseline	0.409	0.375
	Time 1	0.241	0.290
	Time 2	0.217	0.188
150	Baseline	0.807	0.596
	Time 1	0.378	0.501
	Time 2	0.325	0.363

Chapter 5

Conclusion

We have proposed a copula-based approach using max-id copulas for modeling bivariate longitudinal binary data and estimating their dependence structure. Examples of this kind of data include studies in couples whose correlated responses are assessed over time, or longitudinal studies in which multiple correlated outcome assessments are taken on individuals at different intervals. In both cases we are interested in assessing both correlation within an individual or outcome over time, and correlations across members of a pair or cluster. In Chapter 2, we applied this modeling approach to estimate the reliability (dependence) of self-reported, shared behaviors of couples in cross-sectional data using a max-id copula, while constructing the dependence model to explore the influence of additional covariate information on the dependence measure. This approach is useful since it allows estimation of the covariate effect on the marginal probabilities as well as the covariate effect on the dependency. In this context, the estimation of within-couple dependency is a useful proxy for the reliability of subjects' responses. Using a simulation study, we found that the approach performs well. We illustrated our method using data from the Multisite HIV/STD Prevention Trial for African American Couples (AAC) Study to investigate the reliability of couple reports of sexual activity, adjusting for key individual baseline covariates. This approach allows us to model dependence among outcomes, and in addition to model whether patients' characteristics affect the outcomes' level of dependency. The results indicated that the dependency among couples' outcome responses, a good measure of reliability, differs depending on their health insurance status.

In Chapter 3, we extended the max-id copula approach to be applied in bivariate longitudinal data with two assessment times. This approach allows flexibility in the construction of a complex correlation structure via the copula parameter. We performed simulations, evaluating the performance of estimated covariate effects on the marginal probabilities and of copula parameters by summarizing bias and coverage for a number of simulations. In all simulation scenarios, a max-id copula approach performed well in estimating the marginal probabilities of interest and the correlations among outcomes. This approach was also illustrated using couples' longitudinal outcomes from the AAC study. The result showed that the HIV/STD prevention intervention is more effective at reducing risky sexual behavior than the general health promotion intervention for both females and males. The size of the treatment effect seems larger at IPT, and decreases at the 12 month follow-up even though it remains statistically significant. Also, the size of the treatment effect in males was larger than in females. In Chapter 4, we proposed an extended max-id approach for incorporating more repeated measures for bivariate outcomes using joint transition probabilities. To allow a complex correlation structure, the distribution of the current observation of each bivariate outcome was modeled conditional on the previously observed values using a Markov type correlation structure. Conditional probabilities were constructed using a max-id copula. We focused on estimating the covariate effects on the marginal probabilities in this work. We applied this approach to investigate the factors affecting two correlated depression measures assessed simultaneously in patients receiving primary care. Specifically, major depressive disorder (MDD) diagnosis and the Hamilton rating scale for depression (HAMD) were selected as two correlated primary outcomes of interest. An attractive feature of this modeling approach is that it allows us to consider both primary outcomes in a single model, but allows different sets of covariates to be associated with each outcome separately, while easily incorporating the correlation between two outcomes and from repeated measures of the same outcome over time. Several simulation studies were performed to evaluate the performance of this method, and revealed that the proposed copula-based modeling approach produced unbiased estimates of covariate effects on the marginal probabilities of each outcome. The methods described here do not accommodate missing data; this is an interesting topic for future study, but was unfortunately beyond the scope of this dissertation work. Also, incorporating multivariate outcomes rather than bivariate outcomes would be a valuable statistical tool. The model proposed in Chapter 4 can be easily extended to include the dependence model with several sets of covariates of interest. Also, comparing the performance of this approach in modeling multivariate longitudinal data to other common methods such as GEE or generalized linear mixed effect model, would be useful in evaluating this approach. This modeling approach presented in this dissertation provides a useful tool for understanding intervention effects using longitudinal couples' outcomes from HIV prevention trials or investigating the factors affecting correlated outcomes of depression. This approach can be easily applicable in other studies that involve two correlated primary outcomes with repeated measures.

Bibliography

Agresti, A. (1997) A model for repeated measurements of a multivariate binary response.*Journal of the American Statistical Association*; **92**:315-321.

Bahadur,R.R. (1961) A representation of the joint distribution of responses to n dichotomous items. In Solomon,H.(ed.), *Studies in Item Analysis and Prediction*. Stanford University Press: 158-168.

Bellamy, S.L. and NIMH Multisite HIV/STD Prevention Trial for African American Couples Study Group. (2005) A dynamic block-randomization algorithm for group-randomized clinical trials when the composition of blocking factors is not known in advance. *Contemp Clin Trials*; **26**: 469–479.

Carey, V., Zeger, S.L. and Diggle, P. (1993). Modelling Multivariate Binary Data with Alternating Logistic Regressions. *Biometrika*; **80**: 517–526.

Catania, J.A., Gibson, D.R., Chitwood, D.D. and Coates, T.J. (1990). Methodologic problems in AIDS behavioral research: Influences on measurement error and participation bias in studies of sexual behavior.*Psychol Bull*;**108**: 339–362.

