Modeling Nested Copulas with GLMM Marginals for Longitudinal Data

Roba Bairakdar

A Thesis
in
The Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science (Mathematics) at
Concordia University
Montreal, Quebec, Canada

December 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Roba Bairakdar

Entitled: Modeling Nested Copulas with GLMM Marginals for Longitudinal Data

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

______________________________ Examiner
Dr. F. Godin

______________________________ Examiner
Dr. L. Kakinami

______________________________ Co-supervisor
Dr. F. Ducharme

______________________________ Supervisor
Dr. M. Mailhot

Approved by ______________________________
Chair of Department or Graduate Program Director

____
Dean of Faculty

____
Date

Abstract

Modeling Nested Copulas with GLMM Marginals for Longitudinal Data

A flexible approach for modeling longitudinal data is proposed. The model consists of nested bivariate copulas with Generalized Linear Mixed Models (GLMM) marginals, which are tested and validated by means of likelihood ratio tests and compared via their AIC_c and BIC values. The copulas are joined together through a vine structure. Rank-based methods are used for the estimation of the copula parameters, and appropriate model validation methods are used such as the Cramér Von Mises goodness-of-fit test. This model allows flexibility in the choice of the marginal distributions, provided by the family of the GLMM. Additionally, a wide variety of copula families can be fitted to the tree structure, allowing different nested dependence structures. This methodology is tested by an application on real data in a biostatistics study.

Acknowledgments

This thesis reached its final form by the kind support and encouragement of many individuals, to whom I would be eternally grateful. I would like to start by thanking Allah for providing me with the strength and motivation that I required in this journey.

I am extremely thankful to my supervisor Professor Mélina Mailhot for her invaluable direction, trust and encouragement. Her optimistic nature is contagious, and I always look forward to our meetings to get the positive energy boost needed for this research. She has gone out of her way several times to provide me with the best opportunities for this thesis and for my future, and I am forever indebted to her.

I would also like to thank Dr. Francine Ducharme who provided me with the opportunity to collaborate with the research center at Centre Hospitalier Universitaire Sainte-Justine. Her insightful thoughts and massive amount of medical knowledge were extremely useful in this research. Furthermore, I would like to thank Professor Frédéric Godin and Professor Lisa Kakinami for reviewing my thesis and for their comments that helped in improving it. I am thankful to Professor Johanna Nešlehová for allowing me to attend her class to learn from her immense knowledge and for her advice on numerous occasions. I would also like to thank the mathematics and statistics department at Concordia and my supervisor for providing me with financial support during my Masters.

My gratitude goes to all my professors throughout my academic journey that spanned across two continents, specifically professors Zeinab Amin, Aliaa Bassiouny, Frédéric Godin, Ali Hadi, Mélina Mailhot, Étienne Marceau, Magdi Moustafa, Johanna Nešlehová and Lea Popovic, presented alphabetically by their surnames.

A special thank you note goes to my friends, from all over the world, for their moral support, and making me feel home away from home. I would also like to mention my students who brought me here in the first place and always remind me of my sense of purpose. Sincere appreciation goes to my family, specifically my mother, grandmother, aunt and brother. It is painful to be thousands of kilometers away from them, but they always find a way to make that distance emotionally shorter. Finally, I am grateful for my father's support; your memory will always be with me.

To my backbone; my mother, grandmother and aunt:
Lama, Samira and Rama.

Contents

List of Figures

List of Tables

Introduction

Modeling the dependence structure for multivariate longitudinal data is an important challenge in all fields. In the literature, it is usually assumed that the data, or a transformation of the data, is generated from a multivariate normal distribution, with a variance-covariance matrix that explains the dependence between the multiple response variables, and the serial dependence. However, we often come across data that are not normally distributed, and hence, a generalized methodology is needed to fit all distributions. Furthermore, assuming a common distribution for all the responses might not be appropriate. Therefore, in this thesis, we propose a parametric approach for a nested copula model for fitting multiple responses of longitudinal data, where each response is initially modeled by a generalized linear mixed model. This allows for the possibility of using a variety of continuous and discrete distributions. Under this approach, the marginal distributions take into account the dependence between each response and its covariates, over time, while the copula holds the general structure for the dependence between each response. Instead of measuring the linear correlation, we examine a more general and appropriate concept of dependence. The estimates for the marginal distributions are obtained by fitting each response to multiple distributions that fit its characteristics, where afterwards variable and model selection criteria are performed to choose the best fit model. Additionally, the estimates for the dependence parameters of the copula is obtained by maximizing the pseudo log-likelihood by using rank-based methods. An inadequate choice for the dependence parameter and copula may result in unexpected deviations in the response variable, especially when one is provided with a small data set. Therefore, goodness of fit tests are performed to ensure the accuracy of the model. The model can be used for predictive modeling and conditional predictive modeling, where the choice of the conditioning response variable is arbitrary and is chosen based on the context of the data.

Our methodology is applied to real data in biostatistics, provided by the research center of Centre Hospitalier Universitaire Sainte-Justine, Montreal, QC. This data set has been the motivation behind this research.

This thesis is structured as follows. In Chapter 1, we discuss different regression models to fit data with only one outcome variable. The assumptions and properties of each model are explained in details. Chapter 2 introduces multivariate distributions and their link with copulas. Several properties of different copula families are explored. We also explain measures of dependence and how to use copulas for predictions. In Chapter 3, we explain different criteria to choose the model and variables that provide the best fit for any data. A small number of observations can be restrictive in modeling, and special criteria are mentioned to overcome this problem. We propose a model that is appropriate for modeling multivariate longitudinal data in Chapter 4. This model provides flexibility in modeling responses from various distributions, with different pairwise dependence structure. Additionally, in Chapter 5, we provide a real life application in biostatistics for the suggested model and we illustrate the procedure for predictive modeling simulations.

Chapter 1

Univariate Models

Linear regression is used to model the relationship between a response variable, also called the dependent variable, outcome variable, predicted variable or regressand and denoted by y, and a set of predictors, also called the independent variables, explanatory variables, covariates or regressors and denoted by $x_1, x_2, \ldots, x_p$, by assuming a linear relationship between them. If there is only one predictor x_1, this is referred to as Simple Linear Regression, however if there are two or more predictors, it is referred to as Multiple Linear Regression (MLR). The goal of linear regression is to identify the strength of the linear relationship between the response variable and each predictor, identify the predictors that have no effect on the response variable and to predict values for the response variable using any values for the predictors. Given a real data set, we do not know the parameters of the model, but we can explain the relationship between the response variable and the predictors by estimating the model parameters and using them to identify the conditional expectation of the response variable given the predictors. There are several methods to fit linear models, some of which will be explained in this thesis where the availability of more than one predictor (i.e. MLR) is assumed. The discussed methods are Ordinary Least Squares (OLS), Generalized Linear Models (GLM) and Generalized Linear Mixed Models (GLMM). There are also non-linear regression models that assume that the relationship between the dependent variable and the independent variables is non-linear in terms of the regression parameters, but they will not be explored further in this thesis due to their complexity. Linear models are more commonly used as they can be easily modeled and explained.

1.1 Ordinary Least Squares

The earliest method for estimating the parameters of a linear model is the Ordinary Least Squares (OLS) method, which was first used by Gauss and Legendre as explained in Stigler (1981) who applied the model to astronomical data sets. The goal of OLS is to find the linear model that minimizes the square of the prediction error.

1.1.1 The Model

Consider a data set that contains n observations. Each observation i consists of a scalar response variable y_i and a set of p predictors x_{ij} for $j = 1, \ldots, p$. The relationship between the response variable and the predictors for observation i are assumed to be linear in parameters, but not necessarily linear in predictors. This means that, for example, the variables x_{ij} can be to any power, but the parameters of the model have to maintain the linearity assumption. The general form for OLS is

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i,$$

where β_0 is called the model intercept, $\beta_1, \ldots, \beta_p$ are the regression coefficients and ϵ_i is the random error, which is the difference between the actual observed value of the response variable, and the predicted value from using the above model. Each regression coefficient represents an additive change in the expected value of y resulting from a one unit increase in the predictor associated with that regression coefficient. It is assumed that $\epsilon \sim N(0, \sigma^2)$, all ϵ_i's are independent and that the predictors are nearly linearly independent (no strong multi-collinearity). All those assumptions will be discussed later in Section 1.1.3. The above model can be rewritten in matrix notation as follows:

$$\overbrace{Y}^{n\times 1} = \overbrace{\underbrace{X}_{n\times(p+1)} \quad \underbrace{\boldsymbol{\beta}}_{(p+1)\times 1}}^{n\times 1} + \overbrace{\epsilon}^{n\times 1},$$

where all entries in the first column of X are equal to 1. The goal of OLS is to find estimates for the regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize the squared differences between the observed response variable y_i, and the predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}$. The difference $y_i - \hat{y}_i$ is called the regression residuals, and it is represented by the blue lines in Figure 1.1.

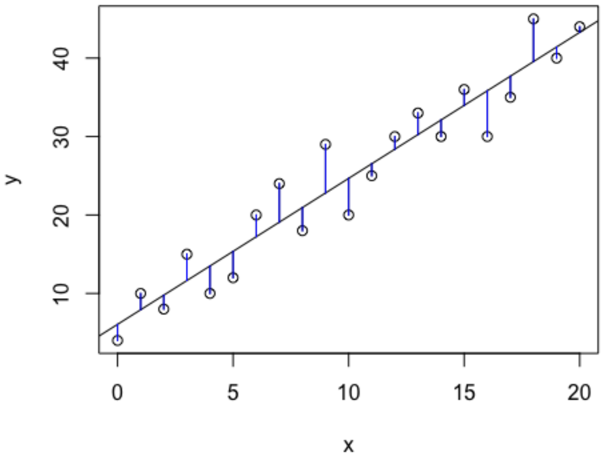

Figure 1.1: Illustration of OLS regression. The straight line minimizes the squared differences between the observed response and the predicted values, indicated by the blue vertical lines.

1.1.2 Estimation of Model Parameters

The regression coefficients of the OLS model can be obtained by solving

$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right) = \underset{(\beta_0,\beta_1,\ldots,\beta_p)}{\arg\min} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2,$$

or its equivalent in matrix notation

$$\hat{\boldsymbol{\beta}} = \underset{\beta \in \mathbb{R}^{(p+1)}}{\arg\min} \|Y - X\boldsymbol{\beta}\|^2. \tag{1.1.1}$$

Let $f\left(\boldsymbol{\beta}\right)$ be the objective function in the optimization of Eq. 1.1.1, then

$$\begin{aligned} f\left(\boldsymbol{\beta}\right) &= \|Y - X\boldsymbol{\beta}\|^2 \\ &= Y^T Y - 2\boldsymbol{\beta}^T X^T Y + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta}, \end{aligned}$$

and

$$\frac{\partial f\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = -2X^T Y + 2X^T X \boldsymbol{\beta} = 0.$$

This leads to the following regression coefficient estimates

$$\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y, \tag{1.1.2}$$

provided that $\left(X^T X\right)^{-1}$ exists.

1.1.3 Assumptions

It is important to validate the required assumptions for fitting OLS regression to the data, otherwise, we can have incorrect and misleading results. Those assumptions are explained in details in Allen (1997) and are summarized in the following points,

Assumption 1. *The linear regression model is linear in parameters.*

The relationship between the response variable Y and the predictors X's is linear in parameters β and not necessarily linear in X's. Therefore, Eq. I and Eq. II are acceptable, but Eq. III is not acceptable;

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \tag{I}$$

$$y_i = \beta_0 + \beta_1 x_{i1}^2 + \ldots + \beta_p x_{ip}^p + \epsilon_i, \tag{II}$$

$$y_i = \beta_0 + \beta_1^2 x_{i1} + \ldots + \beta_p^p x_{ip} + \epsilon_i. \tag{III}$$

Assumption 2. *The observations in the data set are independent and sampled randomly such that the number of observations $y_1, \ldots, y_n$ is bigger than the number of parameters $\boldsymbol{\beta}$.*

Independence of the observations is one of the important assumptions because it assures that we have a model with only fixed effects, and no random effects. Random effects can occur when the observations can be grouped in different categories, such that each category varies uniquely from the mean of the population. Section 1.3 further explains random effects and the changes that occur in the modeling as a result of their presence. In addition, if the number of observations n is equal to the number of parameters p, then we have equal number of equations as unknowns, which can be solved algebraically without the need for OLS. If $n < p$, a unique solution is impossible to find algebraically or by using OLS.

Assumption 3. *Multi-collinearity should be minimized.*

There should be almost no linear relationship between the predictors. If there is a strong linear relationship (Pearson correlation coefficient ρ_p close to ± 1) between the predictors, we should drop some of them such that the chosen model has almost uncorrelated predictors.

Assumption 4. *The predictors are non-random.*

The predictors $x_{i1}, \ldots, x_{ip}$ are assumed to have fixed values such that variation in the predictors causes variation in the outcome y_i. However, changes in the outcome should not imply changes in the predictors. In other words, if we are modeling the amount of auto insurance losses, then it is assumed that the amount of the loss depends on the car type, but the car type does not depend on the amount of the loss.

Assumption 5. $\epsilon \sim N(0, \sigma^2)$.

The error terms should be independent and identically distributed (IID) with mean 0 and constant variance σ^2, and given the previous assumption, this makes the response variable random as well. In addition, there should be no relationship between the predictors and ϵ.

If we consider the case where the response variable Y is Normally distributed such that $Y \sim N(X\beta, \sigma^2 I_n)$, then the maximum likelihood estimates of β results in the same estimates obtained by OLS.

Proof. To obtain the estimates of the parameters β and σ^2, we use the maximum likelihood estimation method. The density function is

$$\phi\left(Y; X\boldsymbol{\beta}, \sigma^2 I_n\right) = (2\pi)^{-n/2} |\sigma^{-2} I_n|^{1/2} \exp\left\{-\frac{1}{2}\left(Y - X\boldsymbol{\beta}\right)^T \sigma^{-2} I_n \left(Y - X\boldsymbol{\beta}\right)\right\},$$

and its log-likelihood function is

$$\begin{aligned} l\left(\boldsymbol{\beta}, \sigma^2; Y, X\right) &= -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\sigma^{-2} I_n| - \frac{1}{2}\left(Y - X\boldsymbol{\beta}\right)^T \sigma^{-2} I_n \left(Y - X\boldsymbol{\beta}\right) \\ &= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log\sigma^{-2} - \frac{1}{2}\sigma^{-2}\left(Y - X\boldsymbol{\beta}\right)^T \left(Y - X\boldsymbol{\beta}\right). \end{aligned}$$

Therefore, the maximum likelihood estimator is obtained by optimizing

$$\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) = \underset{(\boldsymbol{\beta},\sigma^2)\in\mathbb{R}^{(p+1)}\times\mathbb{R}_+}{\arg\min} -l\left(\boldsymbol{\beta}, \sigma^2; Y, X\right).$$

The partial derivatives of the objective function with respect to each parameter is equated to 0 as follows:

$$\begin{aligned} \frac{\partial\left[-l\left(\boldsymbol{\beta}, \sigma^2; Y, X\right)\right]}{\partial\sigma^2} &= \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\left(Y - X\boldsymbol{\beta}\right)^T \left(Y - X\boldsymbol{\beta}\right) = 0 \\ \sigma^2 &= \frac{1}{n}\left(Y - X\boldsymbol{\beta}\right)^T \left(Y - X\boldsymbol{\beta}\right), \end{aligned}$$

and,

$$\frac{\partial\left[-l\left(\boldsymbol{\beta}, \sigma^2; Y, X\right)\right]}{\partial\boldsymbol{\beta},} = \frac{1}{2}\sigma^{-2}\frac{\partial\left[\left(Y - X\boldsymbol{\beta}\right)^T \left(Y - X\boldsymbol{\beta}\right)\right]}{\partial\boldsymbol{\beta}}$$

$$
\begin{aligned}
&= \frac{1}{2}\sigma^{-2}\frac{\partial\left[Y^TY - 2\boldsymbol{\beta}^TX^TY + \boldsymbol{\beta}^TX^TX\boldsymbol{\beta})\right]}{\partial\boldsymbol{\beta}} \\
&= \frac{1}{2}\sigma^{-2}\left(-2X^TY + 2X^TX\boldsymbol{\beta}\right) = 0 \\
\hat{\boldsymbol{\beta}} &= \left(X^TX\right)^{-1}X^TY.
\end{aligned}
$$

Therefore, when the outcome variable is normally distributed, the estimates of the regression coefficients are identical to those obtained by OLS, as shown in Eq. 1.1.2, provided that X^TX is invertible. □

1.1.4 Goodness of Fit Measures

The closer the predicted values obtained from the fitted model $\hat{y}$ are to the observed values from the data y, the better the fit of the model. There are several measures that can be used to asses the accuracy of the fitted model and to compare different models together. They will be discussed in Chapter 3.1.

1.2 Generalized Linear Models

Even though OLS provides a relatively simple method to fit data sets, it has some restrictions that may prevent us from applying it. As explained by McCullagh (1984), Generalized Linear Models (GLMs) are an extension of linear models where the mean of the response variable is linearly related to the predictors via an arbitrary link function and the variance of the response variable depends on the mean. This means that a function of $\mathbb{E}(Y)$ is linearly related to the predictors, rather than the response variable itself being linearly related to the predictors. GLMs also allow us to drop the normality assumption of the error terms under the OLS (i.e. ϵ_i does not have to be normally distributed with zero-mean and constant variance σ^2). In addition, by using GLM, we have the flexibility to model data where the response variable is bounded or discrete. GLMs assumes independence of the observations, and hence there are only fixed effects in the model. Section 1.3 will discuss the modeling performed when there is dependence between the observations. Further assumptions of GLM are discussed in Section 1.2.5.

1.2.1 The Model

Consider a model with response vector $Y = (y_1, \dots, y_n)$, and p predictors arranged in a $n \times p$ matrix X where n represents the number of observations. The responses $y_1, \dots, y_n$ are assumed to be independent and generated from the same exponential family, which is discussed in details in Section 1.2.2. The mean of the response vector Y is assumed to be linearly related to the predictors via an arbitrary link function $g(\cdot)$ as follows:

$$g\left(\mathbb{E}\left[Y\right]\right) = g(\mu) = X\boldsymbol{\beta} = \boldsymbol{\eta},$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients, which is usually estimated using the maximum likelihood method. $\boldsymbol{\eta}$ is called the linear predictor and its components can be defined by

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = X_i^T\boldsymbol{\beta}.$$

The true mean of the response variable can be calculated by taking the inverse of the link function (i.e. $\mathbb{E}\left[Y\right] = g^{-1}\left(X\boldsymbol{\beta}\right) = g^{-1}\left(\boldsymbol{\eta}\right)$) and the variance $Var(Y)$ is a function of the mean, and it is generated from the exponential family chosen for the model.

