Flexible Modeling of non-Gaussian Longitudinal Data: Some Approaches using Copula

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in
Statistics
by

Subhajit Chattopadhyay

Applied Statistics Unit
Indian Statistical Unit, Kolkata
May, 2025

Thesis Title:

Flexible Modeling of non-Gaussian Longitudinal Data: Some Approaches using Copula

A thesis submitted in partial fulfillment of the requirements for the degree of *Doctor of Philosophy* in *Statistics*.

Author: Subhajit Chattopadhyay

Institution: Applied Statistics Unit
Indian Statistical Institute, Kolkata
May, 2025

Supervisor: Prof. Sumitra Purkayastha

I dedicate this dissertation to my parents and grandparents.

List of Research Articles

I declare that the research presented in this dissertation is based on the following list of research articles.

- Chapter 3 of this dissertation is based on the paper:
  S. Chattopadhyay, K. Das and S. Purkayastha, "Skew-elliptical copula based mixed models for non-Gaussian longitudinal data with application to an HIV-AIDS study" (archived).
  Pre-print: https://doi.org/10.48550/arXiv.2402.00651.v3.

- Chapter 4 of this dissertation is based on the paper:
  S. Chattopadhyay, "Factor copula models for non-Gaussian longitudinal data" (under review).
  Pre-print: https://doi.org/10.48550/arXiv.2402.00668

- Chapter 5 of this dissertation is based on the paper:
  S. Chattopadhyay, "Finite mixture copulas for modeling dependence in longitudinal count data" (appeared in *Metron*).
  Article: https://doi.org/10.1007/s40300-025-00292-z

- Chapter 6 of this dissertation is based on the paper:
  S. Chattopadhyay, "Modeling temporal dependence of longitudinal data: use of multivariate geometric skew-normal copula" (appeared in *Journal of Statistical Theory and Practice*).
  Article: https://doi.org/10.1007/s42519-025-00451-5

Declaration about Codes

The programs written and used in this dissertation will be available upon request by email to the following id: subhajitc.stat.isi@gmail.com

Acknowledgement

I begin my expressing my gratitude to Indian Statistical Institute (ISI) for providing me with financial assistance through the research fellowship to support my doctoral studies. I have also used and benefited from infrastructural facilities of both ISI and Applied Statistics Unit (ASU).

I joined ISI as a Junior Research Fellow in July, 2018. After going through the initial course work, I became part of ASU in August, 2019, as one of its Research Scholars. I became attached with Prof. Sumitra Purkayastha in late January, 2020 when he was assigned my supervisor. I began working on a problem jointly with him and Prof. Kalyan Das (former Professor of Dept. of Statistics, Calcutta University who was later associated with Dept. of Mathematics, Indian Institute of Technology, Bombay) during this stage. However, the huge calamity on society and human life induced by Covid did not allow us to develop the kind of collaboration which was needed at that stage. This unforeseen difficulty, however, turned out to be an opportunity in disguise. I began searching for literature related to my work and also started working on certain problems on my own. Prof. Purkayastha encouraged me to pursue my interests; we continued discussing nevertheless. During this phase, I had several valuable discussions with him. In addition, he put me in touch with several internationally acclaimed statisticians who visited ISI for discussion about my work. Also, he encouraged me to contact and solicit views of acclaimed experts related to my research work. All these have helped me in my work. Much of my work has been included in this dissertation. During the last phase of my work when I was preparing the dissertation, there has been huge interaction between me and Prof. Purkayastha–more like a fresh beginning, clarifying many important issues and resulting in number of changes in the dissertation, leading to substantial improvement in presentation. I am grateful to him for all these.

Besides everything mentioned above, I am grateful to Prof. Purkayastha also because he took care of my interests outside academics. In particular, he encouraged me to search for job, and offered very useful suggestions in this regard and put me in touch with several persons. Little did I know at that time that I would be working in an organization where my expertise will be useful. I express my gratitude to Profs. Sumitra Purkayastha and Kalyan Das for allowing me to include our joint work in this dissertation.

I express my gratitude to all my friends with whom I had the pleasure of sharing unforgettable moments at the institute. Only a research scholar will know why such friendships are valuable in his/her life. My special thanks go to Pintu-da (Mr. Pintu Layek), Didi (Ms. Sandipa Roy) and Kaushik-da (Mr. Kaushik Kayal), whose unwavering support and encouragement helped keep me motivated during my challenging times.

I record my deep sense of gratitude to all the staff members of the research scholar hostel, especially to Mr. Madhusudan Pramanik and Mr. Utpal Majumdar, whose support ensured my safe stay and attended to my daily essentials. They have been much more than staff members of the research scholar hostel; I wish I had the right kind of language at my disposal to express my feelings about them and theirs about me. In short, they are of mutual love and respect. One does not usually thank his/her parents for their support. However, in my situation, it would be wrong not to do it. They have stood by me during the last few years which I found difficult. It will take pages if I choose to write everything they did for me during the last few years. With love and respect, I am dedicating this dissertation to them and to my grandparents.

Subhajit Chattopadhyay
Applied Statistics Unit
Indian Statistical Institute
Kolkata 700108, India
May, 2025

Abstract

Longitudinal data are common in medical and biological sciences, where measurements are gathered from subjects over time to explore relationships with explanatory variables (covariates) and to uncover the underlying mechanisms of dependence among these measurements. The responses observed at each instance can be either discrete or continuous. One of the primary challenges in longitudinal data analysis lies in the non-Gaussian nature of the response variables. As a result, there are relatively few multivariate models in the literature that effectively address the specific characteristics observed in such datasets. In this dissertation, we address four problems concerning longitudinal data analysis by developing new statistical models. These models specifically address the time-related relationships found in various types of non-Gaussian longitudinal data by employing suitable classes of parametric copulas.

In the third chapter of this dissertation, we examine a motivating dataset from a recent HIV-AIDS study conducted in Livingstone district, Zambia. The histogram plots of the repeated measurements at each time point reveal asymmetry in the marginal distributions, and pairwise scatter plots uncover non-elliptical dependence patterns. Traditional linear mixed models, typically used for longitudinal data, struggle to capture these complexities effectively. We introduced skew-elliptical copula based mixed models to analyze this continuous data, where we use generalized linear mixed models (GLMM) for the marginals (e.g., Gamma mixed model), and address the temporal dependence of repeated measurements by utilizing copulas associated with skew-elliptical distributions (such as skew-normal/skew-t). The proposed class of copula-based mixed models addresses asymmetry, between-subject variability, and non-standard temporal dependence simultaneously, thereby extending beyond the limitations of standard linear mixed models based on multivariate normality. We estimate the model parameters using the IFM (inference function of margins) method, and outline the procedure for obtaining standard errors of the parameter estimates. To evaluate the performance of this approach under finite sample conditions, rigorous simulation studies are conducted, encompassing skewed and symmetric marginal distributions along with various copula selections. Finally, we apply these models to the HIV dataset and present the insight gained from the analysis.

In the fourth chapter of this dissertation, we introduce factor copula models tailored for unbalanced non-Gaussian longitudinal data. Modeling the joint distribution of such data, where subjects may have varying numbers of repeated measurements and responses can be continuous or discrete, poses practical challenges, especially with numerous measurements per subject. Factor copula models, which are canonical vine copulas, leverage latent variables to elucidate the underlying dependence structure of multivariate data. This approach aids in interpretation and implementation for unbalanced longitudinal

datasets, enhancing our ability to model complex dependencies effectively. We develop regression models for continuous, binary and ordinal longitudinal data, incorporating covariates, using factor copula constructions with subject-specific latent variables. With consideration for homogeneous within-subject dependence, the proposed models enable feasible parametric inference in moderate to high dimensional scenarios, employing a two-stage (IFM) estimation method. We also present a method for evaluating the residuals of factor copula models to visually assess the goodness of fit. The performance of the proposed models in finite samples is assessed through extensive simulation studies. In empirical analyses, we apply these models to analyze various longitudinal responses from two real-world datasets. Furthermore, we compare the performance of these models with widely used random effects models using standard selection techniques, revealing significant improvements. Our findings suggest that factor copula models can serve as viable alternatives to random effect models, offering deeper insights into the temporal dependence of longitudinal data across diverse contexts.

In the fifth chapter of this dissertation, we address the issue of modeling complex and hidden temporal dependence of count longitudinal data. Multivariate elliptical copulas are typically preferred in statistical literature to analyze dependence between repeated measurements of longitudinal data since they allow for different choices of the correlation structure. But these copulas lack in flexibility to model dependence and inference is only feasible under parametric restrictions. In this chapter, we propose the use of finite mixtures of elliptical copulas to enhance the modeling of temporal dependence in discrete longitudinal data. This approach enables the utilization of distinct correlation matrices within each component of the mixture copula. We theoretically explore the dependence properties of finite mixtures of copulas before employing them to construct regression models for count longitudinal data. Inference for this proposed class of models is based on a composite likelihood approach, and we evaluate the finite sample performance of parameter estimates through extensive simulation studies. To validate the fitting of the proposed models, we extend traditional techniques and introduce the t-plot method to accommodate finite mixtures of elliptical copulas. Finally we apply the proposed models to analyze the temporal dependence within two real-world count longitudinal datasets and demonstrate their superiority over standard elliptical copulas.

In the final contributing chapter of this dissertation, we introduce a novel multivariate copula based on the multivariate geometric skew-normal (GSN) distribution. This asymmetric copula serves as an alternative to the skew-normal copula proposed by Azzalini. Unlike the standard skew-normal copula, the multivariate GSN copula retains closure properties under marginalization, which offers computational advantages for modeling multivariate discrete data. In this chapter, we outline the construction of the geometric skew-normal copula and its application in modeling the temporal dependence observed in non-Gaussian longitudinal data. We begin by exploring the theoretical properties of the proposed multivariate copula. Subsequently, we develop regression models tailored for both continuous and discrete longitudinal data using this innovative framework. Notably, the quantile function of this copula remains independent of the correlation matrix of its respective multivariate distribution, offering computational advantages in likelihood inference compared to copulas derived from skew-elliptical distributions pro-

posed by Azzalini. Furthermore, composite likelihood inference becomes feasible for this multivariate copula, allowing for parameter estimation from ordered probit models with the same dependence structure as the geometric skew-normal distribution. We conduct extensive simulation studies to validate the geometric skew-normal copula based models and apply them to analyze the longitudinal dependence of two real-world data sets. Finally, We present our findings in terms of the improvements over regression models based on multivariate Gaussian copulas.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Longitudinal studies play important role in different branches of science; in biomedical, agricultural, health, behavioral sciences, in particular. They also play important role in public health, education, economics etc. Such studies lead to measurements for study participants which are collected over time. Such datasets, also known as repeated measurements, enable scientists and practitioners to study change in a study variable or an outcome over time. The repeated measurements are taken not only on response variables but also on possibly relevant explanatory variables. They allow also to study temporal changes in response variables for individuals, and connect them with factors which affect the changes. For example, the blood pressure of certain patients may be measured repeatedly over time, or multiple assessments may be conducted for students throughout their course of study. In such studies, response variables, along with a set of predictors (covariates), are collected repeatedly over time and meticulously documented. In contrast to cross-sectional studies, which collect data at a single point in time, longitudinal studies enable direct assessment of changes in variables over time. A key characteristic of longitudinal data is that the repeated measurements taken from the same subjects are likely to be correlated, allowing researchers to track trends and patterns more effectively.

The main focus of analyzing longitudinal data is to understand the relationship with explanatory variables (covariates) and to explore the dependence mechanisms among the measurements over time. The responses on a given occasion may be either discrete or continuous. One of the primary assumptions of longitudinal data analysis is that the responses from different subjects are independent of each other. The covariates in these studies may be classified into two categories such as time-dependent covariates and time-independent covariates. Variables that change over time for individuals are referred to as time-dependent covariates. In contrast, time-independent covariates are factors that remain constant, such as an individual's gender, race, and other baseline characteristics. When analyzing longitudinal data, a key interest for statisticians is to examine the strength of dependence among repeated measurements. This phenomenon is often referred to as the temporal dependence in longitudinal data. This helps in developing improved statistical models and ensuring valid inferences. In many cases, response variables exhibit diverse patterns that standard statistical models may not adequately capture. Numerous techniques are available in the literature to address these features; however, within a fully parametric framework, standard approaches typically rely on the assumption of multivariate normality for the re-

sponse variables. The primary goal of this dissertation is to move beyond statistical models that rely on multivariate normality. We contribute to the development of several new statistical models, present their distribution theories, and validate them using various statistical techniques.

1.1 Longitudinal data analysis: brief reviews and our approach

Univariate longitudinal data can be classified into two categories based on the number of measurements per subject: (i) balanced data, where all subjects have the same set of repeated measurements with no missing observations, and (ii) unbalanced data, where the number of time points varies among subjects, resulting in differing numbers of repeated measurements. In this dissertation we cover both of these type of data as well. Due to the unique characteristics of non-Gaussian longitudinal data, statistical methods for analyzing this type of data require special consideration. Many modeling approaches have been developed to analyze longitudinal data in the literature. These can be classified into four broad categories: marginal models, mixed models, transitional models and copula models. For an overview, one can refer to [1], [2], [3], and [4], among others. The important article [5] contains an overview of advances in the analysis of longitudinal data over a span of about twenty years, up to 2009, the year the book [4], in which the article appears, was published. This overview was presented from a historical perspective. The authors began by describing early origins of linear models for longitudinal data analysis. Then they focused on linear mixed-effects model for longitudinal data. This was followed by a discussion on models for non-Gaussian longitudinal data. It's worth noting that the authors' discussion of advances in longitudinal data analysis is based on linear models for continuous responses that are either normally distributed or approximately normal. As the authors noted, new tools were needed for analyzing discrete responses. As we see, while presenting the discussion on non-Gaussian longitudinal data, the authors primarily meant discrete data. They presented their review on models for non-Gaussian longitudinal data focusing mostly on longitudinal binary data. They noted that most of the developments reviewed by them apply also to categorical data and counts as well. They discussed three types of non-Gaussian longitudinal models: (i) marginal or population-averaged models, (ii) random-effects or subject-specific models, and (iii) transition or response-conditional models. Latter in the review article [6] the authors reviewed several approaches to joint modeling of multivariate longitudinal data. The authors noted that the differences among these approaches are similar to those found in their univariate counterparts. They attributed these differences to distinct modeling traditions. Additionally, they observed that the motivations behind each approach become evident in how the models are constructed. In addition, the approaches differ in formal characteristics, such as whether the data are balanced or unbalanced, whether the data are continuous or ordinal or binary, or whether or not latent variables are used to model the association between and across the responses. The authors described four families of models which are based on latent variables. Latent variables are assumed either along the time dimension, the outcome dimension, or both. The models discussed fall into several families: (1) models for the evolution of measured outcomes, including (a) marginal models and (b)

conditional models; (2) models capturing associations between latent evolutions, such as (a) shared parameter models and (b) random-effects models; (3) models for the evolution of latent variables; and (4) models for latent evolutions of latent variables. The authors discussed the strengths and limitations of each of these model types. The review articles cited above give an indication of the development of the theory and practice of longitudinal data in approximately the past three decades prior to publication of the second review. Next we provide a short description of this models for longitudinal data analysis.

On the other hand, yet another type of development have been taking place in the literature on analysis of longitudinal data during little more than the last two decades, maybe even more. Several researchers were becoming interested in flexible modeling of longitudinal data. They were trying to move away from the nature of dependence imposed by the Gaussian assumption or other standard prescriptions, which are dominated by the Gaussian assumption. They were finding copula-based approaches useful. This is yet another arena where the development of computing facilities and statistical computing, in particular, have influenced growth of the discipline of Statistics. In our context, an early copula-based approach in analysis of longitudinal data is [7] which appeared in 2002. The authors of this paper proposed in their work a new model for multivariate non-Gaussian longitudinal data. To begin with, they modeled each longitudinal data series for a given response separately, using a copula to link the marginal distributions of the response across observation times. It may be worthwhile to note here that this paper was cited in the review article by [6]. The authors of this review article observed that to their knowledge, very limited applications of copulas for the analysis of multivariate longitudinal outcomes had been reported and that the only copula-based work cited in their review article is ([7]). A list of other works, by no means exhaustive, relevant in copula-based modeling and its applications, including in analysis of longitudinal data, given by; [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20] and [21]. The paper which appears at the end of this list appeared in 2022. The authors of this paper presented a model based on pair copula construction for bivariate longitudinal mixed ordinal and continuous responses. They separately modeled the temporal association of each response is by using pair copula construction with a D-vine structure and the contemporaneous association of bivariate responses was then joined by a bivariate copula. It is worth while to note here, as [22], has noted, the term pair copula construction was coined in the paper [23] in 2009. Thus, a cursory glance at the content of the abstracts of [7] and [21] may give us a feel for the nature of evolution of copula-based modeling and analysis of longitudinal data. In fact, a careful and critical scrutiny of the references listed above will indicate how the same, i.e., copula-based modeling and analysis of longitudinal data, have evolved over time, over the last few more than two decades, to be more precise. It may be expected that they may also give us a glimpse of the broader picture, if any. Also, importantly, the following important texts, monographs, and proceedings of conference and meeting, played important roles in development of the theory and practice of copula-based models and methods: [24], [25], [26], [27] and [28]. One may also find review of development of certain aspects of the theory and practice of copula-based models and methods in some of these texts. We also note here three recent review articles on copula: [29], [30] and [31]. We wish to submit that the development of the theory and practice of copula-based models and methods

is an evolving area of research. Our work reported in this dissertation is a humble contribution to the literature on copula-based modeling and analysis of longitudinal data. They may be seen against the backdrop of the growth of that literature in the last two decades or so. Motivated by recent research contributions from [32], [33], and [34], we have attempted at developing new statistical models using parametric copulas, both existing and new. As we shall see, we have employed existing approaches also. The broad goal is to capture temporal dependence in non-Gaussian longitudinal data. We have kept the computational issues within our focus.

1.2 Overview of dissertation

This dissertation is divided into seven chapters. Our focus is oriented to (i) statistical modeling, (ii) distribution theory, (iii) simulation and computation and (iv) real data applications. The first chapter, an introductory one, contains, in particular, an overview and organization of the thesis. The second chapter contains some background materials for our work. They contain a brief introduction to key concepts like (i) generalized linear mixed models, (ii) copula functions, (iii) copula models and (iv) dependence properties and dependence measures. Besides appearing repeatedly in this dissertation, these concepts play the role of necessary technical materials for work done in the thesis. Other chapter-specific background materials appear within the chapters. Our research work is divided into the next four chapters; chapters 3 to 6 of the dissertation. For each of the work, we have conducted extensive simulation studies to assess performance under various situations and have also employed them on real-life datasets. Our models seem to demonstrate superiority their over natural competitors, both in simulation studies and analysis of real datasets.

The third chapter of the dissertation (the first contributing chapter) we explore a generalized version of the classical linear mixed model based on multivariate normality. This investigation was motivated by a real-life dataset from an HIV-AIDS study in Livingstone, Zambia, where the marginal distributions of the responses were found to be skewed, and non-elliptical dependence patterns were observed graphically. Thus it was quite natural to extend the distributional assumption of the classical linear mixed effect model. We developed skew-elliptical copula-based mixed models that can accommodate the features present in the dataset, while leveraging generalized linear mixed models (GLMM) in the continuous setting and copulas derived from skew-elliptical distributions (such as skew-normal/skew-t). The proposed copula-based mixed models address asymmetry, between-subject variability, and non-standard temporal dependence simultaneously, thereby extending the capabilities of standard linear mixed models that rely on multivariate normality. We estimate the model parameters using the two-stage Inference Function of Margins (IFM; [35], [36] and [37]) method and provide a procedure for calculating the standard errors of the parameter estimates. To assess the performance of this approach in finite sample settings, we conduct comprehensive simulation studies, including both skewed and symmetric marginal distributions and various copula types. Finally, we apply these models to the HIV dataset and present the insights derived from the analysis.

In the fourth chapter, we introduce factor copula constructions to model unbalanced, non-Gaussian longitudinal data. Modeling the joint distribution of such data, where subjects may have varying numbers of repeated measurements and responses that can be continuous or discrete, presents practical challenges, especially with numerous measurements per subject. Factor copula models, specifically canonical vine copulas, use latent variables to uncover the underlying dependence structure of multivariate data. This approach facilitates interpretation and implementation for unbalanced longitudinal datasets, enhancing our ability to effectively model complex dependencies. We develop regression models for continuous, binary, and ordinal longitudinal data, incorporating covariates, using factor copula constructions with subject-specific latent variables. By considering homogeneous within-subject dependence, the proposed models allow for feasible parametric inference in moderate to high-dimensional scenarios, employing a two-stage Inference Function of Margins (IFM) estimation method. We also propose a method for evaluating the residuals of factor copula models to visually assess the goodness of fit. The performance of these models in finite sample settings is evaluated through extensive simulation studies. In empirical analyses, we apply these models to various longitudinal responses from two real-world datasets. Furthermore, we compare the performance of these models with widely used random effects models through standard selection techniques, demonstrating significant improvements. Our findings suggest that factor copula models can serve as effective alternatives to random effects models, offering deeper insights into the temporal dependence of longitudinal data across various contexts.

In the fifth chapter, we focus on addressing the challenge of modeling the complex and often hidden temporal dependence in count longitudinal data. Longitudinal data frequently exhibit complex dependency structures that traditional methods may fail to capture adequately. Multivariate elliptical copulas have been widely used in the statistical literature to model the dependence between repeated measurements of longitudinal data, as they offer flexibility in specifying various types of correlation structures. However, despite their widespread use, elliptical copulas have notable limitations. Specifically, they lack the flexibility to model more complex dependence patterns, and inference is often restricted by stringent parametric assumptions, making them less adaptable to real-life data with intricate temporal dependencies. To overcome these limitations, we propose the use of finite mixtures of elliptical copulas as an advanced approach to model temporal dependence in discrete longitudinal data more effectively. By employing finite mixtures, we introduce the ability to model distinct correlation structures within each component of the mixture copula, providing greater flexibility and accuracy in capturing the diversity of dependencies observed in longitudinal data. This approach not only enriches the modeling process but also allows for more nuanced and realistic representations of the temporal relationships between observations. Theoretical exploration of finite mixtures of copulas is conducted to understand their dependence properties in detail. We delve into the statistical foundations of this approach, discussing how the components of the mixture interact and contribute to the overall dependence structure. Building on these theoretical insights, we construct regression models specifically tailored for count longitudinal data, which often arise in fields like epidemiology, economics, and social sciences. Inference for the proposed class of finite mixture models is carried out using a composite likelihood approach,

which offers a robust and computationally efficient method for parameter estimation in the presence of complex dependence structures. To assess the performance of the model and its suitability for finite sample conditions, we conduct extensive simulation studies, evaluating the accuracy and precision of parameter estimates under various scenarios. These simulations not only demonstrate the effectiveness of the finite mixture approach but also provide insights into its advantages over traditional methods. In addition to standard diagnostic tools, we extend traditional model validation techniques by introducing the t-plot method, which is specifically designed to assess the goodness-of-fit for finite mixtures of elliptical copulas. This new method enables a more comprehensive evaluation of model performance, providing visual tools to identify potential misfit and refine the model selection process. Finally, we apply the proposed models to real-life count longitudinal datasets, analyzing temporal dependence in two distinct datasets from different fields. The results highlight the superiority of our finite mixture copula models over standard elliptical copulas, demonstrating their enhanced ability to capture the underlying dependence structure and offering a more accurate representation of the temporal dynamics within the data. This chapter underscores the potential of finite mixture models in addressing the complexities of longitudinal count data, offering a powerful tool for researchers and practitioners dealing with similar challenges in various applied fields.

In the final contributing chapter of this dissertation, we introduce a novel multivariate copula based on the multivariate geometric skew-normal (GSN) distribution. This asymmetric copula is proposed as an alternative to the skew-normal copula introduced by Azzalini, addressing some of the key limitations of the existing models. Unlike the standard skew-normal copula, which often faces challenges in terms of computational feasibility when marginalization is involved, the multivariate GSN copula retains closure properties under marginalization. This characteristic provides significant computational advantages, particularly when modeling multivariate discrete data that involve complex dependencies. In this chapter, we begin by constructing the geometric skew-normal copula and delve into its theoretical properties, establishing a strong foundation for its application in modeling the temporal dependence inherent in non-Gaussian longitudinal data. By exploring the properties of this copula, we highlight its flexibility in capturing asymmetric and non-linear relationships, which are often present in longitudinal data but are challenging for standard models to adequately represent. A key innovation of the multivariate GSN copula is that its quantile function remains independent of the correlation matrix of its respective multivariate distribution. This crucial feature offers significant computational advantages in likelihood inference when compared to other copulas, such as those derived from skew-elliptical distributions, which have been widely used in the literature. This independence simplifies the estimation process and enhances the efficiency of parameter inference, especially when working with large and complex datasets. Additionally, we show that composite likelihood inference is feasible for this copula, facilitating parameter estimation in models like ordered probit regressions, while maintaining the same dependence structure as the geometric skew-normal distribution. Building on these theoretical foundations, we develop regression models specifically tailored for both continuous and discrete longitudinal data within this novel framework. These models enable more accurate and flexible modeling of temporal

dependencies, overcoming the limitations of traditional approaches. To validate the effectiveness of the geometric skew-normal copula-based models, we conduct extensive simulation studies, assessing their performance in various settings and comparing them to more conventional models. These simulations provide a robust evaluation of the accuracy and efficiency of parameter estimates, demonstrating the strengths of our proposed method in finite sample conditions. Finally, we apply the proposed models to analyze longitudinal dependence in two real-life datasets, showcasing the utility of the multivariate GSN copula in practical applications. By comparing our models to regression models based on multivariate Gaussian copulas, we highlight the substantial improvements in capturing complex temporal dependencies. These findings suggest that the geometric skew-normal copula offers a promising alternative to existing models, providing deeper insights into the structure of longitudinal data and presenting a powerful tool for researchers working in this area.

A substantial amount of computation was required to develop this dissertation in its entirety. In the preceding four paragraphs, we have provided an outline of the work presented in the dissertation. In the seventh chapter, the last one in the dissertation, we have mentioned the key findings and possible directions of future works. This dissertation has attempted at contributing to the evolving literature on copula-based modeling and analysis of longitudinal data. We believe if researchers take into account the growth of the literature on theory and practice of copula and try to use them in modeling and analysis of longitudinal data, further growth of the literature will take place. Combination of Bayesian analysis and copula-based methods, developing tools of prediction using copulas are some of the areas where one may focus. We hope that copula-based methods will soon become a valuable part of the statistical analysis toolkit.

Chapter 2

Preliminaries

This chapter outlines some of the foundational concepts of longitudinal data analysis and multivariate dependence. To effectively model non-Gaussian longitudinal data, we must address two key challenges simultaneously. First, we need to select appropriate univariate distributions to accurately describe the marginal responses. Second, we must determine the most suitable dependence structure to capture the relationships between the repeated measurements. This chapter introduces key foundational concepts.

2.1 Generalized linear mixed models

One commonly used statistical approach for modeling non-Gaussian data is the generalized linear mixed model (GLMM), which combines the flexibility of generalized linear models with the ability to account for random effects. A generalized linear mixed model has the following form -

$$g(E(Y_{ij}|\mathbf{b}_i)) = \mathbf{x}_{ij}\beta + \mathbf{d}_{ij}\mathbf{b}_i, \tag{2.1}$$

for i-th response of j-th point in time, where β is the fixed effect parameters, $\mathbf{b}_i$ is the vector of random-effects associated with covariates $\mathbf{d}_{ij}$. The vector of random effects $\mathbf{b}_i$ has a certain distribution, which is generally assumed to be normal with mean 0 and variance σ_b^2. Furthermore, we assume that Y_{ij} follows a conditional distribution in the exponential family given the random effects $\mathbf{b}_i$, given as -

$$f(y_{ij}|\mathbf{b}_i, \eta_{ij}, \phi) = \exp[(y_{ij}\eta_{ij} - b(\eta_{ij}))/a(\phi) + c(y_{ij}, \phi)], \tag{2.2}$$

where, $a(.), b(.)$ and $c(.,.)$ are known functions and parameters η_{ij} can be further modeled to accommodate within-subject variability. The ϕ is the dispersion parameter that is known or to be estimated. Generalized linear mixed model simplifies to the classical linear mixed model when the marginal response distribution is normal and the link function is identity. To estimate the parameters, the straight forward way is to use the maximum likelihood estimation (ML) while treating the response for each individuals to be independent. However, computing the likelihood function for GLMMs is often challenging, as it typically involves integrating over random components and is not in closed form in most cases. There are a verity of methods proposed to fit GLMMs in the literature such as Monte-Carlo EM

([38], Laplace approximations ([39]) and penalized quasilikelihood ([40]). A recent alternative indirect strategy for modeling joint distribution of repeated measurements that has attracted the attention of researchers involves using copula functions. Some recent advancements can be found in [11], [8] and [41], among others. The theoretical details in copula construction and discussions of important methodological issues are given in [26] and [28].

2.2 Copula functions

The main focus of this dissertation is on copula-based models for non-Gaussian longitudinal data. These models are flexible because they allow the marginal distributions to be chosen independently, and the dependence structures can be defined using different copula functions. This flexibility offers a better understanding of the relationships within the data. By using these models, it's possible to gain more detailed insights into the dynamics at play, especially in situations where simpler models that assume normality or linear relationships may not capture the full complexity of the data. This work explores different copula-based approaches and their ability to handle non-linear dependencies and variability often seen in longitudinal data.

Definition 2.2.1 *A d-dimensional copula is a cumulative distribution function C of d variables such that the marginal distributions are uniform on $[0, 1]$. Therefore,*

$$C_d(u_1, \ldots, u_d) = P(U_1 \leq u_1, \ldots, U_d \leq u_d), \quad 0 \leq u_1, \ldots, u_d \leq 1, \tag{2.3}$$

where each U_i is uniformly distributed on $[0, 1]$ for $i = 1, \ldots, d$.

The dependence information for the random variables $U_1, \ldots, U_d$ is encoded in the copula C. When C is parameterized by a vector ϕ, this vector ϕ is referred to as the dependence parameter. We use the notation $C(.|\phi)$ to denote a parametric copula function. This parameter captures the structure of dependence between the variables, and its specification is crucial for modeling and understanding how the random variables are related to each other. By adjusting ϕ, one can explore different dependency structures and better fit the model to the observed data.

The condition that C is a distribution function with uniform marginals leads to the following properties ([26]):

(a) $C : [0, 1]^d \to [0, 1]$;

(b) $C(u_1, \ldots, u_d)$ is increasing in each component u_i;

(c) $C(1, \ldots, u_i, \ldots, 1) = u_i$ for all $i = 1, \ldots, d, u_i \in [0, 1]$, and

(d) For any $(a_1, \ldots, a_d), (b_1, \ldots, b_d) \in [0, 1]^d$, with $a_i \leq b_i, \forall i$ we have

$$\sum_{i_1=1}^{1} \cdots \sum_{i_d=1}^{1} (-1)^{i_1+\cdots+i_d} C_d(u_{1i_1}, \ldots, u_{di_d}) \geq 0,$$

where $u_{i0} = a_i$ and $u_{i1} = b_i, \forall i = 1, \ldots, d$. The reason is that the left-hand side represents the probability that $(U_1, \ldots, U_d)$ falls within the interval $(a_1, b_1] \times \cdots \times (a_d, b_d]$.

Conversely every function C with these properties is a copula. The density function of a copula allows maximum likelihood estimation of its dependence parameters.

Definition 2.2.2 *The copula density is a multivariate probability density function which is given by*

$$c_d(u_1, \ldots, u_d) = \frac{\partial^d C_d(u_1, \ldots, u_d)}{\partial u_1, \ldots, \partial u_d}, \quad 0 < u_1, \ldots, u_d < 1, \tag{2.4}$$

when the copula C_d is absolutely continuous.

The foundational theorem of copula functions was established by Sklar in 1959 (see, [42]). This theorem, often referred to as Sklar's Theorem, provides a fundamental result in copula theory, stating that any multivariate distribution can be represented in terms of its marginal distributions and a copula that captures the dependence structure between the variables. Sklar's theorem is central to the use of copulas in statistical modeling, as it allows for the separation of the marginal behavior of the variables from the dependencies between them, offering great flexibility in modeling complex relationships.

Theorem 2.2.1 *(Sklar's Theorem). Let F be a d-dimensional joint distribution function with marginal distribution functions $F_1, \ldots, F_d$. Then there exists a copula $C : [0, 1]^d \to [0, 1]$ such that*

$$F(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d)), \quad (y_1, \ldots, y_d)^\intercal \in \mathbb{R}^d. \tag{2.5}$$

If each F_i is continuous for $i = 1, \ldots, d$, then the copula C is unique. Otherwise, C is uniquely determined on $Range(F_1(y_1)) \times \cdots \times Range(F_d(y_d))$, where $Range(F_i(y_i))$ denotes $\{F_i(y_i), y_i \in R\}$.

Sklar's Theorem guarantees the existence of a copula C, but it does not provide a method for identifying or constructing it. When the marginals $F_i(y_i), i = 1, \ldots, d$ are continuous, $F_i(y_i) \sim U(0, 1)$. Letting $u_i = F_i(y_i)$ in (2.5) one can write

$$C(u_1, \ldots, u_d) = F(F_1^{-1}(y_1), \ldots, F_d^{-1}(y_d)), \quad 0 \leq u_1, \ldots, u_d \leq 1, \tag{2.6}$$

which is the unique copula of F. From any type of multivariate distributions using equation (2.6) one can obtain the respective copula.

2.3 Copula models

Each marginal distribution $F_1(y_i)$ captures the information related to the individual variable Y_i, while the joint distribution $F_d(y_1, \ldots, y_d)$ encompasses both the marginal and joint information. According to Sklar's Theorem, the multivariate distribution can be separated into two components: the set of marginal distributions and the dependence structure, which is defined through its copula. This decomposition is

particularly valuable from a modeling perspective, as it provides a natural framework for constructing multivariate models. A copula based multivariate model can generally takes the form of:

$$f_d(y_1, \ldots, y_d|\theta^*) = c_d(F_1(y_1), \ldots, F_d(y_d)|\phi) \prod_{i=1}^{d} f_i(y_i|\theta), \tag{2.7}$$

where $\theta^* = (\theta^\intercal, \phi^\intercal)^\intercal$ is the set of marginal and dependence parameter respectively. Copulas are highly versatile and have been widely applied across various fields. [43] explore the class of continuous copula models, along with their properties, which can be constructed using elliptical copulas with continuous marginals. In practice, selecting and estimating an appropriate marginal distribution for each variable is often a straightforward task, given the extensive library of univariate distributions available. Additionally, nonparametric methods can be used to estimate the marginal distributions, providing insights into their potential true forms. Once the marginals are determined, copulas can be applied to model the dependence structure between variables. The parameters of a copula reflect the degree of dependence between variables. One can see [28] for an overview of copula based modeling.

2.4 Dependence properties

If the dependence between two variables is such that if one variable increase then the other tends to increase or decrease, then it is referred as monotone association. Kendall's tau and Spearman's rho stand as the most widely utilized measures of monotone association, relying on concordance and discordance. These measures are invariant with respect to the marginal distributions for continuous random variables, i.e. they can be expressed as a function of their copula. Multivariate extensions of such measures have been discussed by several authors as [44] or [45]. But in this dissertation we only focus our attention to the bivariate case. For continuous random variables, the dependence, as quantified by Kendall's tau or Spearman's rho, is solely associated with the copula parameters [26]. Kendall's Tau and Spearman's rho are measures of correlation which can be derived from a copula. These provides the summary of the strength and direction of the dependence between random variables, with values ranging from -1 (perfect negative dependence) to 1 (perfect positive dependence), and 0 indicating independence.

In the context of copulas, Kendall's tau is derived from the copula's dependence structure and reflects how the ranks of two variables are correlated. Before describing the measures, we consider the concepts of concordance and discordance. Given a bivariate random vector $(X_1, X_2)^\intercal$, concordance means that if $X_1 < X_1^1$ then $X_2 < X_2^1$, or if $X_1 > X_1^1$ then $X_2 > X_2^1$ for an independent replication $(X_1^1, X_2^1)^\intercal$. Essentially, both variables move in the same direction. On the contrary, discordance means that if $X_1 < X_1^1$ then $X_2 > X_2^1$, or vice versa, meaning the variables move in opposite directions.

Definition 2.4.1 *Kendall's tau is defined as the probability of concordance minus the probability of discordance of two variables, i.e.*

$$\tau = P[(X_1 - X_1^1)(X_2 - X_2^1) > 0] - P[(X_1 - X_1^1)(X_2 - X_2^1) < 0].$$

For a bivariate copula C, Kendall's tau can also be expressed as

$$\tau = 4\int_0^1 \int_0^1 C(u,v)dC(u,v) - 1.$$

Another important measure of association between two random variables is Spearman's rho. Similar to Kendall's tau, but it is based on the ranks of the variables rather than the concordance and discordance.

Definition 2.4.2 *Spearman's rho is defined as the correlation between two ranked variables, i.e.*

$$\rho = Corr(F_1(X_1), F_2(X_2)).$$

In the context of copulas, Spearman's rho is related to the bivariate copula C by the following relation

$$\rho = 12\int_0^1 \int_0^1 C(u,v)dudv - 3. \tag{2.8}$$

Unlike Kendall's tau, which focuses on pairwise concordance and discordance, Spearman's rho is more sensitive to the overall order of ranks, making it a different but complementary measure of dependence. Both of these measures are non-parametric and do not assume any specific distribution for the variables.

Finally, we conclude this chapter by describing an important property of copula which is the tail dependence. Tail dependence measures how much two variables are related in extreme situations. That is it quantifies the degree of dependence in the joint lower or joint upper tail of a multivariate distribution. Here we will focus on the bivariate tail dependence only, but there are multivariate extensions to the concept in the literature (see, [27]). For a bivariate distribution, tail dependence is defined as the limiting probability of exceeding a certain threshold by one margin given that the other margin has already exceeded that threshold. Both upper and lower tail dependence coefficients are of interest for a bivariate distribution or copula, providing valuable insights into the behavious of extreme events.

Definition 2.4.3 *Let $(X_1, X_2)^\intercal$ be a bivariate random vector with marginal distribution functions F_1, F_2 and copula C. Then the upper and lower tail dependence coefficients are defined by*

$$\lambda_U(C) = \lim_{u\to 1-} \frac{1-2u+C(u,u)}{1-u} = \lim_{u\to 1-} P(U_1 > u | U_2 > u) \quad \textit{and}$$
$$\lambda_L(C) = \lim_{u\to 0+} \frac{C(u,u)}{u} = \lim_{u\to 0+} P(U_1 \leq u | U_2 \leq u) \quad \textit{respectively}$$

where $U_i = F_i(X_i),\ i = \{1,2\}$ provided the above limits exist.

The copula C is said to have upper or lower tail dependence if $\lambda_U, \lambda_L \in (0,1]$. If $\lambda_U = 0$ or $\lambda_L = 0$, we say C has no upper or lower tail dependence, respectively. Tail dependence is important in risk management and extreme value theory, as it helps to quantify the likelihood that two variables will exhibit extreme behavior simultaneously (e.g., both variables reaching high or low values at the same

time). For example, in the case of Gaussian copula both the tail dependence coefficients are 0, meaning this copula can not capture any strong association in the extremes.

In longitudinal data analysis, understanding tail dependence is essential for examining how extreme values in repeated measurements or observations of subjects over time are related. Longitudinal data involves tracking multiple measurements from the same subjects or units over time, often to observe how variables change and interact within individuals or groups. Extreme values in this context refer to unusually high or low measurements, which may indicate rare or significant events. Tail dependence focuses on whether these extreme events—either high or low values—occur simultaneously across multiple variables or over time within the same variable. For example, in a medical study tracking patients' blood pressure and heart rate over time, upper tail dependence measures the likelihood that both blood pressure and heart rate will simultaneously reach high values. This is especially important when assessing rare but critical events, such as heart attacks or strokes, where both elevated blood pressure and heart rate might occur together. Similarly, lower tail dependence in a study on patient health could assess the likelihood of extremely low values in variables like cholesterol levels coinciding with low physical activity levels at the same or subsequent time points. This helps in understanding situations where low values across multiple variables may signal poor health or a decline in condition. In such cases, copulas are used to model the dependency structure between variables across time or between related variables. Copulas provide a flexible way to capture complex dependencies that might not be well-represented by standard linear models, particularly when it comes to tail dependence. Longitudinal data often exhibit temporal dependencies, meaning measurements from the same subject at different time points are correlated. For example, a subject's heart rate at time t_1 might be correlated with their heart rate at time t_2. Copulas can quantify the strength of this dependence, especially for extreme values, helping to understand the likelihood that extreme values in multiple variables will occur simultaneously across time or between different subject variables. By analyzing tail dependence in longitudinal data, researchers can more accurately predict extreme events. For instance, in a longitudinal study of patients, identifying when both blood pressure and cholesterol levels are likely to be extreme can help predict potential health crises. This is why researchers focus on developing new models specifically for longitudinal data analysis using copulas.

Chapter 3

Skew-elliptical copula based GLMMs for continuous longitudinal data

In this chapter, we explore a modified version of the classical linear mixed model. The need for the modification is motivated by a real-life longitudinal dataset. Linear mixed models (LMMs) stand as the cornerstone for analyzing univariate longitudinal data, with the original formulation introduced in the seminal paper by [46]. This framework bridges classical linear models with subject-specific random effects for understanding complex longitudinal dynamics. The foundational work by [46] has since been widely adopted by statisticians, as seen by its extensive application across diverse contexts (see, for example, [2] and [4]). Following the framework in [4], we consider a cohort consisting of m individuals, each contributing to a longitudinal dataset collected at n_i recorded time points and associated predictors. These predictors are structured into a matrix $\mathbf{X}_i$ of dimensions $n_i \times p$, where p represents the number of predictor variables, and β denotes the corresponding $p \times 1$ vector of fixed effects. Each row $\mathbf{x}_{ij}$ of $\mathbf{X}_i$ represents predictor values at a specific time point. The responses for each individual are consolidated into a column vector $\mathbf{Y}_i$ of size $n_i \times 1$, while the random effects are characterized by a design matrix $\mathbf{D}_i$ of dimensions $n_i \times q$ and a vector $\mathbf{b}_i$ of length q.

The standard linear mixed model formulation is expressed as;

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{D}_i\mathbf{b}_i + \epsilon_i, \quad i = 1, \ldots, m, \tag{3.1}$$

where ϵ_i represents the error term of dimension $n_i \times q$. To ensure model identifiability, we assume independence between the collections $\{\mathbf{b}_i\}$ and $\{\epsilon_i\}$. Typically, we assume multivariate normality for both random effects and error terms, as described by

$$\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{\Omega}_b), \quad \epsilon_i \sim N_{n_i}(\mathbf{0}, \mathbf{\Psi}_i), \tag{3.2}$$

where $\mathbf{\Omega}_b$ and $\mathbf{\Psi}_i$ represent the dispersion matrices capturing between-subject and within-subject variability, respectively. However, exploratory data analysis often reveals deviations from normality assumptions, prompting the need for more flexible modeling approaches. Generalized linear mixed models offer one such avenue, accommodating various response distributions from the exponential family. Yet, these models hinge on the conditional independence assumption of response variables given the ran-

dom effects ([47]). In our endeavor to enhance the classical linear mixed model, we turn to the copula framework, aiming to capture dependencies beyond the scope of traditional normality assumptions.

In the literature of statistical methods for analyzing longitudinal data, several copula-based approaches have been developed to address different challenges posed by diverse datasets. For instance, [7] introduced a Gaussian copula-based model for analyzing longitudinal data. This is one of the early work in copula-based research in longitudinal data. Later, [12] studied Gaussian copula-based regression models for non-normal dependent observations. This work contained, in particular, applications in longitudinal studies. In [14], the authors moved beyond Gaussian copula and proposed a generalized linear mixed model which is driven by a skew-normal copula. More recently, [16] explored the use of D-vine copulas to model dependence among repeated measurements in unbalanced longitudinal data. Also, [17] proposed a Gaussian copula-based model that accommodates temporal variability by allowing model parameters to vary with time. This is particularly useful for capturing the evolving nature of mixed longitudinal responses. It is worthwhile to note in this context that in situations where empirical observations deviate from symmetry assumptions, such as in biomedical datasets, skew-elliptical distributions have been suggested to introduce flexibility into the dependence structure ([48]). These distributions can better capture subtle dependencies, including reflection, permutation asymmetry, or tail dependence ([49]).

Motivated by a recent HIV-AIDS CD4 count data from Livingstone district, Zambia, and exhibiting similar lack of symmetry, we aim at developing in this chapter a copula based longitudinal model that accounts for temporal dependence through skew-elliptical copulas. Our work is inspired by [14] but moves beyond that work. We extend multivariate normal linear mixed model as in (3.1) in the sense that we consider GLMMs but describe the dependence structure through some skew-elliptical copulas. In this specific chapter although we exploit the term GLMM, but only consider continuous responses in our application.

Rest of this chapter is organized as follows. In Section 3.1, we describe the data set in details which we have analyzed later in this chapter. We describe our copula-based modeling framework in Section 3.2, and provide an overview of skew-elliptical distributions and their associated copulas in Section 3.3. In Section 3.4, we have discussed the method of estimation of the model parameters using IFM method. In Section 3.5, a relevant result on asymptotic normality has been discussed. In Sections 3.6 and 3.7, we have discussed certain aspects of the numerical work and model comparison which form the subject of the following sections. In Section 3.8, we conduct some simulation studies to study the parametric inference of the proposed class of models using both skewed and symmetric marginal mixed models with different sample sizes. Finally, in Section 3.9, we analyze our data set and report our findings.

3.1 HIV CD4 positive T cell count data

The human immune virus (HIV) is a viral infection that slowly destroys the immune system resulting acquired immunodeficiency syndrome (AIDS). Unfortunately there is no clinically proven vaccine for

this virus till today, so people rely on available antiviral drugs which slow down the viral reproduction. Most important markers for evaluating antiviral therapies are HIV-1 RNA copies and CD4 T^+1 cell counts. Due to the skewed nature of these two markers, researchers prefer modeling these data with skew-elliptical distributions. [50] used multivariate skew-normal mixed model in ACTG 315 study to model these markers. [51] considered linear mixed models replacing the Gaussian assumptions which what are known as SNI (a shortened form of skew normal/independent) distributions. SNI distributions have been nicely discussed in [52]. [53] is a forerunner of this work. The class of SNI distributions is one of asymmetric heavy-tailed distributions that includes the skew-normal, skew-t, skew-slash ([54]), and skew-contaminated normal distributions as special cases. [51] noted that the model they proposed ensures flexibility in capturing the effects of skewness and heavy tail for responses that are either left- or right-censored. They illustrated the procedures which they developed with an HIV case study involving analysis of longitudinal viral loads.

The motivating data set analyzed in this chapter was found in Mendeley about a recent HIV-CD4 study from Livingstone district, Zambia (2016). The data were collected during a survey of antiretroviral (ARV) combination in the treatment of HIV, as part of the work for the PhD thesis of Urban N. Haankuku at the University of South Africa and later made publicly available. They have studied the performance of ARV combinations on HIV naive patients using several different models. According to WHO, HIV-AIDS is a major cause of death in Zambia, with about a million deaths attributed to HIV-AIDS related causes. If the disease left untreated, it can reduce the cluster of CD4 T^+1 cells and increase the HIV viral load. With non permanent cure available till today, the only option is to use antiretroviral drugs to reduce the immune suppression. The CD4 counts of 261 HIV naive patients were measured every twelve weeks from the initial diagnosis for 48 weeks. Three different ARV combinations were given to the patients at first baseline regimen (FBR). Covariates such as gender, age and initial weight of each patient were also reported. In Figure 3.1, the evolution of CD4 T^+1 cell counts over time is shown. Though the mean structure is nearly linear but it reveals substantial variability of between-subject responses. In Figure 3.2 we plot the histograms of each time point in the diagonals and pair-wise scatter plots in the off diagonals of the panel. Based on these initial diagnosis, the marginals appear to be skewed with positive real support and the scatter plots reveals non-elliptical dependence patterns. The same thing can be observed with the normal scores, which suggests reflection and permutation asymmetry and stronger dependence than Gaussian in the joint upper and lower tails. Thus one would require more flexible dependence structure than Gaussian to model the temporal dependence. However, due to the skewed nature of the CD4 counts marker with positive real support, we consider Gamma mixed model for the marginals. For comparison, we also use normal mixed model for the marginals.

3.2 A copula based longitudinal model

As described in the introductory section, let the response variables $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\intercal, i = 1, \ldots, m$, follow an n_i-variate distribution with predefined mean and dispersion matrix. Suppose ob-

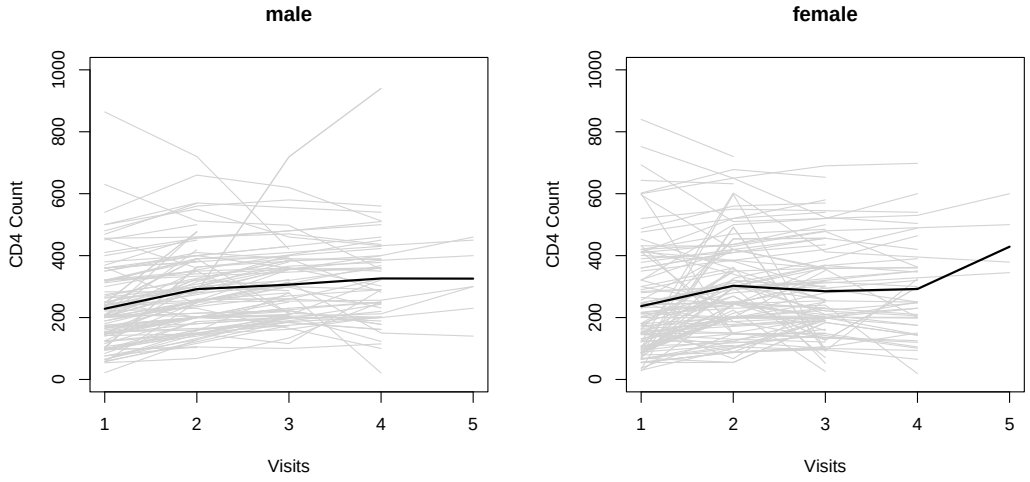

Figure 3.1 Individual and average profiles for male (left panel) and female (right panel) patients over time starting from initial. Black lines represent the mean profiles.

servations from different individuals are independent, and to account for subject's individual effects we consider the distribution of $\mathbf{Y}_i$ conditional on $\mathbf{b}_i$ as

$$\mathbf{Y}_i|\mathbf{b}_i \sim F_{n_i}(\eta(\mathbf{X}_i\beta + \mathbf{D}_i\mathbf{b}_i), \mathbf{\Sigma}(\xi_i, \mathbf{t}_i)), \tag{3.3}$$

where ξ_i is the auto-regressive parameter with respect to the time points, $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})^\intercal$ and $\eta(.)$ is a known link function. $F_n(\eta, \mathbf{\Sigma})$ is an n-variate distribution function with mean η and covariance $\mathbf{\Sigma}$. Furthermore, $\mathbf{X}_i : n_i \times p$ and $\mathbf{D}_i : n_i \times q$ are the known design matrices as described earlier. We assume the marginal densities of $Y_{ij}|\mathbf{b}_i$ are from the exponential family and functions of $\{\mathbf{x}_{ij}, \beta, t_{ij}, \mathbf{d}_{ij}, \mathbf{b}_i\}$ via the same known link $\eta(.)$. In this chapter we assume the random effects are independent and normally distributed, i.e. $\mathbf{b}_i \sim N_q(0, \mathbf{\Omega}_b)$. We model such distribution using a copula based GLMM as

$$F_{n_i}(\mathbf{y}_i|\mathbf{b}_i, \theta_i^*) = C_{n_i}(F(y_{i1}|\mathbf{b}_i, \theta_{i1}), \ldots, F(y_{in_i}|\mathbf{b}_i, \theta_{in_i})|\phi_i) \tag{3.4}$$

where $\theta_i^* = (\theta_i^\intercal, \phi_i^\intercal)^\intercal$ is the set of all vector valued parameters in the conditional model. θ_i accounts for all the parameters present in the marginals and ϕ_i accounts for the dependence parameters. Then the corresponding density function is given by -

$$f_{n_i}(\mathbf{y}_i|\mathbf{b}_i, \theta_i^*) = c_{n_i}(F(y_{i1}|\mathbf{b}_i, \theta_{i1}), \ldots, F(y_{in_i}|\mathbf{b}_i, \theta_{in_i})|\phi_i) \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i, \theta_{ij}). \tag{3.5}$$

We note here from the relations (3.3), (3.4) and (3.5) that the random effects are absorbed only in the marginal GLLMs. They are not part of the specification of the assumed copula.