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

El-Bassell, N., Jemmott, J.B., Wingood, G.M., Wyatt, G.E., Pequegnat, W., Landis, J.R., Bellamy, S.L. and the NIMH Multisite HIV/STD Prevention Trial for African American Couples Group (2010). National Institute of Mental Health Multisite Eban HIV/STD Prevention Intervention for African American HIV Serodiscordant Couples: A Cluster Randomized Trial. *Archives of Internal Medicine*; **170**: 1594–1601.

Emrich, L.J. and Piedmonte, M.R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*; **45**: 302–304.

Escarela, G., Luis Carlos, P.R. and Russell, J.B. (2009). A copula-based Markov chain model for the analysis of binary longitudinal data. *Journal of Applied Statistics*; **36**: 647–657.

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

Gueorguieva, R. (2001) A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*; **1**:177–193.

Heagerty, P.J. (2002) Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*; **58**:342–351.

Ilk, O. and Daniels, M. (2007) Marginalized transition random effects models for multivariate longitudinal binary data. *Canadian Journal of Statistics*; **35**:105–123.

Joe, H. and Hu, T. (1996). Multivariate distributions from mixtures of max-infinitely divisible distributions. *Journal of Multivariate Analysis*; **57**: 240–265.

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

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*; **94**: 401–419.

Lambert, P. and Vandenhende, F. (2002). A copula based model for multivariate non normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine*; **21**: 3197—3217.

Lee, K. and Daniels, M.(2007) A class of markov models for longitudinal ordinal data. *Biometrics*; **63**:1060–1067.

Lee, K. and Daniels, M.(2008) Marginalized models for longitudinal ordinal data with application to quality of life studies. *Statistics in Medicine*; **27**:4359–4380.

Lee, K., Joo, Y., Yoo, J.K. and Lee, J.B. (2009) Marginalized random effects models for multivariate longitudinal binary data *Statistics in Medicine*;**28**:1284–1300.

Liang, K.Y. and Zeger, S.L.(1986). Longitudinal data analysis using generalized linear models.*Biometrika*; **73**: 13–22.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. B*; **54**: 3–40.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1991) Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure of Association. *Biometrika*; **78(1)**: 153–160.

Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J.G., Sinha, D., Parzen, M. and Lipshultz, S. (2009) Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data. *J. R. Statist. Soc. A*; **172**: 3–20.

Little, R.L. and Rubin, D.B. (1990). *Statistical analysis with missing data*. New York: Wiley.

Longford, N.T. (1995). *Random coefficient models*. Oxford: Oxford Science.

Miglioretti, D. and Heagerty, P. (2004) Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* 5:381–398.

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

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

Nikoloulopoulos, A.K. and Karlis, D. (2009) Finite normal mixture copulas for multivariate discrete data modeling. *Journal of Statistical Planning and Inference*; **139(11)**: 3878–3890.

NIMH Multisite HIV/STD Prevention Trial for African American Couples Group(a).(2008) Eban HIV/STD risk reduction intervention: conceptual basis and procedures. *Journal of acquired immune deficiency syndromes*;**49**: Suppl 1:15-27.

NIMH Multisite HIV/STD Prevention Trial for African American Couples Group(b).(2008) Eban health promotion intervention: conceptual basis and procedures. *Journal of acquired immune deficiency syndromes*; **49**: Suppl 1:28-34.

Ochs, E.P. and Binik, Y.M. (1999) The use of couple data to determine the reliability of self-reported sexual behavior. *Journal of Sex Research*; **36**: 374–384.

Qaqish, B.F. (2003) A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*; **90**: 455–463.

Reboussin, B.A., Anthony, J.C. (2001). Latent class marginal regression models for longitudinal data: modeling youthful drug involvement and its suspected influences. *Statistics in Medicine*; **20**: 623–639.

Ribaudo, H.J. and Thompson, S.G. (2002). The analysis of repeated multivariate binary quality of life data: a hierarchical model approach. *Statistical Methods in Medical Research*; **11**:69–83.

Rubin, D.B. (1987). *Multiple imputation for survey nonresponse*. New York: Wiley.

Seal, D.W. (1997). Interpartner concordance of self-reported sexual behavior among college dating couples. *Journal of Sex Research*; **34**:39–55.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de ´linstitut de statistique de ´luniversité de Paris*;**8**: 229–231.

Song, P.X.K. (2000). Multivariate dispersion models generated from Gaussian copulas. *Scandinavian Journal of Statistics*; **27**: 305–320

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

Streiner, D.L. and Norman, G.R. (1995). *Health measurement scales: Practical guide to their development and use (3rd ed.).*Oxford University Press: New York.

Ten Have, T.R., Kunselman, A.R. and Tran, L.A. (1999). A comparison of mixed effects logistic regression models for binary data with two nested levels of clustering.*Statistics in Medicine*; **18**: 947–960.

Ten Have, T.R. and Morabia, A.(1999). Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data. *Biometrics*; **55**: 85–93.

Witte, S.S., El-Bassel, N., Gilbert, L., Wu, E. and Chang, M. (2007). Predictors of discordant reports of sexual and HIV/sexually transmitted infection risk behaviors among heterosexual couples. *Sexually transmitted diseases*; **34**: 302–308.