1.2.2 The Exponential Family

Consider a response vector Y where responses $y_1, \dots, y_n$ are assumed to independent and generated from the same distribution with probability density function

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{1.2.1}$$

where $a_i(\phi), b(\theta_i)$ and $c(y_i, \phi)$ are known functions that differ depending on the chosen distribution. This distribution is referred to as the Exponential Family distribution. The function $a_i(\phi)$ is usually of the form

$$a_i(\phi) = \phi/\omega_i,$$

where ϕ is referred to as the dispersion parameter and is constant over all observations, and ω_i is a known prior weight that differs between observations, but usually equals to 1. The mean and variance of Y are

$$\mathbb{E}[Y_i] = \mu_i = b'(\theta_i) \tag{1.2.2}$$

$$Var[Y_i] = \sigma_i^2 = b''(\theta_i)a_i(\phi) \tag{1.2.3}$$

respectively, where

$$b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}, \text{ and } b''(\theta_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}.$$

Example: Poisson Distribution

Let $Y \sim \text{Poisson}(\lambda)$, where $y \in 0, 1, 2, ...$ and $\lambda > 0$, then the probability mass function of Y is

$$\begin{aligned} f_Y(y) &= \frac{\lambda^y \mathrm{e}^{-\lambda}}{y!} \\ &= \exp\left\{y \log \lambda - \lambda - \log(y!)\right\}. \end{aligned}$$

If we let $\theta = \log \lambda$, $\omega_i = 1$ and $\phi = 1$, then

$$f_Y(y) = \exp\left\{\frac{y\theta - \exp\{\theta\}}{1} - \log(y!)\right\}.$$

Therefore, the Poisson distribution is a member of the exponential family with $b(\theta) = \exp\{\theta\}$ and $a(\phi) = 1$. The mean and variance of Y are obtained by using Eq. 1.2.2 and Eq. 1.2.3

$$\begin{aligned} \mathbb{E}[Y] = \mu = b'(\theta) = \exp\{\theta\} = \exp\{\log \lambda\} = \lambda \qquad (1.2.4) \\ Var[Y] = \sigma^2 = b''(\theta)a(\phi) = \exp\{\log \lambda\} = \lambda. \end{aligned}$$

□

Example: Gamma Distribution

Let $Y \sim \text{Gamma}(\alpha, \beta)$, where $0 < \alpha, \beta, y < \infty$, then the density function of Y is

$$\begin{aligned} f_Y(y) &= \frac{y^{\alpha-1}\beta^{\alpha}\mathrm{e}^{-\beta y}}{\Gamma(\alpha)} \\ &= \exp\left\{-\beta y + \alpha \log \beta + (\alpha - 1)\log y - \log \Gamma(\alpha)\right\} \\ &= \exp\left\{\frac{y(-\beta/\alpha) - [-\log \beta]}{1/\alpha} + (\alpha - 1)\log y - \log \Gamma(\alpha)\right\}. \end{aligned}$$

If we let $\theta = -\beta/\alpha$, $\omega_i = 1$ and $\phi = 1/\alpha$, then

$$\begin{aligned} f_Y(y) &= \exp\left\{\frac{y\theta - [-\log(-\theta\alpha)]}{\phi} + (1/\phi - 1)\log y - \log \Gamma(1/\phi)\right\} \\ &= \exp\left\{\frac{y\theta - [-\log(-\theta)]}{\phi} + \frac{\log \alpha}{\phi} + (1/\phi - 1)\log y - \log \Gamma(1/\phi)\right\}. \end{aligned}$$

Therefore, the Gamma distribution is a member of the exponential family with $b(\theta) = -\log(-\theta)$ and $a(\phi) = \phi = 1/\alpha$. The mean and variance of Y are obtained by using Eq. 1.2.2 and Eq. 1.2.3 as follows:

$$\begin{aligned} \mathbb{E}[Y] = \mu = b'(\theta) = -\frac{1}{\theta} = \frac{\alpha}{\beta} \\ Var[Y] = \sigma^2 = b''(\theta)a(\phi) = \frac{\phi}{\theta^2} = \frac{\alpha}{\beta^2}. \end{aligned} \tag{1.2.5}$$

□

1.2.3 The Link Function

If the distribution from the exponential family is expressed in terms of its mean μ_i, such that $\theta_i = g(\mu_i)$ for a given function $g(\cdot)$, then $g(\cdot)$ is referred to as the canonical link function. The canonical link function is the default link function used in GLMs, but it is not mandatory. Although the canonical link function can provide desirable statistical properties, non-canonical link functions can be used if they provide a better fit for the data or if they can better explain the model and the coefficients. Table 1.1 provides a summary of the canonical link function for some common distributions that are members of the exponential family. The link function is linearly related to the predictors in the model such that

$$g(\mathbb{E}[Y]) = g(\mu) = X\boldsymbol{\beta} = \boldsymbol{\eta}.$$

In order to obtain the mean of the model, one can invert the link function as follows:

$$\mathbb{E}[Y] = g^{-1}(X\boldsymbol{\beta}) = g^{-1}(\boldsymbol{\eta}).$$

Note that with GLMs, one does not transform the response variable, but rather the mean of the response variable. Therefore, a model where $\log Y$ is linearly related to the predictors (Eq. IV) is not the same model where GLM is used with a log link function, where in the latter $\log \mathbb{E}[Y]$ is linear on the predictors (Eq. V).

$$\log y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \tag{IV}$$

$$\log \mathbb{E}[y_i] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i. \tag{V}$$

Table 1.1: Characteristics of some common exponential family members as shown by McCullagh (1984).

Distribution	Support	$g(\mu)$	Canonical link name
Normal	$(-\infty, \infty)$	μ	Identity
Poisson	$0, 1, 2 \ldots$	$\log \mu$	Log
Binomial	$0, 1, 2 \ldots, N$	$\log \frac{\mu}{1-\mu}$	Logit
Gamma	$(0, \infty)$	$1/\mu$	Inverse
Inverse Gaussian	$(0, \infty)$	$1/\mu^2$	Inverse squared

Example: Poisson Distribution

The canonical link function for the Poisson distribution is obtained by finding $g(\cdot)$, where $\theta = g(\mu)$. By observing Eq. 1.2.4, we have that $\theta = \log \mu$, therefore, the canonical link function for the Poisson distribution is $g(\mu) = \log \mu$. □

Example: Gamma Distribution

The canonical link function for the Gamma distribution is obtained by finding $g(\cdot)$, where $\theta = g(\mu)$. By observing Eq. 1.2.5, we have that $\theta = -1/\mu$, therefore the canonical link function for the Gamma distribution is $g(\mu) = -1/\mu$. This canonical link function is equivalent to the inverse link function, which is shown as follows:

$$g(\mu_i) = -\frac{1}{\mu} = \sum_{j=1}^{p} x_{ij}\beta_j.$$

Therefore,

$$\begin{aligned}\frac{1}{\mu} &= \sum_{j=1}^{p} x_{ij}\left(-\beta_j\right)\\ &= \sum_{j=1}^{p} x_{ij}\beta_j^*,\end{aligned}$$

where $\beta_j^* = -\beta_j$. However, this does not enforce positive means for the model. Thus, it is more common to use the log link function $g(\mu) = \log(\mu)$ for data that requires positive values. □

1.2.4 Estimation of Model Parameters

To obtain the model parameters, we can use the maximum likelihood estimation method, where we differentiate the negative log-likelihood with respect to the parameter of interest $\boldsymbol{\beta}$. The likelihood function of a distribution that is a member of the exponential family and has a density function as in Eq. 1.2.1 is defined as

$$L(\boldsymbol{\theta}, \phi; Y) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\},$$

and its log-likelihood is

$$l(\boldsymbol{\theta}, \phi; Y) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right]. \tag{1.2.6}$$

The maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ is obtained by solving the following system of equations

$$\frac{\partial l(\boldsymbol{\theta}, \phi; Y)}{\partial \beta_j} = \frac{\partial l}{\partial \beta_j} = 0.$$

However, our parameter of interest β_j is not explicit in Eq. 1.2.6. But we know the following

$$\mu_i = b'(\theta_i), \qquad \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j, \qquad g(\mu_i) = \eta_i.$$

By applying the chain rule, we have that

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j},$$

where

$$\frac{\partial l}{\partial \theta_i} = -\frac{y_i - b'(\theta_i)}{a_i(\phi)} = -\frac{y_i - \mu_i}{a_i(\phi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = 1/\frac{\partial \mu_i}{\partial \theta_i} = 1/\frac{\partial b'(\theta_i)}{\partial \theta_i} = 1/b''(\theta_i)$$

$$\frac{\partial \mu_i}{\partial \eta_i} = 1/\frac{\partial \eta_i}{\partial \mu_i} = 1/\frac{\partial}{\partial \mu_i} g(\mu_i) = 1/g'(\mu_i)$$

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \sum_{j=1}^{p} x_{ij}\beta_j = x_{ij}.$$

Therefore, we obtain

$$\frac{\partial l}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{a_i(\phi)b''(\theta_i)g'(\mu_i)}.$$

By using Eq. 1.2.3, the above formula can be simplified into

$$\frac{\partial l}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var[\mathrm{Y}_i]g'(\mu_i)}.$$

There are no closed form solutions for all GLM models. Therefore, numerical optimization is performed using computer software. The most common technique is Iterative Weighted Least Squares (IWLS), followed by Fisher scoring method and Newton Raphson method as explained by McCullagh (1984).

The interpretation of model parameters for the GLM models is slightly different than that of the OLS coefficients. Consider a GLM model with a log link function and only one covariate, then it can be expressed as follows:

$$\log \mathbb{E}\left[Y|X\right] = \beta_0 + \beta_1 X \quad \Leftrightarrow \quad \mathbb{E}\left[Y|X\right] = \exp\left(\beta_0 + \beta_1 X\right). \tag{1.2.7}$$

Therefore, if we increase the value of the covariate by 1, the log of $\mathbb{E}\left[Y|X\right]$ increases by β_1 as follows:

$$\begin{aligned} \log \mathbb{E}\left[Y|X+1\right] &= \beta_0 + \beta_1(X+1) \\ \log \mathbb{E}\left[Y|X+1\right] &= \beta_0 + \beta_1 X + \beta_1. \end{aligned}$$

However, we are not interested in the change of the log of the mean of Y, but rather the change of the mean of Y. Therefore, by applying Eq. 1.2.7, we have that

$$\begin{aligned} \log \mathbb{E}\left[Y|X+1\right] = \beta_0 + \beta_1 X + \beta_1 \quad \Leftrightarrow \quad \mathbb{E}\left[Y|X+1\right] &= \exp\left(\beta_0 + \beta_1 X + \beta_1\right) \\ &= \exp\left(\beta_0 + \beta_1 X\right)\exp\left(\beta_1\right) \\ &= \mathbb{E}\left[Y|X\right]\exp\left(\beta_1\right). \end{aligned} \tag{1.2.8}$$

So $\exp(\beta_1)$ is a multiplicative factor that represents the increase due to a 1 unit increase in X, and we refer to it as the transformed parameter estimate.

1.2.5 Assumptions

Similar to OLS, there are some assumptions that should hold in order to use GLMs, or to choose the distribution used in the modeling. Breslow (1996) explained all the assumptions needed for GLMs and they are summarized in this section.

Assumption 6. *Correct choice for all components of the GLM.*

The distribution used in the modeling should be well suited for the data. For example, if we are observing an outcome variable that is positive, continuous and rightly skewed, the Gamma distribution can be a good choice. However, if we are modeling an outcome variable that is discrete and represents count of events, then the Poisson distribution can be a good choice. Additionally, the link function should be chosen to well represent the data. For example, if we are modeling positive outcomes, the link function should be chosen such that its inverse would always result in positive mean values.

Assumption 7. *The observations in the data set are independent and sampled randomly such that the number of observations $y_1, \ldots, y_n$ are bigger than the number of parameters $\boldsymbol{\beta}$.*

Independence of the observations is one of the important assumptions because it assures that we have a model with only fixed effects, and no random effects. We explore the changes in the model for data that has random effects in Section 1.3. In addition, if the number of observations n is equal to the number of parameters p, then we have equal number of equations as unknowns, which can be solved algebraically. If $n < p$, then no unique solution is available.

Assumption 8. *The predictors are non-random.*

The predictors $x_{i1}, \ldots, x_{ip}$ are assumed to have fixed values such that variation in the predictors causes variation in the outcome y_i. However, changes in the outcome should not imply changes in the predictors. In other words, if we are modeling the amount of auto insurance losses, then it is assumed that the amount of the loss depends on the car type, but the car type does not depend on the amount of the loss.

Assumption 9. *Multi-collinearity should be minimized.*

There should be “almost” no linear relationship between the predictors. If there is a strong linear relationship (Pearson correlation coefficient ρ_p close to ± 1) between the predictors, we should drop some of them such that the chosen model has “almost” uncorrelated predictors.

1.2.6 Goodness of Fit Measures

By estimating the parameters of the GLM model, we can obtain the fitted values $\hat{y}$ which are generally not equivalent to the original data values y, with the goal to obtain small differences between them. McCullagh (1984) mentioned that there are several measures used to calculate that difference and they use the log-likelihood function illustrated in Eq. 1.2.6.

Consider the following:

- $l(\hat{\boldsymbol{\theta}}, \phi; Y)$ represents the maximized log-likelihood of the fitted model for a fixed value of the dispersion parameter ϕ,
- $l(\tilde{\boldsymbol{\theta}}, \phi; Y)$ represents the log-likelihood from the saturated model, a hypothetical model, where each observation is perfectly fitted without errors, and
- $l(\boldsymbol{\theta}_0, \phi; Y)$ represents the log-likelihood from the null model, a hypothetical model, with only an intercept value and no predictors such that every observation is estimated by the mean.

The predictions from the saturated model, $\tilde{y}_i$, exactly match the actual observations, y_i. For this model, each observation has its own parameters, i.e. there are n estimates for $\tilde{\boldsymbol{\beta}}$, and therefore, each estimate perfectly predicts the value of the outcome. The saturated model can be observed as the upper bound of the log-likelihood function, because it theoretically provides the best fit, while the null model is the lower bound of the log-likelihood function. The maximized model has a log-likelihood value in between those bounds.

The discrepancy of the fit is measured by obtaining the difference between the saturated model and the fitted model, which gives us a quantity named the scaled deviance, defined as follows:

$$D^*(\hat{\boldsymbol{\theta}}, \phi; Y) = 2\left[l(\tilde{\boldsymbol{\theta}}, \phi; Y) - l(\hat{\boldsymbol{\theta}}, \phi; Y)\right].$$

By using Eq. 1.2.6, and $a_i(\phi) = \phi/\omega_i$ we can rewrite the scaled deviance of the model as

$$\begin{aligned} D^*(\hat{\boldsymbol{\theta}}, \phi; Y) &= 2\left[\sum_{i=1}^{n}\left[\frac{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)}{a_i(\phi)} + c(y_i, \phi)\right] - \sum_{i=1}^{n}\left[\frac{y_i\hat{\theta}_i - b(\hat{\theta}_i)}{a_i(\phi)} + c(y_i, \phi)\right]\right] \\ &= 2\left[\sum_{i=1}^{n}\left[\frac{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)}{\phi/\omega_i} + c(y_i, \phi) - \frac{y_i\hat{\theta}_i - b(\hat{\theta}_i)}{\phi/\omega_i} + c(y_i, \phi)\right]\right] \end{aligned}$$

$$= \frac{2}{\phi} \sum_{i=1}^{n} \omega_i \left[y_i \left(\tilde{\theta}_i - \hat{\theta}_i \right) - \left(b(\tilde{\theta}_i) - b(\hat{\theta}_i) \right) \right], \tag{1.2.9}$$

where the deviance of the model is defined as

$$D(\hat{\boldsymbol{\theta}}, \phi; Y) = 2 \sum_{i=1}^{n} \omega_i \left[y_i \left(\tilde{\theta}_i - \hat{\theta}_i \right) - \left(b(\tilde{\theta}_i) - b(\hat{\theta}_i) \right) \right], \tag{1.2.10}$$

and the quantity $D^*(\hat{\boldsymbol{\theta}}, \phi; Y)$ is simply the deviance scaled by the dispersion parameter ϕ. The values of $D(\hat{\boldsymbol{\theta}}, \phi; Y)$ and $D^*(\hat{\boldsymbol{\theta}}, \phi; Y)$ are always positive since the saturated model has a higher log-likelihood value than any fitted model, and their values will approach 0 when the fitted parameters perfectly explain the model without errors.

Example: Poisson Distribution

The deviance for the Poisson distribution is obtained by using previous results that we obtained in earlier examples. Since $\theta = \log \mu$, and $b(\theta) = \exp\{\theta\}$ for the Poisson distribution and by assuming equal priori weights $\omega_i = 1$, then the deviance from Eq. 1.2.10 becomes

$$\begin{aligned} D(\hat{\boldsymbol{\theta}}, \phi; Y) = D(\hat{\boldsymbol{\mu}}; Y) &= 2 \sum_{i=1}^{n} \omega_i \left[y_i \left(\log y_i - \log \hat{\mu}_i \right) - \left(y_i - \hat{\mu}_i \right) \right] \\ &= 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{\hat{\mu}_i} - \left(y_i - \hat{\mu}_i \right) \right]. \end{aligned}$$

The scaled deviance for the Poisson distribution is the same as the deviance because the dispersion parameter $\phi = 1$. □

Example: Gamma Distribution

Similar to the Poisson distribution, obtaining the deviance for the Gamma distribution relies on previous results that we obtained in earlier examples. Since $\theta = -1/\mu$ and $b(\theta) = -\log(-\theta)$ for the Gamma distribution and by assuming equal priori weights $\omega_i = 1$, then the deviance from Eq. 1.2.10 becomes

$$\begin{aligned} D(\hat{\boldsymbol{\theta}}, \phi; Y) = D(\hat{\boldsymbol{\mu}}; Y) &= 2 \sum_{i=1}^{n} \omega_i \left[y_i \left(\frac{-1}{y_i} - \frac{-1}{\hat{\mu}_i} \right) - \left(\log y_i - \log \hat{\mu}_i \right) \right] \\ &= 2 \sum_{i=1}^{n} \left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log \frac{y_i}{\hat{\mu}_i} \right]. \end{aligned}$$

The scaled deviance for the Gamma distribution is obtained by scaling the deviance by the dispersion parameter ϕ such that $D^*(\hat{\boldsymbol{\theta}}, \phi; Y) = \frac{D(\hat{\boldsymbol{\theta}}, \phi; Y)}{\phi}$. □

It is important to mention that when log-likelihoods or deviance is used to compare models, this comparison is only valid when the compared models are done over the same data set with the same number of observations. This is because the log-likelihood is obtained by summing the log-likelihoods for each observation, and if a model has more observations than another model, then it will have a higher log-likelihood value, which should not be attributed to having a better fit to the data. It is also important to use deviance only in comparing models that have the same distribution and the same dispersion parameter, i.e. everything in the model should be identical except the coefficients. This is because the deviance measures the deviation from the log-likelihood of the saturated model. Therefore changing any assumptions in the model other than the coefficients would change the value of the log-likelihood of the saturated model, not only the fitted model, which makes the comparison between models by using the deviance obsolete. If the distribution of a model is a special case from another model, such as the Poisson distribution being a special case of the Negative Binomial distribution, then it is appropriate to use the deviance as a model selection criteria.

Additionally, a model M_1 is said to be nested of another model M_2 if it uses a subset of the predictors of M_2. If we want to compare the two nested models M_1 and M_2 with p_1 and p_2 number of predictors respectively, such that $p_2 > p_1$, and parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ respectively, we can use the scaled deviance (or deviance) of each model to obtain a likelihood ratio test statistic as follows:

$$D^*(\hat{\boldsymbol{\theta}}_1, \phi; Y) - D^*(\hat{\boldsymbol{\theta}}_2, \phi; Y) = 2\left[l(\hat{\boldsymbol{\theta}}_2, \phi; Y) - l(\hat{\boldsymbol{\theta}}_1, \phi; Y)\right] = 2\ln\frac{l(\hat{\boldsymbol{\theta}}_2, \phi; Y)}{l(\hat{\boldsymbol{\theta}}_1, \phi; Y)}.$$

This statistics asymptotically follows the Chi-Square distribution with degrees of freedom $\nu = p_2 - p_1$.

Frequently, we would like to compare models that are not nested, or from different exponential families, and hence comparing their deviance is not an accurate goodness of fit test because of reasons mentioned earlier. In that case, we will have to refer to other model selection criteria which will be explored later in Chapter 3.1.