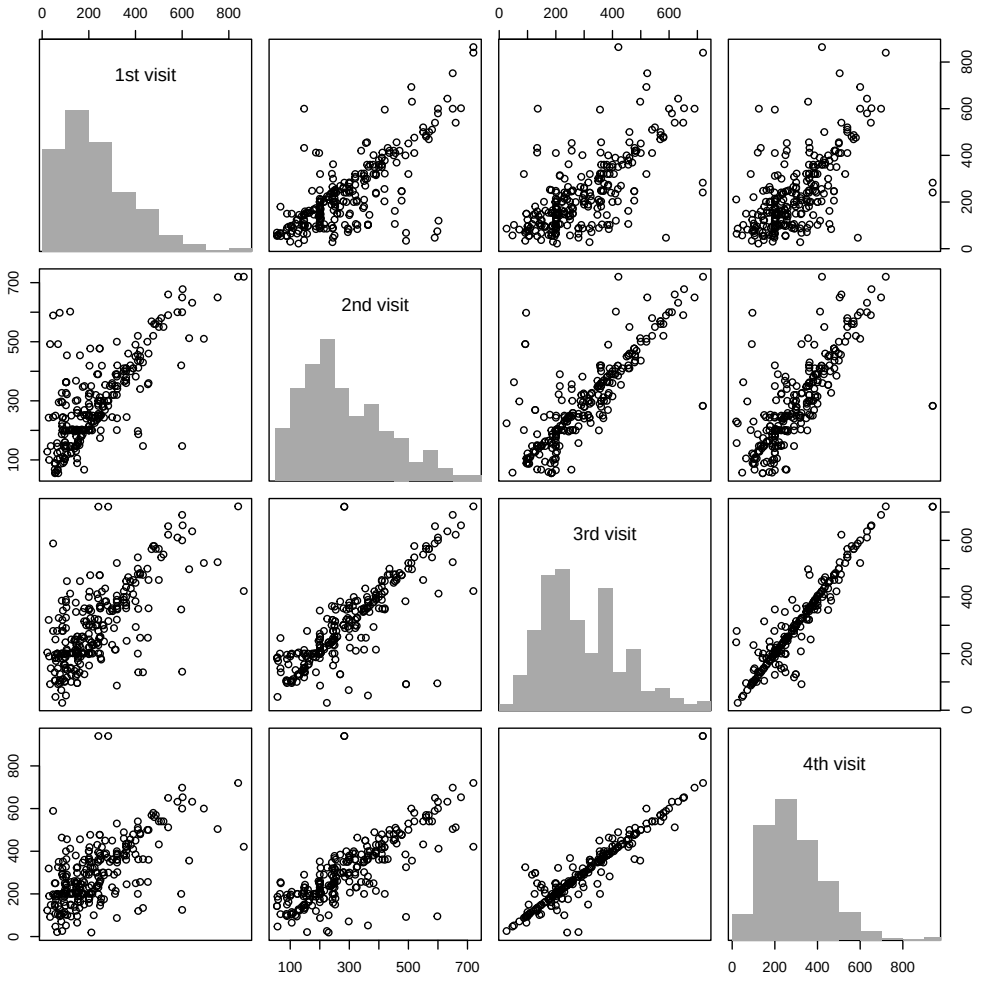

Figure 3.2 Pairwise scatter plots of the CD4 counts of the patients for first 4 visits.

Copula based models are very flexible to analyze temporal dependence of longitudinal data since the responses at each point of time have a predefined marginal distribution. That is, a copula can be viewed as an association function which describes the dependence between separately specified marginals. When the marginals are fixed, many different multivariate models can be obtained from considering different copula functions. The random effects $\mathbf{b}_i$ are interpreted as the unobserved regression parameters for the i-th subject which explains variability between the subjects. The explicit need of random effects in this work to address the between subject variability seen in the profile plots of the considered data set. The next section describes a class of multivariate copulas, which have desirable dependence properties and can be used for our models in 3.4.

3.3 Skew-elliptical distributions and related copulas

Multivariate skew-elliptical distributions have been applied in several bio-medical studies to model non-Gaussian data. They are constructed from (multivariate) elliptical distributions (see e.g., [53], [55], [48], and the bibliographic note in page 175 of [48]). These (multivariate skew-elliptical) distributions constitute a large class. They account for both skewness and a variety of tail properties. In [56], the tail densities of skew-elliptical copulas have been shown to depend only on tail properties of the underlying density generator and conditions on the skewness parameters. Copulas generated from skew-elliptical class of distributions can provide various flexible dependence structures. These multivariate copulas are reflection as well as permutation asymmetric and can be used to describe a general dependence structure of the model. Also, the expressions for the tail densities of skew-normal copula and skew-t copula, as they appear in [56] demonstrate their dependence on the skewness and dependence parameters of the underlying skew-elliptical (multivariate) distributions. [57] also demonstrate similar phenomenon for tail dependence of skew-t copulas. All these facts, put together, make the skew-elliptical copulas attractive tools in application.

Definition 3.3.1 *A random vector* $\mathbf{Z}$ *taking values* $\mathcal{R}^d$ *is said to have a mean zero skew-normal distribution, denoted as* $\mathbf{Z} \sim SN_d(\mathbf{\Sigma}, \lambda)$*, if is continuous with the probability density function,*

$$sn_d(\mathbf{z}|\mathbf{\Sigma}, \lambda) = 2\phi_d(\mathbf{z}|\mathbf{\Sigma})\Phi_1\Big(\lambda^\intercal \mathbf{\Sigma}^{-1/2}\mathbf{z}\Big) \tag{3.6}$$

for $\lambda \in \mathcal{R}^d$ *and* $\mathbf{\Sigma}$ *be a* $d \times d$ *positive definite matrix. For the univariate case*

$$sn_1(z|\sigma^2, \lambda) = 2\phi_1(z|\sigma^2)\Phi_1\big(\frac{\lambda z}{\sigma}\big) \tag{3.7}$$

where $\phi_d(.)$ *and* $\Phi_d(.)$ *are the PDF and CDF of a* d*-dimensional standard multivariate normal variable, respectively.*

Skew-normal copula is obtained from the multivariate skew-normal distribution and the corresponding univariate quantiles using the probability integral transformation method. [58] studied a class of copulas generated from skew-normal distribution.

Definition 3.3.2 *A* d*-dimensional copula is said to be a skew-normal copula if*

$$C_{d,SN}(\mathbf{u}|\mathbf{\Sigma}, \lambda) = SN_d(SN_1^{-1}(u_1|1, \lambda_1^*), \ldots, SN_1^{-1}(u_d|1, \lambda_d^*)|\mathbf{\Sigma}, \lambda) \tag{3.8}$$

where $SN_1^{-1}(u_j|1, \lambda_j^*)$ *denotes the inverse of the CDF of* $Z_j \sim SN_1(1, \lambda_j^*)$ *distribution for* $j = 1, \ldots, d$*. Here the skewness parameters* (λ_j^*) *of the univariate quantiles can be obtained from the multivariate parameters* $\mathbf{\Sigma}, \lambda$ *by*

$$\lambda_j^* = \frac{\delta_j^*}{\sqrt{1 - \delta_j^{*2}}}, \quad \textit{where} \quad \delta^* = \mathbf{\Sigma}^{1/2}\frac{\lambda}{\sqrt{1 + \lambda^\intercal\lambda}}. \tag{3.9}$$

For more details regarding this, one may refer to [59]. The corresponding skew-normal copula density is given by

$$c_{d,SN}(\mathbf{u}|\mathbf{\Sigma},\lambda) = \frac{sn_d(SN_1^{-1}(u_1|1,\lambda_1^*),\ldots,SN_1^{-1}(u_d|1,\lambda_d^*)|\mathbf{\Sigma},\lambda)}{\prod_{j=1}^{d} sn_1(SN_1^{-1}(u_j|1,\lambda_j^*))}. \tag{3.10}$$

Skew-normal copula is exchangeable or permutation symmetric if and only if $\lambda_j = \lambda$ for all $j = 1,\ldots,d$, and all off-diagonal elements of the correlation matrix $\mathbf{\Sigma}$ are equal. Note that Gaussian copula is nested with in the skew-normal copula when $\lambda_j = 0$ for all $j = 1,\ldots,d$.

Remark: Multivariate skew-t distribution is a member of the skew-elliptical family of distribution which is defined as a scale mixture of skew-normal distribution. [60] and [61] discussed the theoretical properties of this distribution.

Definition 3.3.3 *Let $\mathbf{Z}$ taking values in $\mathcal{R}^d$ be a mean zero skew-normal vector and $\mathbf{V}$ be another variable independent with $\mathbf{Z}$ such that, $\mathbf{V} \sim \chi^2_\nu/\nu$. Then $\mathbf{T} = \mathbf{V}^{-1/2}\mathbf{Z}$ follows a mean zero skew-t distribution with the probability density function,*

$$st_d(\mathbf{t}|\mathbf{\Sigma},\lambda,\nu) = 2t_d(\mathbf{t}|\mathbf{\Sigma},\nu)T_1\left(\lambda^\intercal\mathbf{\Sigma}^{-1/2}\mathbf{t}\sqrt{\frac{\nu+d}{\mathbf{Q_t}+\nu}}\Big|\nu+d\right) \tag{3.11}$$

where $\mathbf{Q_t} = \mathbf{t}^\intercal\mathbf{\Sigma}^{-1}\mathbf{t}$, $\lambda \in \mathcal{R}^d$ and $\mathbf{\Sigma}$ be a $d \times d$ positive definite matrix. For the univariate case

$$st_1(t|\sigma^2,\lambda,\nu) = 2t_1(t|\sigma^2,\nu)T_1\Big(\frac{\lambda t}{\sigma}\sqrt{\frac{\nu+1}{Q_t+\nu}}\Big|\nu+1\Big) \tag{3.12}$$

where $t_d(.)$ and $T_d(.)$ are the PDF and CDF of a d-dimensional standard Student-t variable, respectively.

Multivariate skew-t copula can be similarly obtained using the above distribution as -

Definition 3.3.4 *A d-dimensional copula is said to be a skew-t copula if*

$$C_{d,ST}(\mathbf{u}|\mathbf{\Sigma},\lambda,\nu) = ST_d(ST_1^{-1}(u_1|1,\lambda_1^*,\nu),\ldots,ST_1^{-1}(u_d|1,\lambda_d^*,\nu)|\mathbf{\Sigma},\lambda,\nu) \tag{3.13}$$

where $ST_1^{-1}(u_j|1,\lambda_j^*,\nu)$ denotes the inverse of the CDF of $T_j \sim ST_1(1,\lambda_j^*,\nu)$ distribution. The corresponding skew-t copula density is given by

$$c_{d,ST}(\mathbf{u}|\mathbf{\Sigma},\lambda,\nu) = \frac{st_d(ST_1^{-1}(u_1|1,\lambda_1^*,\nu),\ldots,ST_1^{-1}(u_d|1,\lambda_d^*,\nu)|\mathbf{\Sigma},\lambda,\nu)}{\prod_{j=1}^{d} st_1(ST_1^{-1}(u_j|1,\lambda_j^*,\nu))}. \tag{3.14}$$

The skew-normal copula in Definition (3.3.2) captures the non-exchangeable dependence between the variables of interest, with the correlation matrix $\mathbf{\Sigma}$ accounting for association between unobservable or latent variables Z_j in (3.8) and $\lambda = (\lambda_1,\ldots,\lambda_d)^\intercal$ accounting for the differential skewness of the variables involved. Comparing the skew-normal copula with the skew-t copula in Definition (3.3.4) involves knowledge about an extra parameter which is the degrees of freedom (denoted by ν. This

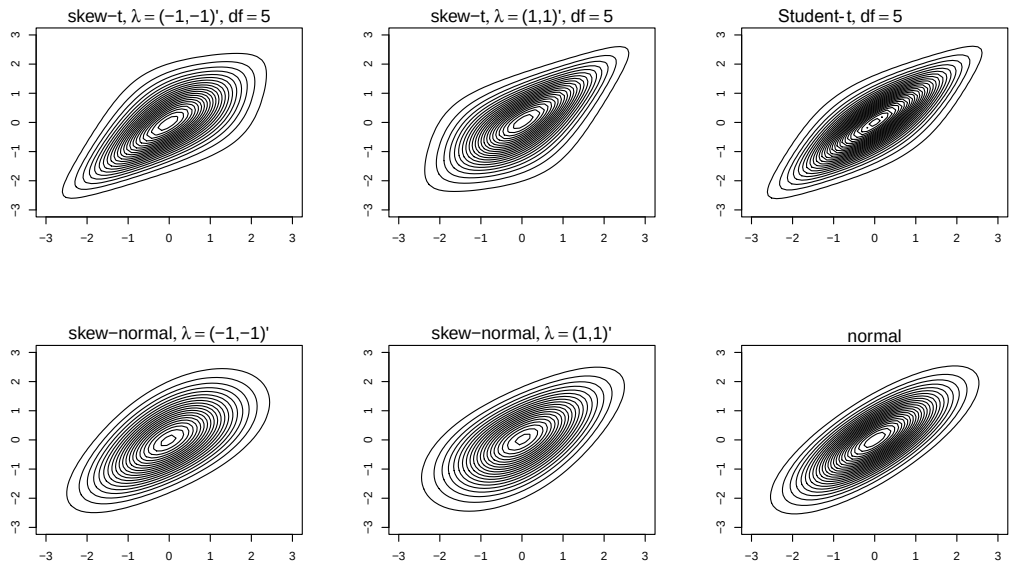

Figure 3.3 Contour plots of bivariate distributions using skew-t copula (on the upper row) (with $\nu = 5$ and $\lambda = \{(-1,-1),(1,1),(0,0)\}$) with corresponding skew-normal copula (on the lower row); the common correlation parameter $\rho = 0.77$ and common marginals are standard normal.

parameter (ν) accounts for possible tail dependence in the data. [62] discussed the applications and maximum likelihood estimation of the skew-t copula while [63] discussed its Bayesian estimation. We note that as the degrees of freedom (ν) approaches ∞ we obtain the skew-normal copula in the limit. Gaussian and Student-t copula are also nested within the skew-t copula. To provide some visual representations, we plot the contours of the joint densities of skew-t, Student-t, skew-normal and normal copula, using standard normal margins in Figure 3.3.

3.4 Parameter estimation

To estimate the parameters for the models in Section 3.2, we use IFM (inference function of margins), a fully parametric method. IFM is particularly useful for models with complex dependence structures like non-exchangeable multivariate copulas, as discussed in Section 3.3. Direct maximum likelihood estimation is computationally demanding due to the time-consuming computation of quantiles for skew-elliptical copulas. [35] proposed a two-stage maximum likelihood estimation method, wherein all parameters are estimated in two steps. [36] further discussed the asymptotic efficiency of this method: the univariate parameters are first estimated using separate univariate likelihoods, followed by estimation of multivariate parameters using the multivariate likelihood with the univariate parameters fixed from the first stage.

For the class of models in (3.4), we first consider the response variables $\mathbf{Y}_i$ are conditionally independent given the random effects $\mathbf{b}_i$, and then the joint density of the i-th response is

$$f_{n_i}(\mathbf{y}_i, \mathbf{b}_i|\theta_i^*) = f_{n_i}(\mathbf{y}_i|\mathbf{b}_i, \theta_i^*)g(\mathbf{b}_i). \tag{3.15}$$

Therefore, we construct the inference functions as

$$\begin{aligned} l_i^*(\theta_i|\mathbf{y}_i) &= \log \int \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i, \theta_{ij})g(\mathbf{b}_i)d\mathbf{b}_i, \\ l_i^*(\theta_i^*|\mathbf{y}_i) &= \log \int c_{n_i}(F(y_{i1}|\mathbf{b}_i, \theta_{i1}), \ldots, F(y_{in_i}|\mathbf{b}_i, \theta_{in_i})|\phi_i) \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i, \theta_{ij})g(\mathbf{b}_i)d\mathbf{b}_i. \end{aligned} \tag{3.16}$$

Now to obtain the IFM estimates based on m independent observations we assume, $\theta_i^* = (\theta_i^\intercal, \phi_i^\intercal)^\intercal$ are all functions of $\theta^* = (\theta^\intercal, \phi^\intercal)^\intercal$ for $i = 1, \ldots, m$. For notational simplicity, we assume that the parameters of the random effects distribution is included in θ and ϕ includes the dependence parameters as in (3.5). The IFM method estimates the model parameters by

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \sum_{i=1}^{m} l_i^*(\theta_i|\mathbf{y}_i), \\ \hat{\phi} &= \arg\max_{\phi} \sum_{i=1}^{m} l_i^*(\hat{\theta}_i, \phi_i|\mathbf{y}_i) \end{aligned} \tag{3.17}$$

and the estimating equations for θ and ϕ based on IFM are

$$\begin{aligned} \Psi_{1m}(\theta) &= \sum_{i=1}^{m} \Psi_{i1m}(\theta) = \sum_{i=1}^{m} \frac{\partial}{\partial\theta} l_i^*(\theta_i|\mathbf{y}_i) = \mathbf{0}, \\ \Psi_{2m}(\phi) &= \sum_{i=1}^{m} \Psi_{i2m}(\phi) = \sum_{i=1}^{m} \frac{\partial}{\partial\phi} l_i^*(\hat{\theta}_i, \phi_i|\mathbf{y}_i) = \mathbf{0} \end{aligned} \tag{3.18}$$

provided the derivatives exist. The IFM estimates are obtained either by numerically maximizing (4.18) or by solving the system of non-linear equations as in (3.18).

3.5 Asymptotic normality

Here we show under some regularity conditions, the IFM estimators for the class of models in (3.4) are consistent and asymptotically normal. We also present the theoretical analysis for independent and non-identically distributed (i.n.i.d) observations including random effects. Let us denote the true value of the parameters $\theta^* = (\theta^\intercal, \phi^\intercal)^\intercal$ by $\theta_0^* = (\theta_0^\intercal, \phi_0^\intercal)^\intercal$ and let $\Psi_{im}^* = (\Psi_{i1m}^\intercal, \Psi_{i2m}^\intercal)^\intercal$ be the stack of the vector valued inference function for the i-th response. To establish the consistency and asymptotic normality, we need the following set of assumptions.

1. The support of $\mathbf{Z} = (\mathbf{Y}^\intercal, \mathbf{b}^\intercal)^\intercal$, $\mathcal{Z}$ does not depend on any $\theta^* \in \mathbf{\Theta}^*$.

2. The partial derivatives $\partial\Psi_m^*/\partial\theta^*$ exist for almost every $\mathbf{z} \in \mathcal{Z}$.

3. (a) For all $\theta^* \in \mathbf{\Theta}^*$,

$$\frac{1}{m}\sum_{i=1}^{m}\Psi_{i1m}(\theta) \xrightarrow{p} \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}E[\Psi_{i1m}(\theta)] = \mathbf{0},$$
$$\frac{1}{m}\sum_{i=1}^{m}\Psi_{i2m}(\phi) \xrightarrow{p} \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}E[\Psi_{i2m}(\phi)] = \mathbf{0}.$$

   (b) For all $\theta^* \in \mathbf{\Theta}^*$,

$$E[\Psi_{im}(\theta^*)\Psi_{im}(\theta^*)^\intercal] \text{ and } \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}E[\Psi_{im}(\theta^*)\Psi_{im}(\theta^*)^\intercal] \text{ exist,}$$
$$\text{and } \frac{1}{m}\sum_{i=1}^{m}\Psi_{im}(\theta^*)\Psi_{im}(\theta^*)^\intercal \xrightarrow{p} \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}E[\Psi_{im}(\theta^*)\Psi_{im}(\theta^*)^\intercal] = M_\Psi(\theta^*)$$

   where $M_\Psi(\theta^*)$ is a positive definite matrix. $E\Big[\frac{\partial}{\partial\theta^*}\Psi_{im}(\theta^*)\Big]$ exist, and

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial\theta^*}\Psi_{im}(\theta^*) \xrightarrow{p} \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}E\Big[\frac{\partial}{\partial\theta^*}\Psi_{im}(\theta^*)\Big] = D_\Psi(\theta^*)$$

   where $D_\Psi(\theta^*)$ is a non-singular matrix.

4. The order of integration and difference can be interchanged as follows

$$\frac{\partial}{\partial\theta^*}\int_{\mathcal{Z}} f^*(z,\theta^*)dz = \int_{\mathcal{Z}}\frac{\partial}{\partial\theta^*}f^*(z,\theta^*)dz.$$

5. For all $\epsilon > 0$ and for any fixed vector $\mathbf{u}$, $(||\mathbf{u}|| \neq 0)$, the following condition is satisfied.

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}E\Big[(\mathbf{u}^\intercal\Psi_{im}(\theta_0^*))^2 I\Big\{|\mathbf{u}^\intercal\Psi_{im}(\theta_0^*)| \geq \epsilon\sqrt{m}\Big\}\Big] = 0.$$

The above mentioned assumptions make $\Psi_m^* = \sum_{i=1}^m \Psi_{im}^*$ a regular inference function vector and consequently we have the following theorem.

Theorem 3.5.1 *Consider the model (3.4) and let $\hat{\theta}^* = (\hat{\theta}^\intercal, \hat{\phi}^\intercal)^\intercal$ denote the IFME of θ^* corresponding to IFM (3.16). Under the above regularity assumptions, $\hat{\theta}^*$ is a consistent estimator of θ^*. Furthermore, as $m \to \infty$, we have asymptotic normality as*

$$\sqrt{m}(\hat{\theta}^* - \theta_0^*) \xrightarrow{d} N(\mathbf{0}, J_\Psi(\theta_0^*)^{-1}), \quad \textit{where} \quad J_\Psi(\theta_0^*) = D_\Psi(\theta_0^*)^\intercal M_\Psi(\theta_0^*)^{-1} D_\Psi(\theta_0^*),$$

$$M_\Psi(\theta_0^*) = \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m} E[\Psi_{im}(\theta_0^*)\Psi_{im}(\theta_0^*)^\intercal] \;\; \textit{and} \;\; D_\Psi(\theta_0^*) = \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m} E\Big[\frac{\partial}{\partial\theta^*}\Psi_{im}(\theta^*)\Big|_{\theta_0^*}\Big].$$

Based on the general discussions in [37], we also provide a simplified proof of the above theorem. It provides a defense for the numerical computations utilized in this chapter.

Proof: Using Taylor's (Lagrange) expansion to the first order, we have

$$\begin{aligned}\Psi_{1m}(\hat{\theta}) &= \Psi_{1m}(\theta_0) + (\hat{\theta}-\theta_0)\frac{\partial}{\partial\theta}\Psi_{1m}(\theta)\Big|_{\theta_1},\\ \Psi_{2m}(\hat{\phi}) &= \Psi_{2m}(\phi_0) + (\hat{\phi}-\phi_0)\frac{\partial}{\partial\phi}\Psi_{2m}(\phi)\Big|_{\phi_1}\end{aligned} \tag{3.19}$$

where θ_1 is some vector value between θ_0 and $\hat{\theta}$, and ϕ_1 is some vector value between ϕ_0 and $\hat{\phi}$, respectively. Note that $\Psi_{2m}(\hat{\phi})$ also depends on the value $\hat{\theta}$. Assumption 3(a) implies

$$E\Big[\frac{\partial}{\partial\theta}\log\int\prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b_i},\theta_{ij})g(\mathbf{b_i})d\mathbf{b_i}\Big]$$

exists and naught for all $i = 1,\ldots,m$. Thus we have

$$\frac{1}{m}\Psi_{1m}(\theta_0) \xrightarrow{p} \mathbf{0}, \;\; \frac{1}{m}\Psi_{2m}(\phi_0) \xrightarrow{p} \mathbf{0}.$$

Also from assumption 3(b), the expectations of

$$\frac{1}{m}\frac{\partial}{\partial\theta}\Psi_{1m}(\theta) \;\text{ and }\; \frac{1}{m}\frac{\partial}{\partial\phi}\Psi_{2m}(\phi)$$

converges to non-zero real vectors almost surely. Since all terms on the right hand side converges to zero, when $\hat{\theta}$ and $\hat{\phi}$ are the solutions of 3.18. Hence we must have

$$\hat{\theta} \xrightarrow{p} \theta_0 \;\text{ and }\; \hat{\phi} \xrightarrow{p} \phi_0.$$

Now we need to derive the asymptotic normality. Let

$$H_m(\theta^*) = \begin{pmatrix} \frac{\partial}{\partial\theta}\Psi_{1m}(\theta) & \mathbf{0} \\ \mathbf{0} & \frac{\partial}{\partial\phi}\Psi_{2m}(\phi)\end{pmatrix} \;\text{ and }\; H_m^1 = \begin{pmatrix} \frac{\partial}{\partial\theta}\Psi_{1m}(\theta)\Big|_{\theta_1} & \mathbf{0} \\ \mathbf{0} & \frac{\partial}{\partial\phi}\Psi_{2m}(\phi)\Big|_{\phi}\end{pmatrix}$$

We rewrite the expression in 3.19 as

$$\sqrt{m}(\hat{\theta}^* - \theta_0^*) = \Big[\frac{1}{m}H_m^1\Big]^{-1}\frac{1}{\sqrt{m}}[-\Psi_m(\theta_0^*)]. \tag{3.20}$$

Since $\hat{\theta}^*$ is a consistent estimator of θ_0^*, from the convergence in probability we have

$$\frac{1}{m}[H_m(\hat{\theta}^*) - H_m(\theta_0^*)] \xrightarrow{p} \mathbf{0}.$$

Now from assumption 3(b) we have,

$$\frac{1}{m}H_m(\theta_0^*) = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^m \frac{\partial}{\partial\theta}\Psi_{i1m}(\theta)\Big|_{\theta_0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{m}\sum_{i=1}^m \frac{\partial}{\partial\phi}\Psi_{i2m}(\phi)\Big|_{\phi} \end{pmatrix} \xrightarrow{p} D_\Psi(\theta_0^*).$$

Thereafter using assumption 3(b) and 4 we get,

$$\begin{aligned}
\frac{1}{m^2}Cov[H_m(\theta_0^*)] &= \frac{1}{m^2}\sum_{i=1}^m Cov\Big[\frac{\partial}{\partial\theta^*}\Psi_{im}(\theta^*)\Big|_{\theta_0^*} \\
&= \frac{1}{m}\Big[\frac{1}{m}\sum_{i=1}^m E\Big[\frac{\partial}{\partial\theta^*}\Psi_{im}(\theta^*)\Big|_{\theta_0^*}\frac{\partial}{\partial\theta^{*\intercal}}\Psi_{im}(\theta^*)^\intercal\Big|_{\theta_0^*}\Big] \\
&= \frac{1}{m}\Big[\frac{1}{m}\sum_{i=1}^m \frac{\partial^2}{\partial\theta^*\partial\theta^{*\intercal}}E\Big[\Psi_{im}(\theta^*)\Big|_{\theta_0^*}\Psi_{im}(\theta^*)^\intercal\Big|_{\theta_0^*}\Big] \quad \to \mathbf{0} \text{ as } m \to \infty.
\end{aligned}$$

Now θ_1^* lies in between $\hat{\theta}^*$ and θ_0^*, thus by weak law of large number

$$\frac{1}{m}H_m^1 - D_\Psi(\theta_0^*) \xrightarrow{p} \mathbf{0}.$$

The final term of the expression 3.20, $\Psi_m(\theta_0^*)$ involves sum of independent terms, which have expectation $\mathbf{0}$ and covariance matrix $Cov[\Psi_{im}(\theta_0^*)] = E[\Psi_{im}(\theta_0^*)\Psi_{im}(\theta_0^*)^\intercal]$ for $i = 1, \ldots, m$. Hence from assumption 5, with direct application of Lindeberg-Feller central limit theorem, for any fixed vector $\mathbf{u}$ we have,

$$\mathbf{u}^\intercal\Big(\frac{\Psi_m(\theta_0^*)}{\sqrt{m}}\Big) \xrightarrow{d} N(0, \mathbf{u}^\intercal M_\Psi(\theta_0^*)\mathbf{u})$$

Combining everything and using Slutsky's theorem we finally have,

$$\sqrt{m}(\hat{\theta}^* - \theta_0^*) \xrightarrow{d} N(\mathbf{0}, J_\Psi(\theta_0^*)^{-1}), \text{ where } J_\Psi(\theta_0^*) = D_\Psi(\theta_0^*)^\intercal M_\Psi(\theta_0^*)^{-1}D_\Psi(\theta_0^*),$$

and that completes the proof. □

Note that in this context, both $M_\Psi(\theta_0^*)$ and $D_\Psi(\theta_0^*)$ are contingent upon the distribution of the random effects $\mathbf{b}_i$. The asymptotic covariance matrix referred to in (3.19) is commonly recognized as the Godambe information matrix within the literature. It encapsulates crucial information regarding the variability and precision of the estimated parameters in the model, particularly with respect to the effects

of the random components. The observed version of this matrix can be numerically obtained by

$$M_\Psi(\hat{\theta}^*) = \frac{1}{m}\sum_{i=1}^{m} \Psi_{im}(\theta^*)\Big|_{\hat{\theta}^*} \Psi_{im}(\theta^*)^\intercal\Big|_{\hat{\theta}^*},$$
$$D_\Psi(\hat{\theta}^*) = \frac{1}{m}\sum_{i=1}^{m} \begin{pmatrix} \frac{\partial}{\partial\theta}\Psi_{i1m}(\theta)\Big|_{\hat{\theta}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial}{\partial\phi}\Psi_{i2m}(\phi)\Big|_{\hat{\phi}} \end{pmatrix}, \tag{3.21}$$

which leads to a straightforward calculation of the standard errors for the parameter estimates based on IFM (3.16), using the square-roots of the diagonals of $J_\Psi(\hat{\theta}^*)^{-1}$. For numerical optimizations, we employ the *optim* ([64]) function in R, utilizing the L-BFGS-B method. We employed numerical derivative methods to obtain the matrices in (3.21) using *numderiv* ([65]) function in R.

3.6 Numerical implementation

In Section 3.2, we have discussed generalized linear mixed models (GLMMs), specifically focusing on their application to the marginals within the class of models outlined in Equation (3.4). During our exploratory analysis, we considered a diverse range of potential distributions and link functions. Among them, the Gamma distribution combined with a skew-elliptical copula emerged as a particularly suitable choice. This combination is especially appropriate for modeling response variables that exhibit asymmetry and are constrained to positive real values, aligning well with the characteristics observed in our data. By modeling the response variable Y_{ij} as Gamma-distributed with a shape parameter κ and mean η_{ij}, we capture the nuances of the data more effectively. Gamma distribution has the density

$$f(y_{ij}|\eta_{ij},\kappa) = \frac{1}{\Gamma(\kappa)}\Big(\frac{\eta_{ij}}{\kappa}\Big)^{-1} y_{ij}^{k-1}\exp\Big(-\frac{y_{ij}\kappa}{\eta_{ij}}\Big). \tag{3.22}$$

For comparison, we also consider a symmetric marginal distribution such as normal with mean η_{ij} and variance σ^2, having the density

$$f(y_{ij}|\eta_{ij},\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{1}{2}\Big(\frac{y_{ij}-\eta_{ij}}{\sigma}\Big)^2\Big). \tag{3.23}$$

Notations are followed from [66]. To evaluate the integrations in (3.16), we use standard Gauss-Hermite quadrature rule with 15 quadrature points. To obtain the asymptotic covariance matrix in (3.21) we used same number of quadrature points to compute the standard errors of the parameter estimates.

The serial correlation among repeated measurements manifests within the correlation matrix of the skew-elliptical copulas. In Equation (3.3), the matrix $\mathbf{\Sigma}$ is construed as a function of both time and the dispersion parameter ξ, reflecting time-varying serial dependence. Aligning with the modeling framework elucidated in Equation (3.4), we presuppose a uniform variance (σ^2 or $1/\kappa$) across units. Furthermore, we adopt an autoregressive order one (AR(1)) structure for the correlation matrix $\mathbf{\Sigma}$ within

the skew-elliptical copulas. This choice of correlation structure is particularly suited for uniformly spaced observations. Within this framework, we implement the exponential autocorrelation function, defined as

$$Corr(Y_{ij}, Y_{ik}) = \exp\left(-\xi|t_{ij} - t_{ik}|\right), \quad \xi \geq 0, \tag{3.24}$$

to construct the correlation matrix $\mathbf{\Sigma}(\xi, \mathbf{t}_i)$ within the class of skew-elliptical copulas, we rely on structural assumptions derived from the sample correlation matrix of our data. Alternatively, one can opt for an exchangeable (EX) correlation matrix tailored to suit the specifics of the dataset. The IFM method, while widely employed for parameter estimation, involves a two-stage process that may not be as efficient as direct maximum likelihood estimation. As noted by [67], a small bias in the initial stage of IFM estimation may influence the accuracy of copula parameter estimation in the subsequent stage. In our analysis, we've observed that leaving the skewness parameter of the skew-normal and skew-t copulas unconstrained often leads to either heavily biased parameter estimates or failure to converge to an optimal solution. Given that longitudinal data consists of repeated observations of the same sample across time, we adopt an equi-skewness approach, assuming a constant skewness parameter λ across all dimensions. This not only reduces the number of estimable parameters in the model but also ensures the positive definiteness of the correlation matrices throughout the estimation process. In the subsequent section, we delve into a comparative analysis of the various models fitted to the data, elucidating their respective strengths and limitations.

3.7 Model comparison

An inherent challenge in analyzing longitudinal data lies in model selection, particularly in determining the most suitable number of components for a given dataset. Within the framework of the models outlined in Equation (3.4), selecting appropriate marginal distributions alongside the multivariate copula becomes paramount. To address this, log-likelihood-based measures such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are commonly employed. These criteria penalize models with a large number of parameters, thus aiding in the selection of parsimonious yet effective models. The AIC and BIC are defined as follows

$$AIC = -2l(\hat{\theta}^*) + 2\dim(\theta^*), \quad BIC = -2l(\hat{\theta}^*) + \log(m)\dim(\theta^*) \tag{3.25}$$

where $\hat{\theta}^*$ represents the maximum likelihood estimates of the model parameters and m denotes the sample size. While the penalty term in AIC depends solely on the number of parameters in the model, BIC additionally considers the sample size. However when we employ two stage maximum likelihood estimation for the class of models these information criteria are modified accordingly in the literature ([68]). Here, we assume that the true data generating model is included among the set of candidate models. [69] demonstrated that if both the marginals and the copula family are correctly specified, then the two-stage AIC converges to the original AIC. Similar justifications are presented in [16] (Remark

1). In our approach, we utilize these adjusted criteria as close approximations of the true AIC and BIC, incorporating two-stage estimates into the evaluated likelihood functions. The use of BIC for model selection in longitudinal data settings remains relatively underexplored.

3.8 Simulation design and analysis

Simulation studies are undertaken to elucidate the efficacy of IFM estimation in the context of the proposed class of multivariate models outlined in (3.4). The primary objective of this investigation is to assess parameter inference using IFM estimation across various combinations of copulas and marginal distributions. To achieve this, we consider two distinct sample sizes, $m = 200$ and 500, while maintaining a fixed dimension of $n_i = 4$ for each response vector. Given that the true parameters are known, the generation of the response variable $\mathbf{Y}_i|\mathbf{b}_i$ entails multiple steps. Initially, we sample from the multivariate copulas discussed in Section 3.3, adjusting the dependence parameters $(\xi, \lambda, \nu)^\intercal$ to mimic real-world scenarios. Subsequently, we employ the probability integral transformation to generate per-unit multivariate response $\mathbf{Y}_i|\mathbf{b}_i$. Throughout these simulations, we assume a normal distribution for the random effects in all cases. However, to streamline computation time, we confine ourselves to models featuring a random intercept structure in the linear predictor. In Equation (3.16), we utilize a straightforward GLMM estimation at the first stage, ensuring comparability across the considered models. While maintaining fixed marginal distributions, we juxtapose parameter estimates across four choices of multivariate copulas. To initiate estimation, we leverage the *nlme* package in R to obtain initial estimates of the marginal parameters. Subsequently, the initial values of the dependence parameters are derived by fitting the rescaled empirical cumulative distribution function (CDF) to the copulas. Notably, for skew-t and Student-t copulas, we adopt fixed integer-valued degrees of freedom parameters ν.

The class of models under consideration aims to examine parameter estimation in scenarios where both binary and categorical variables are included as covariates. This approach is motivated by real-world dataset discussed in subsequent section. The general structure of these models encompasses both fixed and random effects, incorporating time-varying dependence. We denote this structure as -

$$\begin{aligned} \mathbf{Y}_i|\mathbf{b}_i &\sim F_{n_i}(\eta(\mathbf{X}_i\beta + \mathbf{D}_i\mathbf{b}_i), \mathbf{\Sigma}(\xi, \mathbf{t}_i)) \\ \text{with } \; x_{ij}\beta + d_{ij}\mathbf{b}_i &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + t_{ij}\beta_3 + b_i, \end{aligned} \tag{3.26}$$

where F_{n_i} is a multivariate distribution with link function η as described in (3.3). We set here $\beta_0 = 1.5$, $\beta_1 = 0.5$, $\beta_2 = 0.5$ and $\beta_3 = 1.0$ for the fixed effect parameters. Here $x_{1i} = 1 * I(i \leq m/2) + 2 * I(i > m/2)$, and $x_{2i} = 0$ or 1 assigned randomly (m is the sample size). The random effects are set as $b_i \sim N(0, 1)$ and the time points $t_{ij} = j - 2.5$ for $j = 1, \ldots, n_i$. In one scenario we use Gamma distribution with log-link and shape parameter $\kappa = 3$ for the marginals. In the other scenario we use normal distribution with standard deviation $\sigma = 1$. For the correlation matrix in all the multivariate copulas we set the time difference per observation within each response vector to unity. Hence the

off-diagonal entries of the matrix $\mathbf{\Sigma}(\xi)$ are

$$\rho(t_{ij}, t_{ik}) = \exp(-\xi|j-k|), \quad 1 \leq j < k \leq n_i. \tag{3.27}$$

Here we consider skew-t and skew-normal copula along with their symmetric counterparts for the dependence structure of the models. We set $\xi = 0.25$ for all the copulas. For the skew-elliptical ones we set $\bar{\lambda} = 1$ under equi-skewness, and fixed values for the degrees of freedom parameter as $\nu = \{3, 8, 15\}$. For each simulation we generate $N = 500$ Monte Carlo samples and then estimate the parameters and their associated standard errors.

Table 3.1 and 3.2 show parameter estimations of the class of models in (3.26) using skew-elliptical and elliptical copulas and gamma marginals with Gaussian random effects. We present the mean, the biases [$\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_j^* - \theta^*)$], roots of mean square errors [$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_j^* - \theta^*)^2}$], empirical standard errors (denoted as SD) and the standard errors obtained from the asymptotic covariance matrices (denoted as SE), where $\hat{\theta}_j^*$ is the parameter estimates for the j-th sample. In Table 3.3 and 3.4, we provide additional simulation results using normal marginals as well. The results show that parameter estimations tend to have higher accuracies based on larger sample size m with smaller Bias, SE and RMSE. However, there is systematic bias in the estimation of the skewness parameter $\bar{\lambda}$ for all the models, and the shape parameter κ for the Gamma based models. We notice that using IFM method does not change the precision of estimation of the marginal parameters very much. Overall, the estimation of parameters are accurate and SE and SD are relatively close to each other, which shows this estimating approach is viable for drawing inference from real data set.

3.9 Data analysis

Given the skewed nature of the CD4 count marker with positive real support, our aim is to model the marginals using a Gamma mixed model, while capturing temporal dependence through skewed multivariate copulas to elucidate disease progression. Our model focuses on understanding the evolution of CD4 counts over time within individual patients. Despite attempts to alleviate skewness through transformations like $\log$ or square root, we found that these measures do not completely normalize the data. As a result, we opt for a scale transformation of the CD4$^+$ T cell counts by a factor of 100, facilitating easier estimation and interpretation of the coefficients. Notably, due to sparse entries in the 5-th visit column, we exclude it from our analysis. Additionally, some entries were missing in the 4-th visit column, which we address by employing the carry-forward method for imputation ([19]).

Our predictors include age, gender, the first baseline regimen, and the initial weight of each patient. Gender is indicated using binary indicators, with 0 denoting female and 1 denoting male patients. The first baseline regimen (FBR) of antiretroviral (ARV) combination is encoded as $1, 2$ or 5, corresponding to different regimens. With these factors in mind, referencing the models in (3.3), we consider:

$$x_{ij}\beta + d_{ij}\mathbf{b}_i = \beta_0 + \text{gender}_i\beta_1 + \text{age}_i\beta_2 + \text{fbr}_i\beta_3 + \text{weight}_i\beta_4 + t_{ij}\beta_5 + b_i, \tag{3.28}$$

	Parameters	β_0	β_1	β_2	β_3	$V[b]$	κ	ξ	λ
	True Value	1.5	0.5	0.5	1.0	1.0	3.0	0.25	1.0
Copula	Skew-t, $\nu = 3$								
m = 200	Mean	1.4917	0.5054	0.4967	1.0001	0.9890	3.3965	0.3171	0.6465
	Bias	-0.0083	0.0054	-0.0033	0.0001	-0.0110	0.3965	0.0671	-0.3535
	SD	0.2598	0.1542	0.1525	0.0208	0.1820	1.1862	0.1192	0.3294
	SE	0.2800	0.1850	0.1971	0.0317	0.2002	1.9486	0.0423	0.2934
	RMSE	0.2599	0.1543	0.1525	0.0208	0.1823	1.2507	0.1368	0.4832
m = 500	Mean	1.4921	0.5068	0.4953	0.9999	0.9929	3.2168	0.2848	0.7376
	Bias	-0.0079	0.0068	-0.0047	-0.0001	-0.0071	0.2168	0.0348	-0.2624
	SD	0.1712	0.1003	0.1025	0.0117	0.1211	0.7187	0.0738	0.2283
	SE	0.1680	0.0934	0.1107	0.0158	0.1064	0.6521	0.0249	0.2426
	RMSE	0.1714	0.1005	0.1026	0.0117	0.1213	0.7507	0.0816	0.3478
Copula	Skew-t, $\nu = 8$								
m = 200	Mean	1.5115	0.4933	0.5020	1.0024	0.9678	3.3293	0.3307	0.6405
	Bias	0.0115	-0.0067	0.0020	0.0024	-0.0322	0.3293	0.0807	-0.3595
	SD	0.2687	0.1610	0.1574	0.0188	0.1849	1.2072	0.1410	0.3908
	SE	0.2829	0.1336	0.1754	0.0310	0.1458	1.5335	0.0416	0.3185
	RMSE	0.2689	0.1611	0.1575	0.0189	0.1878	1.2513	0.1625	0.5310
m = 500	Mean	1.4945	0.5025	0.4978	1.0008	1.0003	3.2085	0.3087	0.6910
	Bias	-0.0055	0.0025	-0.0022	0.0008	0.0003	0.2085	0.0587	-0.3090
	SD	0.1700	0.1014	0.1003	0.0124	0.1282	0.7294	0.0886	0.2289
	SE	0.1621	0.0773	0.1069	0.0178	0.0804	0.8080	0.0251	0.2005
	RMSE	0.1700	0.1014	0.1004	0.0125	0.1282	0.7587	0.1063	0.3845
Copula	Skew-t, $\nu = 15$								
m = 200	Mean	1.5158	0.4904	0.4935	0.9989	0.9712	3.3751	0.3432	0.6272
	Bias	0.0158	-0.0096	-0.0065	-0.0011	-0.0288	0.3751	0.0932	-0.3728
	SD	0.2655	0.1546	0.1517	0.0198	0.1884	1.2154	0.1563	0.4076
	SE	0.2718	0.1504	0.1751	0.0310	0.1509	1.7044	0.0426	0.3271
	RMSE	0.2659	0.1549	0.1519	0.0198	0.1906	1.2719	0.1810	0.5524
m = 500	Mean	1.4984	0.5003	0.5015	0.9997	0.9994	3.1865	0.3086	0.6926
	Bias	-0.0016	0.0003	0.0015	-0.0003	-0.0006	0.1865	0.0586	-0.3074
	SD	0.1711	0.0981	0.0982	0.0128	0.1272	0.6829	0.0873	0.2690
	SE	0.1573	0.0742	0.1000	0.0158	0.0785	0.6281	0.0248	0.2171
	RMSE	0.1711	0.0981	0.0982	0.0128	0.1272	0.7080	0.1052	0.4085
Copula	Skew-normal								
m = 200	Mean	1.4854	0.5080	0.5098	1.0024	0.9750	3.3101	0.3554	0.5956
	Bias	-0.0146	0.0080	0.0098	0.0024	-0.0250	0.3101	0.1054	-0.4044
	SD	0.2629	0.1546	0.1563	0.0210	0.1843	1.1463	0.1621	0.3980
	SE	0.2920	0.1406	0.2132	0.0387	0.1661	1.5973	0.0472	0.3602
	RMSE	0.2633	0.1548	0.1565	0.0212	0.1860	1.1876	0.1933	0.5674
m = 500	Mean	1.4891	0.5039	0.5017	1.0001	0.9884	3.1932	0.3211	0.6771
	Bias	-0.0109	0.0039	0.0017	0.0001	-0.0116	0.1932	0.0711	-0.3229
	SD	0.1765	0.1009	0.0948	0.0121	0.1249	0.7273	0.1005	0.2788
	SE	0.1576	0.0735	0.0990	0.0157	0.0791	0.7316	0.0258	0.2033
	RMSE	0.1768	0.1009	0.0949	0.0121	0.1255	0.7525	0.1231	0.4266

Table 3.1 Parameter estimation using IFM method when the marginals are distributed as Gamma. Performance for 500 replications with skew-t and skew-normal copula.

	Parameters	β_0	β_1	β_2	β_3	$V[b]$	κ	ξ	λ
	True Value	1.5	0.5	0.5	1.0	1.0	3.0	0.25	-
Copula	Student-t, $\nu = 3$								
m = 200	Mean	1.4864	0.5049	0.5004	1.0002	0.9782	3.5281	0.2811	-
	Bias	-0.0136	0.0049	0.0004	0.0002	-0.0218	0.5281	0.0311	-
	SD	0.2704	0.1567	0.1644	0.0169	0.1831	1.4162	0.1049	-
	SE	0.2963	0.2271	0.2142	0.0441	0.2264	2.1010	0.0243	-
	RMSE	0.2708	0.1568	0.1644	0.0169	0.1844	1.5114	0.1094	-
m = 500	Mean	1.4922	0.4982	0.5029	0.9999	1.0026	3.3119	0.2702	-
	Bias	-0.0078	-0.0018	0.0029	-0.0001	0.0026	0.3119	0.0202	-
	SD	0.1725	0.0997	0.1014	0.0098	0.1354	0.8041	0.0646	-
	SE	0.1945	0.0791	0.1323	0.0144	0.0895	1.3090	0.0153	-
	RMSE	0.1727	0.0997	0.1015	0.0098	0.1355	0.8625	0.0677	-
Copula	Student-t, $\nu = 8$								
m = 200	Mean	1.5173	0.4930	0.4892	0.9996	0.9842	3.4324	0.2822	-
	Bias	0.0173	-0.0070	-0.0108	-0.0004	-0.0158	0.4324	0.0322	-
	SD	0.2819	0.1692	0.1582	0.0160	0.1904	1.2955	0.1074	-
	SE	0.3126	0.1634	0.2097	0.0320	0.1812	2.1569	0.0223	-
	RMSE	0.2824	0.1694	0.1585	0.0160	0.1911	1.3658	0.1122	-
m = 500	Mean	1.5118	0.4953	0.4958	0.9998	0.9939	3.1866	0.2636	-
	Bias	0.0118	-0.0047	-0.0042	-0.0002	-0.0061	0.1866	0.0137	-
	SD	0.1730	0.0979	0.1017	0.0105	0.1411	0.7289	0.0670	-
	SE	0.1649	0.0773	0.1035	0.0138	0.1118	0.6927	0.0135	-
	RMSE	0.1734	0.0980	0.1017	0.0106	0.1413	0.7524	0.0683	-
Copula	Student-t, $\nu = 15$								
m = 200	Mean	1.4785	0.5077	0.5058	1.0004	0.9943	3.5381	0.2953	-
	Bias	-0.0215	0.0077	0.0058	0.0004	-0.0057	0.5381	0.0453	-
	SD	0.2665	0.1613	0.1647	0.0171	0.1843	1.2716	0.1122	-
	SE	0.2860	0.1512	0.1967	0.0343	0.1752	2.2277	0.0223	-
	RMSE	0.2673	0.1615	0.1648	0.0171	0.1844	1.3808	0.1210	-
m = 500	Mean	1.4941	0.5032	0.4931	0.9999	0.9930	3.2570	0.2716	-
	Bias	-0.0059	0.0032	-0.0069	-0.0001	-0.0070	0.2570	0.0216	-
	SD	0.1788	0.1029	0.1065	0.0102	0.1330	0.7906	0.0715	-
	SE	0.1683	0.0818	0.1083	0.0142	0.0837	0.7635	0.0132	-
	RMSE	0.1789	0.1029	0.1067	0.0102	0.1332	0.8313	0.0748	-
Copula	Gaussian								
m = 200	Mean	1.4976	0.4988	0.4987	0.9999	0.9722	3.4854	0.2925	-
	Bias	-0.0024	-0.0012	-0.0013	-0.0001	-0.0278	0.4854	0.0425	-
	SD	0.2643	0.1597	0.1660	0.0163	0.1953	1.3667	0.1264	-
	SE	0.2879	0.1429	0.2174	0.0337	0.1844	1.7209	0.0209	-
	RMSE	0.2643	0.1597	0.1660	0.0163	0.1973	1.4504	0.1333	-
m = 500	Mean	1.4985	0.5006	0.4991	1.0000	0.4937	3.2284	0.2725	-
	Bias	-0.0015	0.0006	-0.0009	0.0000	-0.0063	0.2284	0.0225	-
	SD	0.1780	0.1039	0.1025	0.0096	0.1320	0.7880	0.0778	-
	SE	0.1616	0.0766	0.1030	0.0142	0.0812	0.6222	0.0125	-
	RMSE	0.1781	0.1040	0.1025	0.0096	0.1321	0.8204	0.0809	-

Table 3.2 Parameter estimation using IFM method when the marginals are distributed as Gamma. Performance for 500 replications with Student-t and Gaussian copula.

	Parameters	β_0	β_1	β_2	β_3	$V[b]$	σ	ξ	λ
	True Value	1.5	0.5	0.5	1.0	1.0	1.0	0.25	1.0
Copula	Skew-t, $\nu = 3$								
m = 200	Mean	1.4713	0.5115	0.5102	1.0002	1.0388	0.9768	0.2753	0.6528
	Bias	-0.0287	0.0115	0.0102	0.0002	0.0388	-0.0232	0.0253	-0.3472
	SD	0.3010	0.1882	0.1765	0.0307	0.1674	0.1411	0.0965	0.2841
	SE	0.2722	0.1300	0.1744	0.0308	0.2206	0.1874	0.0694	0.2932
	RMSE	0.3113	0.1886	0.1768	0.0307	0.1736	0.1430	0.1000	0.4486
m = 500	Mean	1.4886	0.5039	0.5071	0.9990	1.0370	0.9771	0.2744	0.6586
	Bias	-0.0114	0.0039	0.0071	-0.0010	0.0374	-0.0229	0.0244	-0.3414
	SD	0.1929	0.1174	0.1139	0.0210	0.1055	0.0756	0.0656	0.1700
	SE	0.1694	0.0818	0.1106	0.0196	0.0947	0.0677	0.0447	0.1784
	RMSE	0.1933	0.1174	0.1141	0.0210	0.1120	0.0720	0.0700	0.3901
Copula	Skew-t, $\nu = 8$								
m = 200	Mean	1.4944	0.5033	0.5101	0.9990	1.0370	0.9771	0.2771	0.6282
	Bias	-0.0056	0.0033	0.0101	-0.0010	0.0370	-0.0229	0.0271	-0.3718
	SD	0.2751	0.1728	0.1711	0.0311	0.1583	0.1320	0.0920	0.3209
	SE	0.2697	0.1278	0.1714	0.0307	0.2076	0.1965	0.0720	0.3146
	RMSE	0.2752	0.1729	0.1714	0.0311	0.1626	0.1340	0.0959	0.4911
m = 500	Mean	1.5046	0.4976	0.5020	1.0002	1.0349	0.9784	0.2764	0.6915
	Bias	0.0046	-0.0024	0.0020	0.0002	0.0349	-0.0216	0.0264	-0.3085
	SD	0.1859	0.1120	0.1107	0.0189	0.1059	0.0989	0.0520	0.1556
	SE	0.1689	0.0812	0.1094	0.0195	0.0943	0.0762	0.0382	0.1449
	RMSE	0.1860	0.1120	0.1107	0.0189	0.1115	0.1012	0.0583	0.3455
Copula	Skew-t, $\nu = 15$								
m = 200	Mean	1.5035	0.4960	0.5011	1.0018	1.0370	0.9773	0.2767	0.6409
	Bias	0.0035	-0.0040	0.0011	0.0018	0.0370	-0.0227	0.0267	-0.3591
	SD	0.2921	0.1747	0.1734	0.0292	0.1594	0.1292	0.0801	0.2943
	SE	0.2663	0.1274	0.1721	0.0311	0.2256	0.1952	0.0670	0.2516
	RMSE	0.2922	0.1747	0.1734	0.0293	0.1636	0.1312	0.0844	0.4643
m = 500	Mean	1.4981	0.5037	0.5009	1.0018	1.0367	0.9775	0.2767	0.6887
	Bias	-0.0019	0.0037	0.0009	0.0010	0.0367	-0.0225	0.0267	-0.3113
	SD	0.1835	0.1102	0.1057	0.0200	0.1042	0.0887	0.0513	0.1672
	SE	0.1695	0.0815	0.1097	0.0195	0.0927	0.0723	0.0363	0.1457
	RMSE	0.1845	0.1105	0.1057	0.0200	0.1105	0.0915	0.0578	0.3534
Copula	Skew-normal								
m = 200	Mean	1.4982	0.4987	0.5097	0.9986	1.0386	0.9775	0.2768	0.6307
	Bias	-0.0018	-0.0013	0.0097	-0.0014	0.0386	-0.0225	0.0268	-0.3693
	SD	0.2889	0.1750	0.1821	0.0319	0.1587	0.1275	0.0791	0.3435
	SE	0.2704	0.1292	0.1723	0.0309	0.2309	0.2162	0.0650	0.2810
	RMSE	0.2889	0.1750	0.1824	0.0319	0.1633	0.1295	0.0835	0.5044
m = 500	Mean	1.4984	0.4990	0.5021	0.9997	1.0381	0.9977	0.2764	0.6963
	Bias	-0.0016	-0.0010	0.0021	-0.0003	0.0381	-0.0223	0.0264	-0.3037
	SD	0.1911	0.1156	0.1109	0.0196	0.0981	0.0859	0.0451	0.1563
	SE	0.1708	0.0839	0.1099	0.0195	0.0818	0.0745	0.0332	0.1509
	RMSE	0.1911	0.1156	0.1109	0.0196	0.1052	0.0887	0.0523	0.3416

Table 3.3 Parameter estimation using IFM method when the marginals are distributed as normal. Performance for 500 replications with skew-t and skew-normal copula.