1.3 Generalized Linear Mixed Models

Sometimes it occurs that the observations in the data are not independent, for example, longitudinal data where repeated observations of the same variables are measured over time for the same individual, or when the data is obtained from groups (countries, hospitals, schools, etc.). This adds a random effect to the model, which makes GLM no longer applicable. Generalized Linear Mixed Models (GLMMs) are an extension of GLMs where random effects are added to the model, in addition to the usual fixed effects available in GLMs. The term "mixed model" implies the use of both fixed and random effects in the modeling. Random effects are always associated with categorical variables, which divides the data into several groups. For example, assume we want to model and make statistical inference about the amount of auto insurance losses in Canada, and the data is collected from several insurance companies. For each company, we will obtain the amount of the losses $\boldsymbol{Y}$, and the predictors $\boldsymbol{X}$ which includes age and profession of the insured, car model, etc. If the companies represent the entire population, (i.e. we collected data from all companies in Canada), then we would want to focus our analysis on the effect of each company and on its impact on the loss amount. Therefore, the parameters for each company are considered model parameters, and not random variables, hence, the company is a fixed effect. However, if the companies represent a sample of the population (i.e. we collected data from some companies in Canada), we will be interested in knowing the trend for the entire population, not just those sampled companies. Therefore, the parameters of those companies are no longer considered fixed model parameters; they are random variables and thus have probability distributions. Hence, we will be interested in knowing the variance between the companies, in order to make a general conclusion about the population. The distinction between the companies being considered a population versus samples is what distinguishes a model with *fixed effects* and *random effects*, respectively.

1.3.1 The Model

Consider a sample of N independent multivariate response $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{in})^T$ such that $i = 1, \ldots, N$, where y_{ij} is the j^{th} response for the i^{th} group/subject. For simplicity of notation, it is assumed that each group has the same number of observations n. We assume that each response y_{ij} depends on a $p \times 1$ vector of fixed predictors $\boldsymbol{x}_{ij}$ associated with a vector of fixed effects coefficients $\boldsymbol{\beta}$ and on a $q \times 1$ vector of fixed predictors $\boldsymbol{z}_{ij}$ associated with a vector of random effects coefficients $\boldsymbol{b}_i = (b_{0i}, b_{1i}, \ldots, b_{qi})^T$. Given the random effect $\boldsymbol{b}$, the mean of the response vector Y is assumed to be related to the predictors via an arbitrary link function $g(\cdot)$ such that

$$g\left(\mathbb{E}\left[Y|\boldsymbol{b}\right]\right) = g(\mu|\boldsymbol{b}) = X\boldsymbol{\beta} + Z\boldsymbol{b}, = \boldsymbol{\eta},$$

where it is assumed that $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G})$, where $\boldsymbol{G}$ is the variance-covariance matrix of the random effects. Note that the random effects help in identifying the variation of each sample/group from the population mean (or the fixed effects), so imposing a mean of zero makes the model unique, and we are interested in estimating the variance. Similar to the GLMs, in order to obtain the mean of the model, one can invert the link function $g(\cdot)$.

To help us explain the above model, we will assume an intercept, only 1 covariate X_{ij} and 3 groups/subject. The expanded matrices become:

$$\begin{bmatrix} \eta_{11} \\ \vdots \\ \eta_{1n} \\ \eta_{21} \\ \vdots \\ \eta_{2n} \\ \eta_{31} \\ \vdots \\ \eta_{3n} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_{11} \\ \vdots \\ x_{1n} \\ x_{21} \\ \vdots \\ x_{2n} \\ x_{3n} \\ \vdots \\ x_{3n} \end{bmatrix} \beta_1 + \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_{01} \\ b_{02} \\ b_{03} \end{bmatrix} + \begin{bmatrix} x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots \\ x_{1n} & 0 & 0 \\ 0 & x_{21} & 0 \\ \vdots & \vdots & \vdots \\ 0 & x_{2n} & 0 \\ 0 & 0 & x_{3n} \\ \vdots & \vdots & \vdots \\ 0 & 0 & x_{3n} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \end{bmatrix},$$

or

$$\begin{bmatrix} \boldsymbol{X}_0 & \boldsymbol{X}_1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z}_0 & \boldsymbol{Z}_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_0 \\ \boldsymbol{b}_1 \end{bmatrix},$$

where $\boldsymbol{X}_0$ is a $(n \times N) \times 1$ vector of ones, $\boldsymbol{X}_1$ is a $(n \times N) \times 1$ vector with elements equal to the covariate X_{ij} for the corresponding observation, β_0 and β_1 are the regression coefficients for $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$, respectively. $\boldsymbol{Z}_0$ is an $(n \times N) \times 3$ matrix whose ij component is 1 if the corresponding observation is in the i^{th} group/subject, and 0 otherwise. $\boldsymbol{Z}_1$ is an $(n \times N) \times 3$ matrix whose elements are X_{ij} if the corresponding observation is from the i^{th} group/subject and 0 otherwise. η_{ij} is represented by

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i} x_{ij}$$

$$= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, x_{ij},$$

where b_{0i} explains the deviation from the intercept, β_0, for the i^{th} group, and b_{1i} is the deviation from the slope of X_1, β_1 for the i^{th} group. In addition,

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_0 \\ \boldsymbol{b}_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I\sigma}_0^2 & \boldsymbol{I\sigma}_{10} \\ \boldsymbol{I\sigma}_{01} & \boldsymbol{I\sigma}_1^2 \end{bmatrix} \right).$$

Conditional on the random effects $\boldsymbol{b}$, the responses $\boldsymbol{Y}$ are assumed to be mutually independent and generated from the same exponential family as explained in Section 1.2.2.

1.3.2 Estimation of Model Parameters

Similar to GLMs, Maximum likelihood estimation is also used to estimate the fixed effects coefficients $\boldsymbol{\beta}$ in GLMMs. In addition, it is also used to estimate the random effects coefficients $\boldsymbol{b}$ and G, the variance of the random effects. Stroup (2012) provided detailed explanation on obtaining the model parameters, and they also confirm on the difficulty of obtaining closed form solutions for the estimates, and hence computer software are used for numerical optimization.

1.3.3 Goodness of Fit Measures

Please refer to Chapter 3.1 for details on assessing the goodness of fit of the model and comparison between models.

Chapter 2

Multivariate Models

The univariate models provide a variety of methods to model data sets that have only one outcome variable Y. However, we often need to model several outcomes and to observe the dependency between them. A multivariate distribution is a distribution that has more than one random variable linked together through a dependence structure. This dependence structure explains if they are independent or dependent on each other, and also the direction and strength of the dependence. This chapter presents the properties of multivariate distribution functions and their relationship with copulas. We also explore the fundamentals of copulas and their use in statistical modeling.

Note that in this section, we work with the assumption that each random variable is a continuous random variable, but some of the notations and properties can be translated to discrete variables by using summands instead of integrands. However, some notations have complicated forms for discrete distributions, especially in copulas.

2.1 Multivariate Distribution Functions

Consider the random vector $\boldsymbol{X}$ which contains n random variables $X_1, \ldots, X_n$ linked together through a joint density function f and joint distribution function F. The support of each random variable X_i is $\mathbb{R}_{X_i} = [L_i, U_i]$, which is the set of values that the random variable can take. For the rest of this chapter, we will assume that the lower limit $L_i = -\infty$ and the upper limit $U_i = \infty, \forall i = 1, \ldots, n$. Consider a set of observations $\{x_1, \ldots, x_n\} \in \mathbb{R}^n$, then their joint distribution function F is defined by

$$F(x_1, x_2, \ldots, x_n) = P\left(X_1 \leq x_1, \ldots, X_n \leq x_n\right).$$

The relationship between the joint probability density function (pdf) and the joint cumulative distribution function (cdf) is defined by

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f\left(x_1, \ldots, x_n\right) \mathrm{d}x_n \cdots \mathrm{d}x_1,$$

and

$$f\left(x_1, \ldots, x_n\right) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, \ldots, x_n). \tag{2.1.1}$$

For a function to be defined as multivariate pdf f, it has to satisfy the following properties:

- $f\left(x_1, \ldots, x_n\right) \geq 0$,

- $\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f\left(x_1, \ldots, x_n\right) \mathrm{d}x_n \cdots \mathrm{d}x_1 = 1$, and

- if $A \subset \mathbb{R}^n$ is a set of values for $\boldsymbol{X}$, then

$$P\left[(X_1, \ldots, X_n) \in A\right] = \int \cdots \int_A f\left(x_1, \ldots, x_n\right) \mathrm{d}x_n \cdots \mathrm{d}x_1.$$

In addition, the multivariate cdf F has the following properties:

- $F(x_1, \ldots, x_n)$ is non-decreasing, i.e. if any of the x_i increases, then $F(x_1, \ldots, x_n)$ also increases,

- If all components approach their maximum attainable value, then the value of the cdf F is equal to 1, i.e.

$$\lim_{x_1, \ldots, x_n \to \infty} F(x_1, \ldots, x_n) = 1,$$

- If one or more components approach their minimum attainable value, then the value of the cdf F is equal to 0, i.e. $, \forall i = 1, \ldots, n$,

$$\lim_{x_i \to -\infty} F(x_1, \ldots, x_n) = 0,$$

- and if $\forall (a_1, \ldots, a_n), (b_1, \ldots, b_n) \in [0, 1]^n$ where $a_i \leq b_i$, we have the rectangle inequality

$$\sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1 + \ldots + i_d} F(x_{1i_1}, \ldots, x_{ni_n}) \geq 0,$$

where $x_{j1} = a_j$ and $x_{j2} = b_j$ $\forall j \in 1, \ldots, n$.

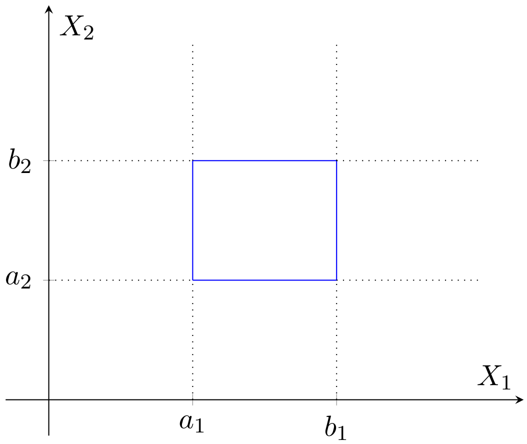

Figure 2.1: Illustration of the rectangle inequality for a bivariate distribution

The last property might not be trivial for a n-dimensional $\boldsymbol{X}$, but it ensures that $P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n)$ is non-negative. A simple example of the rectangle property can be explained by Figure 2.1 which assumes a bivariate cdf, then visualizing a rectangle with vertices $(a_1, a_2), (b_1, a_2), (a_1, b_2)$ and (b_1, b_2) where $0 \leq a_1 \leq b_1 \leq 1$ and $0 \leq a_2 \leq b_2 \leq 1$, then

$$F(b_1, b_2) - F(a_1, b_2) - F(a_2, b_1) + F(a_1, a_2) \geq 0.$$

If one wishes to work with each random variable X_i separately, then we have the marginal pdf and cdf, $f_{X_i}(x)$ and $F_{X_i}(x)$, respectively. They are obtained as follows:

$$\begin{aligned} f_{X_i}(x) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f\left(x_1, \ldots, x_n\right) \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1} \mathrm{d}x_{i+1} \cdots \mathrm{d}x_n, \\ F_{X_i}(x) &= \lim_{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \to \infty} F(x_1, \ldots, x_n). \end{aligned}$$

In addition, the random variables $X_1, \ldots, X_n$ are independent if and only if

$$\begin{aligned} f(x_1, \ldots, x_n) &= f_{X_1}(x_1) \cdots f_{X_n}(x_n), \\ F(x_1, \ldots, x_n) &= F_{X_1}(x_1) \cdots F_{X_n}(x_n). \end{aligned} \tag{2.1.2}$$

The conditional pdf and cdf of X_i given the other variables $\boldsymbol{X}_{i^-}$, where $\boldsymbol{X}_{i^-}$ represents the random vector $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ without the random variable X_i, are given by

$$f_{X_i|\boldsymbol{X}_{i^-}}(x_i|\boldsymbol{x}_{i^-}) = \frac{f(x_1, \ldots, x_n)}{f_{X_i}(x_i)}, \tag{2.1.3}$$

$$F_{X_i|\boldsymbol{X}_{i^-}}(x_i|\boldsymbol{x}_{i^-}) = \frac{F(x_1, \ldots, x_n)}{F_{X_i}(x_i)}. \tag{2.1.4}$$

Additionally, the multivariate survival function $\bar{F}$ is defined by

$$\bar{F}(x_1, x_2, \dots, x_n) = P\left(X_1 > x_1, \dots, X_n > x_n\right).$$

2.2 Copulas

Any multivariate distribution function for a vector of random variables can implicitly describe the marginal distribution functions and their dependence structure. However, with the limited availability of known multivariate distribution functions and the complexity of modeling real data by using them, one is inclined to use copulas. In this section, we define copulas, explain their properties and identify their link with multivariate cdfs. We also provide examples of specific families of copulas. Joe (1997, 2014) and McNeil et al. (2015) provided detailed explanation for copulas and dependence modeling.

Copulas provide a mean to model the dependence relationship between two or more random variables. The n-dimensional copula is a multivariate cdf on $[0,1]^n$ with standard uniform marginal distributions. Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random vector which contains n random variables linked through the cdf F. Set $U_i = F_i(X_i) \sim U(0,1), i = 1, \dots, n$, where $F_i(X_i)$ is the marginal cdf of the random variable X_i. Hence, the copula C is a mapping of the form $C : [0,1]^d \to [0,1]$ and is defined by

$$C(u_1, \dots, u_n) = P\left(U_1 \leq u_1, \dots, U_n \leq u_n\right), \quad u_i \in [0,1], \quad i = 1, \dots, n. \tag{2.2.1}$$

The following properties must hold for any copula C:

- $C(u_1, \dots, u_n)$ is increasing in each of its components, i.e. if any of the u_i increases, then C also increases,

- $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i, \forall i \in 1, \dots, n$. This property holds due to the uniform marginals,

- $C(1, \dots, 1) = 1$, i.e. if all components reach their maximum attainable values, then the value of the copula C is 1,

- $C(u_1, \dots, u_{i-1}, 0, u_{i+1}, \dots, u_n) = 0$, i.e one or more components are at their minimum attainable values, then the value of the copula C is 0, and

- $\forall (a_1, \ldots, a_n), (b_1, \ldots, b_n) \in [0,1]^n$ where $a_i \leq b_i$ we have the rectangle inequality

$$\sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1+\ldots+i_d} C(u_{1i_1}, \ldots, u_{ni_n}) \geq 0,$$

  where $u_{j1} = a_j$ and $u_{j2} = b_j$ $\forall j \in 1, \ldots, n$.

The above properties, except the second one, are the same properties identified for any multivariate cdf as explained in Section 2.1.

Sklar (1959) defined the link between copulas and multivariate cdfs, but the following propositions must be defined first.

Proposition 2.2.1. *Let F be a distribution function and F^{-1} denote its inverse, i.e. $F^{-1}(y) = \inf\{x : F(x) \geq y\}$, then*

1. ***Quantile Transformation.*** *If $U \sim U(0,1)$, then $P\left(F^{-1}(U) \leq x\right) = F(x)$,*
2. ***Probability Transformation.*** *If $X \sim F$ where F is continuous, then $F(X) \sim U(0,1)$.*

This leads us to *Sklar's Theorem*, which proves that all multivariate cdfs can be written in terms of copulas and that copulas can be used with the marginal cdfs to obtain a multivariate cdf.

Theorem 2.2.2. ***Sklar's Theorem*** *Let F be a n-dimensional distribution function with margins $F_1, \ldots, F_n$. Then there exists a coupla $C : [0,1]^n \to [0,1]$ such that $\forall x_1, \ldots, x_n \in \mathbb{R}$*

$$F(x_1, \ldots, x_n) = C\left(F_1(x_1), \ldots, F_n(x_n)\right),$$

where $F_i(x_i)$ is the marginal distribution function of $X_i, \forall i \in 1, \ldots, n$. Conversely, if C is a copula and $F_i(x_i)$ are the marginal distribution function of $X_i, \forall i \in 1, \ldots, n$, then

$$C\left(F_1(x_1), \ldots, F_n(x_n)\right) = F(x_1, \ldots, x_n),$$

where F is a multivariate cdf with margins $F_1, \ldots, F_n$. Additionally, if the margins are continuous, then C is unique; otherwise, if one or more of the marginals is discrete, then C is unique only on Ran $F_1 \times \ldots \times$ Ran F_n, where Ran F_i denotes the range of F_i, and Ran $F_1 \times \ldots \times$ Ran F_n represents the cartesian product of the ranges.

Proof. We provide the proof for the continuous case. For the detailed proof, please refer to Nelsen (1999).

Consider the continuous random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ with a multivariate cdf F, which can be represented as

$$\begin{aligned} F\left(x_1, \ldots, x_n\right) &= P\left(X_1 \leq x_1, \ldots, X_n \leq x_n\right) \\ &= P\left(F_1(X_1) \leq F_1(x_1), \ldots, F_n(X_n) \leq F_n(x_n)\right). \end{aligned} \tag{2.2.2}$$

By using the Probability Transformation defined in Proposition 2.2.1, we have that $F_i(X_i) = U_i \sim U(0,1)$. Then Eq. 2.2.2 corresponds to the cdf of $(F_1(X_1), \ldots, F_1(X_1)) = (U_1, \ldots, U_n)$. We introduce a function C, called a copula, such that

$$P\left(F_1(X_1) \leq F_1(x_1), \ldots, F_n(X_n) \leq F_n(x_n)\right) = C(F(x_1), \ldots, F(x_n))$$

If F is evaluated at the arguments $x_i = F_i^{-1}(u_i), 0 \leq u_i \leq 1, \forall i = 1, \ldots, n$, then,

$$C(u_1, \ldots, u_n) = F\left(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)\right). \tag{2.2.3}$$

Since F is continuous, then Eq. 2.2.3 provides an explicit form for the copula in terms of the cdf F and its margins F_i, which proves it is unique.

Contrarily, assume that C is a copula and that $F_i, \forall i = 1, \ldots, n$ are the univariate cdfs, where $X_i = F_i^{-1}(U_i)$. Let $\boldsymbol{U} \sim C$, then

$$\begin{aligned} F\left(x_1, \ldots, x_n\right) &= P\left(X_1 \leq x_1, \ldots, X_n \leq x_n\right) \\ &= P\left(F_1^{-1}(U_1) \leq x_1, \ldots, F_n^{-1}(U_n) \leq x_n\right) \\ &= P\left(U_1 \leq F_1(x_1), \ldots, U_n \leq F_n(x_n)\right) \\ &= C\left(F_1(x_1), \ldots, F_n(x_n)\right) = F(x_1, \ldots, x_n). \end{aligned}$$

□

The pdf of the copula C can be calculated by using 2.1.1 as follows:

$$c(u_1, \ldots, u_n) = \frac{\partial^n}{\partial u_1 \ldots \partial u_n} C(u_1, \ldots, u_n).$$

Hence, the pdf of the random vector $\boldsymbol{X}$ is

$$\begin{aligned} f(x_1,\ldots,x_n) &= \frac{\partial^n F(x_1,\ldots,x_n)}{\partial x_1\cdots\partial x_n} \\ &= \frac{\partial^n C(F_1(x_1),\ldots,F_n(x_n))}{\partial x_1\cdots\partial x_n} \\ &= \frac{\partial^n C(F_1(x_1),\ldots,F_n(x_n))}{\partial F_1(x_1)\cdots\partial F_n(x_n)}\frac{\partial F_1(x_1)}{\partial x_1}\cdots\frac{\partial F_n(x_n)}{\partial x_n} \\ &= c(F_1(x_1),\ldots,F_n(x_n))f_1(x_1)\cdots f_n(x_n). \end{aligned} \tag{2.2.4}$$

Even though the n-dimensional copula is the general case, to avoid cumbersome notation, we will restrict our discussion to the bivariate random vector $\boldsymbol{X} = (X_1, X_2)$ with observations x_1, x_2, and its associated copula $C\left(F_1(X_1), F_2(X_2)\right)$. All the properties and discussions can be generalized to the n-dimensional copula.

The survival copula is defined by $\bar{C}\left(\bar{F}_1(x_1), \bar{F}_2(x_2)\right)$, where $\bar{F}_i(x_i)$ is the survival function of the random variable X_i, $\forall i \in 1, 2$. $\bar{C}$ is derived as follows:

$$\begin{aligned} \bar{C}\left(\bar{F}_1(x_1), \bar{F}_2(x_2)\right) &= \bar{F}(x_1, x_2) \\ &= 1 - F_1(x_1) - F_2(x_2) + F(x_1, x_2) \\ &= 1 - F_1(x_1) - F_2(x_2) + C(F_1(x_1), F_2(x_2)) \\ &= \bar{F}_1(x_1) + \bar{F}_2(x_2) - 1 + C(1 - \bar{F}_1(x_1), 1 - \bar{F}_2(x_2)). \end{aligned}$$

Therefore,

$$\bar{C}\left(u_1, u_2\right) = C(1 - u_1, 1 - u_2) + u_1 + u_2 - 1. \tag{2.2.5}$$

The conditional copula of U_2 given $U_1 = u_1$ is defined by

$$\begin{aligned} C_{U_2|U_1}(u_2|u_1) &= P(U_2 \leq u_2 \mid U_1 = u_1) \\ &= \lim_{h\to 0} P(U_2 \leq u_2 \mid u_1 \leq U_1 \leq u_1 + h) \\ &= \lim_{h\to 0} \frac{C(u_1 + h, u_2) - C(u_1, u_2)}{P(U_1 \leq u_1 + h) - P(U_1 \leq u_1)} \\ &= \lim_{h\to 0} \frac{C(u_1 + h, u_2) - C(u_1, u_2)}{h} \\ &= \frac{\partial}{\partial u_1} C(u_1, u_2). \end{aligned} \tag{2.2.6}$$

Similarly, $C_{U_1|U_2}(u_1|u_2) = \frac{\partial}{\partial u_2} C(u_1, u_2)$.

Example: FGM Copula

Assume a bivariate distribution $F(X_1, X_2)$, where $X_i \sim \text{Exp}(\beta_i)$ where $\beta_i > 0,\ \forall i = 1, 2$, such that

$$\begin{aligned} F(x_1, x_2) = {} & (1 - e^{-\beta_1 x_1})(1 - e^{-\beta_2 x_2}) \\ & + \theta(1 - e^{-\beta_1 x_1})(1 - e^{-\beta_2 x_2}) e^{-\beta_1 x_1} e^{-\beta_2 x_2}, \end{aligned}$$

with dependence parameter $-1 \leq \theta \leq 1$, to be discussed later in Section 2.3. The corresponding copula C can be obtained by finding the inverse of $F_i(x_i) = 1 - e^{-\beta_i x_i} = u_i$, which is $F_i^{-1}(u_i) = -\frac{1}{\beta_i} \ln(1 - u_i) = x_i$. By replacing each x_i with the inverse in the above bivariate distribution, we obtain

$$C(u_1, u_2) = u_1 u_2 + \theta u_1 u_2 (1 - u_1)(1 - u_2). \tag{2.2.7}$$

The pdf of the copula C is

$$\begin{aligned} c(u_1, u_2) &= \frac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2) \\ &= \frac{\partial^2}{\partial u_1 \partial u_2} u_1 u_2 + \theta u_1 u_2 (1 - u_1)(1 - u_2) \\ &= 1 + \theta(1 - 2u_1)(1 - 2u_2). \end{aligned} \tag{2.2.8}$$

The joint density function is

$$\begin{aligned} f(x_1, x_2) &= c(F_1(x_1), F_2(x_2)) f_1(x_1) f_2(x_2) \\ &= (1 + \theta(1 - 2F_1(x_1))(1 - 2F_2(x_2)))\, \beta_1 e^{-\beta_1 x_1} \beta_2 e^{-\beta_2 x_2} \\ &= \Big(1 + \theta(1 - 2e^{-\beta_1 x_1})(1 - 2e^{-\beta_2 x_2})\Big)\, \beta_1 e^{-\beta_1 x_1} \beta_2 e^{-\beta_2 x_2}. \end{aligned}$$

The survival copula is

$$\begin{aligned} \bar{C}\,(u_1, u_2) &= C(1 - u_1, 1 - u_2) + u_1 + u_2 - 1 \\ &= (1 - u_1)(1 - u_2) + \theta(1 - u_1)(1 - u_2) u_1 u_2 + u_1 + u_2 - 1 \\ &= u_1 u_2 + \theta u_1 u_2 (1 - u_1)(1 - u_2). \end{aligned} \tag{2.2.9}$$

Note that for the FGM, the survival copula $\bar{C}$ is equivalent to the copula C, which makes it a symmetric copula.

The conditional copula of U_2 given $U_1 = u_1$ is

$$\begin{aligned} C_{U_2|U_1}(u_2|u_1) &= \frac{\partial}{\partial u_1} C(u_1, u_2) \\ &= \frac{\partial}{\partial u_1} \left[u_1 u_2 + \theta u_1 u_2 (1 - u_1)(1 - u_2) \right] \\ &= u_2 + \theta u_2 (1 - 2u_2)(1 - u_2). \end{aligned}$$

□

Furthermore, the copula C is bounded as per the following theorem:

Theorem 2.2.3. ***Fréchet-Hoeffding copula bounds*** *For any bivariate copula C and $\mathbf{u} = \{u_1, u_2\} \in [0,1]^2$, we have the following bounds*

$$W(u_1, u_2) \leq C(u_1, u_2) \leq M(u_1, u_2), \tag{2.2.10}$$

where $W(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$ and $M(u_1, u_2) = \min(u_1, u_2)$.

Proof. Upper bound: If C is the cdf of (U_1, U_2), then $C(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2)$. Given that

$$P(U_1 \leq u_1, U_2 \leq u_2) \leq P(U_1 \leq u_1) \text{ and } P(U_1 \leq u_1, U_2 \leq u_2) \leq P(U_2 \leq u_2),$$

then

$$\begin{aligned} P(U_1 \leq u_1, U_2 \leq u_2) &\leq \min(P(U_1 \leq u_1), P(U_2 \leq u_2)) \\ &\leq \min(u_1, u_2). \end{aligned}$$

Lower bound:

$$\begin{aligned} P(U_1 > u_1, U_2 > u_2) &= 1 - P(U_1 \leq u_1) - P(U_2 \leq u_2) + P(U_1 \leq u_1, U_2 \leq u_2) \\ &= 1 - u_1 - u_2 + C(u_1, u_2) \geq 0. \end{aligned}$$

Therefore, by rearranging the above inequality, $C(u_1, u_2) \geq u_1 + u_2 - 1$. □

Note that $M(u_1, \ldots, u_n)$ is a copula for any value of n, however, $W(u_1, \ldots, u_n)$ is only a copula for $n = 2$. M and W are referred to as the comonotonic and countermonotonic copulas, respectively. They will be discussed further in Section 2.2.1.

The empirical estimator of a copula is defined as

$$C_n(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{F_{n1}(X_{i1}) \leq u_1, F_{n2}(X_{i2}) \leq u_2\}}, \tag{2.2.11}$$

where F_{nj} is the empirical cdf of $X_j = (x_{1j}, \ldots, x_{nj})$, such that

$$F_{nj}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x_{ij} \leq t\}}.$$

2.2.1 Families of Copula

In this section, we explore different families of copulas and their properties. The most common families of copulas are

- Perfect dependence and independence copulas,
- Elliptical copulas,
- Archimedean copulas, and
- Extreme-value copulas.

Perfect Dependence and Independence Copulas

Independence Copula

The random variables X_1 and X_2 are independent if and only if

$$C(u_1, u_2) = u_1 u_2.$$

This is equivalent to Eq. 2.1.2. The independence copula has the following notation: $\Pi(u_1, u_2)$.

Comonotonicity Copula

$X_1 = \phi(X_2)$ almost surely (a.s.) for an increasing function $\phi(\cdot)$ if and only if

$$C(u_1, u_2) = \min(u_1, u_2).$$

The comonotonicity copula is the upper Fréchet-Hoeffding copula presented in Theorem 2.2.3. X_1 and X_2 are perfectly positively dependent because they are a.s. strictly increasing functions of each other.

Countermonotonicity Copula

$X_1 = \psi(X_2)$ almost surely (a.s.) for a decreasing function $\psi(\cdot)$, if and only if

$$C(u_1, u_2) = \max(u_1 + u_2 - 1, 0).$$

The countermonotonicity copula is the lower Fréchet-Hoeffding copula presented in Theorem 2.2.3. X_1 and X_2 are perfectly negatively dependent because they are a.s. strictly decreasing functions of each other.

Figure 2.2 illustrates the perspective plots of the cdfs of the dependence copulas and the independence copula. The Fréchet-Hoeffding bounds presented in Theorem 2.2.3 imply that the cdf of all bivariate copulas lie between the surfaces of the Countermonotonicity and Comonotonicity copulas.

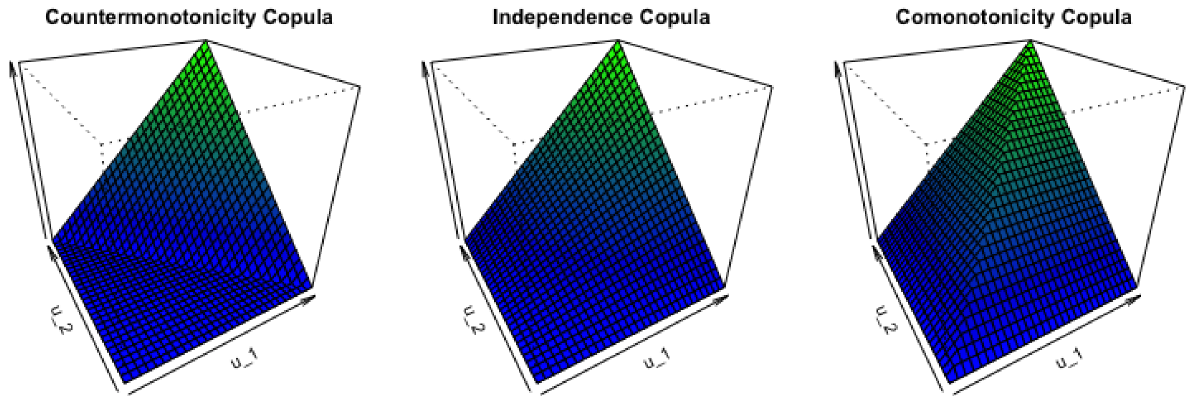

Figure 2.2: Perspective plots of the cdf of the Countermonotonicity copula, Independence copula and Comonotonicity copula.

Elliptical Copulas

An elliptical copula is a generalization of the multivariate Gaussian distribution. They do not have a closed form expression, but they are extracted from the multivariate cdfs by using Sklar's Theorem 2.2.2.

Gauss Copula

If $\boldsymbol{X} = (X_1, X_2)$ follows a standardized bivariate Gaussian distribution, then

$$C_{\rho}^{Gauss}(u_1, u_2) = \Phi_{\rho}\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right),$$

where Φ is the cdf of a standard univariate normal random variable and Φ_{ρ} is the cdf of a bivariate normal random variable with mean 0 and correlation $\rho \in [-1, 1]$. The Gauss copula does not have an explicit form, but it can be represented as the integral over the pdf of $\boldsymbol{X}$ as follows:

$$C_{\rho}^{Gauss}(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2 + y^2 - 2\rho xy}{2(1-\rho^2)}\right\} \mathrm{d}y\mathrm{d}x.$$

Student's t-Copula

If $\boldsymbol{X} = (X_1, X_2)$ follows a bivariate Student's t-distribution, then

$$C_{\nu,\rho}^{t}(u_1, u_2) = \boldsymbol{t}_{\nu,\rho}\left(t_{\nu}^{-1}(u_1), t_{\nu}^{-1}(u_2)\right),$$

where t_{ν} is the cdf of a standard univariate t-distribution with ν degrees of freedom, $\boldsymbol{t}_{\nu,\rho}$ is the cdf of a bivariate t-distribution with mean 0, correlation $\rho \in [0, 1]$ and ν degrees of freedom. ν determines the thickness of the tails of the t-distribution; the more the degrees of freedom, the lighter the tails, and vice verse. The t-copula does not have an explicit form, but it can be represented as the integral over the pdf of $\boldsymbol{X}$ as follows:

$$C_{\nu,\rho}^{t}(u_1, u_2) = \int_{-\infty}^{t_{\nu}^{-1}(u_1)} \int_{-\infty}^{t_{\nu}^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left\{1 + \frac{x^2 + y^2 - 2\rho xy}{\nu(1-\rho^2)}\right\} \mathrm{d}y\mathrm{d}x.$$

We can observe from Figure 2.3 that the t-copula assigns more probability mass to the corners of the unit square, which means they have heavier tails than the Gauss copula. This characteristic for the t-copula can be altered by changing the degrees of freedom ν. Note that if we had assumed no correlation, i.e. $\rho = 0$, then this would result in independence, and hence, there will be no higher mass in the corners.

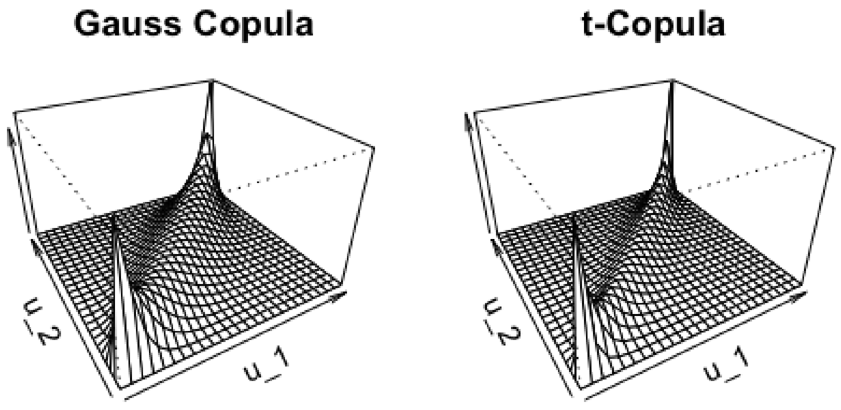

Figure 2.3: Perspective plots of the densities of the bivariate Gauss copula with $\rho = 0.9238795$ and bivariate t-copula with $\rho = 0.9238795$ and $\nu = 2$.

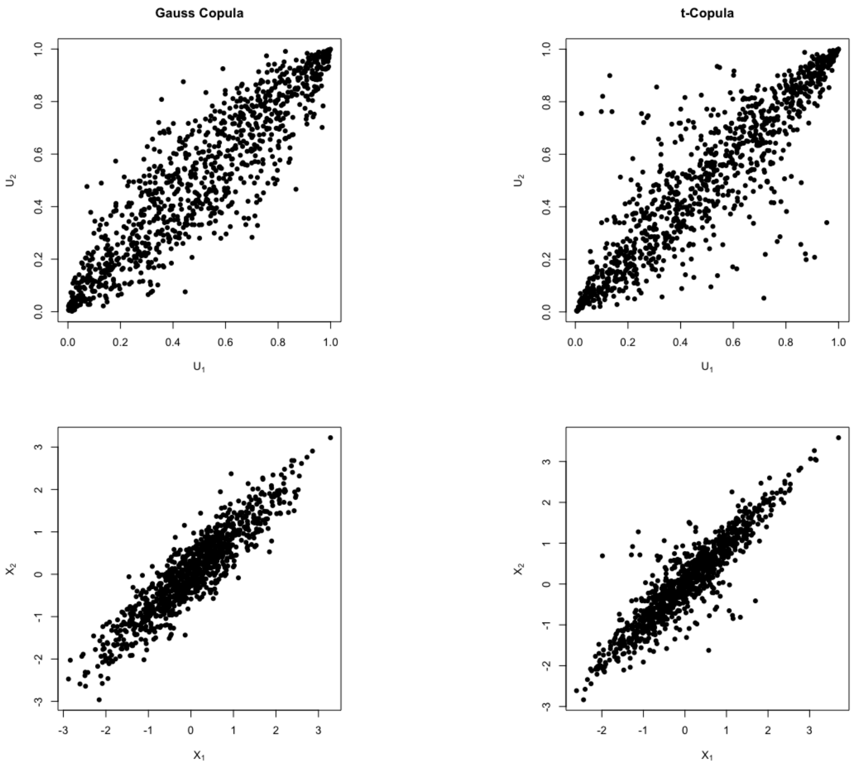

Figure 2.4: Top: One thousand simulated points from the Gaussian copula with $\rho = 0.9238795$ and bivariate t-copula with $\rho = 0.9238795$ and $\nu = 2$.

Bottom: Realizations of X_1 and X_2 by assuming standard normal marginals for the copulas presented on the top row.

The top row of Figure 2.4 represents 1000 simulated points from the Gauss and t-copulas. For the bottom row, we assume that (X_1, X_2) has standard normal marginals, so each simulated point from the copula is transformed component-wise into standard normal. The parameters are chosen such that both copulas have the same value of Kendall's tau, to be discussed in Section 2.3. Other elliptical copulas include the Cauchy copula and the Pearson Type II copula.

Archimedean Copulas

Unlike the elliptical copulas defined in Section 2.2.1, Archimedean copulas have closed forms. In this section, we define bivariate Archimedean copulas and provide examples for it. For multivariate Archimedean copulas, refer to McNeil and Nešlehová (2009) and McNeil et al. (2015).

A bivariate Archimedean copula has the form

$$C_\theta(u_1, u_2) = \phi^{-1}\left[\phi(u_1;\theta) + \phi(u_2;\theta)\ ;\ \theta\right],\ (u_1, u_2) \in [0,1]^2, \theta \in \Theta, \tag{2.2.12}$$

where $\phi : [0,1] \times \Theta \to \mathbb{R}_+$ is a strictly decreasing convex function with dependence parameter θ. The function ϕ is called the generator function of the copula, and its inverse is represented by ϕ^{-1}. In addition, $\phi(0) = \infty$ and $\phi(1) = 0$. Table 2.1 summarizes the generator functions and other details for the most commonly used Archimedean copulas.

Bivariate Clayton Copula

Consider the generator $\phi(t;\theta) = \frac{1}{\theta}\left(t^{-\theta} - 1\right)$, where $\theta \geq -1$ and $t \in [0,1]$. The inverse of the generator is represented by $\phi^{-1}(t;\theta) = (1 + \theta t)^{-1/\theta}$. By using the general form of the Archimedean copula represented in Eq. 2.2.12, we obtain the bivariate Clayton copula

$$C_\theta^{Cl} = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta},\ \theta \geq -1.$$

The bivariate Clayton copula is characterized by the following limiting cases:

- Countermonotonicity copula when $\theta = -1$ (only in the bivariate case),
- Independence copula when $\theta \to 0$, and
- Comonotonicity copula when $\theta \to \infty$.

Bivariate Frank Copula

Consider the generator $\phi(t;\theta) = -\ln\left(\frac{\mathrm{e}^{-\theta t}-1}{\mathrm{e}^{-\theta}-1}\right)$, where $\theta \in \mathbb{R}$ and $t \in [0,1]$. The inverse of the generator is represented by $\phi^{-1}(t;\theta) = -\frac{1}{\theta}\ln\left[1+\mathrm{e}^{-t}\left(\mathrm{e}^{-\theta}-1\right)\right]$. By using the general form of the Archimedean copula represented in Eq. 2.2.12, we obtain the bivariate Frank copula as follows:

$$C_\theta^{Fr} = -\frac{1}{\theta}\ln\left[1+\frac{\left(\mathrm{e}^{-\theta u_1}-1\right)\left(\mathrm{e}^{-\theta u_2}-1\right)}{\mathrm{e}^{-\theta}-1}\right].$$

The bivariate Frank copula is characterized by the following limiting cases

- Countermonotonicity copula when $\theta \to -\infty$ (only in the bivariate case),
- Independence copula when $\theta \to 0$, and
- Comonotonicity copula when $\theta \to \infty$.