	Parameters	β_0	β_1	β_2	β_3	$V[b]$	σ	ξ	λ
	True Value	1.5	0.5	0.5	1.0	1.0	1.0	0.25	-
Copula	Student-t, $\nu = 3$								
m = 200	Mean	1.4715	0.5208	0.4996	1.0014	1.0560	0.9627	0.2855	-
	Bias	-0.0285	0.0208	-0.0004	0.0014	0.0560	-0.0373	0.0355	-
	SD	0.3126	0.1852	0.1884	0.0258	0.1862	0.1124	0.1471	-
	SE	0.3352	0.2044	0.2088	0.0621	0.2129	0.0998	0.0887	-
	RMSE	0.3139	0.1864	0.1884	0.0258	0.1944	0.1184	0.1513	-
m = 500	Mean	1.5017	0.4986	0.4995	0.9989	1.0584	0.9635	0.2837	-
	Bias	0.0017	-0.0014	-0.0005	-0.0011	0.0584	-0.0365	0.0337	-
	SD	0.2007	0.1200	0.1176	0.0158	0.1242	0.0804	0.1020	-
	SE	0.1982	0.0968	0.1307	0.0229	0.1476	0.0507	0.0487	-
	RMSE	0.2007	0.1200	0.1176	0.0158	0.1372	0.0883	0.1074	-
Copula	Student-t, $\nu = 8$								
m = 200	Mean	1.4872	0.5041	0.5054	0.9991	1.0557	0.9630	0.2909	-
	Bias	-0.0128	0.0041	0.0054	-0.0009	0.0557	-0.0370	0.0409	-
	SD	0.3053	0.1881	0.1871	0.0256	0.1716	0.1125	0.1427	-
	SE	0.3361	0.2035	0.2067	0.0608	0.2111	0.1333	0.0888	-
	RMSE	0.3056	0.1881	0.1872	0.0256	0.1804	0.1184	0.1484	-
m = 500	Mean	1.5067	0.4963	0.5041	0.9994	1.0495	0.9633	0.2861	-
	Bias	0.0115	-0.0027	0.0041	-0.0006	0.0495	-0.0361	0.0361	-
	SD	0.1918	0.1211	0.1211	0.0154	0.1214	0.0816	0.0978	-
	SE	0.1985	0.0951	0.1273	0.0241	0.1126	0.0549	0.0427	-
	RMSE	0.1923	0.1212	0.1212	0.0156	0.1311	0.0892	0.1042	-
Copula	Student-t, $\nu = 15$								
m = 200	Mean	1.5067	0.5052	0.4917	0.9998	1.0026	0.9814	0.2643	-
	Bias	0.0067	0.0052	-0.0083	-0.0002	0.0026	-0.0186	0.0143	-
	SD	0.3109	0.1900	0.1893	0.0238	0.1568	0.1200	0.1414	-
	SE	0.3356	0.1810	0.1921	0.0256	0.1869	0.1312	0.0799	-
	RMSE	0.3109	0.1901	0.1895	0.0238	0.1570	0.1214	0.1429	-
m = 500	Mean	1.4939	0.5044	0.5021	0.9998	1.0022	0.9908	0.2567	-
	Bias	-0.0061	0.0044	0.0021	-0.0002	0.0022	-0.0092	0.0067	-
	SD	0.1937	0.1203	0.1146	0.0153	0.1076	0.0784	0.0955	-
	SE	0.1748	0.0843	0.1134	0.0154	0.0889	0.0561	0.0433	-
	RMSE	0.1938	0.1203	0.1147	0.0153	0.1078	0.0786	0.0957	-
Copula	Gaussian								
m = 200	Mean	1.4982	0.5082	0.4927	0.9993	1.0066	0.9836	0.2633	-
	Bias	-0.0018	0.0082	-0.0073	-0.0007	0.0066	-0.0164	0.0133	-
	SD	0.3053	0.1885	0.1860	0.0233	0.1512	0.1102	0.1377	-
	SE	0.3153	0.1315	0.1775	0.0293	0.1732	0.1242	0.0751	-
	RMSE	0.3053	0.1885	0.1861	0.0233	0.1514	0.1114	0.1383	-
m = 500	Mean	1.4988	0.5070	0.5038	1.0010	1.0060	0.9928	0.2557	-
	Bias	-0.0012	0.0070	0.0038	0.0010	0.0060	-0.0072	0.0057	-
	SD	0.2047	0.1255	0.1210	0.0160	0.1037	0.0777	0.0951	-
	SE	0.1754	0.0842	0.1135	0.0153	0.0843	0.0543	0.0447	-
	RMSE	0.2051	0.1256	0.1210	0.0160	0.1039	0.0780	0.0953	-

Table 3.4 Parameter estimation using IFM method when the marginals are distributed as normal. Performance for 500 replications with Student-t and Gaussian copula.

and Y_{ij} is the CD4 count at j-th time point for the i-th patient (normalized by 100). The time variable is rescaled as $t_{ij} = (\text{week} - 18)/12$. We have considered random intercept structure in the models. Based on the sample correlation matrix of this data set, AR(1) structure for the correlation matrices for the multivariate copulas seems to be appropriate. After the rescaling of the time points, the entries of the correlation matrix $\mathbf{\Sigma_i}$, are equivalent to $\rho(t_{ij}, t_{ik}) = \exp(-\xi|j-k|), \;\; 1 \leq k < j \leq n_i$. Considering two marginal mixed models with four multivariate copulas, we estimate the parameters using the method described in Section 3.4. Since we use fixed integer valued ν in the skew-t and Student-t copula, we select the value with in the set $\{3, \ldots, 30\}$ based on the maximum value of the log-likelihood.

	Gamma marginals		Normal marginals	
Parameters	Est.	SE	Est.	SE
β_0	0.2533	0.1558	1.3204	0.4558
β_1	0.0959	0.0539	0.1264	0.1454
β_2	0.0025	0.0019	0.0011	0.0049
β_3	0.0114	0.0154	0.0201	0.0408
β_4	0.0113	0.0015	0.0273	0.0042
β_5	0.0907	0.0103	0.2022	0.0269
$V[b]$	0.0700	0.0258	1.2140	0.3390
κ	5.0979	1.9562	-	-
σ	-	-	0.8890	0.1394

Table 3.5 Marginal parameter estimation of HIV CD4$^+$ T cell count data with model (3.28) using Gamma and normal mixed model.

Gamma marginals					
Copula	degrees of freedom (ν)	3	4	5	6
Skew-t	Log-likelihood	**-1250.84**	-1261.45	-1272.17	-1281.84
Student-t	Log-likelihood	-1288.30	-1304.63	-1338.77	-1347.63
Normal marginals					
Copula	degrees of freedom (ν)	3	4	5	6
Skew-t	Log-likelihood	**-1256.88**	-1271.31	-1285.62	-1296.17
Student-t	Log-likelihood	-1257.02	-1271.50	-1285.89	-1297.65

Table 3.6 Estimation of the degrees of freedom parameter for the skew-t and Student-t copula based on the maximum log-likelihood.

The IFM estimation method yields identical marginal parameter estimates across all multivariate copula-based models. Thus, in Table 3.5, we consolidate the marginal parameter estimates and their corresponding standard errors for the mixed model (3.28) with Gamma and normal marginals, when the copula is Gaussian. In Table 3.6, we present the estimation of the fixed degrees of freedom parameter for the skew-t and Student-t copulas based on maximum log-likelihood. Lastly, Table 3.7 provides the dependence parameter estimates, along with their standard errors, log-likelihood (of the full model), AIC, and BIC, for the skew-t, skew-normal, Student-t, and Gaussian copulas, respectively. We derive the random effects from the posterior modes for each subject under different multivariate models, enabling graphical representation of the models when the unobserved random effects are estimated. Figure 3.4

Model	**Copula**	**Parameters**	Est.	SE	Log-likelihood	AIC	BIC
Gamma	Skew-t, $\nu = 3$	ξ	0.1781	0.0190	-1250.84	**2521.67**	**2557.32**
		$\bar{\lambda}$	1.2765	0.5373			
	Skew-normal	ξ	0.1904	0.0329	-1422.96	2863.92	2895.99
		$\bar{\lambda}$	1.8547	0.4033			
	Student-t, $\nu = 3$	ξ	0.2052	0.0256	-1288.30	2594.60	2626.68
	Gaussian	ξ	0.4525	0.0810	-1468.57	2953.14	2981.65
Normal	Skew-t, $\nu = 3$	ξ	0.2611	0.0285	-1256.88	2535.77	2594.98
		$\bar{\lambda}$	-0.0156	0.0650			
	Skew-normal	ξ	0.3084	0.0481	-1429.61	2879.21	2914.86
		$\bar{\lambda}$	-0.5016	0.0850			
	Student-t, $\nu = 3$	ξ	0.2612	0.0285	-1257.02	**2534.04**	**2569.68**
	Gaussian	ξ	0.5358	0.1113	-1480.53	2979.05	3011.13

Table 3.7 Dependence parameter estimation of HIV CD4$^+$ T cell count data with model (3.28). Maximum log-likelihood value, AIC and BIC for the skew-t, skew-normal, Student-t and Gaussian copula respectively.

showcases histograms of CD4 counts for 261 HIV patients, with dotted lines representing fitted models using two marginals and four multivariate copulas, respectively. Similarly, employing the estimated random effects, we transform the data into normal scores using the cumulative distribution functions of the marginals ($F(.)$). Figure 3.5 provides contour plots of the fitted skew and elliptical copulas to the data of the first two time points using two different marginal distributions.

Through this analysis, we observe that estimates of β_2 are close to zero under both marginal models, indicating that age has a minor impact on disease progression. Conversely, based on the estimate of β_1, gender exerts a significant effect on HIV progression, consistent with observations from the profile plots in 3.1. The significant effect of time as a covariate is evident from the estimate of β_5. Our analysis indicates that, skew-t ($\nu = 3$) copula-based Gamma mixed model appears to provide a fit which is better than the remaining seven candidate models. However, the Student-t ($\nu = 3$) copula based normal mixed model is marginally inferior. On the other hand, this model is computationally easier to work with. A smaller value of the degrees of freedom parameter implies stronger tail dependence across time in the data. Furthermore, estimates of the skewness parameter indicate reflection and permutation asymmetry when the marginals are assumed to be Gamma. However, with normal marginals, these estimates are lower and tend towards negative values. This discrepancy may arise from modeling marginals with symmetric distributions, which could obscure asymmetry in both the marginals and the dependence structure. Thus, when employing copula-based modeling, proper selection of marginals is paramount, balancing computational tractability with capturing underlying asymmetry. Overall, the considered class of copula-based mixed models demonstrates satisfactory performance, accurately estimating 13 coefficients for 261 observations.

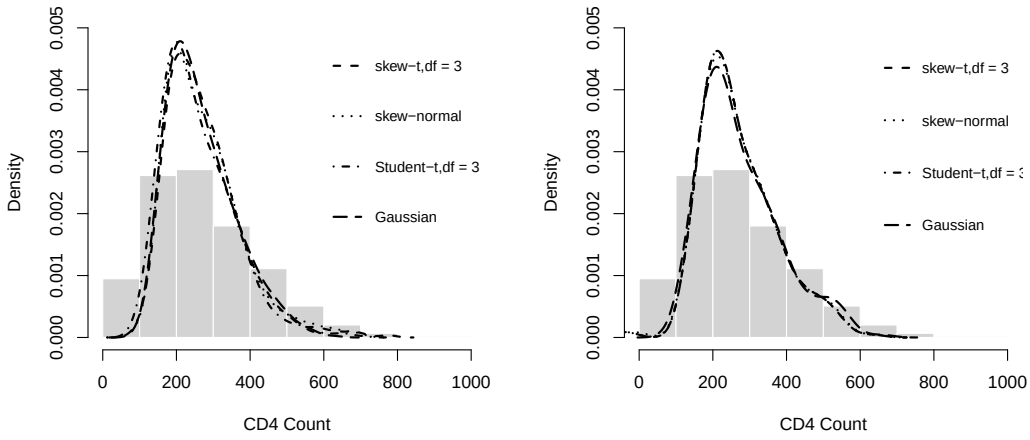

Figure 3.4 Fitting of HIV CD4$^+$ T cell count data with model (3.28) using Gamma marginals (left panel) and normal marginals (right panel). The histograms show the frequency distribution of observed CD4 counts with different dotted lines representing the fitted models.

3.10 Discussion

We have used skew-elliptical copulas, derived from multivariate skew-elliptical distributions, in this chapter. It is natural to try alternatives and explore use of them. One specific alternative is Archimedean copula. As we have asymmetry within our purview, need for use of asymmetric Archimedean copulas ([70], [71]) arise naturally. However, use of asymmetric Archimedean copulas in applied problems is relatively scarce. On the other hand, moving to a point where such copulas can be used in regression analysis, to begin with, and then in linear (mixed) models is a fairly long journey. This is certainly an area where research needs to be pursued. Another alternative may be to try and explore use of D-vine copulas. [28] (page. 144) has observed: "The context of an application can help in choosing a vine, if there are no obvious latent variables to explain the dependence. D-vines are more natural if there is a time or linear spatial order in variables." I submit that this is yet another area where research, both theoretical and applied, needs to be pursued. A few relevant references are [22], [29] and [72]. It has been noted in the introduction that the class of SNI distributions considered in [51] includes the skew-normal, skew-t, skew-slash ([54]), and skew-contaminated normal distributions as special cases. However, unlike the skew-normal and skew-t copulas which have been studied in the literature, skew-slash and skew-contaminated normal copulas do not seem to have been studied. These, therefore, need to be studied. It may also be noted in this context that the papers in [55] may lead to further research on copulas derived from multivariate skew-elliptical families and their applications. Notably, our framework encompasses the standard linear mixed model as a special case. In scenarios where the dataset deviates from normality, our proposed class of models appears to give better fit, effectively capturing reflection, permutation asymmetry, and tail dependence if present in the data. Leveraging

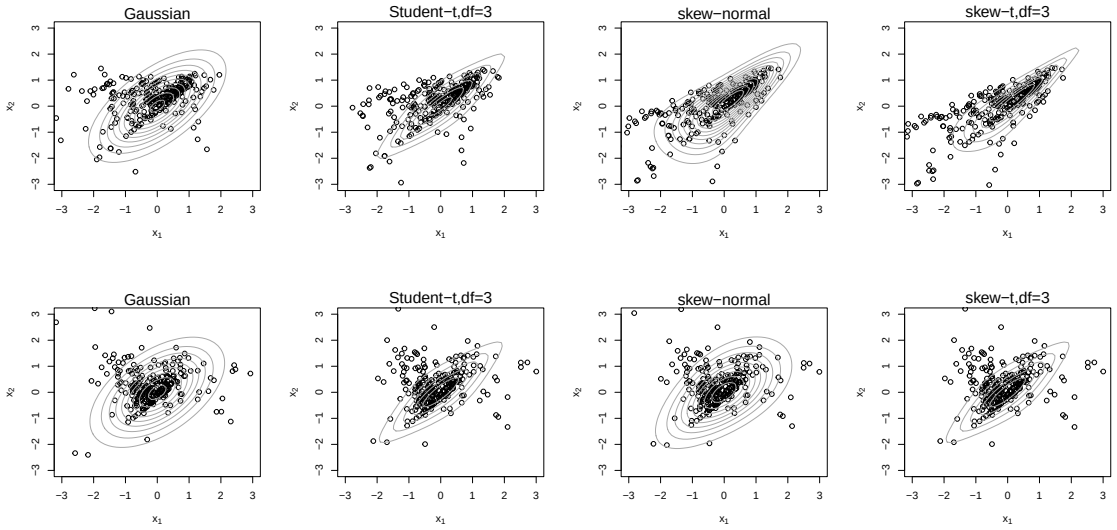

Figure 3.5 Fitting of the copula data (transformed to standard normal margins) using Gamma (upper panel) and normal mixed model (lower panel) of the first two time points, including the contour lines of the fitted skew and elliptical copulas, respectively.

these methods, we applied our framework to model disease progression using CD4 T^+ cell counts from the HIV dataset. The class of models we consider adopts a flexible structure in both the marginals and the dependence component, using skew-elliptical distributions to enable more intuitive interpretation of the dependence parameters. Through our analysis, we observed that the utilization of a skew-t copula-based Gamma mixed model yielded the best fit among the eight candidate models explored. This approach not only enhances our understanding of disease progression dynamics but also underscores the importance of flexible modeling techniques capable of accommodating non-normal data distributions and capturing nuanced dependencies present in real-world datasets.

We obtained the standard errors of the parameter estimates using the corresponding asymptotic covariance matrix, known as the Godambe information matrix, derived from IFM estimation. Our simulation study provides insights into the performance of IFM estimation in capturing model parameters across various combinations of multivariate copulas and marginal distributions. However, it's worth noting that even with IFM estimation, models based on skew-t and skew-normal copulas require significantly more computation time compared to those based on Student-t and Gaussian copulas. While we employ the Gauss-Hermite quadrature rule for numerical integration, the computational burden escalates exponentially with the dimension of the random effects. There are some serious limitations to the IFM procedure implemented in the considered class of models here, such as (a) efficiency loss under strong dependence ([36]), (b) sensitivity to the marginal misspecifications ([73]), (c) asymptotic validity relying on correctly specified margins ([74]), (d) finite sample bias, i.e. underestimating dependence in small/moderate samples ([28]), and challenges with discrete data. Looking ahead, future studies will, therefore, explore alternative estimation methods for our proposed class of models, aiming

for techniques that use some other standard routes applicable here. The following two methods may be tried: (1) canonical maximum likelihood method of estimation ([75]; [76]; [77]; [78]; [79]; [80]) and (2) composite likelihood method of estimation ([81]; [82]; [83]; [84]; [85]; [86]; [87]; [88]; [89]). A crucial assumption about the set-up in which CML is employed is absence of knowledge about marginals. However, in our set-up we are assuming the form of the marginals. So, we are sceptical about the usefulness of CML per se in the situation we are working in. Also, in a recent work ([80]), the authors have noted that estimation using canonical maximum likelihood in parametric copula models might be unstable and have explored robust alternatives. We submit that adequate investigation is needed before we recommend the use of CML in our setting. Some of the work related to composite likelihood method of estimation which have appeared in the past two decades and are relevant to our work include [84], [86], [88], [89]. The last two papers are related to copula-based models. In [88], the authors employed the IFM followed by what they called estimation by pseudo-likelihood based on pairwise margins. They considered copulas with analytically or numerically tractable pairwise margins, such as Archimedean, hierarchical Archimedean, Archimax and hierarchical Archimax copulas. In [89], the author considered the composite likelihood method for the estimation of high-dimensional multivariate normal copula models with discrete responses. He proposed a weighted version of the second method of [84] and an iterative approach to determine good weight matrices. Hence, if we are to use composite likelihood method estimation in our context, we may proceed as follows. First, we estimate the marginal parameters as in IFM described above. In the next stage, instead of considering the 'full likelihood', we have to look for low-dimensional margins which are either either analytically or numerically tractable. This is indeed a combination of IFM and composite likelihood method. We submit that this is a new line of research and that we plan to pursue this in future. Bayesian methods emerge as a promising avenue in this regard, enabling assessment of the copula's impact on the estimation of regression coefficients. Additionally, we aim to incorporate missing data mechanisms into our models, enhancing their generality and versatility. This advancement will enable us to handle real-world datasets more effectively, where missing data is often encountered.

Chapter 4

Factor copula models for non-Gaussian longitudinal data

In this chapter we introduce factor copula constructions for modeling temporal dependence in non-Gaussian longitudinal data. These models, based on canonical vine copulas, utilize latent variables to explain multivariate dependence structures, facilitating interpretation and implementation in unbalanced longitudinal data. In longitudinal studies the repeated measurements can be of different nature (discrete/continuous) depending upon the nature of investigation. Statistical modeling of continuous longitudinal data has been extensively explored, benefiting from a plethora of flexible multivariate distributions. However, when it comes to discrete cases, the available approaches are considerably more limited. The texts [2], [1], and [4] contain overviews and detailed discussion on longitudinal data analysis. Linear mixed models (LMM) are the most popular choice for analyzing continuous longitudinal data, relying on the multivariate normality assumption of repeated measurements [46]. However, various approaches have been proposed to relax distributional assumptions and accommodate non-Gaussian longitudinal data. In this context, the copula framework has been getting increasingly popular in recent times.

[7] presented one of the earlier applications of copulas to longitudinal data analysis. Subsequently, [90] proposed the use of multivariate elliptical copulas for modeling longitudinal data. Later, [91] explored Gaussian copulas for unbalanced longitudinal study designs. More recently, [16] utilized a flexible D-vine copula approach in conjunction with linear mixed models as marginals to model unbalanced continuous longitudinal data. Although D-vine copulas offer flexibility and naturally order repeated measurements, they can be computationally demanding in high dimensions due to numerous parameters involved. Additionally, [92] applied pair-copulas to model dependence in discrete data without covariates, and [93] employed pair-copulas for mixed data modeling. Extending beyond, [20] utilized D-vine copulas to construct multivariate distributions using power series marginal distributions, albeit limited to dimension $d = 3$. Factor copulas, introduced by [94], offer similar flexibility with fewer dependence parameters by explaining dependence through latent variables. This approach was extended by [95] to model discrete-valued item response data, and [96] implemented factor copula models in mixed response type social science data.

Factor copula models offer several advantages, particularly their applicability to response variables of any nature without numerical complications, even in moderate to high dimensions. These models,

characterized by truncated vine-copulas incorporating both observed and latent variables, afford a wide spectrum of asymmetric, tail, and nonlinear dependence patterns. Notably, [96] underscored the interpretability and superior fit of factor copula models compared to vine copula models, highlighting their closure under marginals. This property ensures that lower-order marginals belong to the same parametric family of copulas, and different permutations of observed variables yield identical distributions. In this chapter our primary contribution lies in introducing factor copulas to model temporal dependence in unbalanced non-Gaussian longitudinal data. We propose regression models for continuous, binary, and ordinal longitudinal data with covariates, employing factor copula constructions with subject-specific latent variables. For continuous responses, we employ generalized linear models (GLM) for the marginals, while for binary or ordinal responses, we link factor copulas with an underlying normally distributed latent variable. This framework enables the incorporation of subject-specific covariates, crucial components in longitudinal data analysis. We focus on univariate analysis of longitudinal responses of varying natures, assuming the missing data mechanism to be ignorable. Our models account for homogeneous within-subject dependence, enabling feasible parametric inference in moderate to high-dimensional scenarios using the IFM method (inference function of margins). Comparing factor copula models to random effect models like generalized linear mixed models (GLMM) is another contribution of this chapter, despite potential differences in interpretation under different circumstances. Our inspiration stems from the fact that both factor copula models and random effect models are based on non-latent variable frameworks. We have also showed how residual analysis of factor copula models can be conducted with Rosenblatt's transformation.

This chapter is organized as follows: in Section 4.1, we describe factor copula constructions for (i) continuous responses with generalized linear models and (ii) discrete responses with latent variables. In Section 4.2 we describe the estimation techniques and the computational details. Section 4.7 describes the competing random effect models and some commonly used model selection techniques. In Section 4.3, we present diagnostics of the residuals of fitted factor copula models. Section 4.4 presents extensive simulation studies demonstrating the finite sample performance of our proposed longitudinal models under some parsimonious specifications of factor copulas. We describe two motivating real world data sets in Section 4.5 and present our analysis. Empirical abalysis on these datasets reveals their effectiveness compared to widely used random effect models, suggesting factor copula models are promising alternatives for capturing temporal dependency in longitudinal data. Section 4.6 concludes this chapter and describes some future extensions of the proposed models.

4.1 The factor copula models

By the definition of longitudinal data, the repeated measurement tend to be serially correlated, hence for moderate to high-dimensional situations the use of factor copulas can be quite effective in terms of parametric inference. We assume that for the i-th subject, the responses $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\intercal$ are conditionally independent given p latent variables $V_{i1}, \ldots, V_{ip}$, which are independent and identically

distributed (i.i.d) as $U(0,1)$ ([94]). That is, a factor copula model is a conditional independence model which explains the dependence among the repeated measurements through some latent variables. The responses can be either continuous or discrete in nature. Here, we describe the construction of 1-factor and 2-factor copula models in the context of longitudinal data, with the possibility of extending these models to include up to p factors ($p \geq 3$).

4.1.1 Continuous responses

Consider the marginal density of a continuous response Y_{ij} at the j-th measurement to be $f(y_{ij}|\theta_{ij})$, where θ_{ij} is a parameter which may depend on a vector of covariates $\mathbf{x}_{ij}$. For a 1-factor copula model, let V_{i1} be an $U(0,1)$ variable. Here the suffix 'i' stands for a subject, and hence V_{i1} and V_{i^*1} are considered to be independent ($i \neq i^*$) in a longitudinal setting. The joint distribution of Y_{ij} and V_{i1} can be written in terms of a copula $C_{j,1}(.|\phi_{ij})$ as

$$P(Y_{ij} \leq y_{ij}, V_{i1} \leq v_{i1}) = C_{j,1}(F_{ij}(y_{ij}|\theta_{ij}), v_{i1}|\phi_{ij}), \tag{4.1}$$

where $F_{ij}(.|\theta_{ij})$ is the cumulative distribution function of Y_{ij} and ϕ_{ij} is the dependence parameter associated with the copula $C_{j,1}$. Therefore, we have the conditional distribution of Y_{ij} given V_{i1} as

$$F_{ij|1}(y_{ij}|v_{i1}, \theta_{ij}, \phi_{ij}) = \frac{\partial C_{j,1}(F_{ij}(y_{ij}|\theta_{ij}), v_{i1}|\phi_{ij})}{\partial v_{i1}} = C_{j|1}(F_{ij}(y_{ij}|\theta_{ij})|v_{i1}, \phi_{ij}). \tag{4.2}$$

Then the joint distribution of $\mathbf{Y}_i$ is obtained as

$$F_{n_i}(y_{i1}, \ldots, y_{in_i}|\theta_i^*) = \int_0^1 \prod_{j=1}^{n_i} F_{ij|1}(y_{ij}|v_{i1}, \theta_{ij}, \phi_{ij}) dv_{i1} = \int_0^1 \prod_{j=1}^{n_i} C_{j|1}(F_{ij}(y_{ij}|\theta_{ij})|v_{i1}, \phi_{ij}) dv_{i1} \tag{4.3}$$

with the corresponding density function -

$$f_{n_i}(y_{i1}, \ldots, y_{in_i}|\theta_i^*) = \int_0^1 \prod_{j=1}^{n_i} c_{j,1}(F_{ij}(y_{ij}|\theta_{ij}), v_{i1}|\phi_{ij}) dv_{i1} \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\theta_{ij}) \tag{4.4}$$

where $\theta_i^* = (\theta_i, \phi_i)^\intercal$ contains all the marginal and dependence parameters for the i-th subject and $c_{j,1}$ is the copula density function. Thus in a 1-factor copula model n_i bivariate copulas couple each observed variable to the first latent variable. For a 2-factor copula model we consider 2 latent variables V_{i1} and V_{i2}, i.i.d $\sim U(0,1)$. The joint distribution of Y_{ij} and V_{i2} given V_{i1}, can be written in terms of a copula $C_{j,2}(.|\delta_{ij})$ as

$$P(Y_{ij} \leq y_{ij}, V_{i2} \leq v_{i2}|V_{i1} = v_{i1}) = C_{j,2}(F_{ij|1}(y_{ij}|v_{i1}, \theta_{ij}, \phi_{ij}), v_{i2}|\delta_{ij}), \tag{4.5}$$

where δ_{ij} is the dependence parameter associated with the copula $C_{j,2}$. Similarly, the conditional distribution of Y_{ij} given V_{i1} and V_{i2} is given as

$$\begin{aligned} F_{ij|1,2}(y_{ij}|v_{i1}, v_{i2}, \theta_{ij}, \phi_{ij}, \delta_{ij}) &= \frac{\partial C_{j,2}(F_{ij|1}(y_{ij}|v_{i1}, \theta_{ij}, \phi_{ij}), v_{i2}|\delta_{ij})}{\partial v_{i2}} \\ &= C_{j|1,2}(F_{ij|1}(y_{ij}|v_{i1}, \theta_{ij}, \phi_{ij})|v_{i2}, \delta_{ij}). \end{aligned} \tag{4.6}$$

Therefore, the joint distribution function of $\mathbf{Y}_i$ is obtained as

$$\begin{aligned} F_{ni}(y_{i1}, \ldots, y_{in_i}|\theta_i^*) &= \int_0^1 \int_0^1 \prod_{j=1}^{n_i} F_{ij|1,2}(y_{ij}|v_{i1}, v_{i2}, \theta_{ij}, \phi_{ij}, \delta_{ij}) dv_{i1} dv_{i2} \\ &= \int_0^1 \int_0^1 \prod_{j=1}^{n_i} C_{j|1,2}(F_{ij|1}(y_{ij}|v_{i1}, \theta_{ij}, \phi_{ij})|v_{i2}, \delta_{ij}) dv_{i1} dv_{i2} \end{aligned} \tag{4.7}$$

and the density function, $f_{ni}(y_{i1}, \ldots, y_{in_i}|\theta_i^*) =$

$$\int_0^1 \int_0^1 \prod_{j=1}^{n_i} c_{j,2}(F_{ij|1}(y_{ij}|v_{i1}, \theta_{ij}, \phi_{ij}), v_{i2}|\delta_{ij}) c_{j,1}(F_{ij}(y_{ij}|\theta_{ij}), v_{i1}|\phi_{ij}) dv_{i1} dv_{i2} \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\theta_{ij}). \tag{4.8}$$

In a 2-factor copula model, there are another set of n_i bivariate copulas that link each observed variable to the second latent variable conditioned on the first factor.

4.1.2 Discrete responses

For discrete longitudinal responses (binary or ordinal), copulas are indirectly applied through some latent variables. Generally these variables are assumed to be normally distributed ([97]). Let Y_{ij} represent a categorical response with K possible ordered categories and let Z_{ij} be a normally distributed latent variable underneath Y_{ij}. Let $\gamma(k), 1 < k < K-1$, be ordered thresholds such that: $-\infty = \gamma(0) < \gamma(1) < \cdots < \gamma(K-1) < \gamma(K) = \infty$. Then the discrete response have the stochastic representation as

$$Y_{ij} = k \text{ if } \gamma(k-1) \leq Z_{ij} < \gamma(k), \quad k \in \{1, \ldots, K\}. \tag{4.9}$$

The threshold parameters can be fixed or freely estimated based on the specification of the model. Similar to the continuous case the proxy variables Z_{ij} may depend on a vector of covariates $\mathbf{x}_{ij}$. For 1-factor copula model the conditional distribution of Y_{ij} given V_{i1} is therefore

$$\begin{aligned} P(Y_{ij} = y_{ij}|V_{i1} = v_{i1}) &= P(Z_{ij} < \gamma(y_{ij})|V_{i1} = v_{i1}) - P(Z_{ij} < \gamma(y_{ij} - 1)|V_{i1} = v_{i1}) \\ &= F_{ij|1}(\gamma(y_{ij})|v_{i1}, \theta_{ij}, \phi_{ij}) - F_{ij|1}(\gamma(y_{ij} - 1)|v_{i1}, \theta_{ij}, \phi_{ij}) \\ &= C_{j|1}(F_{ij}(\gamma(y_{ij})|\theta_{ij})|v_{i1}, \phi_{ij}) - C_{j|1}(F_{ij}(\gamma(y_{ij} - 1)|\theta_{ij})|v_{i1}, \phi_{ij}). \end{aligned} \tag{4.10}$$

Hence the joint distribution (pmf) of $\mathbf{Y}_i$ is obtained as

$$\begin{aligned}
&f_{n_i}(y_{i1},\ldots,y_{in_i}|\theta_i^*)\\
&=\int_0^1\prod_{j=1}^{n_i}P(Y_{ij}=y_{ij}|V_{i1}=v_{i1})dv_{i1}\\
&=\int_0^1\prod_{j=1}^{n_i}\Big[C_{j|1}(F_{ij}(\gamma(y_{ij})|\theta_{ij})|v_{i1},\phi_{ij})-C_{j|1}(F_{ij}(\gamma(y_{ij}-1)|\theta_{ij})|v_{i1},\phi_{ij})\Big]dv_{i1}. \qquad (4.11)
\end{aligned}$$

For a 2-factor copula model the conditional distribution of Y_{ij} given V_{i1} and V_{i2} is given by:

$$\begin{aligned}
&P(Y_{ij}=y_{ij}|V_{i1}=v_{i1},V_{i2}=v_{i2})\\
&=P(Z_{ij}<\gamma(y_{ij})|V_{i1}=v_{i1},V_{i2}=v_{i2})-P(Z_{ij}<\gamma(y_{ij}-1)|V_{i1}=v_{i1},V_{i2}=v_{i2})\\
&=C_{j|1,2}(F_{ij|1}(\gamma(y_{ij})|v_{i1},\theta_{ij},\phi_{ij})|v_{i2},\delta_{ij})-C_{j|1,2}(F_{ij|1}(\gamma(y_{ij}-1)|v_{i1},\theta_{ij},\phi_{ij})|v_{i2},\delta_{ij}). \qquad (4.12)
\end{aligned}$$

Hence the joint distribution (pmf) of $\mathbf{Y}_i$ is obtained as

$$\begin{aligned}
f_{n_i}(y_{i1},\ldots,y_{in_i}|\theta_i^*)&=\int_0^1\int_0^1\prod_{j=1}^{n_i}P(Y_{ij}=y_{ij}|V_{i1}=v_{i1},V_{i2}=v_{i2})dv_{i1}dv_{i2}\\
&=\int_0^1\int_0^1\prod_{j=1}^{n_i}\Big[C_{j|1,2}(F_{ij|1}(\gamma(y_{ij})|v_{i1},\theta_{ij},\phi_{ij})|v_{i2},\delta_{ij})\\
&\qquad-C_{j|1,2}(F_{ij|1}(\gamma(y_{ij}-1)|v_{i1},\theta_{ij},\phi_{ij})|v_{i2},\delta_{ij})\Big]dv_{i1}dv_{i2}. \qquad (4.13)
\end{aligned}$$

Here we are making the so called simplifying assumption that the conditional copula for the univariate distributions $F_{ij|1}$ and v_{i2} does not depend on v_{i1}. The choice of bivariate linking copulas in 1-factor and 2-factor copula models is completely arbitrary, but in this chapter we restrict to the elliptical copulas in the sense that the dependence parameter have direct interpretation same as their bivariate distributions. We consider the bivariate Gaussian copula as

$$C(u_1,u_2|\rho)=\Phi_2(\Phi^{-1}(u_1),\Phi^{-1}(u_2)|\rho),\quad -1\le\rho\le 1, \qquad (4.14)$$

where Φ_2 is the bivariate standard normal cdf with correlation parameter ρ, and Φ^{-1} is the quantile of the univariate standard normal. [94] and [95] showed that if all bivariate copulas are normal with all normal marginals then factor copula models are multivariate normal and normal ogive models for the continuous and discrete case, respectively. We also consider the bivariate Student-t copula as

$$C(u_1,u_2|\rho,\nu)=T_2(T^{-1}(u_1|\nu),T^{-1}(u_2|\nu)|\rho,\nu),\quad -1\le\rho\le 1, \qquad (4.15)$$

where T^{-1} is the quantile of the univariate Student-t with ν degrees of freedom, and T_2 is the cdf of a bivariate Student-t distribution with ν degrees of freedom and correlation parameter ρ. This additional

degrees of freedom parameter ν, accounts for possible tail dependence in the data. Small value of ν, such as $2 \leq 5$, leads to a model with more probabilities in the joint upper and joint lower tails compared with the normal copula.

The final aspect in specifying factor copula-based models involves determining the marginals. In accordance with a regression framework, we adopt generalized linear models (GLMs) for the marginals in the case of continuous responses ([47]). Let $\mathbf{x}_{ij}$ denote a p-dimensional vector of covariates, β represent a $p \times 1$ vector of regression coefficients, and $g(.)$ signify a model-specific known link function. Thus, a generalized linear model can be formulated as follows -

$$g(E(Y_{ij})) = \mathbf{x}_{ij}\beta, \quad j = 1, \ldots, n_i. \tag{4.16}$$

For the discrete responses we model the latent variables based on covariate vector $\mathbf{x}_{ij}$ as

$$Z_{ij} = \mathbf{x}_{ij}\beta + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \tag{4.17}$$

where the error term $\epsilon_{ij}(i.i.d) \sim N(0, 1)$. To ensure identifiaility of the ordinal model, we further fix the intercept parameter of β equal to zero.

4.2 Parameter estimation

Two widely employed frequentist techniques for parameter estimation in copula-based models are maximum likelihood estimation (MLE) and inference function of margins (IFM) ([35]; [36]). MLE is acknowledged for its efficiency but can be computationally intensive, particularly for complex dependence structures like factor copula models. In this context, we opt for the IFM method, wherein model parameters are estimated in two distinct stages. Initially, we estimate all marginal parameters under the assumption of independence by maximizing a pseudo-likelihood function of the form -

$$l(\theta|\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log f_{ij}(y_{ij}|\theta_{ij}), \tag{4.18}$$

and then using the parameter estimates from (4.18), we compute the uniform samples: $u_{ij} = F_{ij}(y_{ij}|\hat{\theta}_{ij})$ for the continuous case, and $u_{ij} = F_{ij}(\gamma(y_{ij})|\hat{\theta}_{ij})$, $u_{ij}^{-} = F_{ij}(\gamma(y_{ij} - 1)|\hat{\theta}_{ij})$ for the discrete case, respectively where $i = 1, \ldots, m, j = 1, \ldots, n_i$. Thereafter we estimate the parameters of 1-factor copula models for the continuous case by maximizing the pseudo likelihood -

$$l(\phi|\mathbf{u}) = \sum_{i=1}^{m} \log \int_0^1 \prod_{j=1}^{n_i} c_{j,1}(u_{ij}, v_{i1}|\phi_{ij}) dv_{i1}, \tag{4.19}$$

and for the discrete case by maximizing

$$l(\phi|\mathbf{u}) = \sum_{i=1}^{m} \log \int_0^1 \prod_{j=1}^{n_i} \Big[C_{j|1}(u_{ij}|v_{i1}, \phi_{ij}) - C_{j|1}(u_{ij}^-|v_{i1}, \phi_{ij}) \Big] dv_{i1}. \tag{4.20}$$

Based on the parametric copula family we can estimate the conditional copulas using the corresponding h-functions. For the bivariate Gaussian copula this is

$$C_{j|1}(u_1|u_2, \rho) = \Phi \left(\frac{\Phi^{-1}(u_1) - \rho \Phi^{-1}(u_2)}{\sqrt{1-\rho^2}} \right), \tag{4.21}$$

and for the bivariate Student-t copula we compute

$$C_{j|1}(u_1|u_2, \rho, \nu) = T \left(\frac{T^{-1}(u_1|\nu) - \rho T^{-1}(u_2|\nu)}{\sqrt{\frac{(\nu + T^{-1}(u_2|\nu))^2 (1-\rho^2)}{\nu+1}}} \middle| \nu + 1 \right). \tag{4.22}$$

Similarly in the 2-factor copula models for the continuous case we estimate the dependence parameters by maximizing

$$l(\phi, \delta|\mathbf{u}) = \sum_{i=1}^{m} \log \int_0^1 \int_0^1 \prod_{j=1}^{n_i} c_{j,2}(C_{j|1}(u_{ij}|v_{i1}, \phi_{ij}), v_{i2}|\delta_{ij}) c_{j,1}(u_{ij}, v_{i1}|\phi_{ij}) dv_{i1} dv_{i2}, \tag{4.23}$$

and for the discrete case by maximizing

$$\begin{aligned} l(\phi, \delta|\mathbf{u}) = \sum_{i=1}^{m} \log \int_0^1 \int_0^1 \prod_{j=1}^{n_i} \Big[& C_{j|1,2}(C_{j|1}(u_{ij}|v_{i1}, \phi_{ij})|v_{i2}, \delta_{ij}) \\ & - C_{j|1,2}(C_{j|1}(u_{ij}^-|v_{i1}, \phi_{ij})|v_{i2}, \delta_{ij}) \Big] dv_{i1} dv_{i2}. \end{aligned} \tag{4.24}$$

To perform the numerical integrations in (4.19), (4.20), (4.23), and (4.24), we employ the Gauss-Hermite quadrature rule with 15 quadrature points. The standard errors of the parameter estimates $\hat{\theta}^* = (\hat{\theta}, \hat{\phi})^\intercal$ for the 1-factor and 2-factor copula models can be obtained numerically from the estimated sandwich information matrix (also known as the Godambe information matrix), expressed as -

$$J(\hat{\theta}^*) = D(\hat{\theta}^*)^\intercal M(\hat{\theta}^*)^{-1} D(\hat{\theta}^*) \tag{4.25}$$

where $D(\hat{\theta}^*)$ represents a block diagonal matrix and $M(\hat{\theta}^*)$ denotes a symmetric positive definite matrix. Further details on the estimation of this matrix can be found in [35] or [28]. For parameter estimation, we utilize the *optim* function ([64]), and for estimating the information matrix associated with the parameter estimates, we employ the *numderiv* function ([65]) in R.

4.3 Residual analysis

Here we show Rosenblatt's transformation ([98]) is readily applicable for factor copula models. [12] or [99] previously used this method to validate copula assumptions for multivariate models. This method transforms dependent random variables into independent uniform random variables in the unit interval. Consistent with the setting considered in Section 4.1 let

$$\begin{aligned} w_{i1} &= F(y_{i1}) \\ &\dots \\ w_{in_i} &= F(y_{in_i}|y_{i(n_i-1)}, \dots, y_{i1}, \hat{\theta}^*), \end{aligned} \tag{4.26}$$

which are realizations of n_i uncorrelated uniform variables if the model is correctly specified. For the continuous case under 1-factor copula model we have

$$\begin{aligned} F(y_{ij}|y_{i(j-1)}, \dots, y_{i1}, \hat{\theta}^*) &= \int_0^1 F(y_{ij}|y_{i(j-1)}, \dots, y_{i1}, v_{i1}, \hat{\theta}^*) dv_{i1} \\ &= \int_0^1 F(y_{ij}|v_{i1}, \hat{\theta}^*) dv_{i1} \\ &= \int_0^1 C_{j|1}(\hat{u}_{ij}|v_{i1}, \hat{\phi}_{ij}) dv_{i1}, \end{aligned} \tag{4.27}$$

where $\hat{u}_{ij} = F_{ij}(y_{ij}|\hat{\theta}_{ij})$. Similarly under 2-factor copula model we have

$$\begin{aligned} F(y_{ij}|y_{i(j-1)}, \dots, y_{i1}, \hat{\theta}^*) &= \int_0^1 \int_0^1 F(y_{ij}|y_{i(j-1)}, \dots, y_{i1}, v_{i1}, v_{i2}, \hat{\theta}^*) dv_{i1} dv_{i2} \\ &= \int_0^1 \int_0^1 F(y_{ij}|, v_{i1}, v_{i2}, \hat{\theta}^*) dv_{i1} dv_{i2} \qquad (4.28) \\ &= \int_0^1 \int_0^1 C_{j|1,2}(C_{j|1}(\hat{u}_{ij}|v_{i1}, \hat{\phi}_{ij})|v_{i2}, \hat{\delta}_{ij}) dv_{i1} dv_{i2}. \qquad (4.29) \end{aligned}$$

For the discrete case following [100] or [12], we compute the pseudo residuals by $w^*_{ij} = (w_{ij} + w^-_{ij})/2$ where $w^-_{ij} = F(y^-_{ij}|y_{i(j-1)}, \dots, y_{i1}, \hat{\theta}^*)$ for $j = 1, \dots, n_i$. Therefore, quantiles of the residuals can be plotted against their expected values to graphically visualize the goodness-of-fit of the model.

4.4 Simulation studies

To assess the performance of the estimation methods for our proposed class of models, we conduct simulation studies that replicate the characteristics of the datasets under consideration. We generate datasets from the proposed factor copula-based models and track the parametric inference. We consider two different sample sizes, $m = 200$ and 500, and set the maximum number of longitudinal responses to $d = 10$. To emulate the unbalanced nature of the data, we generate, for each unit,

$n_i \sim$ Tbinomial$(d, p = 0.8)$, for $i = 1, \ldots, m$ (truncated binomial excluding 0), effectively pruning the dataset. Since the true parameters are known, we first sample $\mathbf{U}_i$ from n_i-variate ($n_i \leq d$) 1-factor and 2-factor copulas for $i = 1, \ldots, m$. Subsequently, we employ Probability Integral Transform (PIT) on $\mathbf{U}_i$ to generate per-unit responses $\mathbf{Y}_i$. As the dimension for each response differs, we assume the same bivariate copulas for each factor with exchangeable parameters, resulting in a reduction in the number of estimable parameters from the model. We generate continuous responses using the following model -

$$g(E(Y_{ij})) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + t_{ij}\beta_3, \quad j = 1, \ldots, n_i, \tag{4.30}$$

where we consider two response distributions as Gamma (log-link) and normal (identity-link), respectively (Gamma distribution has asymmetry and positive real support). We assign same values of the regression coefficients for both of this marginals as, $\beta_0 = 1.0, \beta_1 = -0.5, \beta_2 = 0.2, \beta_3 = 0.2$ and set the dispersion parameters to $\kappa = 0.3$ (shape parameter of Gamma) and $\phi = 1.0$ (dispersion parameter of normal). Using the same dependence structure, we generate the binary responses from the following model -

$$\begin{aligned} Y_{ij} &= \begin{cases} 0 & \text{if } Z_{ij} \leq 0 \\ 1 & \text{if } Z_{ij} > 0 \end{cases}, \\ Z_{ij} &= \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + t_{ij}\beta_3 + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \end{aligned} \tag{4.31}$$

where $\epsilon_{ij}(i.i.d) \sim N(0, 1)$. Here we assign the regression coefficients as, $\beta_0 = -0.5, \beta_1 = -0.5, \beta_2 = 0.2, \beta_3 = 0.2$. Finally we generate the ordinal responses from the following model -

$$\begin{aligned} Y_{ij} &= k \text{ if } \gamma(k-1) \leq Z_{ij} < \gamma(k), \quad k = 1, \ldots, 4, \\ Z_{ij} &= x_{i1}\beta_1 + x_{i2}\beta_2 + t_{ij}\beta_3 + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \end{aligned} \tag{4.32}$$

where $\epsilon_{ij}(i.i.d) \sim N(0, 1)$. Here we assign the following regression coefficients: $\beta_1 = -0.5, \beta_2 = 0.2$, and $\beta_3 = 0.2$, along with threshold parameters $\gamma_1 = -1.0$, $\gamma_2 = 1.0$, and $\gamma_3 = 3.0$, respectively. The fixed covariates are generated as follows: $x_{i1} \sim$ Bernoulli$(p = 0.5)$, $x_{i2} \sim$ Uniform$(3, 8)$, and the time points $t_{ij} = j$, for $j = 1, \ldots, n_i$, $i = 1, \ldots, m$, respectively. For the 1-factor copula models, we set the correlation parameter $\rho_1 = 0.5$, and for the 2-factor copula models, we set the two correlation parameters as $\{\rho_1, \rho_2\} = \{0.5, 0.5\}$. Additionally, we set the degrees of freedom parameter $\nu = 4.0$ when the bivariate copulas are assumed to be Student-t. We simulate $N = 500$ Monte Carlo datasets for each model to monitor the performance under IFM estimation.