Bivariate Gumbel Copula

Consider the generator $\phi(t;\theta) = (-\ln t)^\theta$, where $\theta \geq 1$ and $t \in [0,1]$. The inverse of the generator is represented by $\phi^{-1}(t;\theta) = \mathrm{e}^{-t^{1/\theta}}$. By using the general form of the Archimedean copula represented in Eq. 2.2.12, we obtain the bivariate Gumbel copula as follows:

$$C_\theta^{Gu} = \exp\left\{-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta\right]^{1/\theta}\right\}.$$

The bivariate Gumbel copula is characterized by the following limiting cases

- Independence copula when $\theta = 1$, and
- Comonotonicity copula when $\theta \to \infty$.

Figures 2.5 and 2.6 provides an example of each of the Archimdean copulas discussed previously. We can observe that the Clayton copula and the Gumbel copula provide strong lower and upper tail dependence, respectively. However, the Frank copula provides symmetry along both tails.

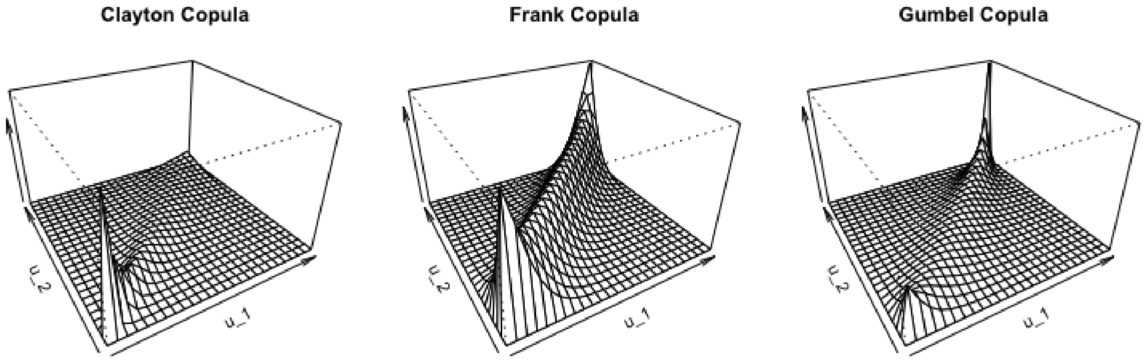

Figure 2.5: Perspective plots of the densities of the bivariate Clayton copula, bivariate Frank copula, and bivariate Gumbel copula. The dependence parameter for each copula is $\theta = 6, 14.1385$ and 4, respectively.

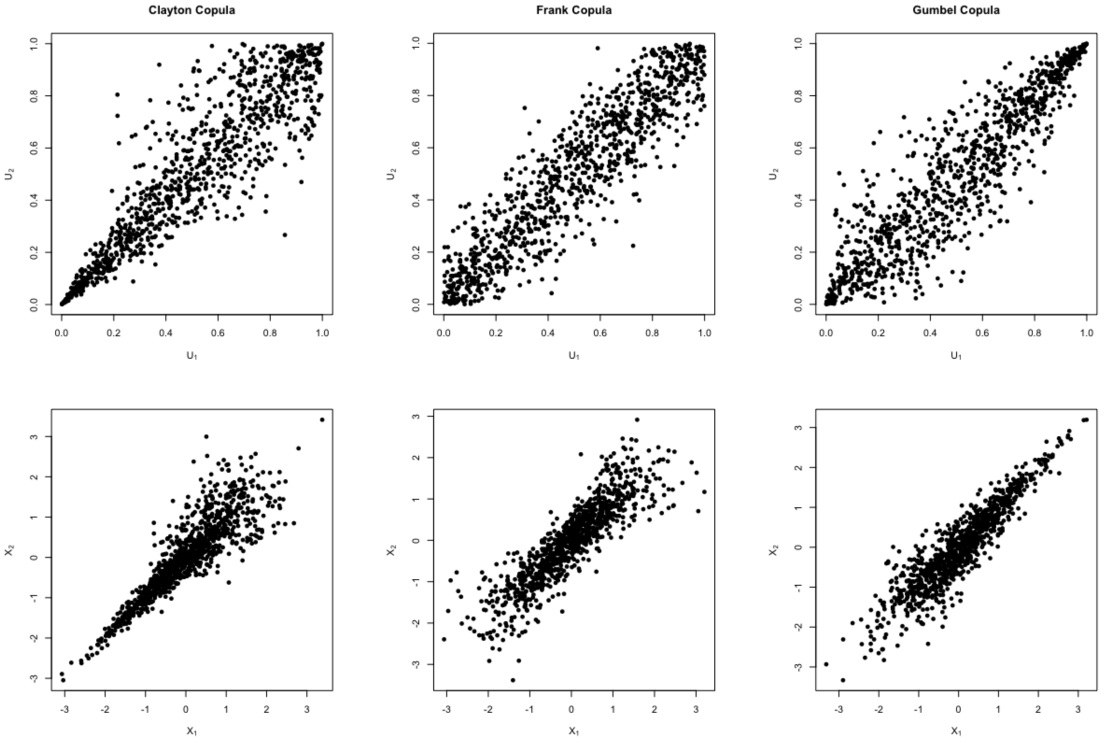

Figure 2.6: Top: One thousand simulated points from the Clayton, Frank and Gumbel copulas with dependence parameter for each copula is $\theta = 6, 14.1385$ and 4, respectively.

Bottom: Realizations of X_1 and X_2 by assuming standard normal marginals for the copulas presented on the top row.

The Archimedean family offers a great deal of flexibility, however they have some limitations that prevent them from modeling asymmetric dependence relationships. Those limitations are

- C is symmetric, i.e. $C(u_1, u_2) = C(u_2, u_1),\ \forall (u_1, u_2) \in [0,1]^2$, and
- C is associative, i.e. $C(C(u_1, u_2), u_3) = C(u_1, C(u_2, u_3)),\ \forall (u_1, u_2, u_3) \in [0,1]^3$.

Table 2.1: Summary of the generators $\phi(t)$, where $t \in [0,1]$, the possible values for the dependence parameter θ, and the limiting cases for some bivariate Archimedean copulas.

Copula	$\phi(t)$	θ	Lower Limit	Upper Limit
$C_\theta^{Cl} = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}$	$\frac{1}{\theta}\left(t^{-\theta} - 1\right)$	$\theta \geq -1$	$W(u_1, u_2)$	$M(u_1, u_2)$
$C_\theta^{Fr} = -\frac{1}{\theta} \ln \left[1 + \frac{\left(\mathrm{e}^{-\theta u_1} - 1\right)\left(\mathrm{e}^{-\theta u_2} - 1\right)}{\mathrm{e}^{-\theta} - 1}\right]$	$-\ln\left(\frac{\mathrm{e}^{-\theta t} - 1}{\mathrm{e}^{-\theta} - 1}\right)$	$\theta \in \mathbb{R}$	$W(u_1, u_2)$	$M(u_1, u_2)$
$C_\theta^{Gu} = \mathrm{e}^{-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta\right]^{1/\theta}}$	$(-\ln t)^\theta$	$\theta \geq 1$	$\Pi(u_1, u_2)$	$M(u_1, u_2)$

Extreme-value Copulas

Rare events need careful modeling because they might have a serious impact on the dependence structure of the distribution. This gives importance to extreme-value copulas. Gudendorf and Segers (2010) provided detailed explanation on the origin and properties of those copulas. Note that extreme-value copulas have complicated forms for dimensions > 2.

Bivariate extreme-value copulas have the form

$$C_A(u_1, u_2) = \exp\left\{\ln(u_1 u_2) A \left[\frac{\ln(u_2)}{\ln(u_1 u_2)}\right]\right\},$$

where $A : [0,1] \to [1/2, 1]$ is a convex mapping such that

$$\max(t, 1-t) \leq A(t) \leq 1,\ \ t \in [0,1].$$

Gumbel's First Asymmetric Model

This is a generalization of the Gumbel Copula from the Archimedean family presented in Section 2.2.1. For this copula, $A(t)$ is given by

$$A(t) = (1-\alpha)t + (1-\beta)(1-t) + \left[(\alpha t)^\theta + (\beta(1-t))^\theta\right]^{1/\theta},\ \ \theta \geq 1, \alpha, \beta \in [0,1].$$

This copula is defined as

$$C_\theta(u_1,u_2) = u_1^{1-\beta} u_2^{1-\alpha} \exp\left\{ - \left[(-\beta \ln u_1)^\theta + (-\alpha \ln u_2)^\theta \right]^{1/\theta} \right\}.$$

Note that if $\alpha = \beta = 1$, we obtain the Gumbel Copula.

Gumbel's Second Model

For this copula, $A(t)$ is given by

$$A(t) = \theta t^2 - \theta t + 1, \ \theta \in [0,1].$$

This copula is defined as

$$C_\theta(u_1,u_2) = u_1 u_2 \exp\left\{ \frac{\ln(u_1)\ln(u_2)}{\ln(u_1)+\ln(u_2)} \right\}.$$

Galambos Asymmetric Copula

For this copula, $A(t)$ is given by

$$A(t) = 1 - \left[(\alpha t)^{-\theta} + (\beta(1-t))^{-\theta} \right]^{-1/\theta}, \ \theta \in [0,\infty)\alpha, \beta \in [0,1].$$

This copula is defined as

$$C_\theta(u_1,u_2) = u_1 u_2 \exp\left\{ - \left[(-\beta \ln u_1)^{-\theta} + (-\alpha \ln u_2)^{-\theta} \right]^{-1/\theta} \right\}.$$

2.2.2 Vine Copula

Vines were originally introduced by Bedford and Cooke (2002) as a graphical model for high dimensional distributions that have conditional dependence. Assuming we have 3 random variables (X_1, X_2, X_3), Figure 2.7 illustrates that $f(x_1|x_2)$ and $f(x_3|x_2)$ are dependent with a conditional correlation coefficient that depends on the value of x_2.

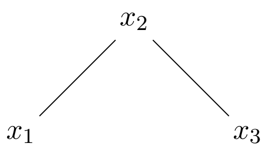

Figure 2.7: A basic vine structure.

Aas et al. (2009) utilized vines and copulas to model high dimensional data using pair-copula by working on two variables at a time. Constructing a vine copula starts with decomposing the joint multivariate distribution function into simple bivariate building blocks, and then combining them together appropriately. This method is a recursive method called pair-copula construction. Each bivariate building block is a two-dimensional copula. In this section, we will go through the process of constructing pair-copulas.

Assume we have a vector of n random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ with a joint distribution function $f(x_1, \ldots, x_n)$. By using 2.1.3 iteratively, $f(x_1, \ldots, x_n)$ can be represented as

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdot f(x_2|x_1) \cdots f(x_n|x_1, \ldots, x_{n-1}). \tag{2.2.13}$$

Let $c_{1,\ldots,n}(\cdot)$ be a copula density for n random variables, and recall from Eq. 2.2.4 that a copula density is represented by

$$f(x_1, \ldots, x_n) = c_{1,\ldots,n}(F_1(x_1), \ldots, F_n(x_n)) f_1(x_1) \cdots f_n(x_n).$$

Example: Bivariate Distribution

If we assume we only have 2 random variables $\boldsymbol{X} = (X_1, X_2)$, then Eq. 2.2.4 can be simplified to

$$f(x_1, x_2) = c_{1,2}\left\{F_1(x_1), F_2(x_2)\right\} f_1(x_1) f_2(x_2),$$

where $c_{1,2}(\cdot, \cdot)$ is the pair-copula density for the pair of transformed random variables $F_1(X_1)$ and $F_2(X_2)$. The conditional density can be represented by using 2.1.3 as follows:

$$\begin{aligned} f_{2|1}(x_2|x_1) &= \frac{f(x_1, x_2)}{f_1(x_1)} \\ &= \frac{c_{1,2}\left\{F_1(x_1), F_2(x_2)\right\} f_1(x_1) f_2(x_2)}{f_1(x_1)} \\ &= c_{1,2}\left\{F_1(x_1), F_2(x_2)\right\} f_2(x_2). \end{aligned}$$

□

Therefore, for any two random variables X_i and X_j, where $i \neq j$ and $i, j = 1, \ldots, n$,

$$f(x_j|x_i) = c_{i,j}\left\{F_i(x_i), F_j(x_j)\right\} f_j(x_j). \tag{2.2.14}$$

Now we will build the 3-dimensional density function.

Example: Trivariate Distribution

By assuming 3 random variables, i.e. $\boldsymbol{X} = (X_1, X_2, X_3)$, the conditional distribution of one variable given the other two can be represented as follows:

$$\begin{aligned} f(x_3|x_1,x_2) &= \frac{f(x_3,x_2,x_1)}{f(x_2,x_1)} \\ &= \frac{f(x_3,x_2,x_1)/f_1(x_1)}{f(x_2,x_1)/f_1(x_1)} \\ &= \frac{f(x_3,x_2|x_1)}{f(x_2|x_1)} \\ &= \frac{c_{(3,2)|1}(F(x_3|x_1), F(x_2|x_1))f(x_3|x_1)f(x_2|x_1)}{f(x_2|x_1)} \\ &= c_{(3,2)|1}(F(x_3|x_1), F(x_2|x_1))f(x_3|x_1), \end{aligned} \tag{2.2.15}$$

where $c_{(3,2)|1}(\cdot,\cdot)$ is the pair-copula density for the pair of transformed random variables $F(X_3|X_1)$ and $F(X_2|X_1)$. Alternatively, $f(x_3|x_1,x_2)$ can also be represented as follows:

$$f(x_3|x_1,x_2) = c_{(3,1)|2}(F(x_3|x_2), F(x_1|x_2))f(x_3|x_2), \tag{2.2.16}$$

where $c_{(3,1)|2}(\cdot,\cdot)$ is the pair-copula density for the pair of transformed random variables $F(X_3|X_2)$ and $F(X_1|X_2)$, and it is different from $c_{(3,2)|1}(\cdot,\cdot)$ in 2.2.15. By using 2.2.14, we can rewrite Eq. 2.2.15 and Eq. 2.2.16 as follows:

$$\begin{aligned} f(x_3|x_1,x_2) &= c_{(3,2)|1}(F(x_3|x_1), F(x_2|x_1))c_{1,3}\left\{F_1(x_1), F_3(x_3)\right\} f_3(x_3) \\ f(x_3|x_1,x_2) &= c_{(3,1)|2}(F(x_3|x_2), F(x_1|x_2))c_{2,3}\left\{F_2(x_2), F_3(x_3)\right\} f_3(x_3), \end{aligned}$$

respectively. □

By generalizing, each term in Eq. 2.2.13 can be represented in terms of a pair-copula and a marginal density function as follows:

$$f(x|\boldsymbol{v}) = c_{(x,v_j)|\boldsymbol{v}_{-j}}\left\{F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j})\right\} f(x|\boldsymbol{v}_{-j}), \forall j = 1, \ldots, d, \tag{2.2.17}$$

where $\boldsymbol{v}$ is a d-dimensional vector of random variables, v_j is one variable chosen from $\boldsymbol{v}$ and $\boldsymbol{v}_{-j}$ is $\boldsymbol{v}$ excluding v_j. Joe (1996) showed that marginal conditional distributions of the copulas can be generalized as follows:

$$F(x|\boldsymbol{v}_j) = \frac{\partial C_{(x,v_j)|\boldsymbol{v}_{-j}}\left\{F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j})\right\}}{\partial F(v_j|\boldsymbol{v}_{-j})}, \forall j = 1, \ldots, d. \tag{2.2.18}$$

For the special case where $\boldsymbol{v}$ is univariate, we have that

$$F(x|v) = \frac{\partial C_{(x,v)}\left\{F(x), F(v)\right\}}{\partial F(v)},$$

which is the bivariate conditional cdf derived in 2.2.6.

Example: Trivariate Distribution continued

By returning to the trivariate distribution example, we have that

$$\begin{aligned}
f(x_1, x_2, x_3) =& [f_1(x_1)]\,[f(x_2|x_1)]\,[f(x_3|x_1, x_2)] \\
=& [f_1(x_1)]\,[c_{1,2}\left\{F_1(x_1), F_2(x_2)\right\} f_2(x_2)]\,\left[c_{(3,2)|1}(F(x_3|x_1), F(x_2|x_1)) f(x_3|x_1)\right] \\
=& [f_1(x_1)]\,[c_{1,2}\left\{F_1(x_1), F_2(x_2)\right\} f_2(x_2)] \\
& \times \left[c_{(3,2)|1}(F(x_3|x_1), F(x_2|x_1)) c_{1,3}\left\{F_1(x_1), F_3(x_3)\right\} f_3(x_3)\right] \\
=& f_1(x_1) f_2(x_2) f_3(x_3) \\
& \times c_{1,2}\left\{F_1(x_1), F_2(x_2)\right\} c_{1,3}\left\{F_1(x_1), F_3(x_3)\right\} \\
& \times c_{(3,2)|1}\left\{F(x_3|x_1), F(x_2|x_1)\right\}.
\end{aligned} \tag{2.2.19}$$

If we assume conditional independence, we will be able to reduce the levels in the pair-copula decomposition. For example, if we assume that X_3 and X_2 are independent, given X_1, then $c_{(3,2)|1}\left\{F(x_3|x_1), F(x_2|x_1)\right\} = 1$, which simplifies Eq. 2.2.19 into

$$\begin{aligned}
f(x_1, x_2, x_3) =& f_1(x_1) f_2(x_2) f_3(x_3) \\
& \times c_{1,2}\left\{F_1(x_1), F_2(x_2)\right\} c_{1,3}\left\{F_1(x_1), F_3(x_3)\right\}.
\end{aligned}$$

□

Hence, a multivariate pdf can be iteratively expressed in terms of pair-copulas and conditional probability distributions, by using Eq. 2.2.17 and Eq. 2.2.18. This can lead to a significant number of possible pair-copulas decompositions for high dimensional distributions. Bedford and Cooke (2001, 2002) have introduced the *regular vine* and the most commonly used from that class are the *canonical vine* and the *D-vine*. A vine copula is the multivariate distribution, and its component bivariate copulas are pair-copulas.

Smith et al. (2010) used vine copulas to model the dependence structure for longitudinal data, where one or more variable of interest is collected over a given period of time. Assuming a univariate longitudinal data $\boldsymbol{X} = (X_1, \ldots, X_t)$ of a continuously distributed data observed at different time points. Therefore, the density function of x_t given all the previous data points can be represented using Eq. 2.2.17 as follows:

$$f(x_t|x_1, \ldots, x_{t-1}) = c_{(t,1)|2,\ldots,(t-1)} \left\{F(x_t|x_2, \ldots, x_{t-1}), F(x_1|x_2, \ldots, x_{t-1})\right\} f(x_t|x_2, \ldots, x_{t-1})$$

By repeatedly applying Eq. 2.2.17, we obtain the following

$$\begin{aligned} f(x_t|x_1, \ldots, x_{t-1}) &= \prod_{j=1}^{t-2} c_{(t,j)|(j+1),\ldots,(t-1)} \left\{F(x_t|x_{j+1}, \ldots, x_{t-1}), F(x_j|x_{j+1}, \ldots, x_{t-1})\right\} f(x_t|x_{t-1}) \\ &= \prod_{j=1}^{t-2} c_{(t,j)|(j+1),\ldots,(t-1)} \left\{F(x_t|x_{j+1}, \ldots, x_{t-1}), F(x_j|x_{j+1}, \ldots, x_{t-1})\right\} \\ &\quad \times c_{t,(t-1)} \left\{F_t(x_t), F_{t-1}(x_{t-1})\right\} f_t(x_t). \end{aligned}$$

2.2.3 Nested Archimedean Copulas

The methodology of nested copulas has been suggested by Joe (1997), and also used in insurance for reserving purposes by Abdallah et al. (2015) and Côté et al. (2016), but not in the field of biostatistics, as far as our knowledge. As shown in Section 2.2.1, the Archimedean copulas have closed forms, and hence Hofert et al. (2011) and Hofert and Pham (2013) showed theoretical properties of the nested Archimedean copula.

As shown in Eq. 2.2.12, a bivariate Archimedean copula with generator ϕ_1 is given by

$$C^{(1)}(u_1, u_2) = \phi_1^{-1} \left[\phi_1(u_1) + \phi_1(u_2)\right], \ (u_1, u_2) \in [0, 1]^2.$$

A n-dimensional copula $C^{(n-1)}$ is called fully nested Archimedean copula with generators $\phi_1, \ldots, \phi_{n-1}$ if it is defined recursively $\forall (u_1, \ldots, u_n) \in [0,1]^n$ as follows:

$$\begin{aligned} C^{(2)}(u_1, u_2, u_3) &= \phi_2^{-1}\left[\phi_2\left\{C^{(1)}(u_1, u_2)\right\} + \phi_2(u_3)\right] \\ \vdots \quad &= \quad \vdots \\ C^{(n-1)}(u_1, \ldots, u_n) &= \phi_{n-1}^{-1}\left[\phi_{n-1}\left\{C^{(n-2)}(u_1, \ldots, u_{n-1})\right\} + \phi_{n-1}(u_n)\right]. \end{aligned}$$

McNeil (2008) showed that this a copula if only if all the generators $\phi_1, \ldots, \phi_n$ are completely monotonic, and the derivative of the composite function $\phi_k \circ \phi_{k-1}^{-1}$ are completely monotonic $\forall k = 2, \ldots, (n-1)$. The copulas can be from different families in the Archimedean family, and they have different dependence parameters. The estimates of the parameters of each copula are obtained sequentially, starting from $C^{(1)}$. The order of which the variables are chosen into the nested structure depends on the strength of the dependence. Let V_1 and V_2 represent the couple of variables that have the strongest dependence. Then they are chosen for $C^{(1)}$ and we define a new pseudo variable $C^{(1)}\{v_1, v_2; \phi_1\}$. We then proceed by considering the remaining variables and the new pseudo variable and choose the couple with the strongest dependence. This process is iterated $(n-1)$ times. Note that the fitting procedure does not require the use of Archimedean copulas, and it can be generalized to any copula family.

2.2.4 Copula Regression

In the concept of regression, each marginal distribution can be a conditional distribution on a vector of covariates. Assume outcome variables Y_1 and Y_2, such that they depend on the random vector of covariates $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. The marginal distributions of the copula are $F_1(Y_1|\boldsymbol{X}_1; \beta_1)$ and $F_2(Y_2|\boldsymbol{X}_2; \beta_2)$, where $\beta_i, i = 1, 2$, represents the vector of fixed regression coefficients. Note that the set of covariates $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can be the same or subsets of each other, and it is assumed that Y_i is independent of $X_j, \forall i, j = 1, 2, i \neq j$. Let $C_\theta(\cdot, \cdot)$ be a copula with parameter θ that captures the degree of dependence between the marginals as follows:

$$F(Y_1, Y_2|\boldsymbol{X}_1, \boldsymbol{X}_2; \beta_1, \beta_2) = C_\theta\left\{F_1(Y_1|\boldsymbol{X}_1; \beta_1), F_2(Y_2|\boldsymbol{X}_2; \beta_2)\right\}.$$

By using Eq. 2.2.4, we can write the joint density function as follows:

$$f(Y_1, Y_2|\boldsymbol{X}_1, \boldsymbol{X}_2; \beta_1, \beta_2) = C_\theta \left\{F_1(Y_1|\boldsymbol{X}_1; \beta_1), F_2(Y_2|\boldsymbol{X}_2; \beta_2)\right\}$$
$$\times f_1(Y_1|\boldsymbol{X}_1; \beta_1) f_1(Y_2|\boldsymbol{X}_2; \beta_2).$$

Using copulas to model GLM marginals has been proposed in the literature by Meester and Mackay (1994) and performed on biostatistics studies by Lambert (1996); Lambert and Vandenhende (2002). Frees and Wang (2005, 2006) were the first to perform similar modeling for insurance claims by using GLMs and copulas. The main advantage of using copulas is that there are no restrictions on the probability distributions or dependence structure used in the model.

2.3 Measures of Dependence

Assume that the random variables X_1 and X_2 are not independent, i.e. Eq. 2.1.2 is not satisfied;

$$F(x_1, x_2) \neq F_{X_1}(x_1) F_{X_2}(x_2),$$

then there are several ways to measure the degree of dependence between the two random variables. In this section, we will explain three methods for measuring dependence, which are:

- Linear correlation: Pearson's correlation coefficient ρ_p,
- Rank correlation: Spearman's rho ρ_S, and
- Rank correlation: Kendall's tau τ.

Each of those measures provide a scalar value for the dependence between X_1 and X_2, however, they differ in their properties and interpretation. The last two are *copula-based* measures, which are used in the parametrization of copula models. Linear correlation depends on the marginal distributions and the joint distribution, however, rank correlations are based on the copula. They are called rank correlation because the empirical estimators are calculated by using the ordering of the data (ranks) for each variable.

Before we explain the above mentioned measures of dependence, we should first define two important terms; comonotonicity and countermonotonicity. The random variables X_1 and X_2 are said to be comonotonic if and only if $X_i = F_{X_i}^{-1}(U)$ where $U \sim U(0,1)$ and $i = 1, 2$. However, the random variables X_1 and X_2 are said to be countermonotonic if and only if $\exists U \sim U(0,1)$ such that $X_1 = F_{X_1}^{-1}(U)$ and $X_2 = F_{X_2}^{-1}(1-U)$. Comonotonicity corresponds to perfect positive dependence, while countermonotonicity corresponds to perfect negative dependence.

Scarsini (1984) mentioned several properties that are desirable for a dependence measure, which are summarized in the below axiom.

Axiom 2.3.1. *Let X_1 and X_2 be two dependent random variables from a copula C and $\pi(X_1, X_2)$ be the dependence measure between them, then $\pi(X_1, X_2)$ is a concordance measure if it satisfies the following properties:*

I Symmetry: $\pi(X_1, X_2) = \pi(X_2, X_1)$,

II Normalization: $-1 \leq \pi(X_1, X_2) \leq 1$,

III Comonotonicity: $\pi(X_1, X_2) = 1$ if and only if X_1 and X_2 are comonotonic,

IV Countermonotonicity: $\pi(X_1, X_2) = -1$ if and only if X_1 and X_2 are countermonotonic,

V Independence: $\pi(X_1, X_2) = 0$ if and only if X_1 and X_2 are independent, and

VI Invariance: for every strictly monotone function $\phi : \mathbb{R} \to \mathbb{R}$, we have

$$\pi(\phi(X_1), X_2) = \begin{cases} \pi(X_1, X_2) & \text{if } \phi \text{ is increasing,} \\ -\pi(X_1, X_2) & \text{if } \phi \text{ is decreasing.} \end{cases}$$

2.3.1 Pearson's Correlation Coefficient ρ_p

The correlation ρ_p between X_1 and X_2 is defined by

$$\rho_p(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}, \tag{2.3.1}$$

where $\text{Var}(X_i)$ is the variance of $X_i, i = 1, 2$, and $\text{Cov}(X_1, X_2)$ is the covariance between the two random variables defined by $\text{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)$. Pearson's correlation coefficient is characterized by the following properties:

- It is a measure of linear dependence,
- $-1 \leq \rho_p(X_1, X_2) \leq 1$, where $|\rho_p(X_1, X_2)| = 1$ implies perfect linear dependence such that $X_2 = \alpha + \beta X_1$ almost surely for some $\alpha \in \mathbb{R}$, and $\beta > 0$ for positive linear dependence or $\beta < 0$ for negative linear dependence, and
- Independence implies $\rho_p(X_1, X_2) = 0$, however, $\rho_p(X_1, X_2) = 0$ does not imply independence.

On the other hand, Pearson's correlation coefficient has some disadvantages that makes it a weak measure for dependence. Those disadvantages are:

- $\rho_p(X_1, X_2)$ depends on the choice of the marginal distributions of X_1 and X_2,
- $\rho_p(X_1, X_2)$ requires finite variances for X_1 and X_2. This can present problems when we deal with heavy-tailed distributions that have infinite second moments, and
- $\rho_p(X_1, X_2)$ only measures *linear* dependence, i.e. $\rho_p(X_1, X_2)$ can be very close to or equal to 0, however, there might be a strong *non-linear* relationship.

Example: Non-linear dependence

Let $X_1 \sim U(-1, 1)$ and $X_2 = X_1^2$, then

$$\mathbb{E}(X_1) = 0, \text{ and } \mathbb{E}(X_1 X_2) = \mathbb{E}(X_1^3) = 0.$$

Therefore,

$$\text{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) = 0, \text{and } \rho_p(X_1, X_2) = 0,$$

however, it is clear that X_2 is a function of X_1. □

Note that the value of $\rho_p(X_1, X_2)$ is bounded depending on the marginal distributions of X_1 and X_2, as illustrated in Figure 2.8. Those bounds are explained in details in McNeil et al. (2015) and summarized in the below theorem.

Theorem 2.3.1. ***Attainable correlations*** *Let X_1 and X_2 be two random variables with finite variances and $Var(X_i) > 0, i = 1, 2$, then the following statements hold:*

1. *The attainable correlations belong to the following interval*

$$\left[\rho_p^{min}, \rho_p^{max}\right] \subseteq [-1, 1],$$

where $\rho_p^{min} < 0 < \rho_p^{max}$*,*

2. ρ_p^{min} *is attained if and only if* X_1 *and* X_2 *are countermonotonic, and* ρ_p^{max} *is attained if and only if* X_1 *and* X_2 *are comonotonic, and*

3. $\rho_p^{min} = -1$ *if and only if* $X_2 = \alpha + \beta X_1$ *where* $\alpha \in \mathbb{R}$*, and* $\beta < 0$*, while* $\rho_p^{min} = 1$ *if and only if* $X_2 = \alpha + \beta X_1$*, where* $\alpha \in \mathbb{R}$*, and* $\beta > 0$*.*

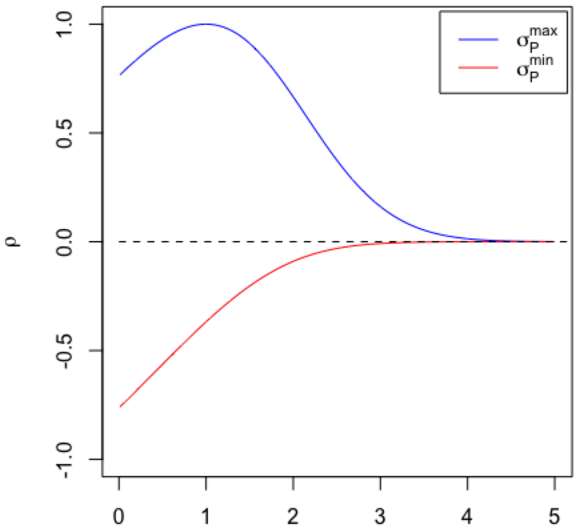

Figure 2.8: The attainable correlations ρ_p^{min} and ρ_p^{max} for $X_1 \sim \text{LogNormal}(0, 1)$ and $X_2 \sim \text{LogNormal}(0, \sigma^2)$, as proved in McNeil et al. (2015).

Therefore, only properties I and II from Axiom 2.3.1 are satisfied for Pearson's correlation coefficient. Note that the invariance property is only satisfied for linear transformations.

2.3.2 Spearman's rho ρ_S

Introduced by Spearman (1904), Spearman's rho is defined as the linear correlation between the marginal cdfs of X_1 and X_2, i.e.

$$\rho_S(X_1, X_2) = \rho_p\left(F_{X_1}(X_1), F_{X_2}(X_2)\right).$$

Recall that $U_i = F_{X_i}(X_i) \sim U(0,1)$ where $i = 1, 2$. Then, $\mathbb{E}(U_i) = \frac{1}{2}$, $\text{Var}(U_i) = \frac{1}{12}$, then by using Eq. 2.3.1, we have

$$\rho_S(X_1, X_2) = \frac{\mathbb{E}(F_{X_1}(X_1)F_{X_2}(X_2)) - \mathbb{E}(F_{X_1}(X_1))\mathbb{E}(F_{X_2}(X_2))}{\sqrt{\text{Var}(F_{X_1}(X_1))\text{Var}(F_{X_2}(X_2))}}$$

$$
\begin{aligned}
&= \frac{\mathbb{E}(U_1U_2) - \mathbb{E}(U_1)\mathbb{E}(U_2)}{\sqrt{\mathrm{Var}(U_1)\mathrm{Var}(U_2)}} \\
&= \frac{\mathbb{E}(U_1U_2) - \frac{1}{2} \times \frac{1}{2}}{\sqrt{\frac{1}{12} \times \frac{1}{12}}} \\
&= 12\mathbb{E}(U_1U_2) - 3 \\
&= -3 + 12 \int_0^1 \int_0^1 u_1u_2c(u_1, u_2)\mathrm{d}u_1\mathrm{d}u_2, \qquad (2.3.2)
\end{aligned}
$$

or equivalently

$$
= -3 + 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{X_1}(x_1)F_{X_2}(x_2)f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2.
$$

Quesada-Molina (1992) generalized an inequality by Hoeffding (1940) and proved that by double partial integrations, Eq. 2.3.2 can be rewritten as follows:

$$
\rho_S(X_1, X_2) = -3 + 12 \int_0^1 \int_0^1 C(u_1, u_2)\mathrm{d}u_1\mathrm{d}u_2.
$$

Therefore, the value of Spearman's rho does not depend on the marginal distribution, but it only depends on the Copula. Ghoudi et al. (1998) proved that for Extreme-Value Copulas, defined in Section 2.2.1, Spearman's rho is defined as

$$
\rho_S(X_1, X_2) = -3 + 12 \int_0^1 \frac{1}{(A(t) + 1)^2}\mathrm{d}t.
$$

We have the following limiting cases for Spearman's rho;

- If X_1 and X_2 are comonotonic, then $U_1 = U_2$. Therefore $\mathbb{E}(U_1U_2) = \mathbb{E}(U_1^2) = \frac{1}{3}$, and

$$
\rho_S(X_1, X_2) = -3 + 12\left(\frac{1}{3}\right) = 1.
$$

- If X_1 and X_2 are independent, then $\mathbb{E}(U_1U_2) = \mathbb{E}(U_1)\mathbb{E}(U_2)$. Therefore,

$$
\rho_S(X_1, X_2) = -3 + 12\left(\frac{1}{2} \times \frac{1}{2}\right) = 0.
$$

- If X_1 and X_2 are countermonotonic, then $U_1 = 1 - U_2$. Therefore, $\mathbb{E}(U_1U_2) = \mathbb{E}(U_1 - U_1^2) = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$, and

$$
\rho_S(X_1, X_2) = -3 + 12\left(\frac{1}{6}\right) = -1.
$$

Example: FGM Copula

Let the pair X_1 and X_2 follow the FGM Copula with the copula density derived in Eq. 2.2.8, then Spearman's rho can be calculated by using Eq. 2.3.2 as follows:

$$\begin{aligned}\rho_S(X_1, X_2) &= -3 + 12 \int_0^1 \int_0^1 u_1 u_2 \left[1 + \theta(1 - 2u_1)(1 - 2u_2)\right] \mathrm{d}u_1 \mathrm{d}u_2 \\ &= \frac{\theta}{3}.\end{aligned}$$

Therefore, for any marginal distributions for X_1 and X_2, $\rho_S(X_1, X_2) = \frac{\theta}{3}$. Note that $\rho_S(X_1, X_2)$ is an increasing function of θ, which is the case for most models. In addition, given that for the FGM Copula, we have $-1 \leq \theta \leq 1$, then $-\frac{1}{3} \leq \rho_S(X_1, X_2) \leq \frac{1}{3}$. □

Figure 2.9 represents the relationship between ρ_S and ρ_p for the Gauss Copula. Spearman's rho for the Gauss Copula is given by $\rho_S = (6/\pi) \arcsin(\rho_p/2)$. We notice that the relationship between them is almost linear. Table 2.2 represents Spearman's rho for some copula families.

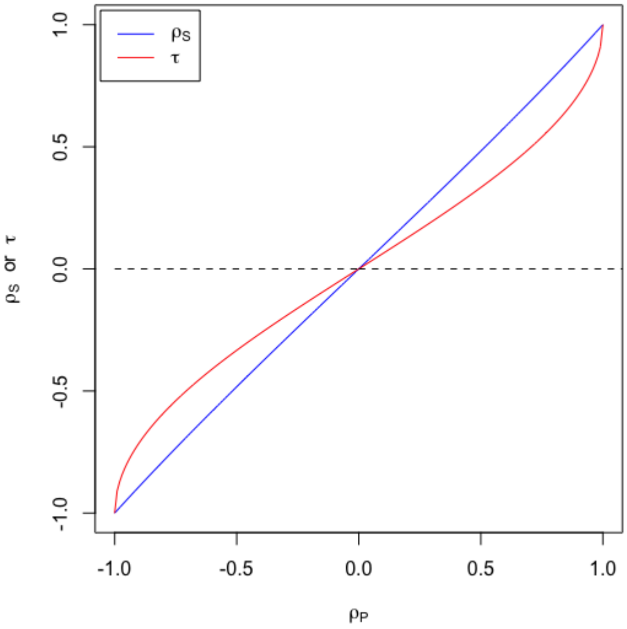

Figure 2.9: Relationship between Spearman's rho ρ_S, Kendall's Tau τ and Pearson's correlation coefficient ρ_p for Gauss Copula.

Note that all the properties in Axiom 2.3.1 are satisfied by Spearman's rho.

2.3.3 Kendall's Tau τ

Kendall (1938) introduced a new measure for rank correlation that measures the concordance between two random variables, X_1 and X_2. Let (x_1, x_2) and $(\tilde{x}_1, \tilde{x}_2)$ be two points in $\mathbb{R}^2$. Then those points are concordant if $(x_1-\tilde{x}_1)(x_2-\tilde{x}_2) > 0$ and discordant if $(x_1-\tilde{x}_1)(x_2-\tilde{x}_2) < 0$. This means that for concordance, we expect X_1 and X_2 to move in the same direction, i.e. increase together, or decrease together. Conversely, we expect the variables to have opposite direction of movement for discordance. Figure 2.10 visually explains the difference between concordance and discordance.

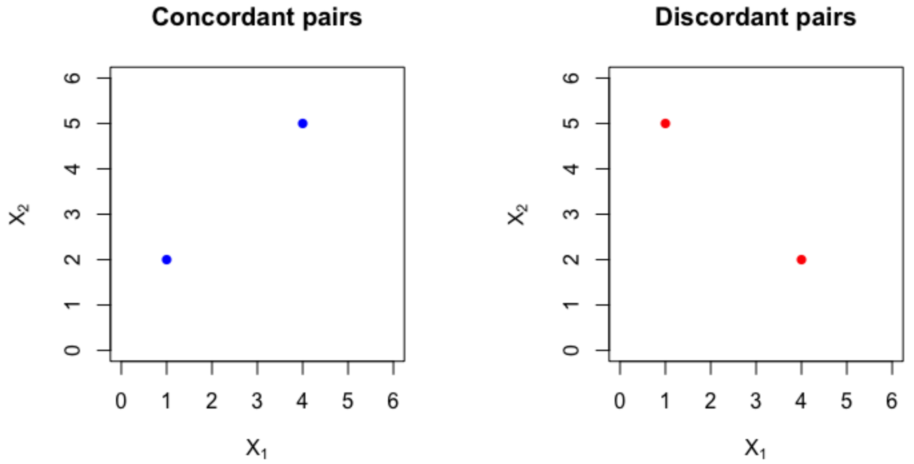

Figure 2.10: On the left, a pair of concordant points, and on the right, a pair of discordant points.