In Table 4.1, 4.2, 4.3, and 4.4, we provide comprehensive summaries of the parameter estimation results. These summaries include the averages of the parameter estimates (denoted as Mean), biases $[\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_j^* - \theta^*)]$, empirical standard deviations (denoted as SD), average standard errors obtained from the asymptotic covariance matrices (denoted as SE), and roots of mean square errors $[\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_j^* - \theta^*)^2}]$, where $\hat{\theta}_j^*$ represents the parameter estimates for the j-th sample. The results

			m = 200					**m** = 500				
Model	**Parameters**	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
Gamma	β_0	1.0	0.9895	-0.0105	0.0952	0.0985	0.0957	0.9875	-0.0125	0.0648	0.0624	0.0686
	β_1	-0.5	-0.5000	0.0000	0.0480	0.0473	0.0480	-0.4964	0.0036	0.0314	0.0300	0.0316
	β_2	0.2	0.2014	0.0014	0.0155	0.0164	0.0156	0.2031	0.0031	0.0105	0.0103	0.0110
	β_3	0.2	0.2006	0.0006	0.0053	0.0053	0.0053	0.2006	0.0006	0.0035	0.0034	0.0035
	κ	3.0	3.0160	0.0160	0.1228	0.1244	0.1238	3.0158	0.0158	0.0731	0.0784	0.0748
	ρ_1	0.5	0.4959	-0.0042	0.0274	0.0225	0.0277	0.4959	-0.0041	0.0169	0.0142	0.0174
Normal	β_0	1.0	1.0152	0.0152	0.1739	0.1730	0.1745	1.0034	0.0034	0.1130	0.1099	0.1131
	β_1	-0.5	-0.5034	-0.0034	0.0841	0.0830	0.0842	-0.4995	0.0005	0.0528	0.0525	0.0528
	β_2	0.2	0.1978	-0.0022	0.0292	0.0287	0.0293	0.1995	-0.0005	0.0190	0.0182	0.0190
	β_3	0.2	0.1998	-0.0002	0.0091	0.0091	0.0091	0.1999	-0.0001	0.0056	0.0058	0.0056
	ϕ	1.0	0.9969	-0.0032	0.0209	0.0208	0.0211	0.9979	-0.0021	0.0132	0.0132	0.0133
	ρ_1	0.5	0.4958	-0.0042	0.0278	0.0226	0.0281	0.4974	-0.0026	0.0169	0.0143	0.0171
Binary	β_0	-0.5	-0.5114	-0.0114	0.2390	0.2318	0.2392	-0.5008	-0.0008	0.1420	0.1462	0.1420
	β_1	-0.5	-0.4968	0.0032	0.1063	0.1119	0.1064	-0.5065	-0.0065	0.0766	0.0708	0.0768
	β_2	0.2	0.2025	0.0025	0.0406	0.0391	0.0406	0.2016	0.0016	0.0234	0.0246	0.0235
	β_3	0.2	0.2011	0.0011	0.0185	0.0185	0.0186	0.2003	0.0003	0.0123	0.0117	0.0123
	ρ_1	0.5	0.4913	-0.0087	0.0536	0.0518	0.0543	0.4949	-0.0051	0.0336	0.0326	0.0340
Ordinal	β_1	-0.5	-0.5022	-0.0020	0.0913	0.0889	0.0914	-0.4976	0.0024	0.0597	0.0563	0.0598
	β_2	0.2	0.1990	-0.0010	0.0312	0.0310	0.0313	0.1998	-0.0002	0.0198	0.0196	0.0198
	β_3	0.2	0.2004	0.0004	0.0117	0.0120	0.0117	0.2002	0.0002	0.0073	0.0076	0.0074
	γ_1	-1.0	-1.0258	-0.0258	0.2197	0.2089	0.2212	-1.0014	-0.0014	0.1300	0.1318	0.1300
	γ_2	1.0	0.9919	-0.0081	0.1907	0.1970	0.1908	1.0027	0.0027	0.1162	0.1097	0.1163
	γ_3	3.0	2.9964	-0.0036	0.2024	0.2124	0.2025	3.0067	0.0067	0.1228	0.1210	0.1230
	ρ_1	0.5	0.4935	-0.0065	0.0337	0.0318	0.0343	0.4977	-0.0023	0.0200	0.0198	0.0202

Table 4.1 Parameter estimation using IFM method for Gaussian 1-factor copula model with continuous and discrete marginals for $N = 500$ simulated data sets.

			m = 200					**m** = 500				
Model	**Parameters**	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
Gamma	β_0	1.0	0.9971	-0.0029	0.1246	0.1182	0.1246	0.9935	-0.0065	0.0757	0.0752	0.0775
	β_1	-0.5	-0.5022	-0.0022	0.0604	0.0572	0.0605	-0.4989	0.0011	0.0359	0.0365	0.0359
	β_2	0.2	0.2000	0.0000	0.0208	0.0199	0.0208	0.2003	0.0003	0.0125	0.0126	0.0127
	β_3	0.2	0.2001	0.0001	0.0050	0.0049	0.0050	0.2004	0.0004	0.0031	0.0031	0.0031
	κ	3.0	3.0410	0.0410	0.1553	0.1576	0.1606	3.0178	0.0178	0.0987	0.1008	0.1003
	ρ_1	0.5	0.5192	0.0192	0.0630	0.0651	0.0659	0.5182	0.0182	0.0508	0.0495	0.0540
	ρ_2	0.5	0.4432	-0.0568	0.1129	0.1091	0.1146	0.4611	-0.0389	0.0822	0.0775	0.0842
Normal	β_0	1.0	1.0133	0.0133	0.2111	0.2078	0.2111	1.0050	0.0050	0.1333	0.1321	0.1334
	β_1	-0.5	-0.5099	-0.0099	0.1000	0.1010	0.1005	-0.5008	-0.0008	0.0644	0.0642	0.0644
	β_2	0.2	0.1980	-0.0020	0.0353	0.0349	0.0354	0.1995	-0.0005	0.0221	0.0222	0.0221
	β_3	0.2	0.2002	0.0002	0.0087	0.0085	0.0087	0.1997	-0.0003	0.0052	0.0054	0.0052
	ϕ	1.0	0.9948	-0.0052	0.0280	0.0264	0.0284	0.9985	-0.0015	0.0170	0.0171	0.0170
	ρ_1	0.5	0.5206	0.0206	0.0625	0.0575	0.0658	0.5164	0.0164	0.0555	0.0512	0.0604
	ρ_2	0.5	0.4472	-0.0528	0.1031	0.1003	0.1051	0.4612	-0.0388	0.0889	0.0803	0.0927
Binary	β_0	-0.5	-0.5038	-0.0038	0.2757	0.2649	0.2758	-0.4972	0.0028	0.1778	0.1699	0.1780
	β_1	-0.5	-0.4971	0.0029	0.1341	0.1310	0.1342	-0.5000	0.0000	0.0890	0.0833	0.0890
	β_2	0.2	0.2005	0.0005	0.0471	0.0455	0.0471	0.1996	-0.0004	0.0291	0.0291	0.0291
	β_3	0.2	0.2012	0.0012	0.0186	0.0184	0.0187	0.2002	0.0002	0.0113	0.0118	0.0113
	ρ_1	0.5	0.5176	0.0176	0.0756	0.0698	0.0778	0.5070	0.0070	0.0472	0.0443	0.0477
	ρ_2	0.5	0.4325	-0.0675	0.1510	0.1501	0.1520	0.4795	-0.0206	0.0848	0.0821	0.0851
Ordinal	β_1	-0.5	-0.5023	-0.0023	0.1034	0.1065	0.1035	-0.5010	-0.0010	0.0685	0.0675	0.0685
	β_2	0.2	0.2009	0.0009	0.0384	0.0370	0.0384	0.2018	0.0018	0.0233	0.0234	0.0234
	β_3	0.2	0.2018	0.0018	0.0119	0.0120	0.0120	0.2012	0.0012	0.0075	0.0076	0.0076
	γ_1	-1.0	-1.0152	-0.0152	0.2490	0.2397	0.2495	-0.9993	0.0007	0.1547	0.1532	0.1547
	γ_2	1.0	1.0127	0.0127	0.2355	0.2208	0.2358	1.0114	0.0114	0.1384	0.1403	0.1388
	γ_3	3.0	3.0241	0.0241	0.2501	0.2376	0.2513	3.0167	0.0167	0.1498	0.1512	0.1507
	ρ_1	0.5	0.4975	-0.0028	0.0310	0.0298	0.0310	0.5036	0.0036	0.0213	0.0210	0.0226
	ρ_2	0.5	0.4917	-0.0083	0.0387	0.0308	0.0388	0.4977	-0.0023	0.0282	0.0273	0.0283

Table 4.2 Parameter estimation using IFM method for Gaussian 2-factor copula model with continuous and discrete marginals for $N = 500$ simulated data sets.

			m = 200					**m** = 500				
Model	**Parameters**	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
Gamma	β_0	1.0	0.9920	-0.0080	0.1064	0.0986	0.1067	0.9941	-0.0059	0.0691	0.0626	0.0695
	β_1	-0.5	-0.5012	-0.0012	0.0451	0.0472	0.0452	-0.4981	0.0019	0.0296	0.0300	0.0296
	β_2	0.2	0.2014	0.0014	0.0172	0.0163	0.0173	0.2021	0.0021	0.0113	0.0104	0.0115
	β_3	0.2	0.1990	-0.0010	0.0055	0.0052	0.0055	0.2003	0.0003	0.0035	0.0034	0.0035
	κ	3.0	3.0293	0.0293	0.1486	0.1545	0.1514	3.0163	0.0163	0.0960	0.0985	0.0973
	ρ_1	0.5	0.4877	-0.0123	0.0293	0.0233	0.0317	0.4905	-0.0095	0.0185	0.0167	0.0208
Normal	β_0	1.0	1.0058	0.0058	0.1775	0.1713	0.1776	1.0045	0.0045	0.1066	0.1097	0.1067
	β_1	-0.5	-0.5067	-0.0067	0.0846	0.0820	0.0848	-0.5020	-0.0020	0.0531	0.0528	0.0531
	β_2	0.2	0.1986	-0.0014	0.0303	0.0284	0.0303	0.1991	-0.0009	0.0179	0.0181	0.0179
	β_3	0.2	0.1995	-0.0005	0.0094	0.0092	0.0094	0.2002	0.0002	0.0056	0.0058	0.0056
	ϕ	1.0	0.9964	-0.0036	0.0268	0.0267	0.0270	0.9998	-0.0002	0.0181	0.0171	0.0181
	ρ_1	0.5	0.4884	-0.0116	0.0295	0.0233	0.0317	0.4926	-0.0074	0.0183	0.0146	0.0197
Binary	β_0	-0.5	-0.5283	-0.0283	0.2423	0.2387	0.2439	-0.5035	-0.0035	0.1452	0.1515	0.1453
	β_1	-0.5	-0.5074	-0.0074	0.1242	0.1161	0.1244	-0.5074	-0.0074	0.0721	0.0738	0.0725
	β_2	0.2	0.2060	0.0060	0.0406	0.0405	0.0407	0.2009	0.0009	0.0252	0.0256	0.0252
	β_3	0.2	0.2028	0.0028	0.0192	0.0191	0.0194	0.2014	0.0014	0.0125	0.0122	0.0125
	ρ_1	0.5	0.4846	-0.0155	0.0638	0.0566	0.0657	0.4953	-0.0047	0.0363	0.0351	0.0366
Ordinal	β_1	-0.5	-0.5036	-0.0036	0.0899	0.0887	0.0900	-0.4987	0.0013	0.0557	0.0562	0.0557
	β_2	0.2	0.2025	0.0025	0.0318	0.0308	0.0319	0.2002	0.0002	0.0187	0.0196	0.0187
	β_3	0.2	0.2009	0.0009	0.0126	0.0125	0.0127	0.2005	0.0005	0.0077	0.0079	0.0077
	γ_1	-1.0	-1.0128	-0.0128	0.2257	0.2142	0.2260	-1.0035	-0.0035	0.1303	0.1352	0.1304
	γ_2	1.0	1.0100	0.0100	0.1907	0.1866	0.1909	1.0019	0.0019	0.1146	0.1084	0.1146
	γ_3	3.0	3.0193	0.0193	0.2099	0.2058	0.2108	3.0043	0.0043	0.1251	0.1211	0.1252
	ρ_1	0.5	0.4943	-0.0057	0.0339	0.0313	0.0344	0.4985	-0.0015	0.0214	0.0196	0.0215

Table 4.3 Parameter estimation using IFM method for Student-t ($\nu = 4$) 1-factor copula model with continuous and discrete marginals for $N = 500$ simulated data sets.

			m = 200					**m** = 500				
Model	**Parameters**	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
Gamma	β_0	1.0	0.9828	-0.0172	0.1227	0.1188	0.1239	0.9852	-0.0148	0.0761	0.0753	0.0776
	β_1	-0.5	-0.4988	0.0012	0.0609	0.0573	0.0609	-0.4999	0.0001	0.0361	0.0364	0.0361
	β_2	0.2	0.2024	0.0024	0.0202	0.0199	0.0203	0.2021	0.0021	0.0127	0.0126	0.0129
	β_3	0.2	0.2003	0.0003	0.0050	0.0049	0.0050	0.2006	0.0006	0.0032	0.0031	0.0032
	κ	3.0	3.0418	0.0418	0.1835	0.1894	0.1882	3.0175	0.0175	0.1179	0.1211	0.1192
	ρ_1	0.5	0.4964	-0.0036	0.0887	0.0737	0.0887	0.5033	0.0033	0.0651	0.0668	0.0652
	ρ_2	0.5	0.4400	-0.0600	0.0897	0.0778	0.0898	0.4680	-0.0320	0.0626	0.0660	0.0630
Normal	β_0	1.0	1.0226	0.0226	0.1970	0.2072	0.1983	0.9981	-0.0019	0.1368	0.1316	0.1368
	β_1	-0.5	-0.5057	-0.0057	0.1046	0.1000	0.1048	-0.5015	-0.0015	0.0642	0.0636	0.0643
	β_2	0.2	0.1970	-0.0030	0.0332	0.0347	0.0333	0.2006	0.0006	0.0235	0.0220	0.0235
	β_3	0.2	0.1997	0.0003	0.0085	0.0085	0.0085	0.1997	-0.0003	0.0054	0.0054	0.0054
	ϕ	1.0	0.9947	-0.0053	0.0336	0.0324	0.0340	0.9991	-0.0009	0.0209	0.0208	0.0209
	ρ_1	0.5	0.4973	-0.0027	0.0854	0.0733	0.0854	0.5108	0.0108	0.0689	0.0521	0.0698
	ρ_2	0.5	0.4502	-0.0498	0.0897	0.0763	0.0898	0.4714	-0.0286	0.0620	0.0544	0.0629
Binary	β_0	-0.5	-0.5050	-0.0050	0.2788	0.2756	0.2789	-0.4989	0.0011	0.1709	0.1729	0.1711
	β_1	-0.5	-0.5121	-0.0121	0.1500	0.1367	0.1504	-0.5094	-0.0094	0.0867	0.0864	0.0872
	β_2	0.2	0.2027	0.0027	0.0491	0.0477	0.0491	0.1990	0.0010	0.0296	0.0302	0.0297
	β_3	0.2	0.2019	0.0019	0.0204	0.0193	0.0205	0.2005	0.0005	0.0127	0.0125	0.0127
	ρ_1	0.5	0.4963	-0.0037	0.0807	0.0791	0.0808	0.4983	-0.0012	0.0603	0.0601	0.0604
	ρ_2	0.5	0.4230	-0.0770	0.1599	0.1529	0.1617	0.4347	-0.0653	0.1245	0.1311	0.1310
Ordinal	β_1	-0.5	-0.5074	-0.0074	0.1078	0.1057	0.1080	-0.5044	-0.0044	0.0691	0.0671	0.0693
	β_2	0.2	0.2008	0.0008	0.0355	0.0367	0.0355	0.2010	0.0010	0.0227	0.0234	0.0227
	β_3	0.2	0.2009	0.0009	0.0122	0.0126	0.0122	0.2004	0.0004	0.0082	0.0080	0.0082
	γ_1	-1.0	-1.0433	-0.0433	0.2554	0.2490	0.2590	-1.0097	-0.0097	0.1574	0.1577	0.1577
	γ_2	1.0	0.9943	-0.0057	0.2163	0.2205	0.2164	1.0020	0.0020	0.1418	0.1398	0.1418
	γ_3	3.0	3.0063	0.0063	0.2385	0.2427	0.2356	3.0045	0.0045	0.1559	0.1543	0.1560
	ρ_1	0.5	0.4912	-0.0088	0.0463	0.0419	0.0463	0.4925	-0.0075	0.0216	0.0199	0.0219
	ρ_2	0.5	0.4831	-0.0169	0.0375	0.0298	0.0377	0.4955	-0.0045	0.0265	0.0249	0.0267

Table 4.4 Parameter estimation using IFM method for Student-t ($\nu = 4$) 2-factor copula model with continuous and discrete marginals for $N = 500$ simulated data sets.

reveal consistent performance of the proposed models with IFM estimation, as evidenced by decreasing biases and roots of mean square errors with increasing sample size. Notably, the average standard errors closely align with the empirical standard deviations for almost all parameters across all models, underscoring the reliability of inference. It's worth mentioning that the standard errors of the parameter estimates for the 2-factor copula based models are slightly larger than those of the 1-factor copula based models. This observation suggests that binary responses yield the least amount of information for parameter estimation, as indicated by the largest standard errors of the regression coefficients among other response models. This may be due to the fact that they represent a crude discretization of the underlying latent continuous vector. Overall, our simulations effectively illustrate the capabilities of factor copula models in moderate to high dimensions under an unbalanced data structure. For specific applications, it may be beneficial to consider bivariate linking copulas other than the elliptical ones to construct the joint probability distribution.

4.5 Applications

In this section, we introduce two real-life longitudinal datasets obtained from mixed-type responses. These datas ets are accessible electronically in the R packages *mixAk* and *lcmm*, respectively. However, our focus in this chapter is to evaluate the temporal dependency of each longitudinal outcome (whether continuous or discrete) separately within a univariate setup.

4.5.1 The PBC 910 data

In clinical practice, multiple markers of disease progression are routinely collected during follow-up to inform future treatment decisions. Motivated by the work of [101], we examine laboratory data on patients with primary biliary cirrhosis (PBC) from a Mayo Clinic trial conducted between 1974 and 1984. PBC is a chronic liver disease characterized by damage to the bile ducts in the liver, leading to bile accumulation and potential liver damage or cirrhosis. Left untreated or in advanced stages, PBC can result in severe complications, including mortality. This longitudinal study, with a median follow-up time of 6.3 years, involved 312 patients randomly assigned to receive either D-penicillamine ($m = 158$) or a placebo ($m = 154$). Various longitudinal biomarkers related to liver function were serially recorded for these patients, alongside baseline covariates such as gender and age. In this dataset, the number and timing of measurements vary considerably within and across subjects, resulting in highly unbalanced longitudinal data. Several authors in the joint modeling literature have analyzed this dataset (e.g., [102], [103], [104]). For our application, we focus on three biomarkers: serum albumin (mg/dl), serum bilirubin (mg/dl), and hepatomegaly (presence of an enlarged liver), comprising two continuous and one binary outcome. Profile plots for each biomarker are depicted in Figure 4.1. Notably, the average profiles for both the control and D-penicillamine-treated patients appear similar for each biomarker, suggesting no therapeutic differences between the two groups. Additionally, subjects exhibit varying numbers of visits, reflecting the unbalanced nature of the data. Serum bilirubin is observed

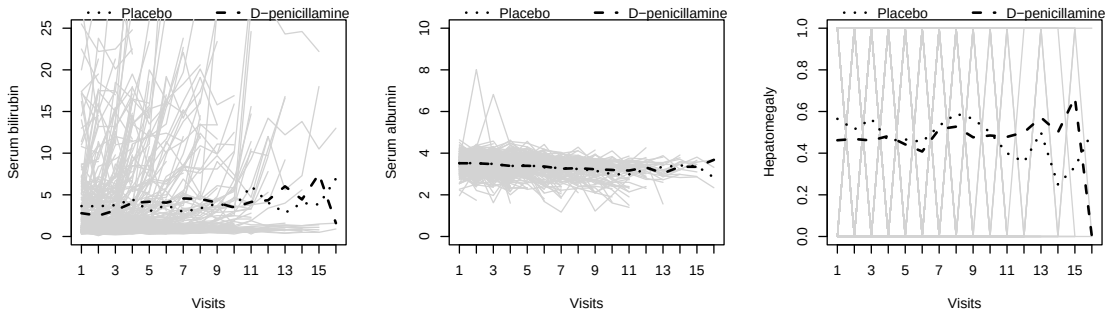

Figure 4.1 Subject-specific profiles over time for (i) Serum bilirubin, (ii) Serum albumin and (ii) Hepatomegaly for PBC 910 data set. The dotted lines show average profiles under placebo and D-penicillamine.

to be right-skewed, indicating that assuming a multivariate Gaussian joint distribution over time may not be reasonable. The maximum number of visits for subjects in this dataset is 16, posing challenges for addressing temporal dependency using standard multivariate copulas within the copula modeling framework.

For this data set we consider same set of covariates for each longitudinal responses. The fixed covariates are considered as sex ($0 =$ male, $1 =$ female), drug ($0 =$ placebo, $1 =$ D-penicillamine) and age along with the time of measurements (rescaled into years). For the two continuous responses we consider GLM for the marginals of the form -

$$g(E(Y_{ij})) = \beta_0 + \text{sex}_i\beta_1 + \text{drug}_i\beta_2 + \text{age}_i\beta_3 + t_{ij}\beta_4, \quad j = 1, \ldots, n_i, \tag{4.33}$$

where observed y_{ij} is the continuous response at the j-th time for subject i. For the serum bilirubin marker, we consider Gamma marginals with $\log$-link and for the serum albumin marker we consider normal marginals with identity link, based on the graphical diagnostics. For the hepatom marker, we consider the latent variable model as

$$Y_{ij} = \begin{cases} 0 & \text{if } Z_{ij} \leq 0 \\ 1 & \text{if } Z_{ij} > 0 \end{cases},$$

$$Z_{ij} = \beta_0 + \text{sex}_i\beta_1 + \text{drug}_i\beta_2 + \text{age}_i\beta_3 + t_{ij}\beta_4 + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \tag{4.34}$$

where $\epsilon_{ij} \sim N(0, 1)$. Note that, unlike the D-vine copula approach in [16], the subjects with recorded measurements only once, still contributes to the density of the factor copula models. We consider the same parsimonious parameterization for the 1-factor and 2-factor copula models described in Section 4.4. We fit our models to each longitudinal responses assuming homogeneous dependence structure for all individuals, $i = 1, \ldots, 312$, over time. We also compare the fittings with respect to the corresponding random effect models as well.

Serum bilirubin (gamma)			Serum albumin (normal)			Hepatom (binary)		
Parameters	Est.	SE	**Parameters**	Est.	SE	**Parameters**	Est.	SE
β_0	1.8223	0.3799	β_0	3.8159	0.1349	β_0	0.1274	0.3174
β_1	-0.2277	0.1745	β_1	-0.0194	0.0757	β_1	-0.3456	0.1495
β_2	-0.0427	0.1372	β_2	0.0369	0.0444	β_2	-0.1194	0.1066
β_3	-0.0083	0.0062	β_3	-0.0061	0.0019	β_3	0.0050	0.0054
β_4	-0.0266	0.0198	β_4	-0.0401	0.0055	β_4	0.0022	0.0158
κ	0.8596	0.0348	-	-	-	-	-	-
-	-	-	ϕ	0.4845	0.0166	-	-	-

Table 4.5 Estimated marginal parameters and their standard errors of 3 considered markers of the PBC910 data using the regression models in (4.33) and (4.34) respectively.

	Copula	**Parameters**	Est.	SE	Log-likelihood	AIC	BIC
Serum bilirubin	Gaussian 1-factor	ρ_1	0.8234	0.0152	-3703.78	7421.56	7447.76
	Student-t 1-factor ($\nu = 6$)	ρ_1	0.8429	0.0125	-3657.41	7330.82	7360.77
	Gaussian 2-factor	ρ_1 ρ_2	0.7234 0.6789	0.0205 0.0349	-3606.78	7229.57	7259.52
	Student-t 2-factor ($\nu = 3$)	ρ_1 ρ_2	0.8545 0.7225	0.0544 0.0619	-3353.31	**6724.62**	**6758.31**
Serum albumin	Gaussian 1-factor	ρ_1	0.6663	0.0205	-1052.75	2119.51	2145.71
	Student-t 1-factor ($\nu = 14$)	ρ_1	0.6810	0.0187	-1047.74	2111.48	2141.42
	Gaussian 2-factor	ρ_1 ρ_2	0.4831 0.5514	0.0151 0.0226	-1043.97	2103.93	2133.88
	Student-t 2-factor ($\nu = 9$)	ρ_1 ρ_2	0.4870 0.5944	0.0183 0.0245	-1032.35	**2082.69**	**2116.38**
Hepatom	Gaussian 1-factor	ρ_1	0.7283	0.0224	-1140.27	2292.53	2314.99
	Student-t 1-factor ($\nu = 30$)	ρ_1	0.7296	0.0223	-1140.69	2295.37	2321.58
	Gaussian 2-factor	ρ_1 ρ_2	0.5821 0.5876	0.0340 0.0368	-1137.54	**2289.08**	**2315.28**
	Student-t 2-factor ($\nu = 30$)	ρ_1 ρ_2	0.5384 0.6282	0.0212 0.0313	-1137.68	2291.37	2321.31

Table 4.6 Estimated dependence parameters and their standard errors of 3 considered markers of the PBC910 data with 1-factor and 2-factor copula models. Maximum log-likelihood value, AIC and BIC for each model are reported.

Serum bilirubin (gamma)			Serum albumin (normal)			Hepatom (binary)		
Parameters	Est.	SE	**Parameters**	Est.	SE	**Parameters**	Est.	SE
β_0	1.2441	0.5374	β_0	3.9450	0.1373	β_0	0.2093	0.4915
β_1	-0.3437	0.2087	β_1	-0.0375	0.0707	β_1	-0.4465	0.2493
β_2	0.0056	0.1776	β_2	0.0289	0.0448	β_2	-0.2412	0.1716
β_3	-0.0039	0.0095	β_3	-0.0081	0.0022	β_3	0.0060	0.0077
β_4	0.0900	0.0082	β_4	-0.0741	0.0030	β_4	0.0505	0.0124
$V[b]$	1.0090	0.0987	$V[b]$	0.1240	0.0125	$V[b]$	1.4673	0.2155
κ	1.1926	0.0367	-	-	-	-	-	-
-	-	-	ϕ	0.3510	0.0062	-	-	-
Log-likelihood -3629.24			Log-likelihood -1033.84			Log-likelihood -1125.45		
AIC **7272.47**			AIC **2084.69**			AIC **2262.90**		
BIC **7298.67**			BIC **2125.70**			BIC **2285.36**		

Table 4.7 Estimated parameters and their standard errors of 3 considered markers of the PBC910 data by adding random intercepts to the regression models in (4.33) and (4.34) respectively. Maximum log-likelihood value, AIC and BIC for each model are reported.

Table 4.5 and 4.6 present the marginal parameter estimates and the dependence parameter estimates of the PBC910 dataset under various factor copula models, respectively. The estimates of β_1 for all models deviate from zero, indicating that male subjects experienced greater liver disease progression than female subjects. Both 1-factor and 2-factor copula models with Student-t bivariate linking copulas outperformed models with Gaussian bivariate linking copulas. Furthermore, 2-factor copula models exhibit superior fits compared to 1-factor copula models with the same bivariate linking copulas, suggesting that a single latent variable is insufficient to describe the underlying dependence structure. The integer-valued degrees of freedom parameter is determined by the maximum value of the log-likelihood over the range of $\{3, \ldots, 30\}$, indicating a higher probability in the joint lower (or upper) tails of the bivariate pairs of each outcome with the corresponding latent variables. The copula parameter estimates indicate strong correlations between the observed responses and the latent variables for each marker. Consequently, by sticking with elliptical copulas, we can compare the dependence between the repeated measurements from the normal factor ([94]) and normal ogive ([95]) models, implying a positive association between the repeated measurements of each marker. In Table 4.7, we provide the parameter estimates with random intercept models. The estimates of the regression coefficients are comparable except for the intercept terms. Factor copula models appear to provide a slightly better model than the competing random effect model for the serum albumin marker, whereas for the hepatom marker we observe the reverse. Also for the serum bilirubin marker factor copula models seem to provide a better fit than the random effect model. Attempts to fit these markers with random intercept and slope models were made, but the likelihood did not converge for all cases. In Figure 4.2, we provide residual plots on a uniform scale for each marker corresponding to their best-fitting model. For the binary marker, being the least informative, no conclusions can be drawn from the residuals.

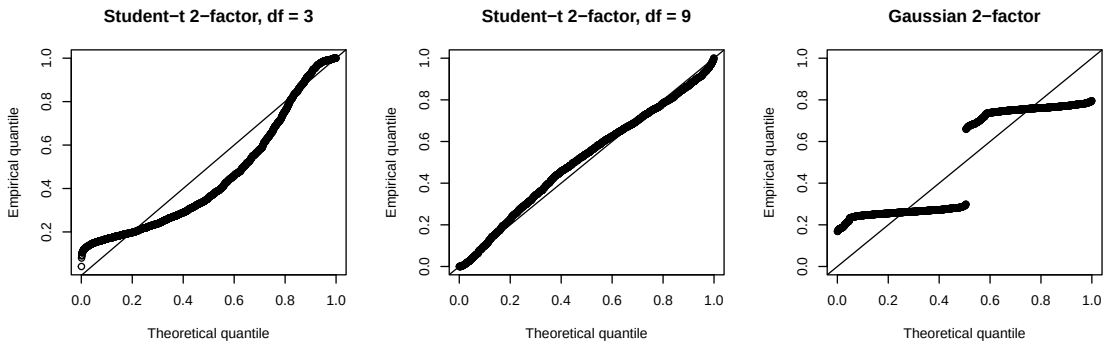

Figure 4.2 Uniform probability plots of the residuals of the best fitting copula models for (i) Serum bilirubin, (ii) Serum albumin and (ii) Hepatomegaly for PBC 910 data set.

4.5.2 The PAQUID data

We also examine the dataset from the French prospective cohort study PAQUID, which commenced in 1988 to investigate normal and pathological aging ([105]). This cohort comprised 3777 individuals aged 65 years and older at the initial visit, who were subsequently followed six times with intervals of 2 or 3 years, undergoing repeated neuropsychological evaluations and clinical diagnoses of dementia—a condition characterized by impaired memory, thinking, or decision-making that hinders daily activities. At each visit, participants completed a battery of psychometric tests, and dementia diagnosis criteria were evaluated. In our analysis, we focus on two psychometric tests: (i) the Mini-Mental State Examination (MMSE), providing an index of global cognitive performance, and (ii) the Benton Visual Retention Test (BVRT), assessing visual memory. Additionally, we include the score of physical dependency (HIER), where lower values indicate more severe impairment. This dataset has been previously examined within a joint modeling framework in the literature (e.g., [106], [107], [108]). We observe two continuous and one ordinal outcome. From the profile plots in Figure 4.3, distinct trajectory differences emerge between individuals diagnosed with dementia and those who were not. For our analysis, we focus on a subsample of size 500, aiming to elucidate the evolution of each longitudinal response. Among these, 128 received a positive dementia diagnosis. These subjects underwent between 1 to 9 measurements per test, with an average of 4.5 measurements.

For this data set we consider the fixed covariates as sex ($1 =$ male, $0 =$ female), dem ($1 =$ diagnosed positive of dementia), cep (educational level, $1 =$ graduated from primary school, $0 =$ otherwise) and age at follow up visits. Following [109], we consider the time covariate as the age minus 65 years per 10 years ($t = \frac{\text{age}-65}{10}$). The MMSE test is skewed to left, hence we apply the $\log$ transformation and consider the marginals to be normal. For the BVRT test, we don't make any transformation and consider normal marginals as well. For the purpose of fitting, we re-scale the score of physical dependency (ordinal response) from $\{0, \ldots, 3\}$ to $\{1, \ldots, 4\}$. Here also for the two continuous responses we consider

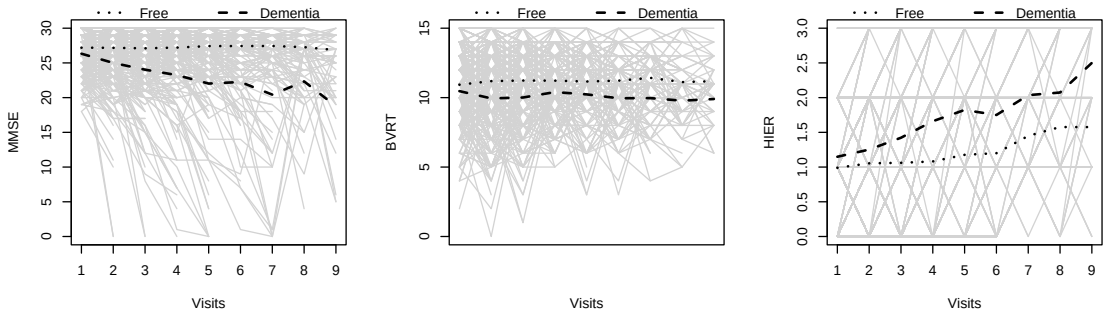

Figure 4.3 Subject-specific profiles over time for (i) MMSE, (ii) BVRT psychometric tests and (iii) HIER for PAQUID data set. The dotted lines show average profiles with free and positive diagnosis of dementia.

log - MMSE (normal)			BVRT (normal)			HIER (ordinal)		
Parameters	Est.	SE	**Parameters**	Est.	SE	**Parameters**	Est.	SE
β_0	3.4208	0.0330	β_0	10.9296	0.1992	-	-	-
β_1	0.0047	0.0173	β_1	0.2297	0.1536	β_1	-0.3077	0.0860
β_2	-0.1758	0.0225	β_2	-0.6483	0.1493	β_2	0.3503	0.0848
β_3	0.0717	0.0214	β_3	1.4134	0.1599	β_3	-0.1579	0.0878
β_4	-0.1240	0.0157	β_4	-0.6678	0.0821	β_4	0.9184	0.0562
ϕ	0.3209	0.0295	ϕ	2.1432	0.0427	-	-	-
-	-	-	-	-	-	γ_1	0.3094	0.1232
-	-	-	-	-	-	γ_2	1.6401	0.1338
-	-	-	-	-	-	γ_3	2.9992	0.1608

Table 4.8 Estimated marginal parameters and their standard errors of 3 considered tests of the PAQUID data using the regression models in (4.35) and (4.36) respectively.

GLM model of the form -

$$g(E(Y_{ij})) = \beta_0 + \text{sex}_i\beta_1 + \text{dem}_i\beta_2 + \text{cep}_i\beta_3 + t_{ij}\beta_4, \quad j = 1, \ldots, n_i, \tag{4.35}$$

where observed y_{ij} is the continuous response at the j-th time for subject i. For the ordinal response we consider the latent variable model as

$$\begin{aligned} &Y_{ij} = k \ \text{ if } \ \gamma(k-1) \leq Z_{ij} < \gamma(k), \ \ k = 1, \ldots, 4, \\ &Z_{ij} = \text{sex}_i\beta_1 + \text{dem}_i\beta_2 + \text{cep}_i\beta_3 + t_{ij}\beta_4 + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \end{aligned} \tag{4.36}$$

where $\epsilon_{ij}(i, i, d) \sim N(0, 1)$. Again, specifications of the factor copula models are similar to Section 4.4. While fitting these models, we also compare them with the corresponding random effect models. Our aim of this analysis is to describe the decline with age of the global cognitive ability measured by these psychometric tests and to evaluate the association within the longitudinal responses.

	Copula	**Parameters**	Est.	SE	Log-likelihood	AIC	BIC
log - MMSE	Gaussian 1-factor	ρ_1	0.7366	0.0339	-455.97	925.93	955.44
	Student-t 1-factor ($\nu = 3$)	ρ_1	0.8488	0.0092	-191.11	398.21	431.93
	Gaussian 2-factor	ρ_1 ρ_2	0.6228 0.5821	0.0719 0.0648	-435.61	887.23	920.94
	Student-t 2-factor ($\nu = 3$)	ρ_1 ρ_2	0.7725 0.7726	0.0472 0.0426	-169.47	**356.94**	**394.87**
BVRT	Gaussian 1-factor	ρ_1	0.5460	0.0249	-4774.56	9563.11	9592.61
	Student-t 1-factor ($\nu = 6$)	ρ_1	0.5678	0.0232	-4766.38	**9548.76**	**9582.47**
	Gaussian 2-factor	ρ_1 ρ_2	0.4036 0.4011	0.0096 0.0376	-4774.65	9565.30	9599.02
	Student-t 2-factor ($\nu = 8$)	ρ_1 ρ_2	0.4209 0.4199	0.0662 0.0696	-4766.33	9550.70	9588.63
HIER	Gaussian 1-factor	ρ_1	0.7282	0.0183	-2154.38	4324.76	4358.48
	Student-t 1-factor ($\nu = 4$)	ρ_1	0.7331	0.0185	-2139.54	4297.07	4335.00
	Gaussian 2-factor	ρ_1 ρ_2	0.5249 0.6192	0.0146 0.0218	-2149.77	4317.54	4355.48
	Student-t 2-factor ($\nu = 3$)	ρ_1 ρ_2	0.6083 0.5611	0.0426 0.0680	-2124.43	**4268.85**	**4311.00**

Table 4.9 Estimated dependence parameters and their standard errors of 3 considered tests of the PAQUID data with 1-factor and 2-factor copula models. Maximum log-likelihood value, AIC and BIC for each model are reported.

log - MMSE (normal)			BVRT (normal)			HIER (ordinal)		
Parameters	Est.	SE	**Parameters**	Est.	SE	**Parameters**	Est.	SE
β_0	3.4308	0.0257	β_0	10.7646	0.1868	-	-	-
β_1	0.0065	0.0185	β_1	0.1048	0.1452	β_1	-0.2736	0.0860
β_2	-0.1804	0.0195	β_2	-0.5950	0.1561	β_2	0.2572	0.0933
β_3	0.0670	0.0202	β_3	1.4516	0.1577	β_3	-0.1860	0.0787
β_4	-0.1346	0.0104	β_4	-0.6690	0.0679	β_4	1.0603	0.0361
$V[b]$	0.0157	0.0024	$V[b]$	0.8749	0.1597	$V[b]$	0.5597	0.0625
ϕ	0.2963	0.0049	ϕ	1.8019	0.0306	-	-	-
-	-	-	-	-	-	γ_1	0.3091	0.1242
-	-	-	-	-	-	γ_2	1.5805	0.1422
-	-	-	-	-	-	γ_3	3.0121	0.1577
Log-likelihood -593.76			Log-likelihood -4777.94			Log-likelihood -2251.08		
AIC **1201.52**			AIC **9569.87**			AIC **4512.17**		
BIC **1231.02**			BIC **9609.89**			BIC **4533.24**		

Table 4.10 Estimated parameters and their standard errors of 3 considered tests of the PAQUID data by adding random intercepts to the regression models in (4.35) and (4.36) respectively. Maximum log-likelihood value, AIC and BIC for each model are reported.

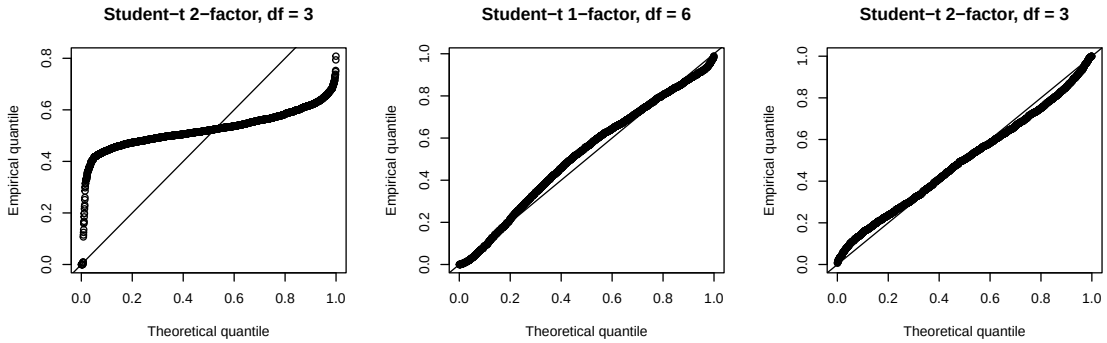

Figure 4.4 Uniform probability plots of the residuals of the best fitting copula models for (i) MMSE, (ii) BVRT psychometric tests and (iii) HIER for PAQUID data set.

Table 4.8 and 4.9 present the marginal parameter estimates and the dependence parameter estimates of the PAQUID dataset under various factor copula models, respectively. Notably, gender is found to be non-significant for MMSE and BVRT tests based on the marginal estimates. As depicted in Figure 4.3 and evident from the estimates of β_2 across all models, individuals with an initial positive dementia diagnosis exhibited poorer performance in two of the psychometric tests, and their score of physical dependency (HIER) increased more compared to others. Moreover, based on estimates of β_3 in all models, individuals with higher educational levels demonstrated higher cognitive ability and lower physical dependency throughout the cohort. Increasing age, as indicated by the estimates of β_4, had a significant impact on psychometric tests. Student-t 2-factor copula models provide the best fit for the MMSE test and the score of physical dependency. For the BVRT test, the 1-factor Student-t copula model appears to best describes the underlying dependence. Strong, positive correlations are observed between the latent variables and each response. Parameter estimates from the considered random effect models are provided in Table 4.10. The estimates of the fixed effects closely resemble those of the marginal models. Notably, factor copula models outperform random effect models for each test in this dataset based on the selection criteria, indicating that the considered factor copula models better explain the temporal dependency. Figure 4.4 displays residual plots on a uniform scale for each psychometric test corresponding to their best-fitting model. While the models for the BVRT test and the score of physical dependency suggest perfect fitting, for the MMSE test, a substantial lack of fit is observed, even though the Student-t 2-factor copula provided the best fit among other competitors. Indeed, mixed models can be influenced by the misspecification of random effects distribution ([110]). Conversely, factor copula models offer a more direct interpretation of the underlying dependence mechanism. The choice between random effects and copula models ultimately depends on the scientific question of interest and what the investigator seeks to uncover.

4.6 Discussion

In this chapter, we introduced factor copula models for arbitrary non-Gaussian longitudinal data, incorporating covariates. Our proposed models demonstrate effectiveness in modeling unbalanced longitudinal data with both discrete and continuous responses. With a parsimonious specification of factor copula parameters, our models are easily scalable to accommodate dependence in moderate to high dimensions without computational obstacles. We employed a two-stage IFM method to estimate the model parameters. Our simulation studies illustrate consistent and reliable estimation of both the marginal and dependence parameters of the models. Furthermore, we compared our proposed models with some widely used random effect models, employing similar specifications of the fixed covariates.

Factor copula models assume a homogeneous dependence structure for all subjects by construction. However, in certain scenarios, this assumption may not be realistic, as measurements taken closer in time are likely to be more dependent than measurements taken farther apart. Therefore, there is scope for further improvement of factor copula models to account for potential time-heterogeneity. A relevant reference is [32]. For purposes of illustration, we considered two real-world datasets that are popular in the joint modeling literature. A natural statistical question would be to evaluate the contemporaneous association between each longitudinal response. This can be addressed in two ways. In the first approach, following [21], we may try to model the temporal association of each response using pair copula construction with a D-vine structure and the contemporaneous association of bivariate responses is then joined by a bivariate copula. In the second approach, following [18], we may try to use a transition model for considering the association of repeated measurements over time. More specifically, in this approach, Gaussian copula is used for considering correlation of multivariate responses for each given time and the transition model is used for considering association of repeated measurements of each individual. The authors ([18]) have noted that this has a flexible and computationally efficient structure for optimizing the likelihood function. This can be addressed in two ways. Similar to these approaches, one can consider multivariate copulas to associate the subject-specific latent variables in the factor copulas, i.e., correlated latent variables on each factor. Therefore, all dependence and association parameters can be estimated from the second stage of IFM estimation using the joint likelihood. A more elegant approach would be to use correlated random effects in the marginals of each longitudinal response to explain contemporaneous association between each response. We propose to investigate this in future. Also, in this regard, some recent developments can be found in [111]. In our future studies we further intend to investigate calibration methods for factor copula models such as PIT histograms and tail dependence checks as described in [112], [113] and [114].

4.7 Appendix

Random effects are commonly employed to incorporate subject-specific effects into the linear predictor of models, addressing within-subject or temporal dependencies (see [13]). Factor copula models can be contrasted with random effect models as both involve unobservable latent variables. Random

effect models adds additive latent variables in the linear predictors but in factor copula models latent variables are used to capture dependence in between variables. In this chapter, we also explore random effect models as competitors for capturing temporal dependencies, extending the marginal models introduced in Section 3.1.

Referring to (4.16), we consider generalized linear mixed models as

$$g(E(Y_{ij}|\mathbf{b}_i)) = \mathbf{x}_{ij}\beta + \mathbf{d}_{ij}\mathbf{b}_i, \quad j = 1, \ldots, n_i, \tag{4.37}$$

where $\mathbf{d}_{ij}$ is the corresponding random effects design vector. Similarly following [115], we extend the latent variable models in (4.17) by

$$Z_{ij} = \mathbf{x}_{ij}\beta + \mathbf{d}_{ij}\mathbf{b}_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i. \tag{4.38}$$

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\intercal$ be the vector of observed responses for the i-th subject, $(i = 1, \ldots, m)$ then the parameters of the random effect models can be estimated from the marginal likelihood as

$$l(\theta|\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i) h(\mathbf{b}_i) d\mathbf{b}_i, \tag{4.39}$$

where $h(.)$ is the distribution of the random effects which is assumed to be normal. For simplicity and relevance to the real-world datasets considered in this chapter, we confine our analysis to random intercept models exclusively. An essential aspect of longitudinal data analysis is model selection, involving the determination of the most appropriate number of components for a given dataset. To this end, we employ two widely utilized tools: AIC (Akaike information criterion) and BIC (Bayesian information criterion), which penalize models with a large number of parameters. These criteria are defined as follows -

$$AIC = -2l(\hat{\theta}^*) + 2\dim(\theta^*), \quad BIC = -2l(\hat{\theta}^*) + \log(m)\dim(\theta^*) \tag{4.40}$$

where $\hat{\theta}^*$ represents the maximum likelihood estimates of the model parameters, and m denotes the sample size. Each criterion comprises two terms: the first term measures the goodness-of-fit, while the second term penalizes model complexity. Although we estimate parameters using the two-stage IFM method, we still employ these criteria as close approximations of the actual ones (see [69] and [16]). Smaller values of AIC and BIC indicate a better-fitting model. We utilize these criteria in our simulations and data analysis to compare different models. However, these selection criteria cannot definitively determine whether the best-fitting model adequately captures the data. For copula-based models, we further validate the fit using residual plots.

Chapter 5

Modeling longitudinal count data using finite mixture copulas

Modeling dependence in multivariate count data has received considerable attention in the literature. The references [116], [117] and [118] contain discussion on development of appropriate models, relevant to analysis of discrete or count data, either in regression set-up or in set-up of longitudinal data. The reference [3] also contains discussion on discrete longitudinal data. These developments have meaningful components on underlying dependence. Multivariate elliptical copulas are often used to analyze dependence in repeated longitudinal data but they are limited in flexibility and require parametric assumptions for inference. In this chapter, we propose using finite mixtures of elliptical copulas to capture complex temporal dependencies in discrete longitudinal data. This approach allows researchers to have distinct correlation matrices in the components of the mixture copula.

The well-known references [119] and [120] have studied longitudinal count data with overdispersion. The reference [119] have used a mixed-effect approach while the reference [120] developed an approach based on generalized estimating equations. This set the stage for subsequent investigations. Following this, we see in [121] a detailed discussion on various estimation techniques related to regression models for discrete longitudinal responses. These techniques are based on generalized estimating equation approach and also on generalized quasi-likelihood estimation approach. The discussion offers practical insights into addressing the challenges of analysis of discrete longitudinal responses. Yet another approach was developed in [122] where the authors proposed a state space model for the analysis of longitudinal count data with serial correlation. They considered both equally or unequally spaced data. However, the computational burden is always an issue from a practitioners point of view. Despite the advancements in discrete longitudinal data analysis, there remains a need to balance methodological sophistication with computational feasibility to ensure practical applicability in real-world scenarios.

Statisticians have explored the use of copulas in the analysis of count longitudinal data. Copulas are valued for their flexibility in modeling multivariate distributions and revealing dependencies among variables. This has enabled researchers to study various dependence structures in regression models while allowing for different marginal distributions. Early contributions by [9] and [10] laid the foundation for applying copulas to multivariate count data. However, applying copulas directly to discrete data can lose desirable properties, though modeling remains feasible through Sklar's theorem. Elliptical copulas, as suggested by [90], are popular for modeling non-Gaussian longitudinal data due to their

ability to capture within-subject dependencies via correlation matrices. However, their parametric restrictions limit their ability to model complex dependence patterns. To overcome this, some researchers have proposed vine copulas, though they require estimating many parameters, presenting computational challenges. Our approach attempts at circumventing these complexities by using a finite mixture of elliptical copulas by extending the concept of mixture distributions. Some recent work have used finite mixture of copulas, enhancing flexibility and in particular, capturing complex dependencies. For example, [123] used a time-varying mixture copula model to analyze stock market data, highlighting efforts to model count longitudinal data effectively while addressing its complexities. The use of copula mixtures is expected to facilitate the modeling and interpretation of diverse dependence mechanisms.

This chapter focuses on modeling the correlation structure of longitudinal count data. Our motivation stems from the recognition that within a mixture of elliptical copulas, diverse correlation structures can be employed across component copulas (see, for example, [124]). The weights assigned to each component copula indicate their relative importance, aiding in understanding data dynamics. We propose a useful technique of parameter estimation methodology which combines the composite likelihood method with a two-stage procedure. Our method captures different aspects of the model.

The contribution in this chapter is two-fold. Firstly, we meticulously derive the dependence properties of mixture copula models under both continuous and discrete setups, filling in certain gaps in the literature. While mixture copulas have been applied to discrete data before, their dependence properties under the discrete setup have not been explicitly derived. The recent work by [125] introduced a population version of Spearman's rho tailored to discrete data and applied various Archimedean copulas to model bivariate count data, laying a foundation for our endeavors reported in this chapter. Secondly, we apply regression models to longitudinal count data, capturing temporal dependencies with a mixture of elliptical copulas. By integrating these sophisticated models, we aim to unravel the intricate interplay between temporal dynamics and count data characteristics, offering practical insights into real-world phenomena. Through empirical analysis and simulation studies, we demonstrate the efficacy and versatility of our proposed methodology, underscoring its potential to enhance understanding and analysis in diverse fields ranging from epidemiology to finance.

Rest of the chapter is organized as follows. In Section 5.1, we review some standard definitions of mixture and elliptical copulas. In Section 5.2, we derive the theoretical properties such as Kendall's tau and Spearman s rho for a general mixture copula under both continuous and discrete case. Section 5.3 describes some standard marginal regression models generally used for modeling longitudinal count data. In Section 5.4 we propose a useful technique of parameter estimation which combines, the composite likelihood method with a two stage procedure. Our method combines different aspect of the proposed models. In Section 5.5, we discuss some standard model validation techniques and also extend the t-plot method of model validation for finite mixture of elliptical copulas. Results of extensive simulation studies are reported in Section 5.6. Their findings help us to see finite sample performances of our proposed class of models under different sample sizes. In Section 5.7 we apply our methods to model the temporal dependency of two real-life data sets. We also compare the fits of mixture copulas with

standard elliptical copulas and demonstrate substantial improvements. Based on the derived expressions of Kendall's tau and Spearman's rho, we estimated the concordance matrices of these two data sets and showed that they are very close to their empirical versions. Finally we conclude this chapter with a general discussion in Section 5.8.

5.1 The K-finite mixture of multivariate copulas

Mixture models have been extensively explored in statistical literature (see, for example, [126] and [127]. The reference [128] contains a comprehensive introduction to mixture modeling and its practical applications. The versatility of mixture models lies in their ability to generate more flexible multivariate distributions by blending different distributions of the same dimension, even if they do not originate from the same family. Similarly, the concept can be extended to copulas, where mixing different copulas allows for the incorporation of diverse dependence characteristics into a statistical model. In many real-world scenarios, a single parametric copula may prove inadequate to capture all pertinent features during analysis. The exploration of finite mixture copula models in the literature has laid a solid foundation for further research in this domain; interested readers may refer to the references [129], [130], and [131] for related discussions.

A K-finite mixture copula is defined as

$$C_{\text{mix},d} = \sum_{l=1}^{K} \pi_l C_{l,d}(.|\phi_l), \quad \sum_{l=1}^{K} \pi_l = 1, \pi_l \geq 0, \quad \forall\, l = 1, \ldots, K \tag{5.1}$$

where $C_{l,d}(.|\phi_l)$ denotes a single d-dimensional multivariate copula component which has a mixture weight π_l, and K is the number of components in the mixture. It is easy to see that the distribution function in (5.1) is a copula. The density function of the mixture copula can be simply obtained as

$$c_{\text{mix},d}(u_1, \ldots, u_d|\eta) = \sum_{l=1}^{K} \pi_l c_{l,d}(u_1, \ldots, u_d|\phi_l) \tag{5.2}$$

where $\eta = \{\pi_l, \phi_l; l = 1, \ldots, K\}$ denotes all dependence parameters, ϕ_l contains the copula parameters of the l-th component and $\mathbf{u} = (u_1, \ldots, u_d) \in (0, 1)^d$. The choice of the copula components in the mixture (5.1) can be arbitrary, but in this chapter we restrict our attention to multivariate elliptical copulas.

5.1.1 Elliptical copulas

Multivariate copulas derived from elliptical distributions, such as multivariate normal or Student-t, have gained prominence in statistics and econometrics due to their simplicity in parametric inference (see, [132], [133]). These copulas offer an elegant framework for modeling dependencies. However, it is essential to acknowledge the discussion by [134], who examined both the dependence structures

generated by elliptical distributions and their associated limitations. This critical assessment sheds light on the nuances of employing such copulas in practical applications, informing researchers and practitioners alike about their strengths and potential pitfalls. For some more insights we refer to [135]. We proceed with some standard definitions as follows.

Definition 5.1.1 *A d-dimensional copula is said to be a Gaussian copula if*

$$C_d(\mathbf{u}|\mathbf{\Sigma}) = \Phi_d(\Phi_1^{-1}(u_1), \ldots, \Phi_1^{-1}(u_d)|\mathbf{\Sigma}), \tag{5.3}$$

where $\Phi_d(.|\mathbf{\Sigma})$ is the CDF of the d-variate normal distribution with standard normal marginals and correlation matrix $\mathbf{\Sigma}$, $\Phi_1^{-1}(\cdot)$ denotes the inverse of the CDF of univariate standard normal distribution.

Definition 5.1.2 *A d-dimensional copula is said to be a Student-t copula if*

$$C_d(\mathbf{u}|\mathbf{\Sigma}, \nu) = T_d(T_1^{-1}(u_1|\nu), \ldots, T_1^{-1}(u_d|\nu)|\mathbf{\Sigma}, \nu) \tag{5.4}$$

where $T_d(.|\mathbf{\Sigma}, \nu)$ is the CDF of the d-variate Student-t distribution with standard-t marginals and scale matrix $\mathbf{\Sigma}$, $T_1^{-1}(\cdot|\nu)$ denotes the inverse of the CDF of univariate standard t-distribution with ν degrees of freedom.

Student-t copula has an additional degrees of freedom parameter ν, which accounts for possible tail dependence in the data. We consider the mixture of Gaussian and Student-t copulas for modeling the temporal dependence of longitudinal count data. As [90] emphasized, elliptical copulas are more useful when the dimension of the data is moderate to high since all lower dimensional sub-copulas stay in the same parametric family ([136]).