Kendall's τ is defined as follows:

$$\begin{aligned}\tau(X_1, X_2) &= P(\text{Concordance}) - P(\text{Discordance}) \\ &= P\left\{(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right\} - P\left\{(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0\right\}, \qquad (2.3.3)\end{aligned}$$

where (X_1, X_2) and $(\tilde{X}_1, \tilde{X}_2)$ are two independent random vectors that have the same distribution. We expect that if X_2 increases as X_1 increases, then the probability of concordance will be high. Conversely, if X_2 decreases as X_1 increases, then the probability of discordance will be high.

Theorem 2.3.2. ***Kendall's τ in terms of Copulas*** *Let (X_1, X_2) and $(\tilde{X}_1, \tilde{X}_2)$ be two continuous independent random vectors that have the same joint distribution function F, and marginals F_1 and F_2. Let C be a copula such that $F(x_1, x_2) = C\left(F_1(x_1), F_2(x_2)\right)$, Then Kendall's tau is given by*

$$\tau(X_1, X_2) = -1 + 4\int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2). \tag{2.3.4}$$

Proof. Note that

$$P(\text{Concordance}) + P(\text{Discordance}) = 1,$$

and

$$P\left\{(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right\} = P(X_1 > \tilde{X}_1, X_2 > \tilde{X}_2) + P(X_1 \leq \tilde{X}_1, X_2 \leq \tilde{X}_2).$$

Then, Eq. 2.3.3 can be rewritten as

$$\begin{aligned}\tau(X_1, X_2) &= P\left\{(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right\} - \left[1 - P\left\{(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right\}\right] \\ &= 2P\left\{(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right\} - 1.\end{aligned}$$

In addition, since the random vectors are continuous, then

$$\begin{aligned}P(X_1 > \tilde{X}_1, X_2 > \tilde{X}_2) &= P(U_1 > \tilde{U}_1, U_2 > \tilde{U}_2), \\ P(X_1 \leq \tilde{X}_1, X_2 \leq \tilde{X}_2) &= P(U_1 \leq \tilde{U}_1, U_2 \leq \tilde{U}_2).\end{aligned}$$

Therefore,

$$\begin{aligned}P(U_1 > \tilde{U}_1, U_2 > \tilde{U}_2) = P(\tilde{U}_1 \leq U_1, \tilde{U}_2 \leq U_2) & \\ &= \int_0^1 \int_0^1 P(\tilde{U}_1 \leq U_1, \tilde{U}_2 \leq U_2 | U_1 = u_1, U_2 = u_2) \mathrm{d}C(u_1, u_2) \\ &= \int_0^1 \int_0^1 P(\tilde{U}_1 \leq u_1, \tilde{U}_2 \leq u_2) \mathrm{d}C(u_1, u_2) \\ &= \int_0^1 \int_0^1 C(u_1, u_2) \mathrm{d}C(u_1, u_2).\end{aligned}$$

Equivalently, $P(U_1 \leq \tilde{U}_1, U_2 \leq \tilde{U}_2) = \int_0^1 \int_0^1 C(u_1, u_2) \mathrm{d}C(u_1, u_2)$. Therefore,

$$\tau(X_1, X_2) = -1 + 2P\left\{(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right\}$$

$$
\begin{aligned}
&= -1 + 2\left[\int_0^1\int_0^1 C(u_1,u_2)\mathrm{d}C(u_1,u_2) + \int_0^1\int_0^1 C(u_1,u_2)\mathrm{d}C(u_1,u_2)\right] \\
&= -1 + 4\int_0^1\int_0^1 C(u_1,u_2)\mathrm{d}C(u_1,u_2).
\end{aligned}
$$

□

Example: FGM Copula

Let the pair X_1 and X_2 follow the FGM Copula derived in Eq. 2.2.7 with copula density defined in Eq. 2.2.8, then Kendall's tau can be calculated by using Eq. 2.3.4 as follows:

$$
\begin{aligned}
\tau(X_1,X_2) &= -1 + 4\int_0^1\int_0^1 C(u_1,u_2)\mathrm{d}C(u_1,u_2) \\
&= -1 + 4\int_0^1\int_0^1 C(u_1,u_2)c(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2 \\
&= -1 + 4\int_0^1\int_0^1 \left[u_1u_2 + \theta u_1u_2(1-u_1)(1-u_2)\right]\left[1+\theta(1-2u_1)(1-2u_2)\right]\mathrm{d}u_1\mathrm{d}u_2 \\
&= \frac{2\theta}{9}.
\end{aligned}
$$

Therefore, for any marginal distributions for X_1 and X_2, $\tau(X_1,X_2) = \frac{2\theta}{9}$. In addition, given that for the FGM Copula, we have $-1 \leq \theta \leq 1$, then $\frac{-2}{9} \leq \tau(X_1,X_2) \leq \frac{2}{9}$. □

As shown in Theorem 2.3.2, the value of Kendall's tau depends solely on the Copula. In fact, Genest and Mackay (1986) proved that for any Archimedean Copula (defined in Section 2.2.1), Kendall's tau is

$$\tau(X_1,X_2) = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}\mathrm{d}t,$$

where $\phi'(t)$ is the first derivative of the generator $\phi(t)$. In addition, Ghoudi et al. (1998) proved that for Extreme-Value Copulas (defined in Section 2.2.1, Kendall's tau is

$$\tau(X_1,X_2) = \int_0^1 \frac{t(1-t)}{A(t)}\mathrm{d}A'(t),$$

where $A'(t)$ is the first derivative of the generator $A(t)$. Table 2.2 represents Kendall's tau for some copula families.

Note that all the copulas represented in Figures 2.4 and 2.6 have Kendall's $\tau = 0.75$. In those figures, if we choose other marginal distributions, the plots in the bottom row will be different, however, the plots in the first row will remain unchanged. Figure 2.9 represents the relationship between τ and ρ_p for the Gauss Copula. Kendall's tau for the Gauss Copula is given by

$\tau = (2/\pi) \arcsin(\rho_p)$. Unlike Spearman's rho, the relationship is not linear.

Note that all the properties in Axiom 2.3.1 are satisfied by Kendall's tau.

Table 2.2: Spearman's rho ρ_S and Kendall's tau τ for the copula families discussed in 2.2.1.

Family	Copula	ρ_S	τ
Elliptical	Gauss	$(6/\pi) \arcsin(\rho_p/2)$	$(2/\pi) \arcsin(\rho_p)$
	t		
Archimedean	Clayton	Complicated	$\theta/(\theta+2)$
	Frank	$1 + \frac{12}{\theta}[D_2(\theta) - D_1(\theta)]$	$1 - \frac{4}{\theta}[1 - D_1(\theta)]$
	Gumbel	No closed form	$1 - 1/\theta$
Extreme Value	Gumbel's First Asymmetric Model	No closed form	$1 - 1/\theta$
	Gumbel's Second Model	Complicated	$\frac{8 \arctan \sqrt{\frac{\theta}{4-\theta}}}{\sqrt{\theta(4-\theta)}} - 2$
	Galambos Asymmetric Copula	No closed form	No closed form

where $D_k(x)$ is the Debye function for any positive integer k and it is given by

$$D_k(x) = \frac{k}{x^k} \int_0^k \frac{t^k}{\mathrm{e}^t - 1} \mathrm{d}t.$$

Chapter 3

Model and Variable Selection Criteria

3.1 Model Selection Criteria

As stated by Box (1976), "All models are wrong, but some are useful". A "true model" does not exist, but some models can be informative and our aim is to find the most accurate approximation of reality. Given several fitted models, we need to identify the model(s) that best explains our data and minimizes the loss of information that occur during modeling. There are several statistics used in model selecting, however, we will only focus on the ones commonly used:

- Coefficient of Determination R^2,
- Adjusted Coefficient of Determination R_a^2,
- Akaike Information Criteria (AIC), and
- Bayesian Information Criteria (BIC).

A general principle is the "law of parsimony", which is originated from Occam's razor principle. This law encourages statisticians to use a simple model to explain their data rather than a complex model, given a certain level of accuracy. This means that the best model to use in fitting the data is the one which provides us with the highest information gain and less complexity. For this reason, several of the model selection criteria discussed below includes a penalty for inclusion of more parameters or variables.

We will assume a simple model in order to explain the model selection criteria. Note that more complicated models can be used, however, we used a simple model to have simplified and easily interpretable results. Consider a data set that contains n observations. Each observation i consists of a scalar response variable y_i and a set of p predictors x_{ij}, for $j = 1, \ldots, p$. We assume a linear relationship between the predictors and the response variable as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i,$$

where β_0 is called the model intercept, $\beta_1, \ldots, \beta_p$ are the regression coefficients and ϵ_i is the random error. The predicted value $\hat{y}_i$ of the model is calculated by using the estimated regression coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_p$, such that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}$. In addition, let $\bar{y}$ be the mean of the n observations of the response variable. We define the following terms:

- Total sum of squares quantifies the variation between the data points y_i and the sample mean $\bar{y}$. It is calculated as follows: $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$,
- Regression sum of squares quantifies the variation between the regression line (i.e. predicted values $\hat{y}_i$) and the sample mean $\bar{y}$. It is calculated as follows: $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, and
- Error sum of squares quantifies the variation between the data points y_i and the predicted values $\hat{y}_i$. It is calculated as follows: $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$,

where $SST = SSR + SSE$, if and only if $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$.

In addition, we define the likelihood function of a given model M, with parameters $\boldsymbol{\theta}$ and data $\boldsymbol{X}$ as $L \coloneqq L(\boldsymbol{\theta}; X) = P(\boldsymbol{X}|\boldsymbol{\theta}, M)$, and the maximized value is $\hat{L} = P(\boldsymbol{X}|\hat{\boldsymbol{\theta}}, M)$ where $\hat{\boldsymbol{\theta}}$ are the parameters that maximize the function.

3.1.1 The Coefficient of Determination R^2

The R^2 measures the proportion of the total variation in the response variable that is accounted for by the predictors in the regression model. This makes it a measure of the success of the predictors in predicting the response variable. In OLS models, it is calculated by

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

R^2 is defined over $[0, 1]$, where

- $R^2 = 0$ means that the response variable cannot be predicted from the predictors,
- $R^2 = 1$ means that the response variable can be predicted without errors from the predictors, and
- $0 < R^2 < 1$ is the percentage by which the variation in the response variable is explained by the variation in the predictors.

In a simple linear regression model (only 1 predictor X), $R^2 = [r(Y, X)]^2$, where $[r(Y, X)]^2 = \rho_p(Y, X)$. However, in multiple linear regression, $R^2 = [r(Y, \hat{Y})]^2$.

One disadvantage of R^2 is that it is a non-decreasing function of the number of predictors. This means that the more predictors are added to the model, the higher the value of R^2 tend to be, even if the additional variables barely contribute to the prediction of the response variable. This makes it extremely difficult to compare models with different sizes, which led researchers to consider the adjusted R^2.

3.1.2 The Adjusted Coefficient of Determination R_a^2

To manage the disadvantage of the R^2, the adjusted R^2, referred to as R_a^2 penalizes the R^2 value based on the number of predictors in the model as follows:

$$R_a^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-p-1}\right).$$

R_a^2 allows us to compare models of different numbers of predictors, and its value will always be less than or equal to R^2. R_a^2 can sometimes hold a negative value if we have a small number of observations and too many predictors. While the value of R^2 can be interpreted, R_a^2 has no interpretation, but it is a statistic used to compare models.

A common disadvantage for the R^2 and the R_a^2 is that it is not defined over all linear models, specifically GLMs, where a pseudo R^2 is calculated using several methodologies as explained in Mittlbock and Heinzl (2004). There are several proposed measures for the pseudo R^2 and pseudo R_a^2 and they are all mostly based on the likelihood value for the fitted model and the null model (a model with only an intercept and no predictors). Given that there are multiple measures for the pseudo R^2 and pseudo R_a^2, this makes them hold a different interpretation than the ones produced from the OLS models, and hence they cannot be directly compared to them. However, the pseudo R^2 and pseudo R_a^2 can be used to compare similar models.

If we wish to compare OLS, GLMs and GLMMs, then the coefficient of determination and the adjusted coefficient of determination are inappropriate statistics to compare our models.

3.1.3 Likelihood Ratio Tests

As discussed earlier in Section 1.2.6, if a model is a special case of another model (i.e. nested models), then using the Likelihood Ratio Test (LRT) becomes appropriate. A model M_1 is said to be nested of another model M_2, if it uses a subset of the predictors of M_2. If we want to compare the two nested models M_1 and M_2 with p_1 and p_2 number of variables respectively, such that $p_2 > p_1$, and parameters $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$, the likelihood ratio test is a conditional test, such that given that model M_2 fits the data, it tests whether the simpler model M_1 also fits the data. Let $\hat{L}_1$ and $\hat{L}_2$ be the likelihood functions for models M_1 and M_2, respectively. The null and alternative hypothesis are defined as follows:

$$H_0 : \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_1,$$
$$H_1 : \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_2.$$

Therefore, obtaining a small p-value makes us reject the simplified model M_1, and a big p-value does not reject that the simplified model is not significantly different from M_2.

The likelihood ratio statistic is defined as

$$D = -2\left[\log(\hat{L}_1) - \log(\hat{L}_2)\right] = -2\log\left(\frac{\hat{L}_1}{\hat{L}_2}\right).$$

This test statistic asymptotically follows the Chi-Square distribution with degrees of freedom $\nu = p_2 - p_1$.

However, frequently we would like to compare models that are not nested. They can be compared by using the Akaike Information Criteria and Bayesian Information Criteria.

3.1.4 Akaike Information Criteria (AIC)

Introduced by Akaike (1973, 1974), Akaike Information Criteria (AIC) is one of the most commonly used methods in model selection. It is used to provide a relative estimate of the lost information by a given model. It is calculated as follows:

$$AIC = -2\log(\hat{L}) + 2k,$$

where k is the number of estimated parameters or coefficients in the model. Since AIC represents the amount of lost information, the best model is the one with the smallest possible AIC value. If we increase the number of variables in the model, we get a better fit to the data, and hence the value of $\hat{L}$ increases. However, this results in an increase in the penalty term (the second term in the formula). Therefore, this penalty is used to restrict overfitting in our model.

However, when the sample size n is small compared to the number of parameters in the model, approximately $n/k < 40$, Hurvich and Tsai (1989) created a corrected measure of AIC which is

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1},$$

that is used to prevent overfitting for small data sets. Burnham and Anderson (2004) suggested that since AIC_c converges to AIC as n gets large, it is always better to use AIC_c for model selection.

The value of AIC (or AIC_c) in itself has no meaning, therefore to have some useful interpretation, it is advisable to calculate

$$\Delta AIC_i = AIC_i - AIC_{min},$$

where AIC_{min} is the smallest AIC (or AIC_c) and ΔAIC_i represents the difference between the AIC value of the i^{th} model and AIC_{min}. This results in having $\Delta AIC_i = 0$ for the best model, and the other models have a positive value. ΔAIC_i represents the information loss if we choose model M_i over the best model M_{min}.

As per Burnham and Anderson (2003, chap. 2.11), there are certain conditions under which it is allowed to use AIC to compare a set of models, which are:

- The models should be for the same data set with the same number of observations,
- The order of calculating AIC over the set of models is insignificant in the comparison, which means that if we are comparing models A and B, it doesn't make a difference if AIC is calculated first for model A and then for model B, or vice versa,
- The models should all have the same response variable. In other words, if a transformation is made on the response variable, the same transformation should be applied over all the other models, and
- AIC is not a hypothesis test. It does not tell the validity or quality of the model.

3.1.5 Bayesian Information Criteria (BIC)

The Bayesian Information Criteria is also a widely used method for model selection, and it is closely related to AIC. Schwarz et al. (1978) provided a Bayesian argument for using BIC and he defined it as

$$BIC = -2\log(\hat{L}) + k\log n.$$

The goal of using BIC is to find a model that maximizes the posterior probability of the model, thus it attempts to find the "true" model, or the one where the posterior probability approaches 1. One of the main assumptions behind BIC is that the "true" model actually exists, and it is included in the set of models being tested, and BIC will converge in probability to the true model as $n \to \infty$.

Following the same methodology as AIC, the BIC is calculated for all models and the model with the smallest value of BIC which is BIC_{min} is chosen. The lower the value of BIC, the higher the probability that this model is the "true" model. For easier interpretation, ΔBIC_i is calculated to be the difference between the BIC value of the i^{th} model and BIC_{min}. In addition, the same conditions mentioned earlier by Burnham and Anderson (2003) apply to BIC.

We can observe that the penalty term (second term) in the BIC formula is larger and more severe than that of the AIC, which makes it choose more parsimonious models.

3.2 Variable Selection Criteria

For every model we attempt to fit, we should identify the set of predictors from all possible predictors $\boldsymbol{X}$, such that we obtain a good fit for the data and maintain a parsimonious model. Our goal is to be able to explain the model in the simplest way. There are several methods that can help us eliminate the redundant predictors that do not add significant information to the model. We will explore the following methods:

- Backward Elimination,
- Forward Selection, and
- Stepwise Selection.

Those methods require calculations of the AIC or AIC_c. Note that for every mention of AIC, it can be replaced by AIC_c. Other criterion can be used instead of AIC, but we use it because it is the most commonly used measure.

3.2.1 Backward Elimination

This method is the simplest of all variable selection procedures. It is usually used when we have a modest number of predictors and we wish to eliminate a few of them. The required steps to perform this method are:

1. Start with a model that includes all predictors $\boldsymbol{X}$,
2. Calculate the AIC of the model,
3. For all predictors included in the model, calculate the AIC if they are individually removed from the model,
4. Remove the predictor that if removed, will provide us with a model with the lowest AIC, and
5. Repeat steps 2, 3 and 4 as long as there is a possibility of having a model with a lower AIC value.

3.2.2 Forward Selection

This is the opposite of backward selection. It is usually used when we have a large number of predictors. The required steps to perform this method are:

1. Start with a model that has no predictors,

2. Calculate the AIC of the model,

3. For all predictors not included in the model, calculate the AIC if they are individually added to the model,

4. Add the variable that if added, will provide us with a model with the lowest AIC, and

5. Repeat steps 2, 3 and 4 as long as there is a possibility of having a model with a lower AIC value.

3.2.3 Stepwise Selection

This is the mixture of backward elimination and forward selection methods. It sometimes provides a more accurate method than the other two measures because sometimes predictors are removed (or added) early in the process, but they prove their importance (or lack of it) later. This way, we can reevaluate removing/adding the predictors at each step. The required steps to perform this method are:

1. Start with a model that includes all predictors $\boldsymbol{X}$,

2. Calculate the AIC of the model,

3. For all predictors included in the model, calculate the AIC if they are individually removed from the model,

4. For all predictors not included in the model, calculate the AIC if they are added to the model,

5. Remove/add the variable that if removed/added, we will have a model with the lowest AIC, and

6. Repeat steps 2, 3, 4 and 5 as long as there is a possibility of having a model with a lower AIC value.

Chapter 4

Modeling GLMMs with Nested Copulas

In this chapter, we will explain the proposed model. Consider a longitudinal data for N participants. Let $t_j : j = 1, \ldots, J$ be the J measurement times and i be the index of the N participants in the study, where $i = 1, \ldots, N$. For each subject i and time t_j, the data has R responses, denoted $y_{ij}^{(r)}$, where $r = 1, \ldots, R$.

We will start by considering each response separately, and then consider the multivariate distribution later. The observations for each participant will be grouped by the participant's ID such that we obtain random effects for the intercept of the model and the time covariate. Each response will be modeled by a GLMM, such that the equation for the j^{th} observation of the i^{th} group/subject for the r^{th} response is

$$\eta_{ij}^{(r)} = (\beta_0 + b_{0i}) + \beta_1 x_{ij}^{(1)} + \ldots + \beta_p x_{ij}^{(p)} + (\beta_t + b_{ti}) t_{ij}.$$

Note that more/less random effects can be incorporated to the model, based on the context of the analysis. An expert's judgment is needed to justify the choice. In our analysis, we use only those two random effects because we assume that the variation between the groups can result from variation in the overall mean for the base scenario (each numerical predictor $x^{(k)} = 0, \forall k = 1, \ldots, p$, and the base case for the categorical predictors), and/or variation at different time points.