5.1.2 Motivation for this proposal

The primary motivation behind this work is to capture the complex dependencies present in count longitudinal data. In many real-world longitudinal datasets, it is often challenging to clearly identify the structure of the underlying correlations (e.g., AR(1), MA, etc.). In this chapter, we address this issue by employing mixture copulas. However, we restrict our attention to the case of $K = 2$ components. One reason that we choose elliptical copulas that they are easy to simulate and have simpler parametric inference using composite likelihood which will be discussed later on.

5.2 Dependence properties

In this section, we present a thorough theoretical analysis of mixture copula models, addressing both continuous and discrete margins. The results we provide are applicable to a general mixture copula model, offering insight into the intricate relationships between variables. For continuous random variables, the dependence structure, often quantified by Kendall's tau or Spearman's rho, depends exclusively on the copula parameters (see, [26]). This means that the dependence parameters η of the mixture

copulas in equation (5.2) can be interpreted directly in terms of these dependence measures, which are constrained within the interval $[-1, 1]$. To simplify the presentation, we assume in this section that the mixture copula, C_{mix}, is bivariate (i.e., $d = 2$). This allows us to focus on bivariate copulas while maintaining generality in the underlying theory. The next step involves deriving the population versions of Kendall's tau and Spearman's rho for continuous random variables, which will provide a clearer understanding of how these measures of dependence relate to the copula parameters in our framework.

Before we proceed to state some of the theoretical results related to mixture copulas under discrete margins, a few relevant facts are in order. Which serve as preamble to the Theorems presented in this section.

Proposition 5.2.1 *Suppose $(X_1, X_2)^\intercal$ is a bivariate random vector with cdf $F(x_1, x_2)$ such that*

$$F(x_1, x_2) = \sum_{i=1}^{K} \pi_i F_i(x_1, x_2),$$

where for $i = 1, \ldots, K$, F_i is a bivariate cdf, and

$$\pi_i \geq 0 \;\; for \;\; i = 1, \ldots, K \;\; and \sum_{i=1}^{K} \pi_i = 1.$$

Suppose that for $i = 1, \ldots, K$, F_{ij} ($j = 1, 2$) denote the marginals associated with F_i. We assume that for every $i = 1, \ldots, K$ and $j = 1, 2$, $F_{ij} = G_j$, a univariate cdf. In other words, each of the two marginals of the components $F_1, \ldots, F_K$, does not depend on the component. This means F is a bivariate cdf which is equal to a mixture of K many bivariate cdf's with equal marginals. Then the univariate marginals of F are given by G_1 and G_2.

Proof: We notice that,

$$F(x_1, \infty) = \sum_{i=1}^{K} \pi_i F_i(x_1, \infty) = \sum_{i=1}^{K} \pi_i F_{i1}(x_1) = \sum_{i=1}^{K} \pi_i G_1(x_1) = G_1(x_1).$$

It follows similarly that $F(\infty, x_2) = G_2(x_2)$, completing the proof of the proposition. □

Remark 5.2.1: It is easy to see that a multivariate version of proposition 5.2.1 is true. We skip the details.

Proposition 5.2.2 *Suppose $(X_1, X_2)^\intercal$ is a bivariate random vector with cdf $F(x_1, x_2)$ as in the preceding proposition. Assume, moreover, that both G_1 and G_2 are continuous. Let, for $i = 1, \ldots, K$, C_i denote the copula corresponding to F_i. Then the copula corresponding to F, denoted by C, is given by*

$$C(u_1, u_2) = \sum_{i=1}^{K} \pi_i C_i(u_1, u_2) = C_{\text{mix}}(u_1, u_2).$$

Remark 5.2.2: Proposition 5.2.2 shows that the copula corresponding to a mixture of cdf's with specified marginals is equal to the same mixture of the copulas corresponding to the component cdf's in the mixture.

Proof of Proposition 5.2.2**:** We have assumed that the univariate marginals of all the F_i's are given by G_1 and G_2. Hence,

$$F(x_1, x_2) = \sum_{i=1}^{K} \pi_i F_i(x_1, x_2) = \sum_{i=1}^{K} \pi_i C_i(G_1(x_1), G_2(x_2)),$$

Moreover, we have seen that the univariate marginals of F also are given by G_1 and G_2. Hence,

$$F(x_1, x_2) = C(G_1(x_1), G_2(x_2)).$$

Hence,

$$C(u_1, u_2) = \sum_{i=1}^{K} \pi_i C_i(u_1, u_2),$$

as asserted. □

Remark 5.2.2: It is easy to see that a multivariate version of proposition 5.2.2 is true. We skip the details.

Theorem 5.2.1 *Let $(X_1, X_2)^\intercal$ be a bivariate continuous random vector having dependence of finite mixture copula C_{mix} defined in (5.1), with marginal distribution functions $F_j, j = 1, 2$. The population version of Kendall's tau for X_1 and X_2 is given by*

$$\tau(C_{mix}) = \sum_{l=1}^{K} \pi_l^2 \tau(C_l) + 2 \sum_{l<m}^{K} \pi_l \pi_m Q_{lm}, \tag{5.5}$$

where $Q_{lm} = Q(C_l, C_m)$ is the concordance function defined for copulas C_l and C_m and $\tau(C_l)$ is the Kendall's tau corresponding to copula C_l, for $l = 1, \ldots, K$. More precisely,

$$Q_{lm} = Q(C_l, C_m) = \int_0^1 \int_0^1 C_l(u, v) dC_m(u, v), \quad 1 \le l, m \le K.$$

Proof: Let $(X_1', X_2')^\intercal$ be an independent copy of $(X_1, X_2)^\intercal$. Using the definition of Kendall's tau we have -

$$\begin{aligned}
\tau(C_{\text{mix}}) &= P(\text{concordance}) - P(\text{discordance}) \\
&= P((X_1 - X_1')(X_2 - X_2') > 0) - P((X_1 - X_1')(X_2 - X_2') < 0) \\
&= 4P(X_1' < X_1, X_2' < X_2) - 1 \\
&= 4 \int_0^1 \int_0^1 C_{\text{mix}}(u_1, u_2) dC_{\text{mix}}(u_1, u_2) - 1.
\end{aligned} \tag{5.6}$$

Therefore, $\int_0^1 \int_0^1 C_{\text{mix}}(u_1, u_2) dC_{\text{mix}}(u_1, u_2)$

$$= \int_0^1 \int_0^1 \sum_{l=1}^{K} \pi_l C_l(u_1, u_2|\phi_l) d \sum_{l=1}^{K} \pi_l C_l(u_1, u_2|\phi_l)$$

$$= \sum_{l=1}^{K} \pi_l^2 \int_0^1 \int_0^1 C_l(u_1, u_2|\phi_l) dC_l(u_1, u_2|\phi_l) + \sum_{m \neq l}^{K} \pi_l \pi_m \int_0^1 \int_0^1 C_l(u_1, u_2|\phi_l) dC_m(u_1, u_2|\phi_m).$$

Now, $\int_0^1 \int_0^1 C_l(u_1, u_2|\phi_l) dC_m(u_1, u_2|\phi_m)$

$$= \int_0^1 \int_0^1 \int_0^{u_1} \int_0^{u_2} c_l(u_1, u_2|\phi_l) c_m(u_1, u_2|\phi_m) du_1^2 du_2^2$$

$$= \int_0^1 \int_0^1 C_m(u_1, u_2|\phi_m) dC_l(u_1, u_2|\phi_l).$$

Since $\sum_{l=1}^{K} \pi_l = 1$, by plugging the values in equation (5.6) we obtain the result. □
Note that the concordance function is symmetric in its arguments for a continuous random vector ([26]; [129]), but this doesn't hold for the discrete case.

Corollary 5.2.1 *For mixture of Gaussian copula, the population version of Kendall's tau can be obtained in a closed form expression as*

$$\tau(C_{mix}) = \frac{2}{\pi} \sum_{l=1}^{K} \pi_l^2 \arcsin(\rho_l) + \frac{4}{\pi} \sum_{l<m}^{K} \pi_l \pi_m \arcsin\Big(\frac{\rho_l + \rho_m}{2}\Big), \tag{5.7}$$

where ρ_l is the correlation parameter of bivariate Gaussian copula, for $l = 1, \ldots, K$.

Proof: A proof follows easily along the line of arguments needed to justify the remark below (Remark 5.2.3). The justification (of Remark 5.2.3) along with necessary reference to the corollary is given in the appendix. See also [43] and [137]. □
Remark 5.2.3: It can be argued that for mixture of Student-t copula, the population version of Kendall's tau can be approximated by a closed form expression as

$$\tau(C_{\text{mix}}) \approx \frac{2}{\pi} \sum_{l=1}^{K} \pi_l^2 \arcsin(\rho_l) + \frac{4}{\pi} \sum_{l<m}^{K} \pi_l \pi_m \arcsin\Big(\frac{\rho_l + \rho_m}{2}\Big). \tag{5.8}$$

Theorem 5.2.2 *Let $(X_1, X_2)^\intercal$ be a bivariate continuous random vector having dependence of finite mixture copula C_{mix} defined in (5.1), with marginal distribution functions $F_j, j = 1, 2$. The population version of Spearman's rho for X_1 and X_2 is given by*

$$\rho(C_{mix}) = \sum_{l=1}^{K} \pi_l \rho(C_l), \tag{5.9}$$

where $\rho(C_l)$ is the Spearman's rho corresponding to copula C_l, for $l = 1, \ldots, K$.

Proof: Let $(X_1^*, X_2^*)^\intercal$ be two independent random variables (i.e. bivariate random vector with independence copula) with same marginal distributions $F_j, j = 1, 2$. Using the definition of Spearman's rho we have -

$$\begin{aligned}
\rho(C_{\text{mix}}) &= 3(P(\text{concordance}) - P(\text{discordance})) \\
&= 3(P((X_1 - X_1^*)(X_2 - X_2^*) > 0) - P((X_1 - X_1^*)(X_2 - X_2^*) < 0)) \\
&= 12P(X_1^* < X_1, X_2^* < X_2) - 3 \\
&= 12 \int_0^1 \int_0^1 C_{\text{mix}}(u_1, u_2) du_1 du_2 - 3. \qquad (5.10)
\end{aligned}$$

$$\text{Therefore, } \int_0^1 \int_0^1 C_{\text{mix}}(u_1, u_2) du_1 du_2$$
$$= \int_0^1 \int_0^1 \sum_{l=1}^{K} \pi_l C_l(u_1, u_2|\phi_l) du_1 du_2 = \sum_{l=1}^{K} \pi_l \int_0^1 \int_0^1 C_l(u_1, u_2|\phi_l) du_1 du_2.$$

Since $\sum_{l=1}^{K} \pi_l = 1$, by plugging the values in equation (5.10) we obtain the result. Expression of Spearman's rho for Gaussian mixture copula can be obtained in a closed form, but for Student-t no closed form is available (see, [28]). □

When dealing with discrete marginal distributions, concordance-based measures are influenced by both the marginal distributions and the copula. [138] and [139] have delved into the population version of Kendall's tau applied to discrete data, which is not distribution-free and typically exhibits a narrower range than the continuous counterpart of $[-1, 1]$. [10] previously derived the population version of Kendall's tau under discrete marginals. More recently, [125] extended this analysis to derive the population version of Spearman's rho, followed by a continuous extension for discrete random variables. Building upon these advancements, we proceed to derive the population versions of these concordance measures for mixture copulas in the discrete case. It's important to note that in the discrete scenario, the probability of tie (i.e., identical observations) is positive, necessitating its consideration in the derivation process. The observation that the Spearman's rho of a convex combination of copulas equals the convex combination of the individual Spearman's rho holds true for both continuous and discrete cases. This finding underscores the influence of marginal distributions on the dependence measures, particularly evident in the discrete case.

Theorem 5.2.3 *Let $(X_1, X_2)^\intercal$ be a bivariate integer valued random vector having dependence of finite mixture copula C_{mix} defined in (5.1), with marginal distribution functions $F_j, j = 1, 2$ and mass functions $f_j, j = 1, 2$. The population version of Kendall's tau for X_1 and X_2 is given by*

$$\tau^*(C_{mix}) = \sum_{l=1}^{K} \pi_l^2 \tau^*(C_l) + \sum_{l \neq m}^{K} \pi_l \pi_m Q_{lm}^*, \tag{5.11}$$

where
$$\begin{aligned} \tau^*(C_l) &= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} h_l(x_1, x_2)\{4C_l(F_1(x_1 - 1), F_2(x_2 - 1)|\phi_l) - h_l(x_1, x_2)\} \\ &\quad + \sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - 1, \end{aligned} \tag{5.12}$$

$$\begin{aligned} Q_{lm}^* &= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} h_m(x_1, x_2)\{4C_l(F_1(x_1 - 1), F_2(x_2 - 1)|\phi_l) - h_l(x_1, x_2)\} \\ &\quad + \sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - 1, \end{aligned} \tag{5.13}$$

and
$$\begin{aligned} h_l(x_1, x_2) &= C_l(F_1(x_1), F_2(x_2)|\phi_l) - C_l(F_1(x_1 - 1), F_2(x_2)|\phi_l) \\ &\quad - C_l(F_1(x_1), F_2(x_2 - 1)|\phi_l) + C_l(F_1(x_1 - 1), F_2(x_2 - 1)|\phi_l). \end{aligned} \tag{5.14}$$

Proof: When $(X_1, X_2)^\intercal$ is integer valued random vector the probability of tie is non-zero and $P(\text{concordance}) + P(\text{discordance}) + P(\text{tie}) = 1$. Therefore we have -

$$\begin{aligned}
\tau^*(C_{\text{mix}}) &= P(\text{concordance}) - P(\text{discordance}) \\
&= 2P(\text{concordance}) + P(\text{tie}) - 1 \\
&= 2P((X_1 - X_1')(X_2 - X_2') > 0) + P(X_1 = X_1' \cup X_2 = X_2') - 1 \\
&= 4P(X_1' < X_1, X_2' < X_2) + P(X_1 = X_1' \cup X_2 = X_2') - 1. \qquad (5.15)
\end{aligned}$$

The last expression is due to the fact that $(X_1', X_2')^\intercal$ and $(X_1, X_2)^\intercal$ are identically distributed.

$$\begin{aligned}
&\text{Now,} \quad P(X_1' < X_1, X_2' < X_2) \\
&= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} P(X_1' \leq x_1 - 1, X_2' \leq x_2 - 1) P(X_1 = x_1, X_2 = x_2) \\
&= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} C_{\text{mix}}(F_1(x_1 - 1), F_2(x_2 - 1)) h_{\text{mix}}(x_1, x_2) \\
&= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} \Big(\sum_{l=1}^{K} \pi_l C_l(F_1(x_1 - 1), F_2(x_2 - 1)|\phi_l) \Big) \Big(\sum_{l=1}^{K} \pi_l h_l(x_1, x_2) \Big) \\
&= \sum_{l=1}^{K} \pi_l^2 \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} C_l(F_1(x_1 - 1), F_2(x_2 - 1)|\phi_l) h_l(x_1, x_2) \\
&+ \sum_{l \neq m}^{K} \pi_l \pi_m \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} C_l(F_1(x_1 - 1), F_2(x_2 - 1)|\phi_l) h_m(x_1, x_2). \qquad (5.16)
\end{aligned}$$

$$\begin{aligned}
&\text{And,} \quad P(X_1 = X_1' \cup X_2 = X_2') \\
&= P(X_1 = X_1') + P(X_2 = X_2') - P(X_1 = X_1', X_2 = X_2') \\
&= \sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} h_{\text{mix}}^2(x_1, x_2) \\
&= \sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - \sum_{l=1}^{K} \pi_l^2 \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} h_l^2(x_1, x_2) \\
&- \sum_{l \neq m}^{K} \pi_l \pi_m \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} h_l(x_1, x_2) h_m(x_1, x_2). \qquad (5.17)
\end{aligned}$$

$$\text{Let,} \quad \tau^*(C_l) = \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} h_l(x_1,x_2)\{4C_l(F_1(x_1-1),F_2(x_2-1)|\phi_l) - h_l(x_1,x_2)\} + \sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - 1,$$

$$Q^*_{lm} = \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} h_m(x_1,x_2)\{4C_l(F_1(x_1-1),F_2(x_2-1)|\phi_l) - h_l(x_1,x_2)\} + \sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - 1.$$

Therefore, using (5.16) and (5.17) in (5.15) we get -

$$\begin{aligned}
\tau^*(C_{\text{mix}}) &= \sum_{l=1}^{K}\pi_l^2\Big(\tau^*(C_l) + \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} h_l^2(x_1,x_2) - \sum_{x_1=0}^{\infty} f_1^2(x_1) - \sum_{x_2=0}^{\infty} f_2^2(x_2) + 1\Big) \\
&+ \sum_{l\neq m}^{K}\pi_l\pi_m\Big(Q^*_{lm} + \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} h_l(x_1,x_2)h_m(x_1,x_2) - \sum_{x_1=0}^{\infty} f_1^2(x_1) - \sum_{x_2=0}^{\infty} f_2^2(x_2) + 1\Big) \\
&- \sum_{l=1}^{K}\pi_l^2\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} h_l^2(x_1,x_2) - \sum_{l\neq m}^{K}\pi_l\pi_m\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} h_l(x_1,x_2)h_m(x_1,x_2) \\
&+ \sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - 1 \\
&= \sum_{l=1}^{K}\pi_l^2\tau^*(C_l) + \sum_{l\neq m}^{K}\pi_l\pi_m Q^*_{lm},
\end{aligned}$$

since $\sum_{l=1}^{K}\pi_l^2 + \sum_{l\neq m}^{K}\pi_l\pi_m - 1 = 0$, and hence the proof is completed. □

Theorem 5.2.4 *Let $(X_1, X_2)^\intercal$ be a bivariate integer valued random vector having dependence of finite mixture copula C_{mix} defined in (5.1), with marginal distribution functions $F_j, j = 1, 2$ and mass functions $f_j, j = 1, 2$. The population version of Spearman's rho for X_1 and X_2 is given by*

$$\rho^*(C_{mix}) = \sum_{l=1}^{K}\pi_l\rho^*(C_l), \tag{5.18}$$

where

$$\rho^*(C_l) = \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} h_l(x_1,x_2)\{6F_1(x_1-1)F_2(x_2-1) + 6(1-F_1(x_1))(1-F_2(x_2)) - 3f_1(x_1)f_2(x_2)\} + 3(\sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - 1) \tag{5.19}$$

and $h_l(x_1, x_2)$ is the joint probability mass function same as defined in (5.14).

Proof: Using the same definition we have -

$$\begin{aligned}
\rho^*(C_{\text{mix}}) &= 3(P(\text{concordance}) - P(\text{discordance})) \\
&= 3(2P(\text{concordance}) + P(\text{tie}) - 1) \\
&= 6P((X_1 - X_1^*)(X_2 - X_2^*) > 0) + 3(P(X_1 = X_1^* \cup X_2 = X_2^*) - 1) \\
&= 6P(X_1^* < X_1, X_2^* < X_2) + 6P(X_1^* > X_1, X_2^* > X_2) \\
&+ 3(P(X_1 = X_1^* \cup X_2 = X_2^*) - 1). \qquad (5.20)
\end{aligned}$$

The last expression is due to the fact that $(X_1^*, X_2^*)^\intercal$ and $(X_1, X_2)^\intercal$ have different joint distribution.

$$\begin{aligned}
&\text{Now,} \quad P(X_1^* < X_1, X_2^* < X_2) \\
&= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} P(X_1^* \leq x_1 - 1, X_2^* \leq x_2 - 1) P(X_1 = x_1, X_2 = x_2) \\
&= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} F_1(x_1 - 1) F_2(x_2 - 1) h_{\text{mix}}(x_1, x_2) \\
&= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} F_1(x_1 - 1) F_2(x_2 - 1) \Big(\sum_{l=1}^{K} \pi_l h_l(x_1, x_2) \Big) \\
&= \sum_{l=1}^{K} \pi_l \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} F_1(x_1 - 1) F_2(x_2 - 1) h_l(x_1, x_2). \qquad (5.21)
\end{aligned}$$

$$\begin{aligned}
&\text{And,} \quad P(X_1^* > X_1, X_2^* > X_2) \\
&= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} P(X_1^* > x_1, X_2^* > x_2 - 1) P(X_1 = x_1, X_2 = x_2) \\
&= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} (1 - F_1(x_1))(1 - F_2(x_2)) h_{\text{mix}}(x_1, x_2) \\
&= \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} (1 - F_1(x_1))(1 - F_2(x_2)) \Big(\sum_{l=1}^{K} \pi_l h_l(x_1, x_2) \Big) \\
&= \sum_{l=1}^{K} \pi_l \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} (1 - F_1(x_1))(1 - F_2(x_2)) h_l(x_1, x_2). \qquad (5.22)
\end{aligned}$$

$$
\begin{aligned}
&\text{And, } P(X_1 = X_1^* \cup X_2 = X_2^*) \\
&= P(X_1 = X_1^*) + P(X_2 = X_2^*) - P(X_1 = X_1^*, X_2 = X_2^*) \\
&= \sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} f_1(x_1) f_2(x_2) h_{\text{mix}}(x_1, x_2) \\
&= \sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - \sum_{l=1}^{K} \pi_l \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} f_1(x_1) f_2(x_2) h_l(x_1, x_2). \qquad (5.23)
\end{aligned}
$$

$$
\begin{aligned}
\text{Let,} \quad \rho^*(C_l) = \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} h_l(x_1, x_2)\{&6F_1(x_1 - 1)F_2(x_2 - 1) + 6(1 - F_1(x_1))(1 - F_2(x_2)) \\
&- 3f_1(x_1)f_2(x_2)\} + 3\Big(\sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - 1\Big).
\end{aligned}
$$

Therefore, using (5.21), (5.22) and (5.23) in (5.20) we get -

$$
\begin{aligned}
\rho^*(C_{\text{mix}}) &= \sum_{l=1}^{K} \pi_l \Big(\rho^*(C_l) + 3 \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} f_1(x_1) f_2(x_2) h_l(x_1, x_2) \qquad (5.24) \\
&\quad - 3\Big(\sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - 1 \Big) \\
&\quad + 3\Big(\sum_{x_1=0}^{\infty} f_1^2(x_1) + \sum_{x_2=0}^{\infty} f_2^2(x_2) - 1 - \sum_{l=1}^{K} \pi_l \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} f_1(x_1) f_2(x_2) h_l(x_1, x_2) \Big) \\
&= \sum_{l=1}^{K} \pi_l \rho^*(C_l),
\end{aligned}
$$

and hence the proof is completed. □

The expressions in (5.5), (5.9), (5.11), and (5.18) reveal that even if the component copulas imply independence, the resulting mixture may still imply dependence. To visually illustrate this phenomenon, we provide plots of Kendall's tau and Spearman's rho for various elliptical mixture copulas and different marginal distributions in Figures 5.1, 5.2, 5.3, and 5.4. These plots offer insight into how the choice of copulas and marginal distributions can impact the resulting dependence measures, highlighting the intricate interplay between different components in mixture copula models. We consider $K = 2$ component mixture and in the first component copula we set the correlation parameter $\rho = 0$ (i.e. independence copula). The correlation parameter of the second copula component is varied over $[-1, 1]$. When the marginal distributions are Poisson with identical parameters for each marginal, we observe that for values greater than 10, the association with the values of Kendall's tau and Spearman's rho becomes negligible. Similarly, when the marginal distributions are Bernoulli with the same parameter for each marginal, the association with the values of Kendall's tau and Spearman's rho is minimized when

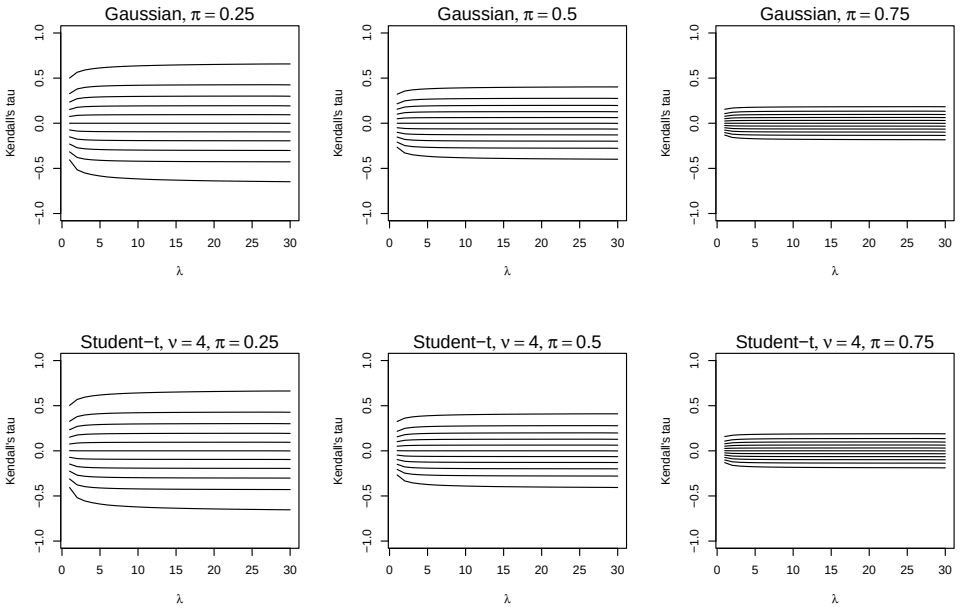

Figure 5.1 Kendall's tau values computed using Gaussian and Student-t ($\nu = 4$) mixture copulas with different mixing proportions and Poisson marginal distributions with the same location parameter $\lambda = 1, \ldots, 30$. Higher curves corresponding to higher values of the copula parameter.

the proportion of success is $p = 0.5$. It's worth noting that elliptical mixture copulas are symmetric by construction, allowing for both negative and positive dependence. As the mixing proportion with the independence copula increases, the values of Kendall's tau and Spearman's rho tend to shrink towards zero, reflecting a decrease in dependence strength.

Finally, let's discuss tail dependence in a mixture copula. Tail dependence measures how much two variables are related in extreme situations. We'll focus on bivariate tail dependence here, but there are also more complex versions discussed in the literature (see, [27]). The following Theorem states the tail behavior of bivariate mixture copulas.

Theorem 5.2.5 *Let $\lambda_U(C_l)$ and $\lambda_L(C_L)$ be the tail dependence coefficients of the component copula C_l, provided these exists and π_l be the mixing proportion for $l = 1, \ldots, K$. Then the upper and lower tail dependence coefficients are given as*

$$\lambda_U(C_{mix}) = \sum_{l=1}^{K} \pi_l \lambda_U(C_l) \quad \textit{and} \quad \lambda_L(C_{mix}) = \sum_{l=1}^{K} \pi_l \lambda_L(C_l), \quad \textit{respectively}. \tag{5.25}$$

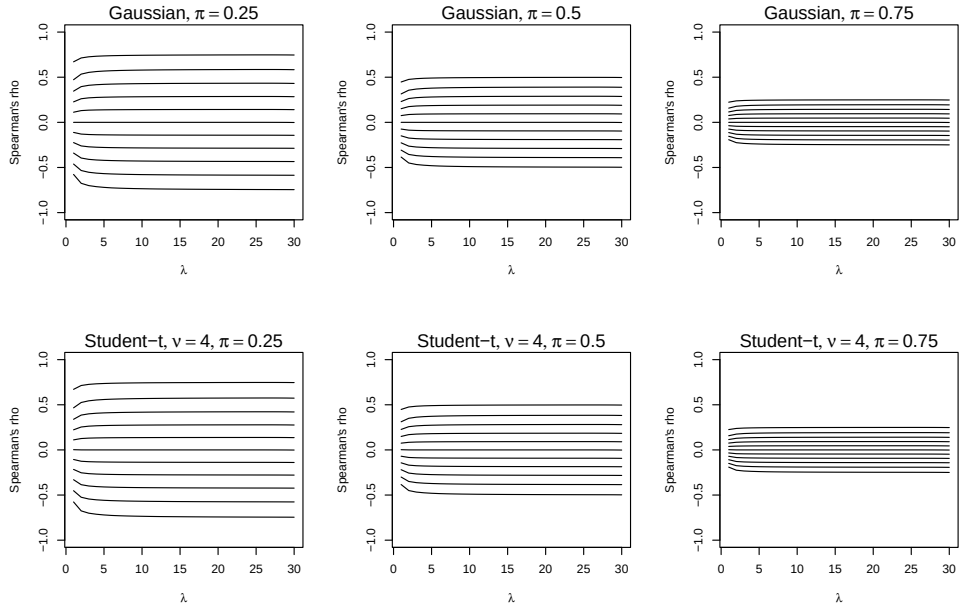

Figure 5.2 Spearman's rho values computed using Gaussian and Student-t ($\nu = 4$) mixture copulas with different mixing proportions and Poisson marginal distributions with the same location parameter $\lambda = 1, \ldots, 30$. Higher curves corresponding to higher values of the copula parameter.

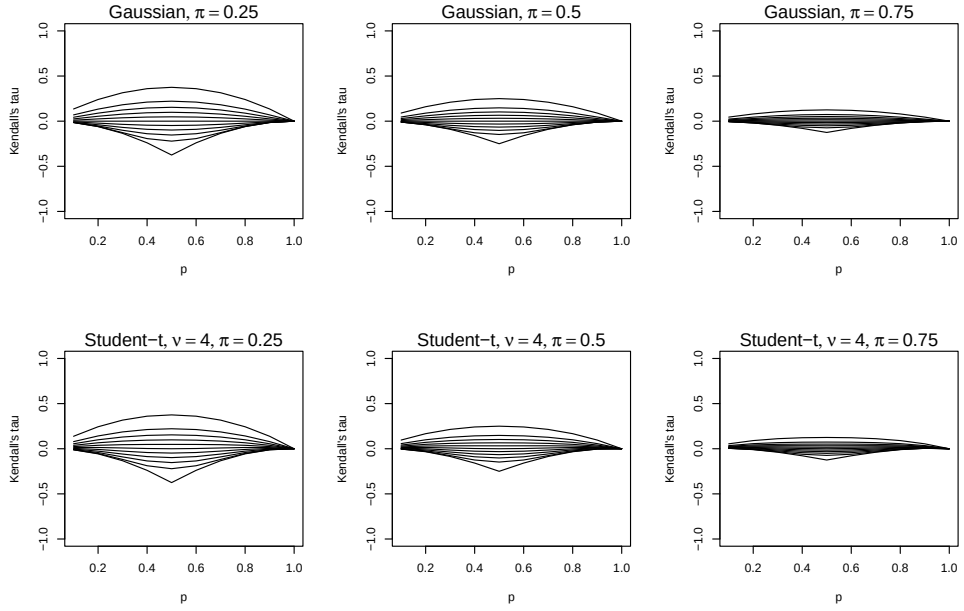

Figure 5.3 Kendall's tau values computed using Gaussian and Student-t ($\nu = 4$) mixture copulas with different mixing proportions and Bernoulli marginal distributions with the same location parameter $p = 0.1, \ldots, 1.0$. Higher curves corresponding to higher values of the copula parameter.

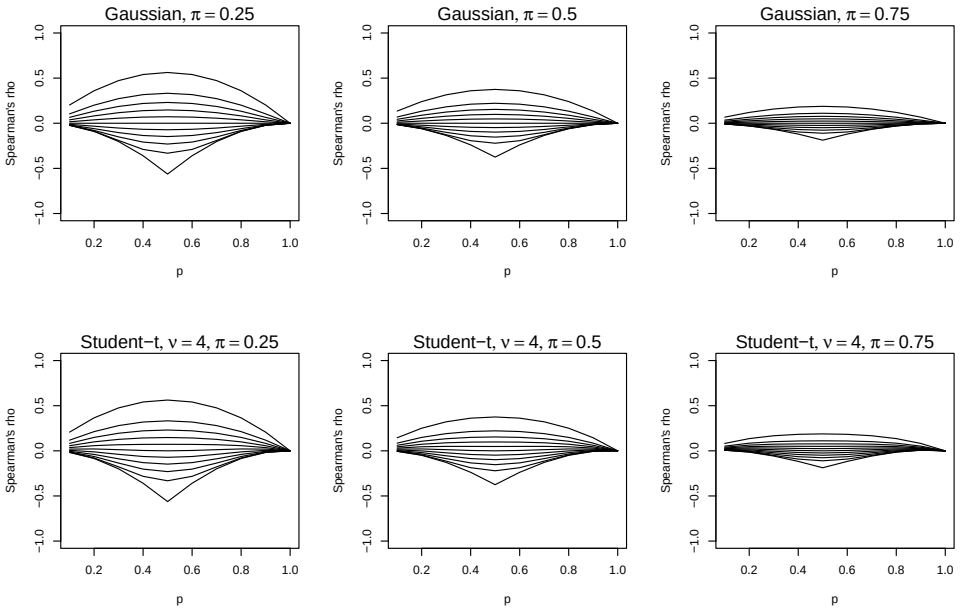

Figure 5.4 Spearman's rho values computed using Gaussian and Student-t ($\nu = 4$) mixture copulas with different mixing proportions and Bernoulli marginal distributions with the same location parameter $p = 0.1, \ldots, 1.0$. Higher curves corresponding to higher values of the copula parameter.

Proof: Straight from the definition we have -

$$\begin{aligned}\lambda_U(C_{\text{mix}}) &= \lim_{u\to 1-} \frac{1-2u+C_{\text{mix}}(u,u)}{1-u} = \lim_{u\to 1-} \frac{1-2u+\sum_{l=1}^{K}\pi_l C_l(u,u)}{1-u} \\ &= \sum_{l=1}^{K}\pi_l \lim_{u\to 1-} \frac{1-2u+C_l(u,u)}{1-u} = \sum_{l=1}^{K}\pi_l \lambda_U(C_l).\end{aligned}$$

$$\begin{aligned}\lambda_L(C_{\text{mix}}) &= \lim_{u\to 0+} \frac{C_{\text{mix}}(u,u)}{u} = \lim_{u\to 0+} \frac{\sum_{l=1}^{K}\pi_l C_l(u,u)}{u} \\ &= \sum_{l=1}^{K}\pi_l \lim_{u\to 0+} \frac{C_l(u,u)}{u} = \sum_{l=1}^{K}\pi_l \lambda_L(C_l). \quad \square\end{aligned}$$

It's straightforward to see that a mixture of Gaussian copulas exhibits zero tail dependence. In the case of a mixture of bivariate Student-t copulas, both upper and lower tail dependence coefficients are identical (see, [133]), i.e.

$$\lambda(C_{\text{mix}}) = 2\sum_{l=1}^{K}\pi_l T\left(-\frac{\sqrt{(\nu+1)(1-\rho_l)}}{\sqrt{1+\rho_l}}\middle|\nu\right). \tag{5.26}$$

5.3 Modeling longitudinal count data

Suppose that $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\intercal$ be a non-negative integer valued dependent random vector for the i-th subject, where Y_{ij} denotes the observation at time t_{ij}. Let $\mathbf{x}_{ij}$ be a p-dimensional vector of covariates observed for the i-th subject at time t_{ij} and β is the $p \times 1$ vector of regression parameters. In the longitudinal set-up, the components of the vector $\mathbf{Y}_i$ are repeated responses, which are likely to be correlated. Since the true correlation structure of these repeated measurements always remains unknown, copula framework is very useful to analyse and investigate the temporal association. Since for discrete random variables the copula is uniquely identified in in the product range of the marginals, we assume restrictively that the marginal models for the repeated responses are correctly specified. The two most commonly applied distributions for count responses are Poisson and Negative Binomial distributions. Poisson regression assumes that the repeated count responses marginally follows a Poisson distribution with CDF

$$F_{ij}(y_{ij}|\mathbf{x}_{ij}, \beta) = \sum_{k=0}^{y_{ij}} \frac{e^{-\mu_{ij}} \mu_{ij}^k}{k!}, \quad j = 1, \ldots, n_i, \tag{5.27}$$

where $E(Y_{ij}) = Var(Y_{ij}) = \mu_{ij} = \exp(\mathbf{x}_{ij}\beta)$. Since poisson distribution doesn't allow for over dispersion in longitudinal count data, we also consider Negative Binomial regression model for the count responses. The CDF of a Negative Binomial marginal can be written as

$$F_{ij}(y_{ij}|\mathbf{x}_{ij}, \beta) = \sum_{k=0}^{y_{ij}} \left(\begin{array}{c} \psi + k - 1 \\ k \end{array} \right) \frac{\psi^\psi \mu_{ij}^k}{(\psi + \mu_{ij})^{\psi+k}}, \quad j = 1, \ldots, n_i, \tag{5.28}$$

where $E(Y_{ij}) = \mu_{ij} = \exp(\mathbf{x}_{ij}\beta)$ and $Var(Y_{ij}) = \mu_{ij} + \mu_{ij}^2/\psi$. The parameter ψ, accounts for the overdispersion in the data. It should be noted that the set of non-negative integers is the support of both these distributions.

Once the marginal distributions have been determined using standard exploratory analysis, an appropriate multivariate copula can be selected to analyze the dependence between repeated measurements. Our approach, employing a mixture of elliptical copulas, offers the flexibility to incorporate different correlation structures into each copula component. This allows us to interpret the data by revealing which correlation structures are present and in what proportion they contribute to the overall dependence structure of the dataset. Common choices for modeling serial dependence in longitudinal data are auto-regressive (AR(1)), moving average (MA(1)) or exchangeable (EX) correlation structures. Under the AR(1) structure, the correlation between the errors on a subject decline exponentially with the distance between the observations. In terms of the correlation matrix $\mathbf{\Sigma}$ in the elliptical copulas, it can be expressed as

$$\rho_{jk} = \exp(-\xi|t_j - t_k|), \xi > 0, \quad 1 \leq j < k \leq n_i, \tag{5.29}$$

where $\xi > 0$ is the auto-regressive parameter. Under the exchangeable structure, the correlation between the errors remains constant and time invariant. Thus, considering elliptical copulas our methods can be extended to the unbalanced longitudinal data where the number of measurements per subject can be different. In the next section we describe the two stage estimation method for our mixture copula based models designed for longitudinal count data.

5.4 Parameter estimation

One of the earliest approaches to estimating Gaussian and Student-t mixture models can be found in [135]. Since the count repeated measurements are discrete, the joint probability mass function of $\mathbf{Y}_i$ involves a $2n_i$-times folded sum, which is difficult to evaluate in practice (see, [8] or [140]). Since for multivariate elliptical copulas all parameters can be identified from its lower dimensional sub-copulas, same holds true for their finite mixtures. This can be shown algebraically from the construction of mixture distributions. To circumvent the computational issues composite likelihood methods (CML) can be employed for the considered class of models. With this method, pseudolikelihoods are constructed and then maximized using the consecutive pairs of observations, and hence also called as pairwise likelihood estimation ([141]). Recent overviews on composite likelihood estimation can be found in [142] and [87]. In contrast to a full likelihood approach, the composite likelihood formulations only requires the specification of bivariate pairs of observations.

To streamline the computational burden associated with estimating the model parameters, we adopt the two-stage estimation procedure of the composite likelihood, as outlined in [84] and [143]. This method, tailored to the copula structure of the dependence models, yields consistent estimates of the model parameters along with their corresponding robust standard errors. Let $f_{ij}(y_{ij}|\theta_{ij})$ denote the marginal probability mass function of the response variable Y_{ij}, and let θ represent the vector of parameters corresponding to the marginals. Then in the first stage, under working independence assumption we estimate the marginal parameters by maximizing

$$l(\theta|\mathbf{y},\mathbf{x}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log f_{ij}(y_{ij}|\theta_{ij}), \tag{5.30}$$

and then using the parameter estimates $\hat{\theta}_{ij}$ from (5.30), we compute the uniform samples: $u_{ij} = F_{ij}(y_{ij}|\hat{\theta}_{ij})$, $u_{ij}^{-} = F_{ij}(y_{ij}-1|\hat{\theta}_{ij})$ where $i = 1,\ldots,m$, $j = 1,\ldots,n_i$. In the second stage, the

estimates $\{u_{ij}, u_{ij}^{-}\}$ are inserted into composite likelihood of the form -

$$
\begin{aligned}
l_c(\eta|\mathbf{u}) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i-1}\sum_{k=j+1}^{n_i} \log P(U_{ij}=u_{ij}, U_{ik}=u_{ik}|\eta_{jk}) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i-1}\sum_{k=j+1}^{n_i} \log\Big[C_{\text{mix},2}(u_{ij},u_{ik}|\eta_{jk}) - C_{\text{mix},2}(u_{ij}^{-},u_{ik}|\eta_{jk}) \\
&- C_{\text{mix},2}(u_{ij},u_{ik}^{-}|\eta_{jk}) + C_{\text{mix},2}(u_{ij}^{-},u_{ik}^{-}|\eta_{jk})\Big] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i-1}\sum_{k=j+1}^{n_i} \log\Big[\sum_{l=1}^{K}\pi_l\Big(C_{l,2}(u_{ij},u_{ik}|\phi_{l,jk}) - C_{l,2}(u_{ij}^{-},u_{ik}|\phi_{l,jk}) \\
&- C_{l,2}(u_{ij},u_{ik}^{-}|\phi_{l,jk}) + C_{l,2}(u_{ij}^{-},u_{ik}^{-}|\phi_{l,jk})\Big)\Big], \qquad (5.31)
\end{aligned}
$$

and then maximized with respect to the set of parameters η, to obtain the estimates of the association parameters. [91] employed a similar estimation strategy to model multilevel insurance claim data using Gaussian copulas. While it's common practice to estimate model parameters of mixture distributions through EM-type algorithms, we opt for numerical maximization of the objective functions in (5.30) and (5.31) with box constraints (see, [144]). This method, alternatively known as the inference function of margins, offers an effective approach to parameter estimation in our context. The standard errors of the parameter estimates $\hat{\theta}^* = (\hat{\theta}, \hat{\eta})^\intercal$ can be numerically obtained from the estimated sandwich information matrix (Godambe information matrix) of the form;

$$
J(\hat{\theta}^*) = D(\hat{\theta}^*)^\intercal M(\hat{\theta}^*)^{-1} D(\hat{\theta}^*), \qquad (5.32)
$$

where $D(\hat{\theta}^*)$ is a block diagonal matrix and $M(\hat{\theta}^*)$ is a symmetric positive definite matrix. The explicit forms of these matrices can be found in [84] or [28]. To estimate the parameters we use *optim* function ([64]), and to estimate the information matrix associted with the parameter estimates we use *numderiv* function ([65]) in R.

5.5 Model Validation

The validation of copula based regression models has been widely discussed in the literature (see, [78] or [69]). To determine which model is best suited for a data set, standard model selection techniques such as AIC (Akaike information criterion) and BIC (Bayesian information criterion) are often used. [145] and [146] modified these two in the case of composite likelihood estimation. These are defined as -

$$
\begin{aligned}
CLAIC &= -2l_c(\hat{\theta}^*) + 2tr(M(\hat{\theta}^*)D(\hat{\theta}^*)^{-1}), \\
CLBIC &= -2l_c(\hat{\theta}^*) + \log(m)tr(M(\hat{\theta}^*)D(\hat{\theta}^*)^{-1}). \qquad (5.33)
\end{aligned}
$$

In our applications, we prefer models with smaller values of these criteria. Since we consider mixture of elliptical copulas to model temporal dependencies of count longitudinal data, additionally we can validate the model assumptions utilizing the t-plot method by [90] and [147].

We propose a straightforward adaptation of the t-plot method for mixture of elliptical copulas, leveraging the definition of the mixture distribution. Originally designed to test the elliptical symmetry of multivariate distributions [148] using invariant statistics under orthogonal transformations, this method has proven effective in validating the elliptical assumption of underlying copulas for unbalanced longitudinal count data [13]. The null hypothesis of a t-plot is that a sample is drawn from an elliptical multivariate distribution. This hypothesis can also be extended and tested for mixture of elliptical distributions. The procedure is outlined as follows -

- For each unit i, transform the count variables on uniform scale by $\hat{u}_{ij} = F_{ij}(y_{ij}|\hat{\theta}_{ij})$ for $j = 1, \ldots, n_i$. Where $\hat{\theta}_{ij}$ is the estimated marginal parameter from the first stage. Now the transformed data can be considered as a realization of the mixture of elliptical copulas.

- Compute the quantiles of $\hat{u}_{ij}$ by $\hat{z}_{ij} = H_j^{-1}(\hat{u}_{ij})$, where $H_j(.)$ denotes the CDF of the j-th marginal associated with the elliptical copula. If the copula is well specified, then the vector $\hat{\mathbf{z}}_i = (\hat{z}_{i1}, \ldots, \hat{z}_{in_i})^\intercal$ would follow a mixture of elliptical distributions.

- Let d_{il} be an indicator representing whether the i-th unit comes from the l-th component of the finite mixture of elliptical distributions as

$$F_{\text{mix},n_i} = \sum_{l=1}^{K} \pi_l F_{l,n_i}(.|\mathbf{\Sigma}_l), \quad \sum_{l=1}^{K} \pi_l = 1, \pi_l \geq 0, \quad \forall\, l = 1, \ldots, K, \tag{5.34}$$

  and w_{il} be the expected value of d_{il}. This can also be interpreted as the posterior probability of i-th unit belonging to the l-th component of the mixture. Hence we estimate w_{il} by

$$\hat{w}_{il} = \frac{\hat{\pi}_l f_{l,n_i}(\hat{\mathbf{z}}_i|\hat{\mathbf{\Sigma}}_l)}{\sum_{l=1}^{K} \hat{\pi}_l f_{l,n_i}(\hat{\mathbf{z}}_i|\hat{\mathbf{\Sigma}}_l)}, \quad \forall\, l = 1, \ldots, K, \tag{5.35}$$

  where $\hat{\mathbf{\Sigma}}_l$ is the estimated value of $\mathbf{\Sigma}_l$. Now the i-th unit is most likely to be coming from the l-th component of the mixture if $\hat{w}_{il} = \max\{\hat{w}_{i1}, \ldots, \hat{w}_{iK}\}$.

- After identifying the l-th component distribution for the i-th unit, we calculate the vector $\hat{\mathbf{z}}_i^* = \hat{\mathbf{\Sigma}}_l^{-1/2}\hat{\mathbf{z}}_i$ and construct the t-statistic as

$$t_i(\hat{\mathbf{z}}_i^*) = \frac{\sqrt{n_i}\bar{\hat{z}}_i^*}{\sqrt{(n_i-1)^{-1}\sum_{j=1}^{n_i}(\hat{z}_{ij}^* - \bar{\hat{z}}_i^*)^2}}, \tag{5.36}$$

  where $\bar{\hat{z}}_i^* = n_i^{-1}\sum_{j=1}^{n_i}\hat{z}_{ij}^*$. Thus $t_i(\hat{\mathbf{z}}_i^*)$ should be from a standard t-distribution with $n_i - 1$ degrees of freedom.

- Repeat the above procedure for all units in the sample and define the transformed variable $v_i = T_1(t_i(\hat{\mathbf{z}}_i^*)|n_i - 1)$ for $i = 1, \ldots, m$. If the copula captures the dependence structure properly, then $\mathbf{v} = (v_1, \ldots, v_m)^\intercal$ should be a random sample from $U(0, 1)$ distribution. Therefore, we plot the sample quantiles of $\mathbf{v}$ against the theoretical quantiles to graphically visualize the goodness-of-fit of the model.

5.6 Simulation studies

In this section, we conduct simulations to evaluate the finite sample performance of the proposed mixture copula models tailored for count longitudinal data. Consistent with our approach in both simulations and applications, we set the number of components of the mixture to $K = 2$. We consider two marginal distributions, as discussed in Section 5.3. The specifications of the marginal models are provided as follows -

$$\mu_{ij} = \exp(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + t_{ij}\beta_3), \quad j = 1, \ldots, 4, \tag{5.37}$$

where for each unit (denoted by $i, i = 1, \ldots, m$), the number of observations $n_i = 4$. We assign the same values of the regression coefficients for both marginal distributions: $\beta_0 = 1.0$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, $\beta_3 = -0.5$. Additionally, we set the overdispersion parameter for Negative Binomial marginals to $\psi = 4.0$. The fixed covariates are generated as follows: $x_{i1} \sim Ber(p = 0.5)$, $x_{i2} \sim$ dUnif$(1, \ldots, 4)$ (discrete uniform distribution), and the time points $t_{ij} = j$ for $j = 1, \ldots, 4$, $i = 1, \ldots, m$, respectively. We consider two different sample sizes in our simulations: $m = 200$ and 500. For the mixture copula models, we designate one component in the mixture of the two copulas as autoregressive and other as exchangeable (parameterized as $\rho = \exp(-\xi)$). The parameters for each multivariate copula are set as $\xi_1 = 0.3$ (under AR(1) structure) and $\xi_2 = 0.7$ (under EX structure). Furthermore, we explore three choices for the mixing proportions: $\pi = \{0.25, 0.50, 0.75\}$. This parameter allows us to quantify the extent of presence of each of the two correlation structures considered here. This 3-set parameterization enables a broad range of dependence in the class of models considered. We generate the datasets using standard probability transformation methods and estimate the model parameters using the two-stage method discussed in Section 5.4. We repeat this entire process for $N = 500$ times and report our findings accordingly.

Here are the summaries of the simulation results for the considered models, presented in Tables 5.1, 5.2, 5.3, and 5.4. Within each table, we provide the mean, biases $[\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_j^* - \theta^*)]$, empirical standard deviations (denoted as SD), average standard errors obtained from the asymptotic covariance matrices (denoted as SE), and roots of mean square errors $[\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_j^* - \theta^*)^2}]$, where $\hat{\theta}_j^*$ represents the parameter estimates for the j-th sample. Across both sample sizes, the average estimates are very close to the corresponding true parameters. The results demonstrate consistent performance of the proposed models with the two-stage composite likelihood estimation, as biases and roots of mean square errors decrease with increasing sample size. The average of the asymptotic standard errors of the parameter es-

		m = 200					**m** = 500				
Parameters	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
π	0.25	0.2555	0.0055	0.1132	0.1333	0.1133	0.2549	0.0049	0.0759	0.0843	0.0761
β_0	1.0	0.9964	-0.0036	0.0893	0.0856	0.0894	1.0006	0.0006	0.0559	0.0543	0.0559
β_1	0.5	0.5011	0.0011	0.0525	0.0506	0.0525	0.5008	0.0008	0.0321	0.0323	0.0322
β_2	0.5	0.5006	0.0006	0.0257	0.0242	0.0257	0.4993	-0.0007	0.0160	0.0154	0.0160
β_3	-0.5	-0.5004	-0.0004	0.0122	0.0132	0.0122	-0.5005	-0.0005	0.0081	0.0084	0.0081
ξ_1	0.3	0.3570	0.0570	0.1687	0.1478	0.1781	0.3302	0.0302	0.1156	0.0958	0.1195
ξ_2	0.7	0.6935	-0.0065	0.1018	0.1075	0.1020	0.6934	-0.0066	0.0755	0.0737	0.0757
π	0.50	0.4884	-0.0116	0.1182	0.1289	0.1188	0.4929	-0.0071	0.0816	0.0819	0.0819
β_0	1.0	0.9966	-0.0034	0.0879	0.0878	0.0880	1.0016	0.0016	0.0544	0.0554	0.0545
β_1	0.5	0.5027	0.0027	0.0526	0.0517	0.0527	0.5017	0.0017	0.0342	0.0328	0.0343
β_2	0.5	0.5003	0.0003	0.0254	0.0249	0.0254	0.4989	-0.0011	0.0155	0.0157	0.0156
β_3	-0.5	-0.5003	-0.0003	0.0135	0.0136	0.0135	-0.4996	0.0004	0.0087	0.0086	0.0087
ξ_1	0.3	0.3179	0.0179	0.0969	0.0721	0.0985	0.3109	0.0109	0.0593	0.0478	0.0603
ξ_2	0.7	0.6862	-0.0138	0.1411	0.1404	0.1417	0.6963	-0.0037	0.1035	0.0994	0.1036
π	0.75	0.7247	-0.0253	0.1119	0.1243	0.1147	0.7502	0.0002	0.0687	0.0772	0.0687
β_0	1.0	1.0023	0.0023	0.0900	0.0888	0.0901	1.0004	0.0004	0.0586	0.0566	0.0586
β_1	0.5	0.4985	-0.0015	0.0531	0.0527	0.0531	0.4997	-0.0003	0.0351	0.0335	0.0352
β_2	0.5	0.4988	-0.0012	0.0253	0.0252	0.0253	0.4998	-0.0002	0.0162	0.0160	0.0162
β_3	-0.5	-0.4993	0.0007	0.0151	0.0139	0.0151	-0.5001	-0.0001	0.0090	0.0089	0.0090
ξ_1	0.3	0.3085	0.0085	0.0532	0.0502	0.0539	0.3077	0.0077	0.0348	0.0302	0.0356
ξ_2	0.7	0.6607	-0.0393	0.1775	0.2173	0.1818	0.6844	-0.0156	0.1434	0.1654	0.1443

Table 5.1 Parameter estimation using two stage composite likelihood method for Gaussian mixture copula model with Poisson marginals for $N = 500$ simulated data sets.

		m = 200					**m** = 500				
Parameters	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
π	0.25	0.2695	0.0195	0.1233	0.1545	0.1249	0.2495	-0.0005	0.0865	0.0960	0.0865
β_0	1.0	0.9958	-0.0042	0.0859	0.0857	0.0860	0.9995	-0.0005	0.0536	0.0541	0.0536
β_1	0.5	0.4981	-0.0019	0.0518	0.0504	0.0518	0.5040	0.0040	0.0319	0.0320	0.0321
β_2	0.5	0.5018	0.0018	0.0236	0.0242	0.0237	0.4990	-0.0010	0.0150	0.0153	0.0150
β_3	-0.5	-0.5004	-0.0004	0.0127	0.0132	0.0127	-0.4997	0.0003	0.0081	0.0084	0.0081
ξ_1	0.3	0.3525	0.0525	0.1959	0.2016	0.2028	0.3232	0.0232	0.1353	0.1457	0.1373
ξ_2	0.7	0.7006	0.0006	0.1291	0.1580	0.1291	0.7047	0.0047	0.1044	0.1166	0.1045
π	0.50	0.5003	0.0003	0.1329	0.1479	0.1329	0.4937	-0.0063	0.0886	0.0920	0.0888
β_0	1.0	0.9998	-0.0002	0.0858	0.0812	0.0858	1.0019	0.0019	0.0578	0.0554	0.0578
β_1	0.5	0.5020	0.0020	0.0516	0.0516	0.0517	0.4968	-0.0032	0.0338	0.0328	0.0340
β_2	0.5	0.4992	-0.0008	0.0240	0.0245	0.0240	0.4998	-0.0002	0.0159	0.0157	0.0159
β_3	-0.5	-0.4994	0.0006	0.0137	0.0136	0.0137	-0.4996	0.0004	0.0082	0.0086	0.0082
ξ_1	0.3	0.3237	0.0237	0.1249	0.1144	0.1271	0.3055	0.0055	0.0682	0.0699	0.0684
ξ_2	0.7	0.6899	-0.0101	0.1666	0.2288	0.1669	0.7042	0.0042	0.1245	0.1580	0.1246
π	0.75	0.7222	-0.0278	0.1223	0.1546	0.1255	0.7320	-0.0180	0.0841	0.0928	0.0860
β_0	1.0	0.9967	-0.0033	0.0942	0.0889	0.0943	1.0041	0.0041	0.0575	0.0563	0.0576
β_1	0.5	0.5033	0.0033	0.0521	0.0524	0.0522	0.4989	-0.0011	0.0318	0.0333	0.0318
β_2	0.5	0.5006	0.0006	0.0260	0.0250	0.0260	0.4988	-0.0012	0.0162	0.0159	0.0162
β_3	-0.5	-0.5010	-0.0010	0.0149	0.0140	0.0149	-0.4996	0.0004	0.0089	0.0089	0.0089
ξ_1	0.3	0.3119	0.0119	0.0627	0.0891	0.0638	0.3057	0.0057	0.0434	0.0512	0.0438
ξ_2	0.7	0.6617	-0.0383	0.1894	0.2880	0.1932	0.6714	-0.0286	0.1535	0.2145	0.1561

Table 5.2 Parameter estimation using two stage composite likelihood method for Student-t ($\nu = 4$) mixture copula model with Poisson marginals for $N = 500$ simulated data sets.

		m = 200					**m** = 500				
Parameters	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
π	0.25	0.2611	0.0111	0.1188	0.1314	0.1193	0.2466	-0.0034	0.0775	0.0832	0.0775
β_0	1.0	1.0117	0.0117	0.1153	0.1212	0.1159	0.9975	-0.0025	0.0771	0.0774	0.0771
β_1	0.5	0.4933	-0.0067	0.0800	0.0808	0.0802	0.5034	0.0034	0.0517	0.0514	0.0518
β_2	0.5	0.4985	-0.0015	0.0366	0.0370	0.0366	0.4992	-0.0008	0.0230	0.0236	0.0230
β_3	-0.5	-0.5017	-0.0017	0.0189	0.0186	0.0190	-0.4997	0.0003	0.0125	0.0117	0.0125
ψ	4.0	4.1369	0.1369	0.5739	0.5530	0.5900	4.0512	0.0512	0.3504	0.3434	0.3541
ξ_1	0.3	0.3555	0.0555	0.1738	0.1348	0.1825	0.3247	0.0247	0.1217	0.0953	0.1242
ξ_2	0.7	0.6994	-0.0006	0.1124	0.1088	0.1124	0.7006	0.0006	0.0785	0.0721	0.0785
π	0.50	0.4910	-0.0090	0.1191	0.1297	0.1195	0.4926	-0.0074	0.0753	0.0813	0.0757
β_0	1.0	1.0040	0.0040	0.1243	0.1236	0.1243	0.9998	-0.0002	0.0788	0.0789	0.0788
β_1	0.5	0.4912	-0.0088	0.0824	0.0822	0.0829	0.5015	0.0015	0.0529	0.0524	0.0529
β_2	0.5	0.5001	0.0001	0.0370	0.0376	0.0370	0.4995	-0.0005	0.0237	0.0240	0.0237
β_3	-0.5	-0.5014	-0.0014	0.0194	0.0192	0.0195	-0.4995	0.0005	0.0127	0.0122	0.0127
ψ	4.0	4.2174	0.2174	0.6299	0.5866	0.6663	4.0615	0.0615	0.3463	0.3565	0.3518
ξ_1	0.3	0.3315	0.0315	0.1139	0.0783	0.1181	0.3087	0.0087	0.0617	0.0488	0.0623
ξ_2	0.7	0.6858	-0.0142	0.1448	0.1460	0.1455	0.7008	0.0008	0.1054	0.1008	0.1054
π	0.75	0.7221	-0.0279	0.1174	0.1228	0.1207	0.7422	-0.0078	0.0745	0.0777	0.0749
β_0	1.0	1.0007	0.0007	0.1322	0.1267	0.1322	0.9969	-0.0031	0.0827	0.0803	0.0828
β_1	0.5	0.4999	-0.0001	0.0832	0.0840	0.0832	0.5005	0.0005	0.0507	0.0533	0.0507
β_2	0.5	0.4980	-0.0020	0.0409	0.0383	0.0409	0.5002	0.0002	0.0260	0.0245	0.0260
β_3	-0.5	-0.5010	-0.0010	0.0191	0.0198	0.0191	-0.4995	0.0005	0.0124	0.0125	0.0124
ψ	4.0	4.1623	0.1623	0.6292	0.5924	0.6498	4.0639	0.0639	0.3554	0.3677	0.3611
ξ_1	0.3	0.3207	0.0207	0.0616	0.0528	0.0650	0.3080	0.0080	0.0401	0.0308	0.0409
ξ_2	0.7	0.6518	-0.0482	0.1770	0.2146	0.1834	0.6816	-0.0184	0.1456	0.1524	0.1467

Table 5.3 Parameter estimation using two stage composite likelihood method for Gaussian mixture copula model with Negative Binomial marginals for $N = 500$ simulated data sets.

		m = 200					**m** = 500				
Parameters	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
π	0.25	0.2651	0.0151	0.1189	0.1413	0.1199	0.2500	0.0000	0.0823	0.0947	0.0823
β_0	1.0	0.9921	-0.0079	0.1229	0.1221	0.1231	0.9956	-0.0044	0.0812	0.0776	0.0814
β_1	0.5	0.4983	-0.0017	0.0851	0.0811	0.0851	0.5018	0.0018	0.0540	0.0515	0.0541
β_2	0.5	0.5012	0.0012	0.0385	0.0372	0.0385	0.5006	0.0006	0.0244	0.0237	0.0244
β_3	-0.5	-0.4991	0.0009	0.0179	0.0184	0.0180	-0.5000	0.0000	0.0121	0.0117	0.0121
ψ	4.0	4.1597	0.1597	0.6878	0.6176	0.7061	4.0586	0.0586	0.3596	0.3803	0.3643
ξ_1	0.3	0.3580	0.0580	0.1988	0.2019	0.2071	0.3168	0.0168	0.1362	0.1314	0.1372
ξ_2	0.7	0.7010	0.0010	0.1237	0.1486	0.1237	0.7053	0.0053	0.0947	0.1018	0.0949
π	0.50	0.5005	0.0005	0.1375	0.1444	0.1375	0.4974	-0.0026	0.0906	0.0909	0.0906
β_0	1.0	0.9911	-0.0089	0.1305	0.1242	0.1308	1.0069	0.0069	0.0811	0.0790	0.0814
β_1	0.5	0.5044	0.0044	0.0836	0.0824	0.0837	0.4966	-0.0034	0.0509	0.0523	0.0510
β_2	0.5	0.5000	0.0000	0.0386	0.0377	0.0386	0.4988	-0.0012	0.0244	0.0241	0.0244
β_3	-0.5	-0.4996	0.0004	0.0184	0.0191	0.0184	-0.5007	-0.0007	0.0119	0.0121	0.0119
ψ	4.0	4.1899	0.1899	0.6372	0.6316	0.6649	4.0948	0.0948	0.3862	0.3930	0.3977
ξ_1	0.3	0.3313	0.0313	0.1208	0.1068	0.1247	0.3120	0.0120	0.0756	0.0692	0.0765
ξ_2	0.7	0.6936	-0.0064	0.1579	0.2037	0.1580	0.7022	0.0022	0.1310	0.1519	0.1310
π	0.75	0.7192	-0.0308	0.1232	0.1462	0.1270	0.7490	-0.0010	0.0748	0.0950	0.0748
β_0	1.0	0.9916	-0.0084	0.1266	0.1258	0.1266	0.9999	-0.0001	0.0782	0.0802	0.0782
β_1	0.5	0.5005	0.0005	0.0793	0.0836	0.0793	0.5002	0.0002	0.0534	0.0534	0.0534
β_2	0.5	0.5023	0.0023	0.0392	0.0384	0.0392	0.5005	0.0005	0.0241	0.0244	0.0241
β_3	-0.5	-0.4999	0.0001	0.0202	0.0196	0.0202	-0.5005	-0.0005	0.0123	0.0125	0.0123
ψ	4.0	4.2224	0.2224	0.6987	0.6535	0.7332	4.0906	0.0906	0.4264	0.4002	0.4359
ξ_1	0.3	0.3180	0.0180	0.0711	0.0824	0.0733	0.3084	0.0084	0.0437	0.0540	0.0445
ξ_2	0.7	0.6743	-0.0257	0.1780	0.2278	0.1799	0.6863	-0.0137	0.1464	0.1606	0.1470

Table 5.4 Parameter estimation using two stage composite likelihood method for Student-t ($\nu = 4$) mixture copula model with Negative Binomial marginals for $N = 500$ simulated data sets.

timates (SE) is comparable with the empirical standard deviation (SD) of point estimates, indicating the accuracy of the uncertainty estimates. Notably, the standard errors of the regression parameter estimates are larger in the Negative Binomial-based models than in the Poisson-based models. Additionally, as the mixing proportion increases for a mixture component, the corresponding bias and RMSE decrease. It is may be to the fact that the Student-t copula implies stronger tail dependence than the Gaussian copula, resulting in increased bias and uncertainty of the estimates for a given sample size (we refer to [10]). Overall, the simulation results suggest that the estimation method for the proposed class of models provides a valid basis for inference. In the next section, we apply these models to analyze two real-life datasets.

5.7 Applications

We illustrate the usefulness of our mixture copula based count regression models by analyzing some real-life data sets (see [149]). These data sets are publicly available in [116] and [119].

5.7.1 The health care utilization data

We analyze a healthcare utilization dataset from the General Hospital of St. John's, Newfoundland, Canada, which records the number of physician visits for 180 individuals from 1985 to 1990 (Appendix $6A$ of [116]). Each individual's data includes covariates: gender, chronic conditions, education level, and age. Given the repeated counts over six years, we account for longitudinal correlation. Our goal is to estimate the temporal dependency of count responses while considering covariate effects. Previous studies, such as [116], explored correlation structures like EX and AR(1), using the generalized quasi-likelihood (GQL) method for parameter estimation. In contrast, our mixture copula-based framework incorporates both AR(1) and EX structures within each copula component. Following Section 5.3 notation, covariates are defined as: sex (0 for male, 1 for female), chronic disease status (0 to 4 active diseases), education level (1 for less than high school, 2 for high school, 3 for university graduate, and 4 for post-graduate), and age (deviated by 50 years). The mean function is considered as follows:

$$\mu_{ij} = \exp(\beta_0 + \text{sex}_i\beta_1 + \text{crn}_i\beta_2 + \text{edu}_i\beta_3 + \text{age}_i\beta_4 + t_{ij}\beta_5), \quad j = 1, \ldots, 6, \tag{5.38}$$

where t_{ij} is the respective year of visit from 1 to 6. From the empirical correlation matrix of the count measurements, EX structure seems to be appropriate for this data set. We apply our methods discussed in Section 5.4 to estimate the regression as well as dependence parameters with 2 choices for the marginal distributions, standard and 2-component mixture of elliptical copulas.

The results are shown in Tables 5.5 and 5.6. For the model with Poisson marginals, the Student-t mixture copula with degrees of freedom (df) $\nu = 17$ appears to provide a model which is better than the remaining three models. In addition, since $\hat{\pi} = 0.5634$, the mixture copula attaches little more weight to the AR(1) structure. A close look at the numbers indicate that those for the Gaussian mixture

Poisson marginals			Negative Binomial marginals		
Parameters	Est.	SE	**Parameters**	Est.	SE
β_0	0.7028	0.1830	β_0	0.6727	0.1903
β_1	0.6179	0.1379	β_1	0.6960	0.1225
β_2	0.2003	0.0480	β_2	0.2106	0.0497
β_3	0.0306	0.0649	β_3	0.0280	0.0632
β_4	0.0076	0.0042	β_4	0.0100	0.0025
β_5	0.0608	0.0165	β_5	0.0619	0.0176
-	-	-	ψ	0.9747	0.0954

Table 5.5 Estimated marginal parameters and their standard errors of the health care utilization data with the model in (5.38) using Poisson and Negative binomial marginals.

copula model are close to those for the Student-t mixture copula model with df $\nu = 17$. This is perhaps expected as with large ν, these two models models are expected to be close to each other. It can also be seen that the Student-t exchangeable copula model with df $\nu = 15$ appears to provide a fit which is slightly better than the Gaussian exchangeable copula model. For the model with Negative Binomial marginals, the Student-t mixture copula with df $\nu = 12$ to provide a model which is better than the remaining three models. In addition, since $\hat{\pi} = 0.4181$, the mixture copula attaches little more weight to the EX structure. Also, the Student-t mixture copula model with df $\nu = 12$ appears to be marginally better than the Gaussian mixture copula model. It can also be seen that the Student-t exchangeable copula model with degrees of freedom $\nu = 10$ appears to provide a fit which is slightly better than the Gaussian exchangeable copula model. For both the models with Student-t mixture copula, the df ($\nu = 17$ corresponding to Poisson marginals and $\nu = 12$ corresponding to Negative Binomial marginals) indicate moderate tail dependence. Among all the eight models considered, the ones with Negative Binomial marginals with (1) Student-t mixture copula with df $\nu = 12$ and (2) Gaussian mixture copula seem to stand out, with the first choice being marginally better than the second. This seems important in view of lesser computational burden associated with the Gaussian mixture copula. The estimates of the parameters of the Negative Binomial marginals are interpreted below.

The positive value of β_1 suggests females visit physicians more than males, while positive values for β_2 and β_4 indicate individuals with chronic diseases or older age groups visit more frequently. The estimate for β_3 suggests education level has a minor impact on physician visits, which contrasts with [121] due to the inclusion of data from all six time points. The positive β_5 indicates that individuals visited physicians more in subsequent years. The mixture copula dependence parameter π suggests the EX structure is more prominent in the data, though the Poisson-based model, which fits worse than the Negative Binomial model, indicates the AR(1) structure. Using the Kendall's tau and Spearman's rho expressions from Section 5.2, we compute the concordance matrices and compare them with the empirical versions. We calculate sample Kendall's tau and Spearman's rho for the residuals from the fitted marginal models. Let $A(\tau)$ and $A(\rho)$ represent the empirical matrices, and using Theorems 5.2.3

Model	Copula	Parameters	Est.	SE	Comp-like	CLAIC	CLBIC
Poisson	Gaussian exchangeable	ξ_2	1.0222	0.0532	-17144.81	34486.16	34808.81
	Gaussian mixture	π	0.5339	0.0386	-16817.61	33839.47	34165.56
		ξ_1	2.1041	0.3020			
		ξ_2	0.2348	0.0259			
	Student-t ($\nu = 15$) exchangeable	ξ_2	1.0291	0.0832	-16932.91	34063.10	34386.36
	Student-t ($\nu = 17$) mixture	π	0.5634	0.0461	-16803.36	**33810.99**	**34137.09**
		ξ_1	1.6801	0.2099			
		ξ_2	0.2653	0.0419			
Negative Binomial	Gaussian exchangeable	ξ_2	0.6181	0.0246	-13010.82	26073.76	26158.89
	Gaussian mixture	π	0.4824	0.0968	-12980.33	26014.73	26101.06
		ξ_1	0.6448	0.1003			
		ξ_2	0.2780	0.0763			
	Student-t ($\nu = 10$) exchangeable	ξ_2	0.6002	0.0255	-12999.87	26061.49	26130.99
	Student-t ($\nu = 12$) mixture	π	0.4181	0.0953	-12973.71	**26010.99**	**26080.55**
		ξ_1	0.3137	0.1472			
		ξ_2	0.5672	0.1812			

Table 5.6 Estimated dependence parameters and their standard errors of the health care utilization data with standard and mixture of elliptical copulas. Maximum composite log-likelihood value, CLAIC and CLBIC for each model are reported.

and 5.2.4, we obtain $A(\hat{\tau})$ and $A(\hat{\rho})$ with the estimated parameters. These are given as

$$A(\tau) = \begin{bmatrix} 1.00 & & & & & \\ 0.44 & 1.00 & & & & \\ 0.41 & 0.40 & 1.00 & & & \\ 0.31 & 0.36 & 0.43 & 1.00 & & \\ 0.30 & 0.37 & 0.39 & 0.50 & 1.00 & \\ 0.27 & 0.27 & 0.33 & 0.43 & 0.48 & 1.00 \end{bmatrix}, A(\hat{\tau}) = \begin{bmatrix} 1.00 & & & & & \\ 0.41 & 1.00 & & & & \\ 0.35 & 0.42 & 1.00 & & & \\ 0.31 & 0.35 & 0.42 & 1.00 & & \\ 0.28 & 0.31 & 0.36 & 0.42 & 1.00 & \\ 0.26 & 0.28 & 0.31 & 0.36 & 0.42 & 1.00 \end{bmatrix},$$

$$A(\rho) = \begin{bmatrix} 1.00 & & & & & \\ 0.60 & 1.00 & & & & \\ 0.56 & 0.55 & 1.00 & & & \\ 0.43 & 0.50 & 0.59 & 1.00 & & \\ 0.42 & 0.52 & 0.53 & 0.67 & 1.00 & \\ 0.38 & 0.40 & 0.46 & 0.60 & 0.60 & 1.00 \end{bmatrix} \text{ and } A(\hat{\rho}) = \begin{bmatrix} 1.00 & & & & & \\ 0.59 & 1.00 & & & & \\ 0.51 & 0.59 & 1.00 & & & \\ 0.45 & 0.51 & 0.59 & 1.00 & & \\ 0.41 & 0.45 & 0.51 & 0.59 & 1.00 & \\ 0.38 & 0.41 & 0.45 & 0.51 & 0.60 & 1.00 \end{bmatrix}.$$

These estimated concordance matrices are very close to their empirical versions and they indicate high longitudinal correlations. To evaluate the goodness-of-fit we implement the modified t-plot discussed in Section 5.5 to the mixture copula models with Negative Binomial marginals in Figure 5.5. We can see linear trend along the 45-degree line for both the plots. But for the Student-t mixture copula the uniform

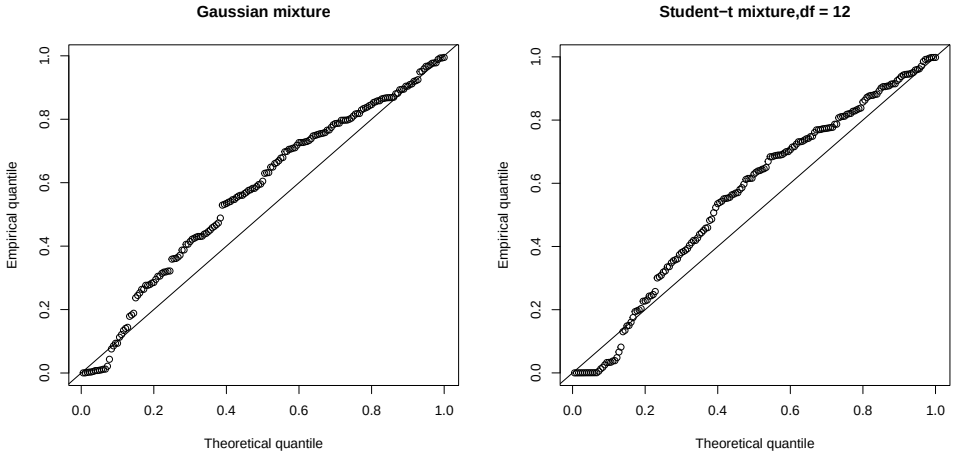

Figure 5.5 t-plots for the mixture of elliptical copulas for the health care utilization data.

quantiles are relatively closer to the line than the Gaussian copula. This suggest that Student-t mixture copula is more suitable for the temporal dependency for this data set.

5.7.2 The epilepsy data

We also examine a well-known longitudinal study on epileptic seizures, previously analyzed by several authors including [119], [120], and [12], among others. This study involved 59 patients suffering from simple or complex partial seizures. Among these patients, 31 were administered the anti-epileptic drug progabide, while the remaining 28 received a placebo randomly. The patients were observed for four successive clinical visits after randomization, during which the number of seizures occurring over the previous two weeks for each individual was recorded. Following the approach of [150], we consider three covariates to analyze this dataset: the logarithm of the baseline seizure count, the logarithm of age in years, and the treatment indicator (0 for placebo, 1 for progabide). From preliminary descriptive analysis, it is evident that the data exhibit overdispersion. We consider the mean function as

$$\mu_{ij} = \exp(\beta_0 + \text{lbase}_i\beta_1 + \text{trt}_i\beta_2 + \text{lage}_i\beta_3 + t_{ij}\beta_5), \quad j = 1, \ldots, 4, \tag{5.39}$$

where t_{ij} is the respective visit number from 1 to 6. For this data set also EX structure seems to be appropriate based on the empirical correlation matrix. We analyze this data set using previously described methodology as well.

The results we have obtained are summarized in Tables 5.7 and 5.8. For the model with Poisson marginals, the Student-t mixture copula model with df $\nu = 8$ appears to provide a model which is better than the remaining three models. In addition, since $\hat{\pi} = 0.2496$, the mixture copula attaches little more weight to the EX structure. Also, in terms of the values of CLAIC and CLBIC, the Gaussian mixture copula model and the Student-t mixture copula with df $\nu = 8$ are close to each other, but the

Poisson marginals			Negative Binomial marginals		
Parameters	Est.	SE	**Parameters**	Est.	SE
β_0	-3.7090	0.9579	β_0	-2.3136	0.8877
β_1	1.2199	0.1543	β_1	1.0536	0.1084
β_2	-0.0397	0.1869	β_2	-0.2421	0.1545
β_3	0.5184	0.2377	β_3	0.3034	0.2312
β_4	-0.0574	0.0350	β_4	-0.0542	0.0360
-	-	-	ψ	2.6101	0.5975

Table 5.7 Estimated marginal parameters and their standard errors of the epilepsy data with the model in (5.39) using Poisson and Negative binomial marginals.

estimates of π are very different; $\hat{\pi} = 0.5603$ for the Gaussian mixture copula model and $\hat{\pi} = 0.2496$ for the Student-t mixture copula model with df $\nu = 8$. A close look at the numbers indicate that those for the Gaussian mixture copula model are close to those for the Student-t mixture copula model with degrees of freedom $\nu = 17$. This is perhaps expected as with large ν, these two models models are expected to be close to each other. It can also be seen that the Student-t exchangeable copula model with degrees of freedom $\nu = 15$ appears to provide a fit which is slightly better than the Gaussian exchangeable copula model. For the model with Negative Binomial marginals, the Student-t mixture copula with degree of freedom $\nu = 22$ to provide a model which is better than the remaining three models. In addition, since $\hat{\pi} = 0.6976$, the mixture copula attaches little more weight to the AR(1) structure. A close look at the numbers indicate that those for the Gaussian mixture copula model are close to those for the Student-t mixture copula model with degrees of freedom $\nu = 22$. This is perhaps expected as with large ν, these two models models are expected to be close to each other. It can also be seen that the fit provided by Student-t exchangeable copula model with degrees of freedom $\nu = 21$ is very close to that provided by the Gaussian exchangeable copula model. This also is not unexpected. For both the models with Student-t mixture copula, the degrees of freedom ($\nu = 8$ corresponding to Poisson marginals and $\nu = 22$ corresponding to Negative Binomial marginals) indicate moderate tail dependence. Among all the eight models considered, the ones with Negative Binomial marginals with (1) Student-t mixture copula with df $\nu = 22$ and (2) Gaussian mixture copula seem to stand out, with the first choice being marginally better than the second. This seems important in view of lesser computational burden associated with the Gaussian mixture copula. The estimates of the parameters of the Negative Binomial marginals are interpreted below.

The estimate of β_2 is negative, implying that patients under the progabide treatment group experience lower seizure counts compared to the control group. Additionally, the estimate of β_3 suggests a positive relationship between age and seizure rate. Interestingly, the estimate of β_1 indicates that patients who start with a high seizure rate tend to maintain this rate consistently. The estimated standard errors of the marginal parameters are relatively large, which can be attributed to the low sample size in this dataset. Examining the estimated dependence parameters of the Negative Binomial-based models, we find that Student-t and Gaussian mixture copulas do not significantly differ for this dataset based on the estimated value of the degrees of freedom parameter and the selection criteria. Interestingly, based

Model	Copula	Parameters	Est.	SE	Comp-like	CLAIC	CLBIC
Poisson	Gaussian exchangeable	ξ_2	1.1832	0.0875	-2245.24	4636.15	4789.72
	Gaussian mixture	π	0.5603	0.0899	-2221.13	4590.23	4743.93
		ξ_1	2.8997	0.5365			
		ξ_2	0.3485	0.0850			
	Student-t ($\nu = 8$) exchangeable	ξ_2	1.2254	0.1191	-2219.88	**4587.65**	**4742.23**
	Student-t ($\nu = 8$) mixture	π	0.2496	0.4007	-2219.67	4588.32	4743.08
		ξ_1	2.7921	1.3011			
		ξ_2	0.9632	0.4314			
Negative Binomial	Gaussian exchangeable	ξ_2	0.9550	0.1141	-1928.56	3889.84	3924.25
	Gaussian mixture	π	0.6836	0.1174	-1919.69	3873.65	3909.26
		ξ_1	1.4546	0.4483			
		ξ_2	0.1195	0.0799			
	Student-t ($\nu = 21$) exchangeable	ξ_2	0.9172	0.1168	-1924.85	3881.29	3916.95
	Student-t ($\nu = 22$) mixture	π	0.6976	0.1244	-1919.48	**3872.90**	**3908.15**
		ξ_1	1.3930	0.4360			
		ξ_2	0.1261	0.0882			

Table 5.8 Estimated dependence parameters and their standard errors of the epilepsy data with standard and mixture of elliptical copulas. Maximum composite log-likelihood value, CLAIC and CLBIC for each model are reported.

on the estimate of π, it appears that the AR(1) structure is more prominent. However, the relatively high value of ξ_1 indicates that the first component copula is quite close to the independence copula. To further analyze the epilepsy dataset, we obtain the concordance matrices and compare them with their empirical versions. Let $A(\tau)$, $A(\hat{\tau})$, $A(\rho)$ and $A(\hat{\rho})$ be the matrices described in Sub-section 5.7.1 given as

$$A(\tau) = \begin{bmatrix} 1.00 & & & \\ 0.34 & 1.00 & & \\ 0.21 & 0.28 & 1.00 & \\ 0.26 & 0.40 & 0.28 & 1.00 \end{bmatrix}, A(\hat{\tau}) = \begin{bmatrix} 1.00 & & & \\ 0.30 & 1.00 & & \\ 0.21 & 0.30 & 1.00 & \\ 0.19 & 0.21 & 0.30 & 1.00 \end{bmatrix},$$

$$A(\rho) = \begin{bmatrix} 1.00 & & & \\ 0.47 & 1.00 & & \\ 0.30 & 0.39 & 1.00 & \\ 0.37 & 0.53 & 0.40 & 1.00 \end{bmatrix} \text{ and } A(\hat{\rho}) = \begin{bmatrix} 1.00 & & & \\ 0.43 & 1.00 & & \\ 0.30 & 0.42 & 1.00 & \\ 0.27 & 0.30 & 0.42 & 1.00 \end{bmatrix}.$$

Despite the relatively low sample size of this dataset, we observe that the best-fitting mixture copula closely captures the longitudinal correlation. The modified t-plot for the epilepsy dataset is shown in Figure 5.6, demonstrating that both of these elliptical mixture copulas effectively capture the temporal dependency. Overall, our models exhibit substantial improvements in capturing longitudinal correlation

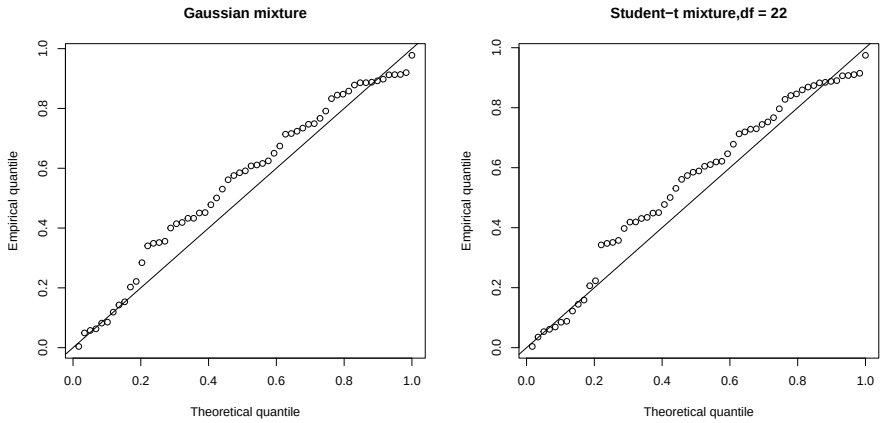

Figure 5.6 t-plots for the mixture of elliptical copulas for the epilepsy data.

compared to alternative approaches. Although our parametric results differ from those in [121] or [120] due to our different covariate setup, the conclusions are somewhat similar. It's worth noting that widely used random effect models can be challenging to interpret and may pose computational challenges in estimating model parameters due to multidimensional integrations. Copula-based models serve as viable alternatives, as they can be estimated using efficient one-stage or two-stage procedures with valid standard errors.

5.8 Discussion

Modeling the correlation structure of discrete longitudinal data is crucial, given the lack of proper multivariate distributions for such data. Copula approaches have emerged as primary techniques for modeling the dependence structure of multivariate data across various fields. However, the application of copulas in discrete data modeling has been relatively unexplored due to theoretical limitations. Drawing inspiration from the literature on finite mixture models, this chapter introduces a novel approach: the use of mixture of elliptical copulas to model the temporal dependency of longitudinal count data. We derive the dependence properties of finite mixture copulas for both continuous and discrete cases. Under the full parametric setup, we estimate the parameters of our proposed models using a two-stage composite likelihood method and validate them through extensive simulation studies. In our future works we also want to investigate resampling based methods to obtain the standard errors of the parameter estimates of the proposed class of models. Furthermore, we extend and modify the t-plot method for model validation specifically for elliptical copulas. By employing different correlation structures in each copula component, we demonstrate that our mixture copula models offer improved insights into the dependence structure of count longitudinal data. While we focus on balanced cases in this article, our approach can readily be extended to handle unbalanced longitudinal data, as long as mixture of elliptical copulas are utilized.

As noted in section 5.7.1 (The health care utilization data) of the dissertation, this data-set has been analyzed earlier by [116]. [116] considered three kinds of correlation scenarios: (i) AR(1), (ii) MA(1), and (iii) exchangeable (equicorrelation) structure (EX in our notation, and EQC in the notation of [116]). In section 6.7, [116] (pp. $217-219$) has tried all three models in fitting the health care utilization data. In [116] (pp. $218-219$), we see the following: "As far as the selection of a model from these three lower-order models is concerned, we have computed the fitted residual squared distance GM ... under all three models and reported them in the ... Table 6.13. As the GM statistic has the lowest value 14.20 under the AR(1) structure, we chose the AR(1) model to interpret the estimates." [116] (p. 219) has also argued in favour of the AR(1) model using findings from simulation reported in an earlier section. The computations show that the GM statistic for the remaining two models are 15.34 (for EX) and 20.46 (for MA(1)). One also notices in the Table 6.13 ([116] p. 218) the closeness of the estimates of the regression and the correlation parameters for the AR(1) and the EX choices, in comparison to the MA(1) choices. These findings encouraged us to try a mixture of two elliptical copulas generated from the AR(1) and EX models. We wish to submit that for this data-set, this appears to be meaningful, and introducing more components in the mixture does not appear, in our humble judgement, to be appealing.

Besides the above, while suggesting a mixture copula we have generally kept in our consideration the underlying issue of computational burden in the regression setting. We notice in this context that an early reference where a 2-component mixture model was used in copula based analysis for the purposes of clustering is [129]. However, in a recent work by [151], the authors have proposed a conditional mixture copula which is a weighted average of several individual conditional copulas. In their work, they have allowed both the weights and copula parameters to vary with a covariate so that the conditional mixture copula offers additional flexibility and accuracy in describing the dependence structure. They have proposed a two-step semi-parametric estimation method and have develop asymptotic properties of the estimators. Moreover, they have introduced model selection procedures to select the component copulas of the conditional mixture copula model. Their work has enabled them to choose the number of components as well. We think this is a meaningful approach. Another recent work in modeling via mixture of copulas is [123]. In our future work, we wish to pursue a comprehensive investigation taking into account these recent researches and also our own.

5.9 Appendix

We present below a justification of Remark $5.2.3$. A proof of Corollary $5.2.1$ also follows along the same line. With some details, this also is mentioned at the end. To begin with, we note that if (X, Y) has a bivariate Student-t distribution with degrees of freedom ν and pdf given by

$$t_{2,\nu}(x,y) = \frac{1}{\sqrt{1-\rho^2}} \frac{\Gamma((\nu+2)/2)}{\Gamma(\nu/2)(\nu\pi)} \left[1 + \frac{x^2+y^2-2xy}{2(1-\rho^2)}\right]^{-(\nu+2)/2}, \quad (x,y) \in \mathcal{R}^2,$$

then $(X, Y) \stackrel{d}{=} W/\sqrt{V}$, where $W \sim N_2(0, 0, 1, 1, \rho), V \sim \chi^2_\nu/\nu$ and W and V are independent. ([28], p. 44) In other words, (X, Y) can be recognized as scale mixture of a bivariate normal distribution with both marginals being standard normal and with correlation ρ.
For $i = 1, 2$, we consider $(X_i, Y_i) \stackrel{d}{=} W_i/\sqrt{V_i}$ where $W_i \sim N_2(0, 0, 1, 1, \rho_i), V_i \sim \chi^2_\nu/\nu$, and W_1, W_2, V_1, V_2 are independent.
Next, we let $p_C := P[(X_1 - X_2)(Y_1 - Y_2) > 0]$, the probability of concordance for the independent pairs (X_1, Y_1) and (X_2, Y_2). It can be proved that

$$p_C = \frac{1}{2} + \frac{E[\arcsin(\rho(V_1, V_2)]}{\pi} \quad \text{where } \rho(V_1, V_2) := \frac{V_1^{-1}\rho_1 + V_2^{-1}\rho_2}{V_1^{-1} + V_2^{-1}}.$$

We notice that $E(\rho(V_1, V_2)) = (\rho_1 + \rho_2)/2$ and propose a first-order approximation for p_C and τ (analogue of Kendall's tau ([26] p. 159)) as follows:

$$p_C \approx \frac{1}{2} + \frac{\arcsin(E[\rho(V_1, V_2)])}{\pi} = \frac{1}{2} + \frac{\arcsin\left(\frac{\rho_1+\rho_2}{2}\right)}{\pi},$$
$$\tau = 2p_C - 1 \approx \frac{\arcsin\left(\frac{\rho_1+\rho_2}{2}\right)}{\pi}.$$

We propose to carry out a detailed investigation of this approximation later. Finally, we notice that if we take both of V_1 and V_2 above to be degenerate at 1, Corollary 5.2.1 follows immediately.

Chapter 6

Modeling longitudinal data using geometric skew-normal copula

In the previous chapters, we have discussed and relied on dependence properties of several multivariate elliptical copulas. In this final chapter, we introduce a copula which is derived from the multivariate geometric skew normal (MGSN), introduced in [152]. Prior to this, univariate version of MGSN distribution was introduced in [153]. The reference [153] has discussed several attractive properties of the MGSN distribution. In particular, it has discussed why it can be and is an alternative to Azzalini's mulivariate skew-normal distribution due to Azzalini and Valle ([154]). This serves as one of the motivations to develop and study the copula based on the MGSN distribution.

The multivariate copula we are going to introduce in this chapter will be called geometric skew-normal (GSN) copula. While the multivariate Gaussian copula is commonly used to model temporal dependence in non-Gaussian longitudinal data, it may fail to capture non-exchangeable dependence or tail dependence. To address this, we propose the GSN copula as a flexible alternative, which is crucial for reliable inference in such data. We begin by exploring the theoretical properties of the GSN copula and then develop regression models for both continuous and discrete longitudinal data. One advantage of our proposed copula is that its quantile function is independent of the correlation matrix, offering computational benefits for likelihood inference compared to skew-elliptical copulas like those from [154]. Additionally, composite likelihood inference allows parameter estimation using an ordered probit model with the same dependence structure as the GSN distribution. Many longitudinal models rely on the multivariate normal distribution, but normality is often unsuitable, especially when graphical diagnostics show asymmetry in the marginals or temporal dependence. Our model provides a flexible, computationally tractable alternative for both continuous and discrete longitudinal data, offering useful dependence properties and simple interpretations of the marginals.

Elliptical copulas (multivariate Gaussian or Student-t) are popular in the literature of multivariate dependence for their simplicity in terms of parametric inference (see, [132], [12] and [90]). These can be used to construct parametric models for continuous data with arbitrary marginal distributions and also for discrete data under latent variable formulation. The references [24], [26], [27] and [28] contain elaborate descriptions on copula models and their dependence properties. It has been argued in [155] and [156] that Gaussian copulas fail to capture dependence between non-exchangeable variables and the extreme ones. The reference [133] proposed using multivariate Student-t and related copulas to address

tail dependence in correlated data. Usually copula based models do not have closed form expressions, and requires numerical computations. Most of the commonly used multivariate copulas satisfy the assumption of exchangeability which means the copula function is permutation-invariant. However, when the extent of influence of some of the components on others is asymmetric, this assumption may not be appropriate. An alternative is to construct multivariate copulas using various skew-elliptical distributions by [48]. However, they have certain limitations in applications due to unavailability of algebraically tractable form, strictly positive support and linear relationship between parameters and dimension. The reference [58] used skew-normal copula to capture non-exchangeable dependence with block coordinate algorithm for parameter estimation. The reference [63] discussed Bayesian inference and applications of skew-t copula. The numerical difficulties to obtain the maximum likelihood estimates of skew-t copula in high dimension have been discussed in [62]. Another limitation of these multivariate copulas is that their parameters can not be identified from their lower dimensional densities.

Recently, [153] proposed an alternative skew-normal distribution with multivariate extension in [152], of which multivariate normal distribution is a special case. This distribution is obtained as a geometric sum of independent and identically distributed normal random variables, and hence is called the geometric skew-normal (GSN) distribution. Unlike Azzalini's skew-normal distribution, this distribution can be multi-modal and take different shapes depending on the three-set parameter values. The author of [153] and [152] has developed several interesting properties of this distribution and demonstrated computational advantage of this distribution in the multivariate setup. Being relatively new, not too many applications of this distribution seem to be known in the literature. However, [157] discussed a class of power series skew-normal distributions by generalizing the GSN distribution. Also, [158] proposed Bayesian model-based clustering based on geometric skew-normal distribution and validated the performance through some simulation studies.

In this chapter we develop an alternative asymmetric multivariate copula constructed from geometric skew-normal distribution to model temporal dependence of longitudinal data. First we derive the theoretical properties of the proposed copula and then develop appropriate dependence models for continuous and ordinal data. For the continuous repeated measurements we use generalized linear models for the marginals, and for the ordinal responses we use latent variable formulation. The quantile function of this multivariate copula is independent of the correlation matrix of its respective multivariate distribution, which provides computational advantages in terms of parametric inference. Another interesting advantage over Azzalini's skew-normal copula is that multivariate GSN copula is closed under marginalization that is all its lower dimensional sub-copulas belong to the same parametric family. That is why composite likelihood inference is possible for this multivariate copula, which facilitates to estimate parameters from ordered probit models with dependence structure of geometric skew-normal distribution.

Rest of the chapter is organized as follows. In Section 6.3, details of construction of the multivariate geometric skew-normal copula are described. This is followed by discussion on construction of the GSN copula which is presented in Section 6.1 . Section 6.3 elaborates the dependence properties of the

GSN copula. The details of maximum likelihood estimation for unrestricted GSN copula using block-coordinate ascent algorithm is described in Section 6.4. In Section 6.5 , we develop regression models for continuous and ordinal longitudinal data and describe their parametric inference. In Section 6.6, we describe some standard model evaluation methods. Section 6.7 presents the finite sample performance of our proposed models using some simulated data sets. Thereafter in Section 6.8, we analyze two real-life data sets and compare the fits with corresponding Gaussian copula based models. Section 6.9 concludes this chapter with a general discussion.

6.1 Multivariate geometric skew-normal distribution

This section is mainly a review from the previous works of [153] and [152]. Multivariate geometric skew-normal variable can be expressed as a geometric random sum of Gaussian random variables. A d-variate MGSN distribution is defined as follows.

Definition 6.1.1 *Suppose $N \sim GE(p)$, and $\{\mathbf{X}_i; i = 1, 2 \dots\}$ are i.i.d. $N_d(\mu, \mathbf{\Sigma})$ random vectors. It is assumed that N and $\mathbf{X}_i$'s are independently distributed. Then the random variable $\mathbf{X}$, where*

$$\mathbf{X} \stackrel{dist}{=} \sum_{i=1}^{N} \mathbf{X}_i \tag{6.1}$$

is said to have a d-variate geometric skew-normal distribution with parameters p, μ and $\mathbf{\Sigma}$. p be the success probability of Geometric distribution for $0 < p \leq 1$. We denote this distribution by $MGSN_d(p, \mu, \mathbf{\Sigma})$.

From [152], if $\mathbf{X} \sim MGSN_d(p, \mu, \mathbf{\Sigma})$, then the CDF and PDF of $\mathbf{X}$ take the following forms -

$$F_{d,GSN}(\mathbf{x}|\mu, \mathbf{\Sigma}, p) = \sum_{k=1}^{\infty} p(1-p)^{k-1} \Phi_d(\mathbf{x}|k\mu, k\mathbf{\Sigma}) \quad \text{and} \tag{6.2}$$

$$\begin{aligned} f_{d,GSN}(\mathbf{x}|\mu, \mathbf{\Sigma}, p) &= \sum_{k=1}^{\infty} p(1-p)^{k-1} \phi_d(\mathbf{x}|k\mu, k\mathbf{\Sigma}) \\ &= \sum_{k=1}^{\infty} \frac{p(1-p)^{k-1}}{(2\pi k)^{d/2}\sqrt{|\mathbf{\Sigma}|}} e^{-\frac{1}{2k}(\mathbf{x}-k\mu)^{\intercal}\mathbf{\Sigma}^{-1}(\mathbf{x}-k\mu)}, \end{aligned} \tag{6.3}$$

respectively. Here $\Phi_d(\mathbf{x}|k\mu, k\mathbf{\Sigma})$ and $\phi_d(\mathbf{x}|k\mu, k\mathbf{\Sigma})$ denote the CDF and PDF of a d-variate normal distribution respectively, with mean vector $k\mu$ and dispersion matrix $k\mathbf{\Sigma}$. The PDF of standard $MGSN_d(p)$ distribution where $\mu = 0$ and $\mathbf{\Sigma} = I$ is symmetric and unimodal, for all values of d and p. However, depending on the parameter values the PDF of $MGSN_d(p, \mu, \mathbf{\Sigma})$ can be skewed and multimodal as well. Note that for $p = 1$ it reduces to $N_d(\mu, \mathbf{\Sigma})$ distribution. Similarly for the univariate case we have

the CDF as

$$F_1(x|\mu,\sigma,p)=\sum_{k=1}^{\infty}p(1-p)^{k-1}\Phi\left(\frac{x-k\mu}{\sigma\sqrt{k}}\right), \tag{6.4}$$

where X follows $GSN(\mu,\sigma,p)$ when $d=1$. Generation of random samples from multivariate geometric skew-normal distribution is simple with two steps. Now to construct and study the GSN copula the following results are needed.

Result 6.1.1 *Let* $\mathbf{X}$ *be partitioned with the corresponding partitions for* μ *an* $\mathbf{\Sigma}$ *as*

$$\mathbf{X}=\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix},\ \ \mu=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\ \textit{and}\ \ \mathbf{\Sigma}=\begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{pmatrix}, \tag{6.5}$$

where it corresponds to vector of parameters of dimension $(h,d-h)$*. If* $\mathbf{X}\sim MGSN_d(p,\mu,\mathbf{\Sigma})$ *and* $\mathbf{X}_1\sim MGSN_h(p,\mu_1,\mathbf{\Sigma}_{11})$ *then* $\mathbf{X}_2\sim MGSN_{d-h}(p,\mu_2,\mathbf{\Sigma}_{22})$.

Result 6.1.2 *If* $\mathbf{X}\sim MGSN_d(p,\mu,\mathbf{\Sigma})$*, then* $\mathbf{Z}=\mathbf{D}\mathbf{X}\sim MGSN_s(p,\mathbf{D}\mu,\mathbf{D}\mathbf{\Sigma}\mathbf{D}^{\intercal})$*, where* $\mathbf{\Sigma}$ *is a non-singular matrix of dimension* $d\times d$ *and* $\mathbf{D}$ *is a* $s\times d$ *matrix of rank* $s\leq d$*.*

Remark 6.1.1 It can be seen that Result 6.1.1 can be easily derived from Result 6.1.2.

Result 6.1.3 *If* $\mathbf{X}\sim MGSN_d(p,\mu,\mathbf{\Sigma})$*, and if we denote the mean vector and dispersion matrix of* $\mathbf{X}$ *as* $\mu_{\mathbf{X}}$ *and* $\mathbf{\Sigma}_{\mathbf{X}}$*, respectively, then we obtain the following relation;*

$$\mu_{\mathbf{X}}=\frac{\mu}{p}\ \textit{and}\ \mathbf{\Sigma}_{\mathbf{X}}=\frac{\mathbf{\Sigma}}{p}+\frac{1-p}{p^2}\mu\mu^{\intercal}. \tag{6.6}$$

Result 6.1.4 *If* $\mathbf{X}\sim MGSN_d(p,\mu,\mathbf{\Sigma})$*, then* $\det(\mathbf{\Sigma}_{\mathbf{X}})$ *can be obtained as*

$$\det(\mathbf{\Sigma}_{\mathbf{X}})=\frac{|\mathbf{\Sigma}|}{p^d}\Big[1+\frac{1-p}{p}\mu^{\intercal}\mathbf{\Sigma}^{-1}\mu\Big]. \tag{6.7}$$

Remark 6.1.2 It is easy to see from the definition of $MGSN_d(p,\mu,\mathbf{\Sigma})$ that for p close to 1, this distribution, as expected, is close to $N_d(\mu,\mathbf{\Sigma})$. Also, when p is close to 1, the dispersion matrix $\mathbf{\Sigma}_{\mathbf{X}}$ of $\mathbf{X}$ is close to $\mathbf{\Sigma}$ and the generalized variance $\det(\mathbf{\Sigma}_{\mathbf{X}})$ of $\mathbf{X}$ is close to $\det(\mathbf{\Sigma})$.

Remark 6.1.3 The generalized variance $\det(\mathbf{\Sigma}_{\mathbf{X}})$ of $\mathbf{X}$ depends on μ through $\mu^T\mathbf{\Sigma}\mu$.

Remark 6.1.4 Result 6.1.3 can be used to estimate the parameters via method of moments which will be discussed latter on. It is to be noted from Result 6.1.4 that when $p\to 1$ it reaches to multivariate normal distribution. These are similar to the multivariate normal distribution and provide the marginals of a MGSN distribution. The joint to marginal relationship here is much simpler than Azzalini's skew-elliptical class of distributions. To construct the copula from (6.1.1) and (6.1.2), without loss of generality we can take $\mathbf{\Sigma}$ to be a correlation matrix by putting $\mathbf{D}=diag(\sigma_{11}^{-1/2},\ldots,\sigma_{dd}^{-1/2})$.

6.2 Construction of the GSN copula

In this section we discuss in details the construction of multivariate geometric skew-normal copula. [159] discussed the construction of copulas from skew-elliptical class of distributions. The geometric skew-normal copula (GSN) copula can be seen as an one-to-one transformation from $\mathbf{X}$ having geometric skew-normal distribution. If $\mathbf{\Sigma}$ is a correlation matrix then the j-th marginal distribution of the d-variate $MGSN_d(p, \mu, \mathbf{\Sigma})$ distribution is $GSN(p, \mu_j, 1)$. Now we propose the geometric skew-normal copula as follows.

Definition 6.2.1 *A d-dimensional copula $C_{d,GSN}$ is called a GSN copula with parameters $\mu, \mathbf{\Sigma}$ ($\mathbf{\Sigma}$ be the correlation matrix) and p if*

$$C_{d,GSN}(\mathbf{u}|\mu, \mathbf{\Sigma}, p) = F_{d,GSN}(F_1^{-1}(u_1|\mu_1, 1, p), \ldots, F_1^{-1}(u_d|\mu_d, 1, p)|\mu, \mathbf{\Sigma}, p) \tag{6.8}$$

where $F_1^{-1}(u_j|\mu_j, 1, p)$ denotes the inverse of the CDF of the $GSN(p, \mu_j, 1)$ distribution and $\mathbf{\Sigma}$ denote a correlation matrix. The corresponding geometric skew-normal copula density is given by

$$c_{d,GSN}(\mathbf{u}|\mu, \mathbf{\Sigma}, p) = \frac{f_{d,GSN}(F_1^{-1}(u_1|\mu_1, 1, p), \ldots, F_1^{-1}(u_d|\mu_d, 1, p)|\mu, \mathbf{\Sigma}, p)}{\prod_{j=1}^{d} f_{1,GSN}(F_1^{-1}(u_j|\mu_j, 1, p))} \tag{6.9}$$

where the multivariate density $f_{d,GSN}(.)$ is given in (6.3) and $f_{1,GSN}$ is the marginal density of a geometric skew-normal variable $X_j \sim GSN(p, \mu_j, 1)$.

It is direct from (6.3) that Gaussian copula is nested in (6.8) when $p = 1$. Thus we derive the dependence properties when $0 < p < 1$. Unlike the skew-elliptical copulas, we have no such slant parameter here and the parametric relations to the marginal distributions are simpler. To illustrate the dependence shapes imposed by the bivariate GSN copula model, in Figure 6.1 we provide the contour plots of the densities with $N(0, 1)$ marginals. As we can see, the asymmetry in this copula comes from the location parameter μ of the bivariate normal component. Following [160], we can also graphically represent the imposed dependence of the bivariate GSN copula using regression curves based on conditional expectation. In Figure 6.2 we plot

$$E(U|V = v) = \int_0^1 uc(u|v)du = \int_0^1 uc(u, v)du, \tag{6.10}$$

where U and V are uniformly distributed random variables on the interval $(0, 1)$, for various values of p and $(\mu_1, \mu_2)^\intercal$ taking values in the set $\{-1, 0, 1\}$. The shape of the conditional expectation depends on the magnitude and the sign of the location parameter of the bivariate normal component and the mean of the geometric random variable. This shows a wide range of different relationships can be modeled by the GSN copula. It also reveals how shifting the value of p, from 0 to 1 the dependence tend to be linear which is same as the Gaussian copula. The generation of random samples from MGSN distribution is very simple and so is from the copula. A random sample from $C_{d,GSN}(p, \mu, \mathbf{\Sigma})$ can be obtained using the following algorithm.

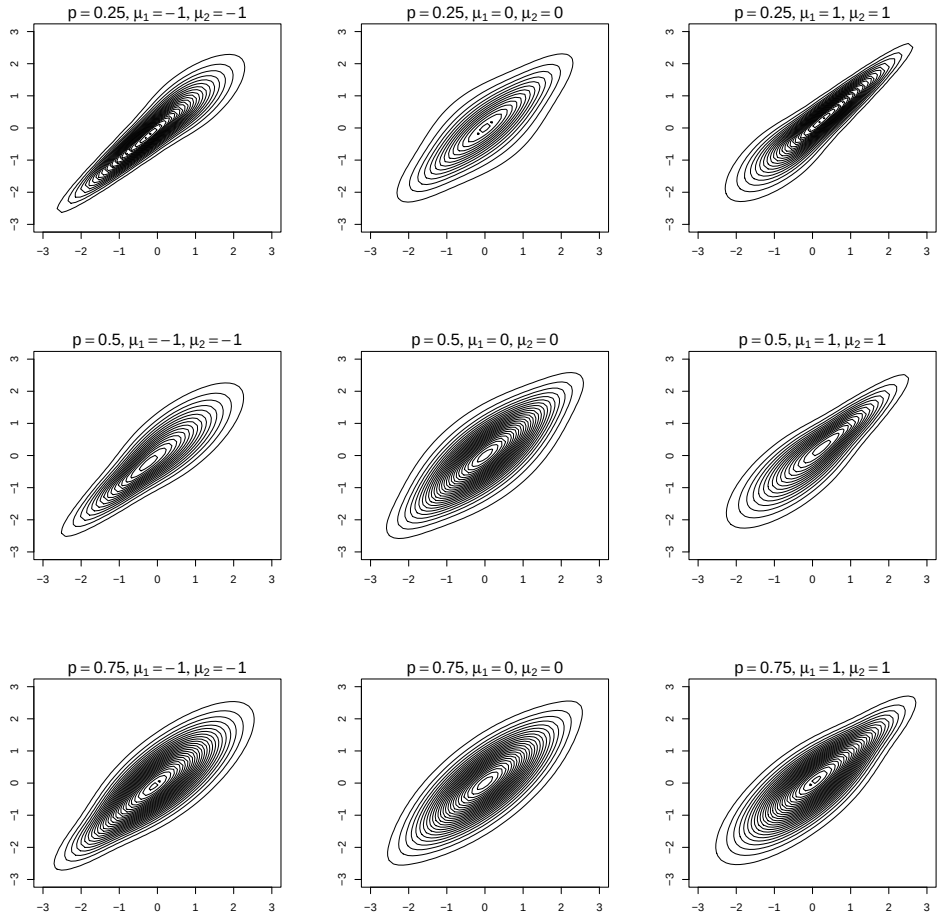

Figure 6.1 Contour plots of bivariate geometric skew-normal copula using standard normal marginals. The values of the parameters are used as $p = \{0.25, 0.5, 0.75\}$, $\mu = \{(-1,-1),(0,0),(1,1)\}$ and the common parameter $\rho = 0.77$.

Algorithm 6.2.1 *(Sampling from the GSN copula.)*

- *Step 1: Generate* $n \sim GE(p)$.
- *Step 2: Generate* $X \sim N_d(n\mu, n\mathbf{\Sigma})$.
- *Step 3: Set* $U_j = F_1(X_j|\mu_j, 1, p)$ *for* $j = 1, \ldots, d$.

Note that, Step 1 and 2 of Algorithm 6.2.1 involves simulating a random vector from multivariate geometric skew-normal distribution. In the earlier chapters, we have relied on dependence properties of several multivariate elliptical copulas. Next we describe the dependence properties.

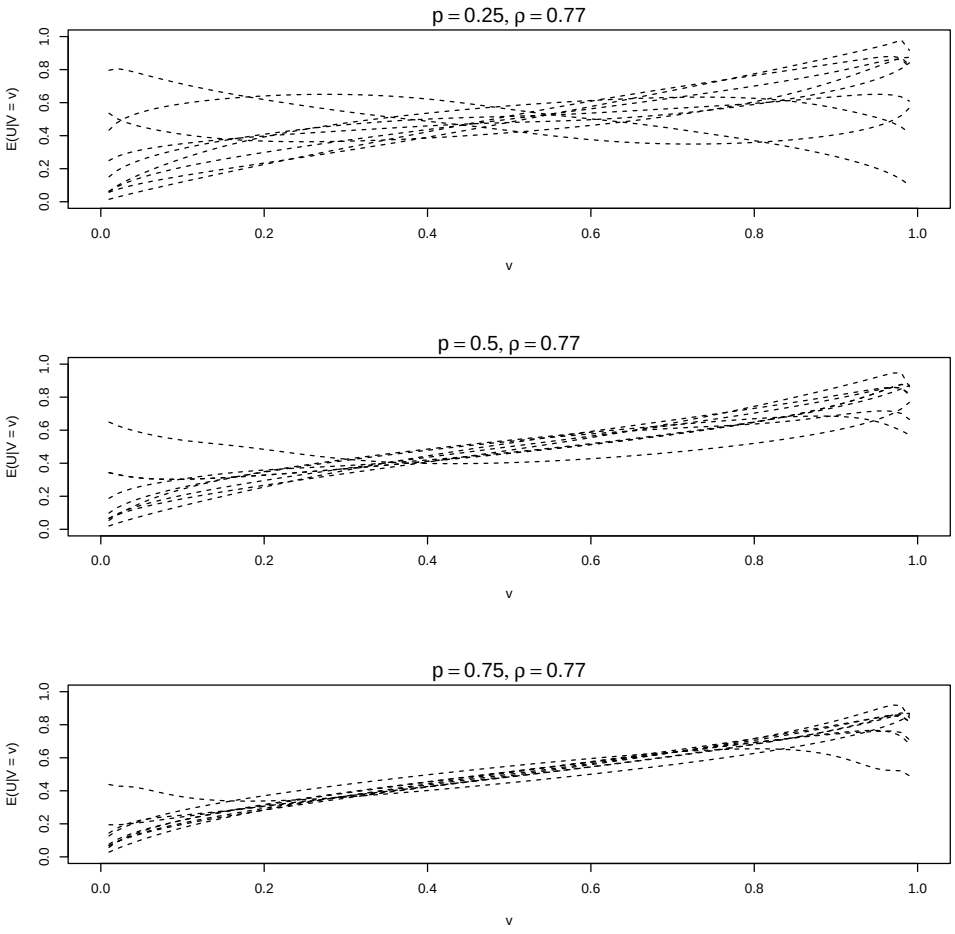

Figure 6.2 Regression curves of bivariate geometric skew-normal copula. The values of the parameters are used as $p = \{0.25, 0.5, 0.75\}, \mu = \{(-1,-1),(0,0),(1,1)\}$ and $\rho = 0.77$.

6.3 Dependence properties

In this section we discuss different properties of a GSN copula. But first we need to look at the correlation structure of the MGSN distribution as discussed in [152], derived from its moment generating function. If $\mathbf{X} = (X_1, \ldots, X_d)^\intercal \sim MGSN_d(p, \mu, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a correlation matrix then

$$Corr(X_i, X_j) = \frac{p\rho_{ij} + \mu_i\mu_j(1-p)}{\sqrt{p + \mu_i^2(1-p)}\sqrt{p + \mu_j^2(1-p)}}. \tag{6.11}$$

Hence from (6.11) the correlation between X_i and X_j for $i \neq j$ depends on ρ_{ij} as well as μ_i and μ_j. It follows that when $\mu_i = \mu_j = \rho_{ij} = 0$ for fixed p, $Corr(X_i.X_j) = 0$. Therefore, in this case although X_i and X_j are uncorrelated, they are not independent. But the Pearson's correlation coefficient is a symmetric measure of dependence which doesn't tell which variable have more influence on other. Hence we need to investigate other dependence measures of the GSN copula. Firstly, let us consider the dependence measure Kendall's tau. For a bivariate GSN copula we obtain τ as follows.

Proposition 6.3.1 *Let $\mathbf{X}, \mathbf{X}' \sim MGSN_2(p, \mu, \mathbf{\Sigma})$ are independent and $\mathbf{\Sigma}$ is a bivariate correlation matrix. Then Kendall's tao is given by*

$$\begin{aligned}\tau(C_{GSN}) &= 4p^2 \sum_{n=1}^{\infty}\sum_{n'=1}^{\infty}(1-p)^{n+n'-2}\Phi_2\Big(\frac{n'-n}{\sqrt{n+n'}}\mathbf{\Sigma}^{-1/2}\mu\Big) - 1 \qquad (6.12)\\ &= 4p^2 \sum_{n=1}^{\infty}\sum_{n'=1}^{\infty}(1-p)^{n+n'-2}\int_{-\infty}^{\frac{(n'-n)\mu_1}{\sqrt{n+n'}}}\Phi\Big(\frac{(n'-n)\mu_2}{\sqrt{(n+n')(1-\rho^2)}} - \frac{\rho y}{\sqrt{1-\rho^2}}\Big)\phi(y)dy - 1.\end{aligned}$$

Proof: From the definition in 6.1, we have $\mathbf{X} \overset{d}{=} \sum_{i=1}^{N}\mathbf{Y_i}$ and $\mathbf{X}' \overset{d}{=} \sum_{i=1}^{N'}\mathbf{Y_i}'$ where $N, N' \sim GE(p)$ and $\mathbf{Y_i}, \mathbf{Y_i}' \sim N_2(\mu, \mathbf{\Sigma})$ are all independent. From the expression in (2.8) we obtain Kendall's tao by

$$P(\mathbf{X} - \mathbf{X}' < 0) = \sum_{n=1}^{\infty}\sum_{n'=1}^{\infty}P(\mathbf{X} - \mathbf{X}' < 0 | N = n, N' = n') \cdot P(N = n) \cdot P(N' = n').$$

Now, $\mathbf{Z} := \mathbf{X} - \mathbf{X}' | N = n, N' = n' \sim N_2\big((n-n')\mu, (n+n')\mathbf{\Sigma}\big)$ and hence

$$\begin{aligned}P(\mathbf{Z} < 0 | N = n, N' = n') &= \int_{-\infty}^{0}\int_{-\infty}^{0} f(z_2|z_1)f(z_1)dz_1dz_2\\ &= \int_{-\infty}^{0}\Phi\Big(\frac{(n'-n)\mu_2 - \rho(z_1 - (n-n')\mu_1)}{\sqrt{(n+n')(1-\rho^2)}}\Big)\phi\Big(\frac{z_1 - (n-n')\mu_1}{\sqrt{n+n'}}\Big)dz_1\\ &= \int_{-\infty}^{\frac{(n'-n)\mu_1}{\sqrt{n+n'}}}\Phi\Big(\frac{(n'-n)\mu_2}{\sqrt{(n+n')(1-\rho^2)}} - \frac{\rho y}{\sqrt{1-\rho^2}}\Big)\phi(y)dy,\end{aligned}$$

since $Z_2|Z_1 = z_1, N = n, N' = n' \sim N\big((n-n')\mu_2 + \rho(z_1 - (n-n')\mu_1), (n+n')(1-\rho^2)\big)$.

Proposition 6.3.2 *Let* $\mathbf{X} \sim MGSN_2(p, \mu, \mathbf{\Sigma})$ *and* $\mathbf{X}^* \sim MGSN_2(p, \mu, \mathbf{I})$ *are independent where* $\mathbf{\Sigma}$ *is a bivariate correlation matrix and* I *is the identity matrix. Then Spearman's rho is given by*

$$\rho(C_{GSN}) = 12p^2 \sum_{n^*=1}^{\infty}\sum_{n=1}^{\infty}(1-p)^{n^*+n-2}\Phi_2\Big((n-n^*)(n^*\mathbf{I}+n\mathbf{\Sigma})^{-1/2}\mu\Big) - 3 = \tag{6.13}$$

$$12p^2 \sum_{n^*=1}^{\infty}\sum_{n=1}^{\infty}(1-p)^{n^*+n-2}\int_{-\infty}^{\frac{(n-n^*)\mu_1}{\sqrt{n^*+n}}}\Phi\Big(\frac{(n-n^*)\mu_2\sqrt{n^*+n}}{\sqrt{(n^*+n)^2-n^2\rho^2}} - \frac{n\rho y}{\sqrt{(n^*+n)^2-n^2\rho^2}}\Big)\phi(y)dy - 3.$$

Proof: In the similar manner, we can calculate Spearman's rho for a bivariate GSN copula.

$$P(\mathbf{X}^* - \mathbf{X} < 0) = \sum_{n^*=1}^{\infty}\sum_{n=1}^{\infty} P(\mathbf{X}^* - \mathbf{X} < 0 | N^* = n^*, N = n) \cdot P(N^* = n^*) \cdot P(N = n).$$

Here, $\mathbf{Z}^* := \mathbf{X}^* - \mathbf{X} | N^* = n^*, N = n \sim N_2\big((n^*-n)\mu, n^*\mathbf{I} + n\mathbf{\Sigma}\big)$ and hence

$$\begin{aligned}
P(\mathbf{Z}^* < 0 | N^* = n^*, N = n) &= \int_{-\infty}^{0}\int_{-\infty}^{0} f(z_2|z_1)f(z_1)dz_1dz_2 \\
&= \int_{-\infty}^{0}\Phi\Big(\frac{(n-n^*)\mu_2\sqrt{n^*+n} - \frac{n\rho(z_1-(n^*-n)\mu_1)}{\sqrt{n^*+n}}}{\sqrt{(n^*+n)^2-n^2\rho^2}}\Big)f(z_1)dz_1 \\
&= \int_{-\infty}^{\frac{(n-n^*)\mu_1}{\sqrt{n^*+n}}}\Phi\Big(\frac{(n-n^*)\mu_2\sqrt{n^*+n}}{\sqrt{(n^*+n)^2-n^2\rho^2}} - \frac{n\rho y}{\sqrt{(n^*+n)^2-n^2\rho^2}}\Big)\phi(y)dy,
\end{aligned}$$

since $Z_2^*|Z_1^* = z_1, N^* = n^*, N = n \sim N\big((n^*-n)\mu_2 + \frac{n\rho(z_1-(n^*-n)\mu_1)}{n+n^*}, n + n^* - \frac{n^2\rho^2}{n+n^*}\big)$.

It is important to point out that when $\mu = (\mu_1, \mu_2) = 0$, we find $\tau(C_{GSN}) = (2/\pi)\arcsin\rho$, which is same as the bivariate Gaussian copula. But that is not the case with $\rho(C_{GSN})$ as it simplifies to

$$\rho(C_{GSN}) = 12p^2 \sum_{n^*=1}^{\infty}\sum_{n=1}^{\infty}(1-p)^{n^*+n-2}\arcsin\frac{np}{n^*+n}.$$

One can obtain the values of (6.12) and (6.13) by a numerical approximation of the infinite series up to some finite terms and a numerical integration.

Many of the most widely used copulas in applied research, such as Archimedean and meta-elliptical copulas, typically assume that the dependence structure between the variables of interest is symmetric. This symmetry assumption, while useful in many contexts, may not always be appropriate for capturing real-world dependencies. Asymmetry, which refers to differences in the joint behavior of the upper and lower tails of multivariate distributions, has gained significant attention in the copula literature. As noted by [28], asymmetry is crucial for accurately modeling extreme dependencies, particularly when the tail behavior of variables differs. The concept of asymmetry in copulas is thus defined by the contrasting behaviors of these tails, which can indicate the presence of different strengths of dependence at the extreme ends of the distribution. Below, we provide a more formal definition of the symmetry properties

of copulas and discuss how these properties influence their ability to model asymmetric dependence structures.

Definition 6.3.1 *A d-dimensional copula C is exchangeable or permutation symmetric if it is the distribution function of an uniform vector $\mathbf{U} = (U_1, \ldots, U_d)^\intercal$ satisfying*

$$C(u_1, \ldots, u_d) = C(u_{r(1)}, \ldots, u_{r(d)})$$

for any permutation $r \in \Gamma$, where Γ denotes the set of all permutations on the set $\{1, \ldots, d\}$.

Note that a d-dimensional continuous random vector $\mathbf{X}$ is exchangeable if and only if the marginal CDFs are identical and the copula is exchangeable. Most of the commonly used two-parameter bivariate copula families are exchangeable.

Definition 6.3.2 *A d-dimensional copula C is reflection symmetric if $\mathbf{U}$ has the same distribution as $\mathbf{1} - \mathbf{U}$ where $\mathbf{1} - \mathbf{U} = (1 - U_1, \ldots, 1 - U_d)^\intercal$.*

The definition of reflection or central symmetry is that a d-dimensional random vector $\mathbf{X}$ as centrally symmetric about $\mathbf{a} = (a_1, \ldots, a_d)^\intercal$ if, and only if each X_i is marginally symmetric about a_i and the corresponding copula C is reflection symmetric. If $\mathbf{U} \sim C$ and $\mathbf{1} - \mathbf{U} \sim \hat{C}$ then $\hat{C}$ is called a reflected copula of C. [26] called the condition of reflection symmetry, $C \equiv \hat{C}$ as radial symmetry. If the copula density as in (2.4) exists, then reflection symmetry implies

$$c(u_1, \ldots, u_d) = c(1 - u_1, \ldots, 1 - u_d), \quad \mathbf{u} \in [0, 1]^d. \tag{6.14}$$

The simplest one-parameter bivariate copula families are exchangeable but not necessarily reflection symmetric. For practical situations these assumptions on copulas are too restrictive and they need to be relaxed for flexible dependence modeling. We can see that the GSN copula in (6.8) is asymmetric in general. The following theorem state that it can be symmetric under certain situations. We continue with the assumption, $p \in (0, 1)$.

Theorem 6.3.1 *A d-dimensional $C_{d,GSN}(p, \mu, \mathbf{\Sigma})$ copula is exchangeable if the correlation matrix $\mathbf{\Sigma}$ is exchangeable and $\mu_j = \mu \in \mathcal{R}$ for all $j = 1, \ldots, d$. Moreover, it is radially symmetric when $\mu = 0$.*

Proof: Let $\mathbf{U} = (U_1, \ldots, U_d)^\intercal \sim C_{d,GSN}(p, \mu, \mathbf{\Sigma})$ and take $x_j = F(u_j|\mu_j, i, p)$, for $j = 1, \ldots, d$. Then

$$\mathbf{X} = (X_1, \ldots, X_d)^\intercal \sim MGSN_d(p, \mu, \mathbf{\Sigma}).$$

Therefore, for any permutation $r \in \Gamma$ we have

$$\mathbf{X}_r = (X_{r(1)}, \ldots, X_{r(d)})^\intercal \sim MGSN_d(p, \mu_r, \mathbf{\Sigma}_r)$$

where μ_r and $\mathbf{\Sigma}_r$ are the corresponding rearrangements of μ and $\mathbf{\Sigma}$ respectively. The moment generating function of $\mathbf{X}_r$ is given as

$$M_{\mathbf{X}_r}(\mathbf{t}) = \frac{p \exp\left[\mu_r^\intercal \mathbf{t} + \frac{1}{2}\mathbf{t}^\intercal \mathbf{\Sigma}_r \mathbf{t}\right]}{1-(1-p)\exp\left[\mu_r^\intercal \mathbf{t} + \frac{1}{2}\mathbf{t}^\intercal \mathbf{\Sigma}_r \mathbf{t}\right]}.$$

Since $\mathbf{\Sigma}$ is a correlation matrix, the quadratic form $\mathbf{t}^\intercal \mathbf{\Sigma}_r \mathbf{t} = \mathbf{t}^\intercal \mathbf{\Sigma} \mathbf{t}$ for all $\mathbf{t}$ and hence $M_{\mathbf{X}_r}(\mathbf{t}) = M_{\mathbf{X}}(\mathbf{t})$, if and only if $\mu_j = \mu \in \mathcal{R}$ for all $j = 1, \ldots, d$. Now the univariate geometric skew-normal density follows

$$f_X(x|\mu, 1, p) = f_X(-x| - \mu, 1, p), \quad \mu \in \mathcal{R}, 0 < p \leq 1.$$

Hence, $X_i \overset{d}{=} -X_i$ if and only if $\mu = 0$ and that completes the proof. □

6.4 Maximum likelihood estimation

Let us assume that the data have been transformed into m independent vector valued observations, $\mathbf{u}_i \in [0,1]^d, i = 1, \ldots, m$ using some parametric or non-parametric distribution function. Then the set of observations $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ is called a 'pseudo sample', which provide only the 'dependence' information of the data. The log-likelihood function for the copula parameters based on a sample $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ of size m from $C_{d,GSN}(p, \mu, \mathbf{\Sigma})$, using (6.9) is given as

$$\begin{aligned} l(p, \mu, \mathbf{\Sigma}|\mathbf{u}_1, \ldots, \mathbf{u}_m) &= \sum_{i=1}^{m} l_i(p, \mu, \mathbf{\Sigma}|\mathbf{u}_i) \\ &= \sum_{i=1}^{m} \Big[\log\Big(\sum_{k=1}^{\infty} \frac{p(1-p)^{k-1}}{(2\pi k)^{d/2}\sqrt{|\mathbf{\Sigma}|}} e^{-\frac{1}{2k}(\mathbf{x}_i - k\mu)^\intercal \mathbf{\Sigma}^{-1}(\mathbf{x}_i - k\mu)} \Big) \\ &\qquad - \sum_{j=1}^{d} \log\Big(\sum_{k=1}^{\infty} \frac{p(1-p)^{k-1}}{\sqrt{2\pi k}} e^{-\frac{1}{2k}(x_{ij} - k\mu_j)^2} \Big) \Big], \end{aligned} \quad (6.15)$$

where $x_{ij} = F_1^{-1}(u_{ij}|\mu_j, 1, p)$ denotes the corresponding quantiles. The maximum likelihood estimators (MLEs) can be obtained by maximizing (6.15) with respect to the unknown parameters. The main advantage of the GSN copula lies in its efficiency in the parameter estimation as the quantile function is independent of the correlation matrix $\mathbf{\Sigma}$. Direct maximization of (6.15) is quite a challenging issue since it involves solving a $(d + 1 + d(d+1)/2)$ dimensional optimization problem. The problem becomes more severe when d is large. [152] pointed out that the density function of a geometric skew-normal distribution can be multimodal and so is the log-likelihood. He provided an EM algorithm to maximize the log-likelihood of an MGSN distribution but that is not applicable for the GSN copula. The effective solution to the maximization problem here is to decompose it into simpler sub-problems.

[161] introduced globally convergent block-coordinate techniques for unconstrained optimization. The core idea is to break down a complex optimization problem into two simpler estimation subproblems. Rather than estimating all the parameters at once from the objective function, the parameter set can be divided into two disjoint blocks. This allows one block to be optimized at a time while keeping the other block fixed. The authors demonstrate that, under suitable convergence conditions—defined by the operations applied to each block and appropriate (sequential or parallel) coordination rules—the two-block decomposition algorithm converges globally to stationary points, even without assuming convexity or uniqueness. Since the quantile functions in the GSN copula only involves the parameters $\{p, \mu\}$, two block-coordinate ascent algorithm provides with very efficient estimation of the parameters. It doesn't require additional restrictions to ensure the positive definiteness of the correlation matrix $\mathbf{\Sigma}$. Also the quantile function can be computed relatively fast, with the infinite sum in the distribution function is approximated upto some finite values. Here we apply the Newton's method to obtain the quantiles of the GSN copula. Here the parameters involved in (6.15) are partitioned into $\theta = \{\theta_1, \theta_2\}$, where $\theta_1 = \{p, \mu\}$ and $\theta_2 = \mathbf{\Sigma}$. Starting with some initial approximations, the following algorithm iteratively updates the parameters over one of the blocks by maximizing (6.15), while keeping the other block fixed at their current values.

Algorithm 6.4.1 *(Two block-coordinate ascent algorithm for the GSN copula)*

- *Step 1: Start with some initial approximations of $\hat{\theta}_1^0$ and $\hat{\theta}_2^0$.*
- *Step 2: At the r-th iteration, update the estimate $\hat{\theta}_1^r$ by maximizing $l(\theta_1, \theta_2)$ over θ_1 when θ_2 is fixed at $\hat{\theta}_2^{r-1}$, i.e.*
$$\hat{\theta}_1^r := \arg\max_{\theta_1}\{l(\theta_1, \hat{\theta}_2^{r-1})\}.$$
- *Step 3: At the r-th iteration, update the estimate $\hat{\theta}_2^r$ by maximizing $l(\theta_1, \theta_2)$ over θ_2 when θ_1 is fixed at $\hat{\theta}_1^r$, i.e.*
$$\hat{\theta}_2^r := \arg\max_{\theta_2}\{l(\hat{\theta}_1^r, \theta_2)\}.$$
- *Step 4: Repeat Steps 2 and 3 until the algorithm converges.*

Standard numerical optimization method with box constraints, L-BFGS-B can be used to find the maximum likelihood estimates in Steps 2 and 3 with bounds for the parameter p as $(0, 1)$ and for the correlation parameters $\{\rho_{ij}; 1 \leq i < j \leq d\}$ as $(-1, 1)$, respectively. Algorithm 6.4.1 is also applicable to estimate the parameters from the log-likelihood of MGSN distribution. The initial approximations of $\hat{\theta}_1^0$ and $\hat{\theta}_2^0$ can be chosen by the combination of method of moments estimation and profile likelihood based on the observed data which results in faster convergence of the proposed algorithm. For a fixed value of p, the MOM estimates of μ and $\mathbf{\Sigma}$ are given as (referring to Result 6.1.3)

$$\tilde{\mu} = p\bar{\mathbf{X}} \text{ and } \tilde{\mathbf{\Sigma}} = p\mathbf{S}_{\mathbf{X}} - p(1-p)\bar{\mathbf{X}}\bar{\mathbf{X}}^{\intercal}. \tag{6.16}$$

Therefore, the MLE of p, denoted by $\tilde{p}$ can be obtained by maximizing the profile log-likelihood function of MGSN distribution with known μ and $\mathbf{\Sigma}$, i.e. $l(p, \tilde{\mu}, \tilde{\mathbf{\Sigma}})$, with respect to p. Finally the initial estimates of μ and $\mathbf{\Sigma}$ become

$$\tilde{\mu} = \tilde{\mu}(\tilde{p}) \text{ and } \tilde{\mathbf{\Sigma}} = \tilde{\mathbf{\Sigma}}(\tilde{p}), \text{ respectively.} \tag{6.17}$$

The above algorithm is applicable for the general structure of GSN copula for moderate to high dimensions, which is verified in one of our simulation study with unrestricted μ and $\mathbf{\Sigma}$. In this setting, the observed information matrix for (6.15) can be numerically obtained as

$$I_m(\theta) := \sum_{i=1}^{m} \frac{\partial}{\partial \theta} l_i(\theta|\mathbf{u_i}) \frac{\partial}{\partial \theta^\intercal} l_i(\theta|\mathbf{u}_i), \tag{6.18}$$

which can be used to get the standard errors of the parameter estimates.

The implementation of Algorithm 6.4.1 leverages the structural elegance of the Geometric Skew-Normal copula, where the quantile functions depend solely on the parameters $\theta_1 = \{p, \mu\}$, effectively decoupling them from the correlation structure $\theta_2 = \mathbf{\Sigma}$. By partitioning the optimization into these two disjoint blocks, the algorithm addresses the computational burden of a $(d + 1 + d(d + 1)/2)$ dimensional problem, transforming a potentially divergent high-dimensional search into simpler, manageable sub-problems. This specific partitioning exploits the near block-diagonal structure of the information matrix, which reduces the 'zig-zagging' inefficiency often associated with coordinate descent in poorly specified models ([132]. Furthermore, while the GSN log-likelihood (6.15) is known to be potentially multimodal, the global convergence properties established by [162] and [163] provide a rigorous safeguard; they ensure that the iterative updates between the skewness-location block and the dependence block converge to a stationary point even in the absence of global convexity. The integration of L-BFGS-B for constrained sub-steps, combined with MOM initializations (6.17–6.18), further mitigates the risk of local optima. Consequently, for the moderate-to-high dimensional settings investigated, the BCA approach offers a robust, scalable, and numerically stable alternative to direct maximization, ensuring high-fidelity parameter estimation and reliable standard errors via the observed information matrix (6.18).

6.5 Regression models for longitudinal data

In longitudinal studies, repeated measurements are collected over time to assess the evolution of the responses with respect to some covariates. Suppose $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\intercal$ be a vector of n_i dependent responses for i-th subject. The marginal cumulative distribution of a single variable Y_{ij} is denoted by $F(Y_{ij}|\mathbf{x}_{ij}, \beta)$ and depends on a s-dimensional vector of covariates $\mathbf{x}_{ij}$ and a regression parameter β (we use here s here for the number of regression parameters). Our scientific objective is to evaluate how the distribution of Y_{ij} varies according to the changes in a vector of s covariates $\mathbf{x}_{ij}$ as well as the dependence among $\mathbf{Y}_i$. When $\mathbf{Y}_i$ is continuous, we consider the marginals to follow a generalized

linear model as

$$g(E(Y_{ij}|\mathbf{x}_{ij})) = \mathbf{x}_{ij}\beta, \quad j = 1, \ldots, n_i, \tag{6.19}$$

where $g(.)$ is a suitable link function and β is an $s \times 1$ vector of regression coefficients. However, under the copula framework any kind of marginals can be used other than those belonging to the exponential family. Then the joint distribution function of $\mathbf{Y}_i$ given $\mathbf{x}_i$ can be expressed as

$$F_{n_i}(y_{i1}, \ldots, y_{in_i}|\mathbf{x}_i) = C_{n_i}(F(y_{i1}|\mathbf{x}_{i1}), \ldots, F(y_{in_i}|\mathbf{x}_{in_i})|\phi_i), \tag{6.20}$$

where $C_{n_i}(.|\phi_i)$ is a n_i-dimensional copula with parameter vector ϕ_i. The corresponding density function is given by

$$f_{n_i}(y_{i1}, \ldots, y_{in_i}|\mathbf{x}_i) = c_{n_i}(u_{i1}, \ldots, u_{in_i}|\phi_i) \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{x}_{ij}), \tag{6.21}$$

where $u_{ij} = F(y_{ij}|\mathbf{x}_{ij})$. The copula identifies a regression model constructed in way to (i) preserve the marginal univariate distributions and (ii) have separate dependence structure. For a fixed set of marginals different multivariate models can be constructed by various choices of the copula function. Based on m independent observations, the log-likelihood of (6.20) is given as

$$l(\beta, \phi|\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{m} \log c_{n_i}(u_{i1}, \ldots, u_{in_i}|\phi_i) + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log f(y_{ij}|\mathbf{x}_{ij}). \tag{6.22}$$

We note that the copula can be uniquely identified for continuous dependent random variables, but that is not the case with discrete variables. Hence latent variable formulation ([97]) can be used to construct ordered probit models for ordinal data. Let Y_{ij} represent a categorical response with K possible ordered categories and let Z_{ij} be a normally distributed latent variable underneath Y_{ij}. Let $\gamma(k),\ 1 < k < K-1$, be ordered thresholds such that: $-\infty = \gamma(0) < \gamma(1) < \cdots < \gamma(K-1) < \gamma(K) = \infty$. Then the ordinal variable have the stochastic representation as

$$Y_{ij} = k \ \text{ if } \ \gamma(k-1) \leq Z_{ij} < \gamma(k), \ \ k \in \{1, \ldots, K\}. \tag{6.23}$$

The threshold parameters can be fixed or freely estimated based on specification of the model. Note that the monotonic increasing nature of the thresholds accounts for the ordered nature of the observed outcomes. We model the latent variable Z_{ij}, based on covariate vector $\mathbf{x}_{ij}$ as

$$Z_{ij}|\mathbf{x}_{ij} = \mathbf{x}_{ij}\beta + \epsilon_{ij}, \ \ j = 1, \ldots, n_i, \tag{6.24}$$

where β is a vector of regression coefficients and ϵ_{ij} is the error term. To ensure the identifiability of the model we assume $\epsilon_{ij} \sim N(0, 1)$ and the intercept of β equal to zero. Therefore the dependence structure of the observed response vector $\mathbf{Y}_i$ is explained through the dependence of the latent vector

$\mathbf{Z}_i$ (see, [164]). Then the joint probability mass function can be written as

$$\begin{aligned} &P(Y_{i1} = y_{i1}, \ldots, Y_{in_i} = y_{in_i}|\mathbf{x}_i) \\ &\qquad = P(\gamma(y_{i1}-1) \leq Z_{i1} < \gamma(y_{i1}), \ldots, \gamma(y_{in_i}-1) \leq Z_{in_i} < \gamma(y_{in_i})) \\ &= \int_{\gamma(y_{i1}-1)}^{\gamma(y_{i1})} \cdots \int_{\gamma(y_{in_i}-1)}^{\gamma(y_{n_i})} c_{n_i}(F(z_{i1}|\mathbf{x}_{i1}), \ldots, F(z_{in_i}|\mathbf{x}_{in_i})|\phi_i) \prod_{j=1}^{n_i} f(z_{ij}|\mathbf{x}_{ij}) dz_{i1} \ldots dz_{in_i}. \end{aligned} \quad (6.25)$$

The joint PMF in (6.25) involves n_i-dimensional integral which can be obtained as a finite difference of the CDF of $\mathbf{Z}_i$ as noted by [165] and [140]. But as the dimension n_i increases, evaluating the rectangular probability becomes computationally infeasible as one need to consider summation of $2n_i$ many terms ([9]).

To circumvent the computational issues associated with discrete Gaussian copula regression models composite likelihood methods (CML) are often employed (see, [142] and [87]). These pseudo-likelihood methods are useful when all the multivariate parameters can be identified from lower dimensional marginals. Similar to multivariate Gaussian or Student-t copula, our proposed GSN copula permits to construct composite likelihood combining likelihoods for pairs of observation, because of the description of the quantile function in (6.8). The pair-wise likelihood approach that we implement here, involves only 2-dimensional integrals. This is a major computational advantage, if compared with the skew-elliptical copulas derived from [154]. Based on m independent observations the pairwise log-likelihood can be written as

$$\begin{aligned} &l_c(\beta, \phi|\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} \log P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik}|\mathbf{x}_{i(j,k)}) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} \log \Big[C_2(u_{ij}, u_{ik}|\phi_{jk}) - C_2(u_{ij}^-, u_{ik}|\phi_{jk}) - C_2(u_{ij}, u_{ik}^-|\phi_{jk}) + C_2(u_{ij}^-, u_{ik}^-|\phi_{jk}) \Big], \end{aligned} \quad (6.26)$$

where $u_{il} = \Phi(\gamma(y_{il}) - \mathbf{x}_{il}\beta)$ and $u_{il}^- = \Phi(\gamma(y_{il}-1) - \mathbf{x}_{il}\beta)$, for $l = j, k$ respectively. We use geometric skew-normal copula to construct flexible dependence models for continuous and discrete longitudinal data. [15] discussed Gaussian copula regression models along with their computational implementations in R. In order to account for the within-subject dependency or temporal dependency, appropriate structures for the correlation matrix $\mathbf{\Sigma}$ can be considered in our GSN copula based regression models. In particular, we implement the following structures as

- Exchangeable (EX): $\rho_{jk} = \rho \in (-\frac{1}{n_i-1}, 1), \quad 1 \leq j < k \leq n_i$;
- AR(1) with exponential decay : $\rho_{jk} = \exp(-\xi|t_j - t_k|), \xi > 0, \quad 1 \leq j < k \leq n_i$.

Additionally, we assume equal value of the parameter μ for each dimension, i.e. $\mu_j = \mu, \;\; j = 1, \ldots, n_i$. That leads to reduction in the number of estimable parameters from the model, for moderate to high

dimensional longitudinal data. The parameters of the considered regression models can be obtained from the likelihood and pseudo-likelihood functions in (6.22) and (6.26). But under the regression setup, direct maximization of these is still computationally challenging, especially for complex dependence structure of the GSN copula, since we have additional marginal parameters in the regression models. To obtain valid parameter estimates, we employ the two-stage estimation method often known as inference function for margins (IFM) by [35] and [36]. Under this, we estimate the marginal parameters, say θ, from marginal likelihoods assuming independence. Then at the second step, the copula parameters, say ϕ, are estimated from the multivariate likelihood or the composite likelihood with univariate parameter estimates held fixed. The standard errors of the parameter estimates $\hat{\theta}^* = (\hat{\theta}, \hat{\phi})^\intercal$ can be numerically obtained from the observed sandwich information matrix (Godambe information matrix) as

$$J(\hat{\theta}^*) = D(\hat{\theta}^*)^\intercal M(\hat{\theta}^*)^{-1} D(\hat{\theta}^*),$$

where $D(\hat{\theta}^*)$ is a block diagonal matrix and $M(\hat{\theta}^*)$ is a symmetric positive definite matrix. The explicit forms of these can be found in [84] or [28]. To estimate the parameters we use *optim* ([64]) function, and to estimate the information matrix associted with the parameter estimates we use *numderiv* ([65]) function in R.

6.6 Model comparison

For our proposed GSN copula based regression models we wish to compare the fits with the corresponding Gaussian copula based regression models with the same structure of the correlation matrix and investigate for improvements, if any. For this purpose we consider Akaike information criterion and one of its modified version, evaluated at the parameter estimates $\hat{\theta}^*$ as

- AIC under correct specification of the copula and the marginals by [69]:

$$AIC = -2l(\hat{\theta}^*) + 2\dim(\theta^*), \tag{6.27}$$

- Composite likelihood version of AIC, as given in [145]:

$$CLAIC = -2l_c(\hat{\theta}^*) + 2tr(M(\theta^*)D(\theta^*)^{-1}), \tag{6.28}$$

for the continuous and ordinal regression models respectively. The smaller values of these criteria leads to better fitting regression models. Note that the matrices used in CLAIC is not same as in 6.27. But we use the exact form of CLAIC as mentioned in the previously cited paper by assuming the two step estimates are very close to the estimates if one step estimation of the composite likelihood was obtained. But with this we can compare different models as will be described next.

In addition, we will also use Voung's test ([166]) to show if a GSN copula based model provides a better fit than Gaussian copula model with same structure of the correlation matrix. Voung's test is

the sample version of the difference in Kullback-Leibler divergence and sample size to differentiate two models which could be non-nested. This test has been used extensively in the copula literature to compare vine copula models (e.g., [167]; [28] or [168]). Here we provide the details in a general context. Assume that we have two models M_1 and M_2, with parametric densities $f_{\mathbf{y}}^{(1)}$ and $f_{\mathbf{y}}^{(2)}$ respectively, we can compare

$$\begin{aligned}\Delta_{1f} &= \frac{1}{m}\Big[\sum_{i=1}^{m}\Big\{E_f \log f_{\mathbf{y}}(\mathbf{y}_i) - E_f \log f_{\mathbf{y}}^{(1)}(\mathbf{y}_i|\theta_1^*)\Big\}\Big],\\ \Delta_{2f} &= \frac{1}{m}\Big[\sum_{i=1}^{m}\Big\{E_f \log f_{\mathbf{y}}(\mathbf{y}_i) - E_f \log f_{\mathbf{y}}^{(2)}(\mathbf{y}_i|\theta_2^*)\Big\}\Big],\end{aligned} \tag{6.29}$$

where θ_1, θ_2 are the parameters in models M_1 and M_2, respectively, that lead to the closest Kullback-Leibler divergence to the true $f_{\mathbf{y}}$; equivalently, they are the limits in probability of the ML estimates based on models M_1 and M_2, respectively. Model M_1 is closer to the true $f_{\mathbf{y}}$, i.e., it is the better-fitting model if $\Delta_{12} = \Delta_{1f} - \Delta_{2f} < 0$, and Model M_2 is the better-fitting model if $\Delta_{12} > 0$. The sample version of Δ_{12} with ML estimates $\hat{\theta}_1^*, \hat{\theta}_2^*$ is

$$\bar{D}_{12} = \frac{1}{m}\sum_{i=1}^{m} D_i, \quad \text{where} \quad D_i = \log \frac{f_{\mathbf{y}}^{(2)}(\mathbf{y}_i|\hat{\theta}_2^*)}{f_{\mathbf{y}}^{(1)}(\mathbf{y}_i|\hat{\theta}_1^*)}. \tag{6.30}$$

In our setup we use two stage estimates assuming they are very close to the true ML estimates. For non-nested or nested models where $f_{\mathbf{y}}^{(1)}(\mathbf{y}_i|\theta_1^*)$ and $f_{\mathbf{y}}^{(2)}(\mathbf{y}_i|\theta_2^*)$ are not the same density, a large sample 95% confidence interval (CI) for the parameter Δ_{12} is

$$\bar{D}_{12} \pm 1.96 \times \frac{\bar{s}_{12}}{\sqrt{m}}, \quad \text{where} \quad \bar{s}_{12} = \frac{1}{m-1}\sum_{i=1}^{m}(D_i - \bar{D}_{12})^2. \tag{6.31}$$

If the interval in (6.31) contains 0, models M_1 and M_2 would not be considered significantly different, which can be used as a diagnostic in our comparison. Point to be noted that, Voung's test is applicable for both likelihood and pseudo-likelihood based methods, but the condition is the expectations defined in (6.29), should consistently estimate the model parameters. However, Voung's test does not evaluate whether any of the models provide a sufficiently good fit, specifically in terms of how closely the model approximates the true data-generating mechanism.

6.7 Simulation studies

In this section we investigate the finite sample performance of the parametric inference of the proposed GSN copula and the considered regression models in Section 6.5. We consider 3 simulation studies. We generate random data sets form the respective models and then estimate the parameters using the methods described in Section 6.4 and 6.5. For each simulations we consider two different

sample sizes as $m = 200$ and 500 and for the first simulation we consider the number of samples to be 200 and in the subsequent 2 simulations we consider the replication number to be 500.

For the unrestricted case of the MGSN distribution and the GSN copula we consider the following set of parameters -

$$p = 0.5, \quad \mu = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & 0.6 & 0.4 & 0.2 \\ 0.6 & 1 & 0.2 & 0.4 \\ 0.4 & 0.2 & 1 & 0.2 \\ 0.2 & 0.4 & 0.2 & 1 \end{pmatrix}. \tag{6.32}$$

The data sets, say $\mathcal{X}$ (MGSN distribution) and $\mathcal{U}$ (GSN copula) of size m are generated using Algorithm 6.2.1. Then we calculate the MLEs of the parameters using Algorithm 6.4.1, for the distribution and the copula data, respectively. The maximization of the log-likelihood of the GSN copula takes significantly more time than the log-likelihood of the MGSN distribution, since it involves computation of the quantiles in each block. Two block-coordinate ascent algorithm converges within 5 iterations for the MGSN distribution and 15 iterations for the GSN copula respectively. The starting values of the parameters are obtained by the methods discussed in Section 4.

Next we consider regression models for continuous data with generalized linear model for the marginals including a continuous time-varying covariate and GSN copula. We take structured correlation matrix $\mathbf{\Sigma}$ and equal value of μ denoted as $\bar{\mu}$ for the GSN copula. The response distribution for all the marginals are taken as Gamma with $\log$ link function. First we sample the copula data and then use PIT to generate response variables from the model -

$$g(E(Y_{ij})) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + t_{ij}\beta_3, \quad j = 1, \ldots, 4, \tag{6.33}$$

where the response distribution is Gamma ($\log$-link). For values of the marginal parameters we set $\beta_0 = 1.0$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, $\beta_3 = 1.0$ and the shape parameter $\kappa = 3$. The covariates are generated as $x_{i1} \sim Ber(p = 0.5)$, $x_{i2} \sim N(5, 4)$ and the time points $t_{ij} = j$ for $j = 1, \ldots, 4$. We consider exchangeable (EX) and AR(1) correlation structure for the matrix $\mathbf{\Sigma}$, and in both the scenarios we set the autocorrelation parameter $\xi = 0.50$. Finally for other two parameters of the reduced GSN copula we set $p = 0.5$ and $\bar{\mu} = 1.0$, respectively.

Finally we consider regression models for ordinal data with latent latent variable formulation including a continuous time-varying covariate and the dependence structure is framed through GSN copula. The parametric structure of the GSN copula is same as the previous study. Here we consider the ordered probit model as -

$$\begin{aligned} &Y_{ij} = k \text{ if } \gamma(k-1) \leq Z_{ij} < \gamma(k), \quad k = 1, \ldots, 4, \\ &Z_{ij} = x_{i1}\beta_1 + x_{i2}\beta_2 + t_{ij}\beta_3 + \epsilon_{ij}, \quad j = 1, \ldots, 4, \end{aligned} \tag{6.34}$$

		m = 200					**m** = 500				
Parameters	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
MGSN Distribution											
p	0.5	0.5172	0.0172	0.0676	0.0681	0.0698	0.5092	0.0092	0.0495	0.0491	0.0504
μ_1	0.0	-0.0035	-0.0035	0.0526	0.0525	0.0527	0.0027	0.0027	0.0289	0.0287	0.0289
μ_2	0.0	-0.0060	-0.0060	0.0534	0.0536	0.0538	0.0035	0.0035	0.0331	0.0330	0.0333
μ_3	1.0	1.0285	0.0285	0.1376	0.1380	0.1405	1.0170	0.0170	0.0996	0.0990	0.1011
μ_4	1.0	1.0254	0.0254	0.1421	0.1432	0.1443	1.0136	0.0136	0.1043	0.1041	0.1052
ρ_{12}	0.6	0.5955	-0.0045	0.0444	0.0446	0.0446	0.5945	-0.0055	0.0248	0.0247	0.0252
ρ_{13}	0.4	0.3895	-0.0105	0.0801	0.0818	0.0833	0.3928	-0.0072	0.0544	0.0538	0.0561
ρ_{14}	0.2	0.1932	-0.0068	0.0998	0.1008	0.1011	0.1937	-0.0063	0.0606	0.0605	0.0610
ρ_{23}	0.2	0.1947	-0.0053	0.1106	0.1117	0.1122	0.1950	-0.0050	0.0578	0.0572	0.0581
ρ_{24}	0.4	0.3963	-0.0037	0.0886	0.0888	0.0891	0.3965	-0.0035	0.0605	0.0602	0.0608
ρ_{34}	0.2	0.2179	0.0179	0.0943	0.0950	0.0956	0.2123	0.0123	0.0667	0.0663	0.0679
GSN Copula											
p	0.5	0.5266	0.0266	0.1051	0.1062	0.1084	0.5019	0.0019	0.0630	0.0628	0.0631
μ_1	0.0	-0.0272	-0.0272	0.1325	0.1333	0.1353	-0.0036	-0.0036	0.0767	0.0762	0.0768
μ_2	0.0	-0.0285	-0.0285	0.1482	0.1490	0.1509	0.0074	0.0074	0.0751	0.0749	0.0754
μ_3	1.0	1.1805	0.1805	0.4956	0.5117	0.5274	1.0697	0.0697	0.2124	0.2112	0.2235
μ_4	1.0	1.0704	0.0704	0.3294	0.3312	0.3368	1.0550	0.0550	0.2098	0.2090	0.2169
ρ_{12}	0.6	0.5992	-0.0008	0.0477	0.0482	0.0478	0.5976	-0.0024	0.0277	0.0275	0.0286
ρ_{13}	0.4	0.4133	0.0133	0.0904	0.0911	0.0914	0.4063	0.0063	0.0612	0.0610	0.0615
ρ_{14}	0.2	0.2144	0.0144	0.1059	0.1063	0.1069	0.1979	-0.0021	0.0694	0.0690	0.0697
ρ_{23}	0.2	0.2004	0.0004	0.1253	0.1258	0.1253	0.1974	-0.0026	0.0600	0.0599	0.0604
ρ_{24}	0.4	0.4106	0.0106	0.0909	0.0912	0.0915	0.3976	-0.0024	0.0639	0.0632	0.0642
ρ_{34}	0.2	0.1907	-0.0093	0.1189	0.1192	0.1193	0.1893	-0.0107	0.0826	0.0821	0.0841

Table 6.1 Parameter estimation for multivariate geometric skew-normal distribution and geometric skew-normal copula for $N = 200$ simulated data sets with two different sample sizes.

		m = 200					**m** = 500				
Parameters	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
Exchangeable											
β_0	1.0	1.0163	0.0163	0.1114	0.1079	0.1126	1.0144	0.0144	0.0744	0.0692	0.0750
β_1	0.5	0.4995	-0.0005	0.0752	0.0744	0.0752	0.5015	0.0015	0.0487	0.0473	0.0487
β_2	0.5	0.4974	-0.0026	0.0188	0.0185	0.0189	0.4968	-0.0032	0.0120	0.0118	0.0124
β_3	1.0	0.9977	-0.0023	0.0096	0.0082	0.0099	0.9983	-0.0017	0.0060	0.0052	0.0062
κ	3.0	2.9994	-0.0006	0.2380	0.2227	0.2380	2.9962	-0.0038	0.1630	0.1411	0.1637
p	0.5	0.5095	0.0095	0.1236	0.1079	0.1240	0.4967	-0.0033	0.0744	0.0646	0.0744
ξ	0.5	0.5113	0.0113	0.1024	0.1001	0.1030	0.5023	0.0023	0.0607	0.0596	0.0607
$\bar{\mu}$	1.0	1.0469	0.0469	0.2962	0.2848	0.2999	1.0081	0.0081	0.1687	0.1697	0.1689
Autoregressive											
β_0	1.0	1.0064	0.0064	0.1171	0.1072	0.1173	1.0027	0.0027	0.0724	0.0684	0.0753
β_1	0.5	0.5041	0.0041	0.0745	0.0718	0.0746	0.4957	-0.0043	0.0456	0.0458	0.0458
β_2	0.5	0.4980	-0.0020	0.0199	0.0178	0.0200	0.4976	-0.0024	0.0117	0.0114	0.0120
β_3	1.0	0.9996	-0.0004	0.0125	0.0123	0.0125	0.9989	-0.0011	0.0086	0.0078	0.0088
κ	3.0	3.0124	0.0124	0.2363	0.2283	0.2366	2.9831	-0.0169	0.1483	0.1307	0.1498
p	0.5	0.5049	0.0049	0.1114	0.0960	0.1115	0.5041	0.0041	0.0803	0.0606	0.0804
ξ	0.5	0.5106	0.0106	0.0810	0.0764	0.0817	0.4968	-0.0032	0.0529	0.0468	0.0530
$\bar{\mu}$	1.0	1.0355	0.0355	0.2740	0.2466	0.2763	1.0226	0.0226	0.1833	0.1612	0.1847

Table 6.2 Parameter estimation for geometric skew-normal copula model with Gamma marginals for $N = 500$ simulated data sets. Exchangeable and autoregressive correlation structures are considered.

where $\epsilon_{ij}(i.i.d) \sim N(0,1)$. Here we set the marginal parameters, $\beta_1 = 0.5$, $\beta_2 = 0.5$, $\beta_3 = 1.0$ and the threshold parameters $\gamma_1 = 2.0$, $\gamma_2 = 4.0$, $\gamma_3 = 6.0$, respectively. The covariates are generated in the similar format of the previous study. First we generate the copula data using same set of parameters as previous, then use (6.34) to obtain the ordinal response variables.

Tables 5.1, 5.2, and 5.3 present the simulation results for the models considered in this study. Specifically, we report the mean estimates, biases [$\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_j^* - \theta^*$], empirical standard deviations (denoted as SD), average standard errors derived from the asymptotic covariance matrices (denoted as SE), and the root mean square errors (RMSE) [$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_j^* - \theta^*)^2}$], where $\hat{\theta}_j^*$ represents the parameter estimates for the j-th sample. The results show that the mean estimates closely align with the true parameter values, and the RMSEs decrease as the sample size increases, indicating the robustness of the estimation process. In Table 6.1, we observe that the standard errors of the parameters $\mu_j; j = 1, \ldots, 4$ are slightly larger when estimated from the GSN copula likelihood. This increase can be attributed to the additional complexity introduced by the quantile calculation involved in the GSN copula estimation. In Tables 6.2 and 6.3, we note that the bias and RMSE for the autocorrelation parameter ξ are slightly higher when using the EX correlation matrix Σ compared to the AR(1) correlation matrix. Despite these minor differences, the standard errors and empirical standard deviations are consistent across all models, supporting the validity of the proposed methods. Overall, the encouraging results from these simulations suggest that our models are promising candidates for application to real-world datasets.

6.8 Applications

In this section we illustrate the flexibility of the regression models described in this paper through some examples, and compare the fits with the corresponding Gaussian copula based models. The datasets considered are publicly available in several R packages such as *qrLMM* or *mixor*.

6.8.1 Framingham heart study

This is a benchmark data set in longitudinal studies, which was previously analyzed by several authors, such as [169] and [170]. The data set provides cholesterol levels over time, age at baseline and gender for 200 randomly selected patients, measured at the beginning of the study and every two years for a total of 10 years. The primary objective is to model the change of cholesterol levels over time within patients. However, we apply KNN method to impute the missing entries in this data set beforehand. Since the cholesterol levels observed to be positively skewed, we consider Gamma distribution (log-link) for the marginals. We adopt the following model -

$$g(E(Y_{ij})) = \beta_0 + \text{sex}_i\beta_1 + \text{age}_i\beta_2 + t_{ij}\beta_3, \quad j = 1, \ldots, 6, \tag{6.35}$$

where observed y_{ij} is cholesterol level divided by 100 at the j-th time for subject i. The available covariates are: $t_{ij} = (\text{time} - 5)/10$ (time measured in years), sex (0 = female, 1 = male) and age at

		m = 200					**m** = 500				
Parameters	**True Value**	Mean	Bias	SD	SE	RMSE	Mean	Bias	SD	SE	RMSE
Exchangeable											
β_1	0.5	0.5048	0.0048	0.1348	0.1388	0.1348	0.5027	0.0027	0.0841	0.0878	0.0841
β_2	0.5	0.5081	0.0081	0.0423	0.0420	0.0431	0.5030	0.0030	0.0264	0.0264	0.0271
β_3	1.0	1.0144	0.0144	0.0531	0.0536	0.0550	1.0039	0.0039	0.0354	0.0338	0.0356
γ_1	2.0	2.0300	0.0300	0.2563	0.2450	0.2580	2.0149	0.0149	0.1618	0.1550	0.1625
γ_2	4.0	4.0600	0.0600	0.2795	0.2853	0.2858	4.0270	0.0270	0.1845	0.1801	0.1865
γ_3	6.0	6.0865	0.0865	0.3530	0.3563	0.3634	6.0334	0.0334	0.2344	0.2249	0.2368
p	0.5	0.5092	0.0092	0.1217	0.1360	0.1229	0.5079	0.0079	0.1033	0.1064	0.1094
ξ	0.5	0.5174	0.0174	0.1762	0.1800	0.1801	0.5041	0.0041	0.1245	0.1283	0.1251
$\bar{\mu}$	1.0	1.0789	0.0789	0.4466	0.5140	0.4662	1.0244	0.0244	0.3587	0.3616	0.3631
Autoregressive											
β_1	0.5	0.4909	-0.0091	0.1292	0.1335	0.1296	0.5003	0.0003	0.0792	0.0743	0.0792
β_2	0.5	0.5083	0.0083	0.0412	0.0404	0.0420	0.5015	0.0015	0.0253	0.0225	0.0254
β_3	1.0	1.0132	0.0132	0.0531	0.0546	0.0547	1.0025	0.0025	0.0344	0.0334	0.0345
γ_1	2.0	2.0350	0.0350	0.2358	0.2448	0.2383	2.0063	0.0063	0.1538	0.1536	0.1539
γ_2	4.0	4.0615	0.0615	0.2749	0.2808	0.2817	4.0119	0.0119	0.1771	0.1768	0.1775
γ_3	6.0	6.0835	0.0825	0.3389	0.3475	0.3490	6.0168	0.0168	0.2190	0.2187	0.2197
p	0.5	0.5130	0.0130	0.1333	0.1402	0.1365	0.5032	0.0032	0.0983	0.0959	0.0997
ξ	0.5	0.5120	0.0120	0.1642	0.1756	0.1672	0.5049	0.0049	0.1074	0.1062	0.1084
$\bar{\mu}$	1.0	1.0585	0.0585	0.4344	0.4597	0.4642	1.0405	0.0405	0.3349	0.3325	0.3412

Table 6.3 Parameter estimation for geometric skew-normal copula based ordered probit models for $N = 500$ simulated data sets. Exchangeable and autoregressive correlation structures are considered.

Marginal					
Parameters	β_0	β_1	β_2	β_3	κ
Est.	0.5861	-0.0061	0.0063	0.1181	33.0399
SE	0.0652	0.0225	0.0016	0.0091	2.8327
Copula	GSN			Gaussian	
Parameters	p	ξ	$\bar{\mu}$	ξ	
Est.	0.7185	0.3250	-0.1577	0.3284	
SE	0.0647	0.0214	0.1514	0.0172	
Likelihood	-116.47			-129.61	
AIC	**248.95**			271.22	
D_{12} = 0.0657, 95% CI = (0.0140,0.1174), p-val = **0.0127**					

Table 6.4 Fitting of cholesterol data under model (6.35) with GSN and Gaussian copula. Observed log-likelihoods, AICs and the summary of Voung's statistic are reported.

baseline. We consider GSN and Gaussian copula to model the temporal dependency of the cholesterol levels. The sample correlation matrix of the responses suggests that exchangeable correlation structure for the matrix Σ is adequate, which we reparameterize by $\rho = \exp(-\xi), \xi > 0$.

Table 6.4 presents the parameters estimates, standard errors for model (6.35) with the GSN and the Gaussian copula. It also displays the observed log-likelihoods, AICs and the summary of Voung's statistic for comparison of two models. It is evident that GSN copula better explains the temporal dependence of the cholesterol levels, since confidence interval of D_{12} does not contain zero and also the observed AIC is minimum for this model. The value of correlation parameter is very close under both the copulas, which validates the assumed exchangeable correlation structure for $\mathbf{\Sigma}$. The estimates of the regression parameters β_1 and β_2 are close to zero, suggesting that patient's gender and age have insignificant effect on the change in cholesterol levels. Moreover, the underlying copula is negatively skewed, as shown by the value of $\bar{\mu}$. Note that in this case, the estimate of p is not close to one, a scenario in which D_{12} is expected to effectively distinguish between the two models. On the other hand, the estimate of p is not too far away from one. Thus, the facts that the quantities $D_{12} = 0.0657$, despite being away from zero, is not too far away from zero, the 95% CI $= (0.0140, 0.1174)$ is not too far away from zero, and the **p-val** $= \mathbf{0.0127}$ lend moderate support to the GSN copula based model in comparison with the Gaussian copula model. These findings seem to motivate the need of more detailed modeling and/or use of other techniques of model selection.

6.8.2 Schizophrenia collaborative study

This data set is from the National Institute of Mental Health Schizophrenia Collaborative Study, previously analyzed by [171] or [172]. Patients were randomly assigned to receive one of four medications, either placebo or one of thee different anti-psychotic drugs (chlorpromazine, fluphenazine or thioridazine). Here we analyze the outcome variable *imps79o*, which is an ordinally scaled version of the original variable *imps79*. This scaling was done in [172] to retain more information about the response but to ensure each response category has a relatively large number of respondents, since some

Marginal					
Parameters	β_1	β_2	γ_1	γ_2	γ_3
Est.	-0.4082	-0.6584	-2.6283	-1.4228	-0.5880
SE	0.1207	0.0354	0.1243	0.1203	0.1196
Copula	GSN			Gaussian	
Parameters	p	ξ	$\bar{\mu}$	ξ	
Est.	0.8616	0.6843	1.4479	0.5423	
SE	0.0957	0.0959	0.5369	0.0339	
Comp-like	-4156.80			-4161.32	
CLAIC	**8339.26**			8343.17	
D_{12} = 0.0147, 95% CI = (-0.0081,0.0375), p-val = **0.2075**					

Table 6.5 Fitting of schizophrenia data under model (6.36) with GSN and Gaussian copula. Observed composite log-likelihoods, CLAICs and the summary of Voung's statistic are reported.

response categories had relatively small number of subjects compared to others. The ordinal response variable has the following interpretation: 1 = not ill or borderline; 2 = mildly or moderately ill; 3 = markedly ill; and 4 = severely or most extremely ill. Here we perform complete data analysis for 308 patients who were evaluated at weeks $0, 1, 3$ and 6 to assess severity of illness. The covariates are taken as treatment (0 = placebo, 1 = drug) and the square root of the time variable (measured in weeks). Based on the available covariates, we adopt the following model -

$$\begin{aligned} Y_{ij} &= k \ \text{ if } \ \gamma_{k-1} \leq Z_{ij} < \gamma_k, \ \ k = 1, \ldots, 4, \\ Z_{ij} &= \text{treat}_i \beta_1 + t_{ij} \beta_2 + \epsilon_{ij}, \end{aligned} \tag{6.36}$$

where $\epsilon_{ij}(i.i.d) \sim N(0, 1)$, and $t_{ij} = \sqrt{time_{ij}}$. Similarly we consider to copulas to model the dependency across time of the ordinal response variables. Here we consider AR(1) structure of the correlation matrix $\mathbf{\Sigma}$ and equal value of μ across time points.

Results are presented in Table 6.5 for model (6.36) with the GSN and the Gaussian copula. The observed composite log-likelihoods, CLAICs and the summary of Voung's statistic for comparison of two models, are also displayed. Though the CLAIC of the GSN copula based model is less than that of the Gaussian copula based model but the confidence interval of D_{12} contains zero. Hence the two models are not significantly different for this data set. It may be related to the fact that the estimate of p is close to one. Both of the covariates of this model are significant. The patients under the treatment group in general had improvements over the study period from response 4 to response 1, which is shown by the negative value of β_1. The copula underneath this ordinal longitudinal data is positively skewed as shown by the estimate of $\bar{\mu}$.

6.9 Discussion

In recent time modeling dependency across time of repeated measurement data has become an important area of research. In this chapter we developed a new asymmetric multivariate multivariate copula,

using multivariate geometric skew-normal distribution by [152]. Unlike the skew-normal copula from Azzalini's skew-normal distribution, the parametric structure of geometric skew-normal copula is much simpler and it is closed under marginalisation. Multivariate Gaussian copula can be considered as a special case of the GSN copula. We established the dependence properties of the proposed geometric skew-normal copula. We have also shown the explicit forms of the standard dependence measures such as Kendall's tau and Spearman's rho. For moderate to high dimensional data, estimation of the parameters of the unrestricted GSN copula is a challenging issue, due to the numerical instabilities during likelihood optimization. For such situation we proposed to use block-coordinate ascent algorithm to compute the MLEs of the unknown parameters of the MGSN distribution as well as the GSN copula. We observed that the proposed algorithm works efficiently under both the cases.

The second contribution to this chapter is to construct regression models for continuous and discrete longitudinal data. Utilizing the marginalisation property of the GSN copula we constructed composite likelihood in order to estimate parameters from an ordered probit model where the temporal dependency is described by the GSN copula. We examined our approaches with some rigorous simulation studies and examined the fits with two real world data sets. We found that the GSN copula provides a better fit compared to the Gaussian copula. The geometric skew-normal distribution has a great potential in a variety of applications in statistical modeling. We have seen that the GSN copula lacks in non-zero tail dependence, which is similar to Azzalini's skew-normal copula. It addresses only the asymmetry and does not address tail dependence. Hence, it is important to develop the skew-t extension for this distribution as well as for the copula. We plan to conduct some misspecification studies related to the models developed from geometric skew-normal copula in our future works. This will help in numerous applications in finance and risk management. It will be interesting to see a Bayesian procedure for estimating the parameters of the GSN copula. This constitute to our future works regarding dependence modeling of multivariate data. A meaningful question is to explore robustness of the GSN under misspecified marginals or copulas. We note that we are considering the copula of a compound distribution itself unlike the introduction of copula while modeling the dependence between the number of summands and the $i.i.d$ components. See, e.g., [173]. In other words, MGSN copula appears to be structurally new. This leads to challenging questions, for both a theoretician and a practitioner. One would, we believe, encounter difficult computational issues as well. We submit that studying the robustness of the MGSN copula is an interesting question. We plan to undertake this study in future.

Chapter 7

Conclusions and Some Directions of Future Work

In this dissertation, we have developed several new and flexible statistical models to account for potential temporal dependence in longitudinal data in a number of settings. The proposed approaches are designed to be both flexible and computationally efficient, while retaining ease of interpretation. In Chapter 3, we explored an extended class of mixed models where the repeated measurements exhibit non-normal behavior in both the marginal distributions and the dependence structures. We introduced skew-elliptical copula-based mixed models, incorporating copulas derived from skew-elliptical distributions to capture the temporal dependence. To illustrate the practical applicability of our models, we applied them to analyze recent HIV-AIDS data from a study conducted in the Livingstone district of Zambia, focusing on identifying key epidemiological patterns, assessing the effectiveness of intervention strategies, and uncovering potential areas for targeted public health initiatives.

While the models in chapter 3 were estimated using a two-step Inference Functions of Margins (IFM) procedure, we plan to further investigate the use of Bayesian estimation techniques in future research to improve model estimation and inference. This future work will provide additional insight into the potential advantages of a Bayesian framework for estimating these complex models in longitudinal data analysis paradigm.

In Chapter 4, we introduced factor copula models to capture temporal dependence in unbalanced longitudinal data. The primary innovation of this chapter is the use of latent variables to model the joint distribution of longitudinal data. We developed regression models for continuous, binary, and ordinal longitudinal data, which were estimated using the IFM procedure. Additionally, we compared the performance of these models to traditional random effects models, demonstrating significant improvements in model fit. We also discussed a method for evaluating the proposed models through Rosenblatt-type residual analysis.

However, an important limitation of factor copula models, as currently constructed and also assumed in chapter 4, is the assumption of homogeneous dependence mechanism across subjects, which may not always hold in real-life applications. As a result, there is potential for further refinement by incorporating time-dependent covariates into the dependence structure of factor copula models. Currently, we are extending these models to analyze both longitudinal and time-to-event data, with a particular focus on improving dynamic predictions of survival probabilities. Our comparative analysis with random

effects-based joint models reveals that factor copula models offer substantially better predictive performance for survival probabilities, highlighting their potential for more accurate and flexible modeling in longitudinal and time-to-event data analysis.

We have used the method of inference function for margins (IFM) in both chapters 3 and 4. A few words regarding its robustness are in order. We see in [24], (sections 10.1 and 10.2), the IFM method presupposes a particular structure of the model under consideration. Letting m denote the dimension of the observations, $F_1, \ldots, F_m$ denote the marginal cdf's, and $\alpha_1, \ldots, \alpha_m$ denote the parameters corresponding to the marginals ([24], pp. 299-301). The parameter α_i is assumed to be associated with F_i only ($i = 1, \ldots, m$). [24] has also discussed (in section 10.2) how IFM can be used in situations where (1) covariates are present and (2) parameters are common to more than one margin. In the following brief discussion on possible robustness studies relevant to our context, we restrict our attention only to the following scenario.

Let $\mathbf{Y} = (Y_1, \ldots, Y_m)^\intercal$ denote an observation with pdf

$$f(\mathbf{y}|\alpha_1, \ldots, \alpha_m, \theta) = c(F_1(y_1, \alpha_1), \ldots, F_m(y_1, \alpha_m); \theta) \prod_{j=1}^{m} f_j(y_j, \alpha_j)$$

where c is the underlying copula density and $f_j(y_j, \alpha_j)$ is the pdf corresponding to $F_j(y_j, \alpha_j)$. We assume that some or all of the marginals f_j (or, equivalently F_j) are misspecified and for each of the misspecified marginals, we work with a new and incorrect marginal. For robustness studies, this is a fairly complicated scenario even when the copula is held fixed as the distribution of $\mathbf{Y}$ is no longer same as the actual one. Also, a possible route of inference is the following. We may use some robust method of estimation (like weighted likelihood or divergence based method) for estimating the α's. Then we proceed, as in IFM, for estimating θ. However, outcome of such a study is not easy to see beforehand unless this has been studied thoroughly. We are not aware of any such study. A study close in spirit is [80]. We submit that this issue is both important and interesting. However, this is a fairly long journey. We plan to pursue this in future.

In Chapter 5, we employed finite mixtures of elliptical copulas to model the hidden temporal dependence in count longitudinal data. While mixture copulas have been previously applied in modeling discrete longitudinal data, their dependence properties within the discrete setup were not fully explored in the existing literature. Therefore, our primary contribution in this chapter was to derive the dependence properties of mixture copulas in the context of discrete longitudinal data. Building on these properties, we developed regression models tailored for count longitudinal data applications, demonstrating significant improvements in model fit compared to traditional multivariate elliptical copulas. To further validate the performance of the proposed models, we utilized a modified t-plot method, which provided additional support for the robustness and accuracy of the fitting procedure. Overall, the results highlight the potential of finite mixture copulas in capturing complex temporal dependencies in count data, offering a more flexible alternative to conventional approaches.

As a follow-up of the development presented in chapter 5, we may eplre feasibility of using mixtures of families other than those considered in that chapter. It is also worthwhile to develop methods when the number of components in the mixture is not known a priori as in [123].

In the final contributing chapter of this dissertation, we introduced a novel multivariate copula derived from the multivariate geometric skew-normal distribution. Unlike the skew-normal copula proposed by Azzalini, this new multivariate copula is closed under marginalization, which helps alleviate the dimensionality problem often encountered when modeling dependence in discrete longitudinal data. This feature makes it more flexible and efficient for practical applications. Similar to the multivariate Gaussian copula, the geometric skew-normal copula is both mathematically and computationally tractable, offering an advantage in terms of ease of implementation. One of the key strengths of this copula is its ability to take various shapes based on its parameters, making it particularly beneficial in modeling complex dependencies, including those encountered in financial applications.

As a follow-up of the development presented in chapter 6, it is worthwhile to study the properties, other than those studied in that chapter, of the proposed copula in detail; given its contribution to the existing literature. It is also worthwhile to develop a Bayesian route to the questions of inference related to geometric skew-normal copula and its associated models, which will further enhance the model's flexibility and robustness. We submit that the MGSN copula studied in chapter 7 appears to be structurally new in the context of compound distributions. To make our point clear, we consider, for instance, the set-up considered in [173]. While describing the random sum (a compound distribution) $S_N := \sum_{i=1}^{N} X_i$, [173] have brought in copula for modeling the dependence between N and the $i.i.d$ X_i's. Whereas, in our situation, we have studied the copula for S_N itself, assuming the entire collection $\{N, X_i : i \leq 1\}$ to be independent, besides assuming independence of the X_i's. We are not aware of any work where the copula of a compound distribution as we have done has been pursued. We believe this opens a new area of research, from the standpoint of both theoreticians and practitioners. One tentative route may be to approximate the distribution of S_N and study the copula of the approximation.

Longitudinal data have attracted huge attention by scientists of different disciplines and practitioners. The nature of the data generated by longitudinal studies have posed numerous challenging questions for statisticians. It appears from the available literature on longitudinal data that considerable care is needed to model a set or repeated measurements. This means keeping in mind the genesis of data. We submit that this should be applied to choice of copulas as well, while working on a practical problem. This may require creative use of existing copulas or existing methods of construction of copulas, or may lead to construction of new copulas. Such work will, in particular, require a careful and critical look at temporal dependence, taking into account their nature of development. The associated computational burden of such research work also needs to be kept in mind in such an exploration. We also submit that there is a pressing need to develop theoretically sound methods for addressing missing data issues within the copula framework, particularly in the context of longitudinal data analysis. Such methods will be crucial for improving the reliability and applicability of copula-based models in real-life settings. One key challenge in this area is addressing temporal dependence, which plays a crucial role in understanding

how past values influence future outcomes. The temporal dependence and the prediction of an upcoming event or outcome are closely linked, as the conditional expectation of a future response depends on its previous values and the associated conditional copula. This means that if we fail to properly account for temporal dependence—such as by misspecifying the underlying copula; the accuracy of our predictions can be severely compromised. While this phenomenon is not directly addressed in this dissertation, it has been discussed in our other works. We notice that in each chapter, wherever we could, we have undertaken comparison of the copula based methods developed and studied. However, we have not undertaken any study of comparison with other existing non-copula methods. This issue of comparison with non-copula models, raised by the reviewer, is too important to not be pursued further. In fact, one major lesson from the dissertation, we submit with due humility, is that besides the usefulness of copulas-based modeling they also call for further scrutiny in terms of comparison with other existing non-copula methods. We fully agree with this issue raised by the reviewer. We plan to undertake this in future. Other meaningful areas of future research are combination of Bayesian methods and copula-based methods, development of tools of prediction using copulas etc. Another relevant and worthwhile issue to pursue would be to develop goodness-of-fit tests in the context of longitudinal data, especially in the context of the models developed in this dissertation. A considerable amount of work on goodness-of-fit tests in copula based models have appeared in the literature. Here is a snapshot of some of these work, listed chronologically; [174], [78], [175], [176], [177], [178], [179], [180], [181] and [182]. This dissertation has attempted at contributing to the evolving literature on copula-based modeling and analysis of longitudinal data. Our humble belief is that the methods developed in the preceding four chapters will fill certain gaps in the existing literature on copula-based longitudinal data analysis, or more generally, to the literature on longitudinal data analysis. We believe if researchers take into account the growth of the literature on theory and practice of copula and try to use them in modeling and analysis of longitudinal data, further growth of the literature will take place. We hope that copula-based methods will find their place in toolkits of statistical analysis in near future.

Bibliography

[1] C. S. Davis *et al.*, *Statistical Methods for the Analysis of Repeated Measurements*. Springer, New York, 2002.

[2] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*. Springer, New York, 1997.

[3] R. E. Weiss, *Modeling Longitudinal Data*. Springer, USA, 2005.

[4] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, *Longitudinal Data Analysis (Handbook of Mordern Statistical Methods)*. Chapman & Hall / CRC Press, 2008.

[5] ——, "Advances in longitudinal data analysis: An historical perspective," in *Longitudinal Data Analysis*. Chapman & Hall / CRC Press, 2008, pp. 17–42.

[6] G. Verbeke, S. Fieuws, G. Molenberghs, and M. Davidian, "The analysis of multivariate longitudinal data: a review," *Statistical methods in medical research*, vol. 23, no. 1, pp. 42–59, 2014.

[7] P. Lambert and F. Vandenhende, "A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant," *Statistics in Medicine*, vol. 21, no. 21, pp. 3197–3217, 2002.

[8] P. X.-K. Song, M. Li, and Y. Yuan, "Joint regression analysis of correlated data using gaussian copulas," *Biometrics*, vol. 65, no. 1, pp. 60–68, 2009.

[9] A. K. Nikoloulopoulos and D. Karlis, "Modeling multivariate count data using copulas," *Communications in Statistics-Simulation and Computation*, vol. 39, no. 1, pp. 172–187, 2010.

[10] ——, "Regression in a copula model for bivariate count data," *Journal of Applied Statistics*, vol. 37, no. 9, pp. 1555–1568, 2010.

[11] A. R. de Leon and B. Wu, "Copula-based regression models for a bivariate mixed discrete and continuous outcome," *Statistics in Medicine*, vol. 30, no. 2, pp. 175–185, 2011.

[12] G. Masarotto and C. Varin, "Gaussian copula marginal regression," *Electronic Journal of Statistics*, vol. 6, pp. 1517–1549, 2012.

[13] P. Shi and E. A. Valdez, "Longitudinal modeling of insurance claim counts using jitters," *Scandinavian Actuarial Journal*, vol. 2014, no. 2, pp. 159–179, 2014.

[14] K. Das, M. Elmasri, and A. Sen, "A skew-normal copula-driven GLMM," *Statistica Neerlandica*, vol. 70, no. 4, pp. 396–413, 2016.

[15] G. Masarotto and C. Varin, "Gaussian copula regression in R," *Journal of Statistical Software*, vol. 77, no. 8, pp. 1–26, 2017.

[16] M. Killiches and C. Czado, "A D-vine copula-based model for repeated measurements extending linear mixed models with homogeneous correlation structure," *Biometrics*, vol. 74, no. 3, pp. 997–1005, 2018.

[17] E. Kürüm, J. Hughes, R. Li, and S. Shiffman, "Time-varying copula models for longitudinal data," *Statistics and its Interface*, vol. 11, no. 2, pp. 203–221, 2018.

[18] T. Baghfalaki and M. Ganjali, "A transition model for analyzing multivariate longitudinal data using gaussian copula approach," *AStA Advances in Statistical Analysis*, vol. 104, no. 2, pp. 169–223, 2020.

[19] K. Suresh, J. M. Taylor, and A. Tsodikov, "A gaussian copula approach for dynamic prediction of survival with a longitudinal biomarker," *Biostatistics*, vol. 22, no. 3, pp. 504–521, 2021.

[20] S. Sefidi, M. Ganjali, and T. Baghfalaki, "Pair copula construction for longitudinal data with zero-inflated power series marginal distributions," *Journal of Biopharmaceutical Statistics*, vol. 31, no. 2, pp. 233–249, 2021.

[21] ——, "Analysis of ordinal and continuous longitudinal responses using pair copula construction," *Metron*, vol. 80, no. 2, pp. 255–280, 2022.

[22] C. Czado, *Analyzing dependent data with vine copulas*. Springer, Chambridge, 2019.

[23] K. Aas, C. Czado, A. Frigessi, and H. Bakken, "Pair-copula constructions of multiple dependence," *Insurance: Mathematics and economics*, vol. 44, no. 2, pp. 182–198, 2009.

[24] H. Joe, *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall / CRC Press, 1997.

[25] C. M. Cuadras, J. Fortiana, and J. A. Rodriguez-Lallena, *Distributions with given marginals and statistical modelling*. Springer, Dordrecht, 2002.

[26] R. B. Nelsen, *An Introduction to Copulas*, 2nd ed. Springer, New York, 2006.

[27] P. Jaworski, F. Durante, W. K. Hardle, and T. Rychlik, *Copula Theory and its Applications*. Springer, Hindenberg, 2010, vol. 198.

[28] H. Joe, *Dependence Modeling with Copulas*. Chapman & Hall / CRC Press, 2014.

[29] C. Czado and T. Nagler, "Vine copula based modeling," *Annual Review of Statistics and Its Application*, vol. 9, no. 1, pp. 453–477, 2022.

[30] J. Größer and O. Okhrin, "Copulae: An overview and recent developments," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 14, no. 3, p. e1557, 2022.

[31] C. Genest, O. Okhrin, and T. Bodnar, "Copula modeling from abe sklar to the present day," *Journal of Multivariate Analysis*, vol. 201, p. 105278, 2024.

[32] M. E. Hoque, *Accounting for heterogeneity in the dependence mechanism of longitudinal data*. University of Manitoba at Winnipeg, Canada, 2023.

[33] H. S. Katesari, *Bayesian Dynamic Factor Analysis and Copula-Based Models for Mixed Data*. Southern Illinois University at Carbondale, 2021.

[34] G. Popovic, "Covariance modelling and inference for multivariate discrete data in ecology," Ph.D. dissertation, UNSW Sydney, 2017.

[35] H. Joe and J. J. Xu, "The estimation method of inference functions for margins for multivariate models," *Technical Report, University of British Columbia*, vol. 166, p. 22, 1996.

[36] H. Joe, "Asymptotic efficiency of the two-stage estimation method for copula-based models," *Journal of Multivariate Analysis*, vol. 94, no. 2, pp. 401–419, 2005.

[37] J. J. Xu, "Statistical modelling and inference for multivariate and longitudinal discrete response data," Ph.D. dissertation, University of British Columbia, 1996.

[38] R. McCulloch and P. E. Rossi, "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, vol. 64, no. 1-2, pp. 207–240, 1994.

[39] N. E. Breslow and X. Lin, "Bias correction in generalised linear mixed models with a single component of dispersion," *Biometrika*, vol. 82, no. 1, pp. 81–91, 1995.

[40] N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American statistical Association*, vol. 88, no. 421, pp. 9–25, 1993.

[41] D. M. Zimmer and P. K. Trivedi, "Using trivariate copulas to model sample selection and treatment effects: application to family health care demand," *Journal of Business & Economic Statistics*, vol. 24, no. 1, pp. 63–76, 2006.

[42] M. Sklar, "Fonctions de repartition an dimensions et leurs marges," *Publ. inst. statist. univ. Paris*, vol. 8, pp. 229–231, 1959.

[43] H.-B. Fang, K.-T. Fang, and S. Kotz, "The meta-elliptical distributions with given marginals," *Journal of M'1ultivariate analysis*, vol. 82, no. 1, pp. 1–16, 2002.

[44] R. B. Nelsen, "Concordance and copulas: A survey," *Distributions with Given Marginals and Statistical Modelling*, pp. 169–177, 2002.

[45] H. Joe, "Multivariate concordance," *Journal of Multivariate Analysis*, vol. 35, no. 1, pp. 12–30, 1990.

[46] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, pp. 963–974, 1982.

[47] C. E. McCulloch, "Generalized linear mixed models," *NSF-CBMS Regional Conference Series in Probability and Statistics*, vol. 7, p. 84, 2003.

[48] A. Azzalini and A. Capitanio, *The Skew-normal and Related Families (IMS Monographs)*. Cambridge University Press, 2013, vol. 3.

[49] B. Chang and H. Joe, "Copula diagnostics for asymmetries and conditional dependence," *Journal of Applied Statistics*, vol. 47, no. 9, pp. 1587–1615, 2020.

[50] T.-I. Lin and W.-L. Wang, "Multivariate skew-normal at linear mixed models for multi-outcome longitudinal data," *Statistical Modelling*, vol. 13, no. 3, pp. 199–221, 2013.

[51] D. Bandyopadhyay, V. H. Lachos, L. M. Castro, and D. K. Dey, "Skew-normal/independent linear mixed models for censored responses with applications to hiv viral loads," *Biometrical Journal*, vol. 54, no. 3, pp. 405–425, 2012.

[52] V. H. Lachos, P. Ghosh, and R. B. Arellano-Valle, "Likelihood based inference for skew-normal independent linear mixed models," *Statistica Sinica*, vol. 20, pp. 303–322, 2010.

[53] M. D. Branco and D. K. Dey, "A general class of multivariate skew-elliptical distributions," *Journal of Multivariate Analysis*, vol. 79, no. 1, pp. 99–113, 2001.

[54] J. Wang and M. G. Genton, "The multivariate skew-slash distribution," *Journal of Statistical Planning and Inference*, vol. 136, no. 1, pp. 209–220, 2006.

[55] M. G. Genton, *Skew-elliptical distributions and their applications: A journey beyond normality*. Chapman & Hall / CRC, Boca Raton, 2004.

[56] H. Joe and H. Li, "Tail densities of skew-elliptical distributions," *Journal of Multivariate Analysis*, vol. 171, pp. 421–435, 2019.

[57] T. Kollo, G. Pettere, and M. Valge, "Tail dependence of skew t-copulas," *Communications in Statistics-Simulation and Computation*, vol. 46, no. 2, pp. 1024–1034, 2017.

[58] Z. Wei, S. Kim, and D. Kim, "Multivariate skew normal copula for non-exchangeable dependence," *Procedia Computer Science*, vol. 91, pp. 141–150, 2016.

[59] A. Azzalini and A. Capitanio, "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 2, pp. 367–389, 2003.

[60] A. Gupta, "Multivariate skew t-distribution," *Statistics: A Journal of Theoretical and Applied Statistics*, vol. 37, no. 4, pp. 359–363, 2003.

[61] S. K. Sahu, D. K. Dey, and M. D. Branco, "A new class of multivariate skew distributions with applications to bayesian regression models," *Canadian Journal of Statistics*, vol. 31, no. 2, pp. 129–150, 2003.

[62] T. Yoshiba, "Maximum likelihood estimation of skew-t copulas with its applications to stock returns," *Journal of Statistical Computation and Simulation*, vol. 88, no. 13, pp. 2489–2506, 2018.

[63] M. S. Smith, Q. Gan, and R. J. Kohn, "Modelling dependence using skew-t copulas: Bayesian inference and applications," *Journal of Applied Econometrics*, vol. 27, no. 3, pp. 500–522, 2012.

[64] P. Cortez and Cortez, *Modern Optimization with R*. Springer, 2014.

[65] P. Gilbert, R. Varadhan, and M. P. Gilbert, "Package 'numderiv'," *Differential Equations*, vol. 3, pp. 203–267, 2009.

[66] J. D. Singer and J. B. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford university press, 2003.

[67] G. Kim, M. J. Silvapulle, and P. Silvapulle, "Comparison of semiparametric and parametric methods for estimating copulas," *Computational Statistics & Data Analysis*, vol. 51, no. 6, pp. 2836–2850, 2007.

[68] L. A. Jordanger and D. Tjøstheim, "Model selection of copulas: AIC versus a cross validation copula information criterion," *Statistics & Probability Letters*, vol. 92, pp. 249–255, 2014.

[69] V. Ko and N. L. Hjort, "Copula information criterion for model selection with two-stage maximum likelihood estimation," *Econometrics and Statistics*, vol. 12, pp. 167–180, 2019.

[70] E. Liebscher, "Construction of asymmetric multivariate copulas," *Journal of Multivariate analysis*, vol. 99, no. 10, pp. 2234–2250, 2008.

[71] Z. Guo, "Asymmetric multivariate archimedean copula models and semi-competing risks data analysis," Ph.D. dissertation, New Jersey Institute of Technology, 2021.

[72] T. Nagler, "Simplified vine copula models: state of science and affairs," *Risk Sciences*, p. 100022, 2025.

[73] J. H. Shih and T. A. Louis, "Inferences on the association parameter in copula models for bivariate survival data," *Biometrics*, pp. 1384–1399, 1995.

[74] H. Tsukahara, "Semiparametric estimation in copula models," *Canadian Journal of Statistics*, vol. 33, no. 3, pp. 357–375, 2005.

[75] D. Oakes, "Multivariate survival distributions," *Journaltitle of Nonparametric Statistics*, vol. 3, no. 3-4, pp. 343–354, 1994.

[76] C. Genest, K. Ghoudi, and L.-P. Rivest, "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions," *Biometrika*, vol. 82, no. 3, pp. 543–552, 1995.

[77] B. J. Werker, "Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models," *Distributions with given marginals and statistical modelling*, p. 103, 2002.

[78] C. Genest, B. Rémillard, and D. Beaudoin, "Goodness-of-fit tests for copulas: A review and a power study," *Insurance: Mathematics and economics*, vol. 44, no. 2, pp. 199–213, 2009.

[79] C. Genest, A. Carabarín-Aguirre, and F. Harvey, "Copula parameter estimation using blomqvist's beta," *Journal de la Société Française de Statistique*, vol. 154, no. 1, pp. 5–24, 2013.

[80] P. Alquier, B.-E. Chérief-Abdellatif, A. Derumigny, and J.-D. Fermanian, "Estimation of copulas via maximum mean discrepancy," *Journal of the American Statistical Association*, vol. 118, no. 543, pp. 1997–2012, 2023.

[81] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 192–225, 1974.

[82] B. G. Lindsay[1], "Composite likelihood methods," in *Statistical Inference from Stochastic Processes: Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference Held August 9-15, 1987, with Support from the National Science Foundation and the Army Research Office*, vol. 80. American Mathematical Soc., 1988, p. 221.

[83] D. R. Cox and N. Reid, "A note on pseudolikelihood constructed from marginal densities," *Biometrika*, vol. 91, no. 3, pp. 729–737, 2004.

[84] Y. Zhao and H. Joe, "Composite likelihood estimation in multivariate data analysis," *Canadian Journal of Statistics*, vol. 33, no. 3, pp. 335–356, 2005.

[85] C. Varin, "On composite marginal likelihoods," *Asta advances in statistical analysis*, vol. 92, no. 1, pp. 1–28, 2008.

[86] H. Joe and Y. Lee, "On weighting of bivariate margins in pairwise likelihood," *Journal of Multivariate Analysis*, vol. 100, no. 4, pp. 670–685, 2009.

[87] C. Varin, N. Reid, and D. Firth, "An overview of composite likelihood methods," *Statistica Sinica*, vol. 21, no. 1, pp. 5–42, 2011.

[88] J. Górecki and M. Hofert, "Composite pseudo-likelihood estimation for pair-tractable copulas such as archimedean, archimax and related hierarchical extensions," *Journal of Statistical Computation and Simulation*, vol. 93, no. 13, pp. 2321–2355, 2023.

[89] A. K. Nikoloulopoulos, "Efficient and feasible inference for high-dimensional normal copula regression models," *Computational Statistics & Data Analysis*, vol. 179, p. 107654, 2023.

[90] J. Sun, E. W. Frees, and M. A. Rosenberg, "Heavy-tailed longitudinal data modeling using copulas," *Insurance: Mathematics and Economics*, vol. 42, no. 2, pp. 817–830, 2008.

[91] P. Shi, X. Feng, and J.-P. Boucher, "Multilevel modeling of insurance claims using copulas," *The Annals of Applied Statistics*, vol. 10, no. 2, pp. 834 – 863, 2016.

[92] A. Panagiotelis, C. Czado, and H. Joe, "Pair copula constructions for multivariate discrete data," *Journal of the American Statistical Association*, vol. 107, no. 499, pp. 1063–1072, 2012.

[93] A. A. Zilko and D. Kurowicka, "Copula in a multivariate mixed discrete–continuous model," *Computational Statistics & Data Analysis*, vol. 103, pp. 28–55, 2016.

[94] P. Krupskii and H. Joe, "Factor copula models for multivariate data," *Journal of Multivariate Analysis*, vol. 120, pp. 85–101, 2013.

[95] A. K. Nikoloulopoulos and H. Joe, "Factor copula models for item response data," *Psychometrika*, vol. 80, no. 1, pp. 126–150, 2015.

[96] S. H. Kadhem and A. K. Nikoloulopoulos, "Factor copula models for mixed data," *British Journal of Mathematical and Statistical Psychology*, vol. 74, no. 3, pp. 365–403, 2021.

[97] A. Agresti, *Analysis of Ordinal Categorical Data*, 2nd ed. John Wiley & Sons, New York, 2010, vol. 656.

[98] M. Rosenblatt, "Remarks on a multivariate transformation," *The annals of mathematical statistics*, vol. 23, no. 3, pp. 470–472, 1952.

[99] M. Hofert and M. Mächler, "A graphical goodness-of-fit test for dependence models in higher dimensions," *Journal of Computational and Graphical Statistics*, vol. 23, no. 3, pp. 700–716, 2014.

[100] W. Zucchini and I. L. MacDonald, *Hidden Markov Models for Time Series: An Introduction using R*. Chapman & Hall / CRC Press, 2009.

[101] E. R. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy, "Prognosis in primary biliary cirrhosis: model for decision making," *Hepatology*, vol. 10, no. 1, pp. 1–7, 1989.

[102] A. Komárek and L. Komárková, "Clustering for multivariate continuous and discrete longitudinal data," *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 177 – 200, 2013.

[103] E.-R. Andrinopoulou and D. Rizopoulos, "Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures," *Statistics in Medicine*, vol. 35, no. 26, pp. 4813–4823, 2016.

[104] G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "Joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes," *BMC Medical Research Methodology*, vol. 18, no. 50, pp. 1–14, 2018.

[105] L. Letenneur, D. Commenges, J.-F. Dartigues, and P. Barberger-Gateau, "Incidence of dementia and Alzheimer's disease in elderly community residents of south-western France," *International Journal of Epidemiology*, vol. 23, no. 6, pp. 1256–1261, 1994.

[106] C. Proust, H. Jacqmin-Gadda, J. M. Taylor, J. Ganiayre, and D. Commenges, "A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data," *Biometrics*, vol. 62, no. 4, pp. 1014–1024, 2006.

[107] C. Proust-Lima, L. Letenneur, and H. Jacqmin-Gadda, "A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome," *Statistics in Medicine*, vol. 26, no. 10, pp. 2229–2245, 2007.

[108] C. Proust-Lima, V. Philipps, and J.-F. Dartigues, "A joint model for multiple dynamic processes and clinical endpoints: Application to Alzheimer's disease," *Statistics in Medicine*, vol. 38, no. 23, pp. 4702–4717, 2019.

[109] C. Proust-Lima, H. Amieva, and H. Jacqmin-Gadda, "Analysis of multivariate mixed longitudinal data: A flexible latent process approach," *British Journal of Mathematical and Statistical Psychology*, vol. 66, no. 3, pp. 470–487, 2013.

[110] S. Litière, A. Alonso, and G. Molenberghs, "Type I and type II error under random-effects misspecification in generalized linear mixed models," *Biometrics*, vol. 63, no. 4, pp. 1038–1044, 2007.

[111] P. Krupskii, B. R. Nasri, and B. N. Rémillard, "On factor copula-based mixed regression models," *Electronic Journal of Statistics*, vol. 19, no. 1, pp. 1133–1173, 2025.

[112] H. Jin, "Copula theory on the histogram-valued data," Ph.D. dissertation, University of Georgia, 2021.

[113] J. F. Ziegel and T. Gneiting, "Copula calibration," *Electronic Journal of Statistics*, vol. 8, no. 2, pp. 2619–2638, 2014.

[114] F. Durante, J. Fernandez-Sanchez, and R. Pappada, "Copulas, diagonals, and tail dependence," *Fuzzy Sets and Systems*, vol. 264, pp. 22–41, 2015.

[115] D. Hedeker and R. D. Gibbons, "A random-effects ordinal regression model for multilevel analysis," *Biometrics*, pp. 933–944, 1994.

[116] B. C. Sutradhar, *Dynamic Mixed Models for Familial Longitudinal Data*. Springer, New York, 2011.

[117] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*. Cambridge university press, 2013, no. 53.

[118] J. M. Hilbe, *Negative binomial regression*. Cambridge University Press, 2011.

[119] P. F. Thall and S. C. Vail, "Some covariance models for longitudinal count data with overdispersion," *Biometrics*, vol. 46, no. 6, pp. 657–671, 1990.

[120] V. Jowaheer and B. C. Sutradhar, "Analysing longitudinal count data with overdispersion," *Biometrika*, vol. 89, no. 2, pp. 389–399, 2002.

[121] B. C. Sutradhar, "An overview on regression models for discrete longitudinal responses," *Statistical Science*, vol. 18, no. 3, pp. 377–393, 2003.

[122] S. Xu, R. H. Jones, and G. K. Grunwald, "Analysis of longitudinal count data with serial correlation," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 49, no. 3, pp. 416–428, 2007.

[123] B. Yang, Z. Cai, C. M. Hafner, and G. Liu, "Time-varying mixture copula models with copula selection," *Statistica Sinica*, vol. 32, no. 2, pp. 1049–1077, 2022.

[124] Y. Liu, D. Xie, D. A. Edwards, and S. Yu, "Mixture copulas with discrete margins and their application to imbalanced data," *Journal of the Korean Statistical Society*, vol. 52, no. 4, pp. 878–900, 2023.

[125] H. Safari-Katesari, S. Y. Samadi, and S. Zaroudi, "Modelling count data via copulas," *Statistics*, vol. 54, no. 6, pp. 1329–1355, 2020.

[126] D. M. Titterington, A. F. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester, 1985.

[127] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching models*. Springer, New York, 2006.

[128] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annual Review of Statistics and its Application*, vol. 6, pp. 355–378, 2019.

[129] V. Arakelian and D. Karlis, "Clustering dependencies via mixtures of copulas," *Communications in Statistics-Simulation and Computation*, vol. 43, no. 7, pp. 1644–1661, 2014.

[130] I. Kosmidis and D. Karlis, "Model-based clustering using copulas with applications," *Statistics and computing*, vol. 26, pp. 1079–1099, 2016.

[131] H. Zhuang, L. Diao, and Y. Y. Grace, "A Bayesian nonparametric mixture model for grouping dependence structures and selecting copula functions," *Econometrics and Statistics*, vol. 26, no. 5, pp. 172–189, 2022.

[132] P. Xue-Kun Song, "Multivariate dispersion models generated from gaussian copula," *Scandinavian Journal of Statistics*, vol. 27, no. 2, pp. 305–320, 2000.

[133] S. Demarta and A. J. McNeil, "The t-copula and related copulas," *International Statistical Review*, vol. 73, no. 1, pp. 111–129, 2005.

[134] G. Frahm, M. Junker, and A. Szimayer, "Elliptical copulas: applicability and limitations," *Statistics & Probability Letters*, vol. 63, no. 3, pp. 275–286, 2003.

[135] G. J. McLachlan, D. Peel, K. E. Basford, and P. Adams, "The emmix software for the fitting of mixtures of normal and t-components," *Journal of Statistical Software*, vol. 4, no. 2, 2000.

[136] E. W. Frees and P. Wang, "Copula credibility for aggregate loss models," *Insurance: Mathematics and Economics*, vol. 38, no. 2, pp. 360–373, 2006.

[137] F. Lindskog, A. McNeil, and U. Schmock, "Kendall's tau for elliptical distributions," in *Credit risk: Measurement, evaluation and management*. Springer, 2003, pp. 149–156.

[138] M. Denuit and P. Lambert, "Constraints on concordance measures in bivariate discrete data," *Journal of Multivariate Analysis*, vol. 93, no. 1, pp. 40–57, 2005.

[139] M. Mesfioui and A. Tajar, "On the properties of some nonparametric concordance measures in the discrete case," *Nonparametric Statistics*, vol. 17, no. 5, pp. 541–554, 2005.

[140] L. Madsen and Y. Fang, "Joint regression analysis for discrete longitudinal data," *Biometrics*, vol. 67, no. 3, pp. 1171–1175, 2011.

[141] S. le Cessie and J. Van Houwelingen, "Logistic regression for correlated binary data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 43, no. 1, pp. 95–108, 1994.

[142] C. Varin and C. Czado, "A mixed autoregressive probit model for ordinal longitudinal data," *Biostatistics*, vol. 11, no. 1, pp. 127–138, 2010.

[143] L. Diao and R. J. Cook, "Composite likelihood for joint analysis of multiple multistate processes via copulas," *Biostatistics*, vol. 15, no. 4, pp. 690–705, 2014.

[144] I. L. MacDonald, "Numerical maximisation of likelihood: A neglected alternative to EM ?" *International Statistical Review*, vol. 82, no. 2, pp. 296–308, 2014.

[145] C. Varin and P. Vidoni, "A note on composite likelihood inference and model selection," *Biometrika*, vol. 92, no. 3, pp. 519–528, 2005.

[146] X. Gao and P. X.-K. Song, "Composite likelihood bayesian information criteria for model selection in high-dimensional data," *Journal of the American Statistical Association*, vol. 105, no. 492, pp. 1531–1540, 2010.

[147] P. Shi, "Multivariate longitudinal modeling of insurance company expenses," *Insurance: Mathematics and Economics*, vol. 51, no. 1, pp. 204–215, 2012.

[148] R.-Z. Li, K.-T. Fang, and L.-X. Zhu, "Some QQ probability plots to test spherical and elliptical symmetry," *Journal of Computational and Graphical Statistics*, vol. 6, no. 4, pp. 435–450, 1997.

[149] R. Horta, "The city of boca raton: A case study in water utility cybersecurity," *Journal-American Water Works Association*, vol. 99, no. 3, pp. 48–50, 2007.

[150] F. Ahmmed and A. R. Jamee, "Generalized quasi-likelihood estimation procedure for non-stationary over-dispersed longitudinal counts," *Journal of Statistical Computation and Simulation*, vol. 91, no. 9, pp. 1802–1814, 2021.

[151] G. Liu, W. Long, B. Yang, and Z. Cai, "Semiparametric estimation and model selection for conditional mixture copula models," *Scandinavian Journal of Statistics*, vol. 49, no. 1, pp. 287–330, 2022.

[152] D. Kundu, "Multivariate geometric skew-normal distribution," *Statistics*, vol. 51, no. 6, pp. 1377–1397, 2017.

[153] ——, "Geometric skew normal distribution," *Sankhya B*, vol. 76, no. 2, pp. 167–189, 2014.

[154] A. Azzalini and A. D. Valle, "The multivariate skew-normal distribution," *Biometrika*, vol. 83, no. 4, pp. 715–726, 1996.

[155] A. Ang and J. Chen, "Asymmetric correlations of equity portfolios," *Journal of Financial Economics*, vol. 63, no. 3, pp. 443–494, 2002.

[156] A. J. Patton, "Modelling asymmetric exchange rate dependence," *International Economic Review*, vol. 47, no. 2, pp. 527–556, 2006.

[157] R. Roozegar and S. Nadarajah, "The power series skew normal class of distributions," *Communications in Statistics-Theory and Methods*, vol. 46, no. 22, pp. 11 404–11 423, 2017.

[158] E. Redivo, H. D. Nguyen, and M. Gupta, "Bayesian clustering of skewed and multimodal data using geometric skewed normal distributions," *Computational Statistics & Data Analysis*, vol. 152, pp. 107 040–107 056, 2020.

[159] T. Kollo, A. Selart, and H. Visk, "From multivariate skewed distributions to copulas," in *Combinatorial matrix theory and generalized inverses of matrices*. Springer, 2013, pp. 63–72.

[160] A. K. Nikoloulopoulos and D. Karlis, "Finite normal mixture copulas for multivariate discrete data modeling," *Journal of Statistical Planning and Inference*, vol. 139, no. 11, pp. 3878–3890, 2009.

[161] L. Grippof and M. Sciandrone, "Globally convergent block-coordinate techniques for unconstrained optimization," *Optimization methods and software*, vol. 10, no. 4, pp. 587–637, 1999.

[162] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of optimization theory and applications*, vol. 109, no. 3, pp. 475–494, 2001.

[163] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *The annals of applied statistics*, vol. 5, no. 1, p. 232, 2011.

[164] C. R. Bhat, C. Varin, and N. Ferdous, "A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered-response model," in *Maximum Simulated Likelihood Methods and Applications*. Emerald Group Publishing Limited, 2010, vol. 26, pp. 65–106.

[165] X.-K. S. Peter and K. Song, *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer, 2007.

[166] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica: Journal of the Econometric Society*, vol. 57, no. 6, pp. 307–333, 1989.

[167] E. C. Brechmann, C. Czado, and K. Aas, "Truncated regular vines in high dimensions with application to financial data," *Canadian Journal of Statistics*, vol. 40, no. 1, pp. 68–85, 2012.

[168] A. K. Nikoloulopoulos, "A vine copula mixed effect model for trivariate meta-analysis of diagnostic test accuracy studies accounting for disease prevalence," *Statistical Methods in Medical Research*, vol. 26, no. 5, pp. 2270–2286, 2017.

[169] D. Zhang and M. Davidian, "Linear mixed models with flexible distributions of random effects for longitudinal data," *Biometrics*, vol. 57, no. 3, pp. 795–802, 2001.

[170] R. Arellano-Valle, H. Bolfarine, and V. Lachos, "Skew-normal linear mixed models," *Journal of Data Science*, vol. 3, no. 4, pp. 415–438, 2005.

[171] R. D. Gibbons, D. Hedeker, C. Waternaux, and J. Davis, "Random regression models: A comprehensive approach to the analysis of longitudinal psychiatric data," *Psychopharmacology Bulletin*, vol. 24, no. 3, pp. 438–443, 1988.

[172] R. D. Gibbons and D. Hedeker, "Application of random-effects probit regression models." *Journal of Consulting and Clinical Psychology*, vol. 62, no. 2, pp. 285–296, 1994.

[173] P. Shi and Z. Zhao, "Regression for copula-linked compound distributions with applications in modeling aggregate insurance claims," *The Annals of Applied Statistics*, vol. 14, no. 1, pp. 357–380, 2020.

[174] P. K. Andersen, C. T. Ekstrøm, J. P. Klein, Y. Shu, and M.-J. Zhang, "A class of goodness of fit tests for a copula based on bivariate right-censored data," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 47, no. 6, pp. 815–824, 2005.

[175] M. Lakhal-Chaieb, "Copula inference under censoring," *Biometrika*, vol. 97, no. 2, pp. 505–512, 2010.

[176] T. Emura, C.-W. Lin, and W. Wang, "A goodness-of-fit test for archimedean copula models in the presence of right censoring," *Computational Statistics & Data Analysis*, vol. 54, no. 12, pp. 3033–3043, 2010.

[177] A. Wang, "Goodness-of-fit tests for archimedean copula models," *Statistica Sinica*, pp. 441–453, 2010.

[178] C. Genest, I. Kojadinovic, J. Nešlehová, and J. Yan, "A goodness-of-fit test for bivariate extreme-value copulas," *Bernoulli*, vol. 17, no. 1, pp. 253–275, 2011.

[179] I. Kojadinovic, J. Yan, and M. Holmes, "Fast large-sample goodness-of-fit tests for copulas," *Statistica Sinica*, pp. 841–871, 2011.

[180] S. Zhang, O. Okhrin, Q. M. Zhou, and P. X.-K. Song, "Goodness-of-fit test for specification of semiparametric copula dependence models," *Journal of Econometrics*, vol. 193, no. 1, pp. 215–233, 2016.

[181] A. Prokhorov, U. Schepsmeier, and Y. Zhu, "Generalized information matrix tests for copulas," *Econometric Reviews*, vol. 38, no. 9, pp. 1024–1054, 2019.

[182] T. Sun, Y. Cheng, and Y. Ding, "An information ratio-based goodness-of-fit test for copula models on censored data," *Biometrics*, vol. 79, no. 3, pp. 1713–1725, 2023.