Now, we explore the dependence structure between the responses, which is represented by the following joint cdf

$$P\left(Y_{ij}^{(1)} \le y_{ij}^{(1)}, \dots, Y_{ij}^{(R)} \le y_{ij}^{(R)}\right) = C\left(F_1(y_{ij}^{(1)}), \dots, F_R(y_{ij}^{(R)})\right). \tag{4.0.1}$$

We propose a nested copula structure to model the dependence between the responses. The model is obtained recursively as follows:

$$\begin{aligned} C^{(1)}(u_1, u_2) &= C_1^{(\theta_1)}(u_1, u_2), \\ C^{(2)}(u_1, u_2, u_3) &= C_2^{(\theta_2)}\left(C_1^{(\theta_1)}(u_1, u_2), u_3\right), \\ \vdots \quad &= \quad \vdots \\ C^{(R-1)}(u_1, \dots, u_R) &= C_{R-1}^{(\theta_{R-1})}\left(C_{R-2}^{(\theta_{R-2})}(u_1, u_2, \dots, u_{R-1}), u_R\right), \end{aligned}$$

for $R \ge 2$, and $C_0^{(\theta_0)}(u_1) \coloneqq u_1$, where the parameters of the copulas are estimated sequentially, rather than jointly. We will explain below how to obtain values for $u_i, \forall i = 1, \dots, R$.

Figure 4.1 represents the suggested structure of the model, assuming 4 responses. It can be extended to a higher number or responses, if needed. The model requires estimation of $R-1$ bivariate copulas.

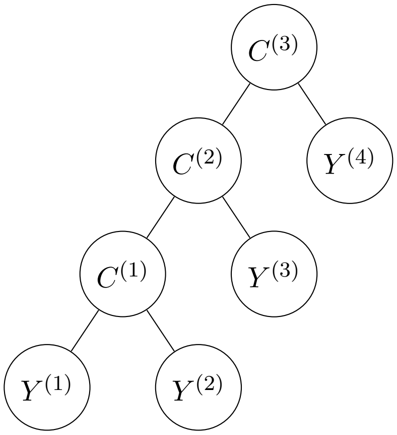

Figure 4.1: The tree structure used for modeling longitudinal data with 4 responses.

This proposed method provides flexibility by allowing different choices for the copulas $C^{(1)}, C^{(2)}$, and $C^{(3)}$, independently. This becomes very relevant in situations where pairs of variables behave significantly different from other pairs. For example, if we observe strong upper tail dependence between responses $Y^{(1)}$ and $Y^{(2)}$, while this dependence structure is not present in the other responses. In this situation, it is reasonable to choose different copulas for the pairs, instead of being restricted to modeling the four responses with one copula.

Shi and Frees (2011) suggested to model the standardized residuals from each regression model in order to remove the effects of the covariates. The standardized residuals are not obtained by the common way of subtracting the mean and dividing by the standard deviation, but rather by a specific formula for each distribution used. The goal is to form a sample that is independent and identically distributed, therefore, each vector of residuals can be standardized by using some of the estimated parameters of the distribution from the GLMM model, namely the location and scale parameters. Assume that $Y^{(1)}$ was fitted by using a GLMM with a normal distribution and identity link function, such that $y_{ij}^{(1)} \sim N(\mu_{ij}^{(1)}, \sigma^{(1)})$ and $\eta_{ij}^{(1)} = \mu_{ij}^{(1)}$. Therefore, the standardized residual vector for $Y^{(1)}$, namely $\hat{\boldsymbol{\epsilon}}^{(1)}$, is defined as

$$\hat{\epsilon}_{ij}^{(1)} = \frac{y_{ij}^{(1)} - \hat{\mu}_{ij}^{(1)}}{\hat{\sigma}^{(1)}}, \tag{4.0.2}$$

such that $\hat{\epsilon}_{ij}^{(1)} \sim N(0,1)$.

Similarly, if we assume that $Y^{(2)}$ was fitted by using a GLMM with a Poisson distribution and log link function, such that $y_{ij}^{(2)} \sim \text{Poisson}(\lambda_{ij}^{(2)})$ and $\eta_{ij}^{(2)} = \log \mu_{ij}^{(2)}$. Therefore, the standardized residual vector for $Y^{(2)}$, namely $\hat{\boldsymbol{\epsilon}}^{(2)}$, is defined as

$$\hat{\epsilon}_{ij}^{(2)} = \frac{y_{ij}^{(2)}}{\hat{\lambda}_{ij}^{(2)}}, \tag{4.0.3}$$

such that $\hat{\epsilon}_{ij}^{(2)} \sim \text{Poisson}(1)$. Additionally, assume that the response $Y^{(3)}$ was modeled by a GLMM with a gamma family and log link function, such that $y_{ij}^{(3)} \sim \text{Gamma}(\alpha^{(3)}, \beta_{ij}^{(3)})$ and $\eta_{ij}^{(3)} = \log \mu_{ij}^{(3)}$. Therefore, the standardized residual vector for $Y^{(3)}$, namely $\hat{\boldsymbol{\epsilon}}^{(3)}$, is defined as

$$\hat{\epsilon}_{ij}^{(3)} = \frac{y_{ij}^{(3)}}{\hat{\beta}_{ij}^{(3)}}, \tag{4.0.4}$$

such that $\hat{\epsilon}_{ij}^{(3)} \sim \text{Gamma}(\alpha^{(3)}, 1)$. Note that for the Gamma distribution, the shape parameter α is the inverse of the dispersion parameter ϕ, that is used in the GLMM model. Refer to Section 1.2.2 for further details on the parameters of the Exponential family.

In this situation, the pair $\left(\hat{\epsilon}_{ij}^{(1)}, \hat{\epsilon}_{ij}^{(2)}, \hat{\epsilon}_{ij}^{(3)}\right)$ are a pseudo-random sample from a copula C with marginals that are approximately $N(0,1)$, Poisson(1) and Gamma$(\alpha^{(3)}, 1)$, respectively.

Let $\hat{U}_{ij}^{(r)}$ represent the standardized ranks of $\hat{\boldsymbol{\epsilon}}^{(r)}$, represented by

$$\hat{U}_{ij}^{(r)} = \frac{R_{ij}^{(r)}}{n+1},$$

where $R^{(r)}$ is the ranks of $\hat{\epsilon}_{ij}^{(r)}, \forall r = 1, \ldots, R$, and we divide by $(n+1)$ to ensure that all standardized ranks lie strictly between 0 and 1.

The estimate of the dependence parameter $\hat{\theta}_1$ of $C_1^{(\theta_1)}$ is obtained by maximizing the pseudo log-likelihood of the copula density function from Eq. 2.2.4 as follows:

$$l(\theta_1) = \sum_{i=1}^{N}\sum_{j=1}^{J} \log c_1^{(\theta_1)}\left(\hat{U}_{ij}^{(1)}, \hat{U}_{ij}^{(2)}\right) + \sum_{i=1}^{N}\sum_{j=1}^{J} \log f_1\left(\epsilon_{ij}^{(1)}\right) + \sum_{i=1}^{N}\sum_{j=1}^{J} \log f_2\left(\epsilon_{ij}^{(2)}\right), \tag{4.0.5}$$

where $c_1^{(\theta_1)}$ is the density of $C_1^{\theta_1}$, and f_1 and f_2 are the density functions of $\hat{\boldsymbol{\epsilon}}^{(1)}$ and $\hat{\boldsymbol{\epsilon}}^{(2)}$, respectively.

Note that only the first term in Eq. 4.0.5 has θ_1, therefore, the pseudo log-likelihood function of $\hat{\boldsymbol{\epsilon}}^{(1)}$ and $\hat{\boldsymbol{\epsilon}}^{(2)}$ is reduced to

$$l(\theta_1) = \sum_{i=1}^{N}\sum_{j=1}^{J} \log c_1^{(\theta_1)}\left(\hat{U}_{ij}^{(1)}, \hat{U}_{ij}^{(2)}\right).$$

Similarly, the estimate of the dependence parameter $\hat{\theta}_2$ of $C_2^{(\theta_2)}$ is obtained by maximizing the pseudo log-likelihood

$$l(\theta_2) = \sum_{i=1}^{N}\sum_{j=1}^{J} \log c_2^{(\theta_2)}\left\{C_{n1}^{(\theta_1)}\left(\hat{U}_{ij}^{(1)}, \hat{U}_{ij}^{(2)}\right), \hat{U}_{ij}^{(3)}\right\},$$

where $c_2^{(\theta_2)}$ is the density of $C_2^{(\theta_2)}$ and $C_{n1}^{(\theta_1)}$ represents the empirical copula of $C_1^{(\theta_1)}$. This procedure is iterated for as many response variables as needed, namely $R-1$ times.

The adequacy of the fit of a copula C is tested by 1000 bootstrap iterations for the Cramér Von Mises statistic, defined as

$$S_n = \int_0^1 \int_0^1 \left\{C_n(u_1, u_2) - C_{\theta_n}(u_1, u_2)\right\}^2 \mathrm{d}u_1 \mathrm{d}u_2,$$

where C_n represents the empirical copula and C_{θ_n} represent the fitted copula with the rank-estimate of the dependence parameter. The null hypothesis of the test is defined as $H_0 : C \in C_{\theta_n}$, and it is compared to the significance level α. Therefore, for a p-value $> \alpha$, we do not reject the null hypothesis, and a p-value $< \alpha$ results into rejection of H_0. Further details on the goodness-of-fit procedure is explained in Genest et al. (2009).

Note that fitting the linear models is done by using the *stats* and *lme4* R package. Stepwise variable selection for the GLM were done by using a function that we have built, because R's function performs it based on the AIC criteria, so we amended the function to do the procedure based on the AIC_c criteria. Note that the copula fitting procedure and goodness of fit tests are done by using the *copula* R package.

Chapter 5

Application

This chapter provides details about the study that motivated the work performed in this thesis. Two randomized, placebo-controlled, double-blinded studies, referred to as pilot and bridge, have been performed on children aged between 1 and 6 years who have been previously diagnosed with recurrent, moderate or severe asthma. The main goal of the studies was to observe if supplementation of vitamin D can decrease the number of asthma exacerbations that require the use of rescue oral corticosteroids (OCS). The patients of the pilot study were recruited between November 2013 to February 2014. They received either a single oral dose of 100,000 International Unit (IU) of vitamin D, or a placebo. In addition, all participants took a daily dose 400 IU of vitamin D for the duration of the study. Further details of the study and its outcome are explained in Jensen et al. (2016). However, the patients for the bridge study received either 2 oral doses of 100,000 IU of vitamin D, or placebo. The doses were taken 3.5 months apart, beginning in Fall of 2016. Unlike the pilot study, the participants in the bridge study did not take an additional daily dose of vitamin D. Under both studies, the participants attended 3 clinical visits; at baseline, i.e. $t = 0$, at 3 months (or 3.5 for bridge) and at 6 months (or 7 for bridge). In our analysis, we assume that the visits were at time t-months, where $t = 0, 3, 6$. Blood samples from the patients were collected at each visit, in addition to demographic and medical characteristics.

The goal of our analysis is to identify the dependence structure between the change in the amount of vitamin D in the blood, $Y^{(1)}$, and the number of asthma attacks that require the use of rescue OCS, $Y^{(2)}$. This is done by initially modeling the marginal distribution of each outcome, and then finding their joint distribution by using copulas. Note that we have 3 time points, namely $t = 0, 3, 6$, however, we only have two intervals, namely $[0, 3]$, $[3, 6]$, for the change in vitamin D and the number of asthma attacks.

5.1 Data Analysis

In our analysis, we combine patients from the pilot and the bridge study together. Table 5.2 provides the baseline patients' characteristics that we used in our analysis. The following data manipulations were performed:

- At each time point t, we removed the observations that belong to patients that dropped out of the study prior to time t. At baseline, i.e. $t = 0$, 1 patient from the bridge study dropped out prior to taking the required blood sample, and hence, he was removed. Similarly, a total of 6 patients are removed at $t = 3$ and 11 patients are removed at $t = 6$,
- Missing values were imputed by using the mean for numerical variables and the median for categorical variables,
- The Z-score of the BMI per patient was calculated as per the World Health Organization (WHO) standards that are presented in Who et al. (2006),
- 1 outlier in the Z-score of the BMI was replaced by the data for the same patient at the following visit. The outlier was due to a typo in the patient's weight,
- The Fitzpatrick scale of skin color is grouped such that each group has at least 5 observations to maintain good credibility in the results,
- The daily 400 IU of vitamin D for the pilot participants is added to the Daily Dietary vitamin D intake. Additionally, the variable is categorized into 4 quartiles, namely Q1 - Q4. The split of the quartiles is based on the data of daily dietary and supplementary vitamin D at baseline,
- The minutes spent in the sun variable is split into two categories, namely $\leq$ or $>$ the median (60 minutes per day),
- Body coverage of SPF variable is removed due to its high correlation with SPF usage, and

- Since none of the patients took a sunny vacation at $t = 6$, the corresponding variable is removed from the analysis only at this time point. Similarly, since the blood samples at baseline were taken before randomization, the placebo/treatment and pilot/bridge variables are not included in the analysis at baseline.

Initially, we start with modeling the amount of vitamin D in the blood at each visit to have a better understanding of the data. Note that this is different from $Y^{(1)}$, as $Y^{(1)}$ represents the change in vitamin D between visits. This is a preliminary analysis and the estimates obtained from the model will not be used in further steps. Since our outcome is a continuous variable that is strictly positive, we fit GLM models with Gamma and Inverse Gaussian families and with log link functions. Stepwise variable selection by using the AIC_c values is performed in order to obtain parsimonious models that explain our data well. In addition, LRTs are performed such that the original model with all variables is compared to its nested model that is obtained from the stepwise process. The p-values in Table 5.1 show that the parsimonious models are not significantly different from the original models (with all predictors) at the level of $\alpha = 0.01$, and hence, they are not rejected. In addition, we compare the Gamma and the Inverse Gaussian models by comparing their AIC_c values. As shown in Table 5.3, the Gamma models outperform the Inverse Gaussian models. Note that the BIC values of the models confirmed the selection of the Gamma over the Inverse Gaussian models.

Table 5.1: p-value of the LRTs on the GLM models for the amount of vitamin D at each visit. The LRTs compare the model with all predictors and the parsimonious model obtained from the stepwise variable selection.

GLM model	$t = 0$	$t = 3$	$t = 6$
Gamma	0.63	0.89	0.83
Inverse Gaussian	0.56	0.87	0.67

Table 5.2: Patients Characteristics at each visit. Numerical variables are associated with the mean and the 95% confidence interval.

Variable	Categories	$t = 0$ $(n = 68)$	$t = 3$ $(n = 63)$	$t = 6$ $(n = 58)$
Study	Pilot	22	21	18
	Bridge	46	42	40
Group	Placebo	35	34	32
	Treatment	33	29	26
Vit. D in blood	-	71 $(37, 105)$	77 $(44, 110)$	80 $(46, 114)$
Age	-	2.9 $(0.8, 4.9)$	3.1 $(1.1, 5.2)$	3.4 $(1.3, 5.4)$
Gender	Male	36	32	30
	Female	32	31	28
Z-score of BMI	-	0.6 $(-0.4, 2.7)$	0.6 $(-1.4, 2.5)$	0.5 $(-1.5, 2.6)$
Fitzpatrick Scale	$[1, 2]$	41	37	35
	$[3, 4]$	22	21	18
	$[5, 6]$	5	5	5
Season	Fall	54	-	-
	Winter	14	48	-
	Spring	-	15	37
	Summer	-	-	21
Asthma Severity	Persistent	46	42	31
	Episodic	22	21	27
Daily Dietary and Supplementary Vit. D	Q1: $[0 - 165]$	18	12	11
	Q2: $(165 - 215]$	16	8	14
	Q3: $(215 - 356]$	17	17	11
	Q4: $(356 - \infty)$	17	26	22
Daily inhaled corticosteroids (ICS) intake	No	21	11	10
	Yes	47	52	48
Minutes spent in the sun per day in past 3 months	$\leq$ median (60 mins)	24	50	23
	$>$ median (60 mins)	44	13	35
SPF usage in past 3 months	< 30	18	55	10
	≥ 30	50	8	48
SPF coverage	Minimal coverage	29	57	14
	Good coverage	39	6	44
Sunny Vacation in past 3 months	No	57	56	58
	Yes	11	7	0

Table 5.3: AIC_c of the optimal GLM models for the amount of vitamin D at each visit, obtained by stepwise variable selection.

GLM model	$t = 0$	$t = 3$	$t = 6$
Gamma	572.89	530.41	488.46
Inverse Gaussian	590.60	532.56	490.52

Additionally, we explore the possibility of including interaction between the variables, but LRTs confirm that the model with interaction terms does not add significant value to the model, and hence they are rejected. Furthermore, the models are fitted without imputing the missing values. We observe that we obtain the same variables by using stepwise variable selection and that the coefficients of the variables are very close to those of the imputed models. Therefore, we accept that the imputation of the data was robust and did not result in significant inaccuracy.

The estimates of the parameters are transformed by using the inverse of the link function, as explained in Eq. 1.2.7. Table 5.4 provides the transformed parameters estimates of the chosen models (GLM with Gamma family and log link function) for each time point $t = 0, 3, 6$, along with their 95% confidence intervals. As explained in Eq. 1.2.8, the transformed coefficients provide a multiplicative factor for the change in the mean of Y due to a 1 unit increase in the corresponding variable.

We can observe that the variables obtained by stepwise variable selection for the three models are somehow consistent, i.e. same direction for the coefficients and common variables across the models. As expected, the patients in the treatment group attain higher levels of vitamin D in their blood. In addition, baseline patients tend to have lower vitamin D in their blood as they grow older, which is explained by the fact that nursing mothers are recommended to take supplementary vitamin D, hence, when the child stops breastfeeding, his dietary intake of vitamin D reduces. However, we believe that this is not significant in later months due to the supplementary vitamin D intake for all patients. Moreover, dark skinned patients tend to have lower vitamin D in their body, which can be explained by slower creation of vitamin D by their skin from sun exposure. We also observe lower vitamin D levels in winter and higher levels in the summer, this is due to the less exposure to the sun in winter than summer. As expected, patients in the higher quartiles of dietary and supplementary vitamin D intake have higher amounts of vitamin D in their blood. Finally, we noticed that the daily use of the medication ICS tends to increase the amount of vitamin D.

Table 5.4: Transformed parameter estimates and their 95% confidence intervals for the Gamma GLM models for the amount of vitamin D at each visit.

Variable	Categories	$t = 0$	$t = 3$	$t = 6$
Intercept	-	$92.40\ (78.42, 108.97)$	$61.04\ (51.20, 73.08)$	$66.48\ (57.73, 77.39)$
Group	Placebo	-	reference	reference
	Treatment	-	$1.17\ (1.05, 1.30)$	$1.12\ (1.01, 1.24)$
Age	-	$0.93\ (0.89, 0.98)$	-	-
Fitzpatrick Scale	[1,2]	reference	reference	reference
	[3,4]	$0.87\ (0.77, 0.97)$	$0.89\ (0.79, 0.99)$	$0.85\ (0.76, 0.94)$
	[5,6]	$0.91\ (0.74, 1.12)$	$0.89\ (0.74, 1.08)$	0.85 (0.71 - 1.02)
Season	Fall	reference	-	-
	Winter	$0.87\ (0.77, 1.00)$	-	-
	Spring	-	-	reference
	Summer	-	-	$1.07\ (0.96, 1.21)$
Daily Dietary and Supplementary Vit. D	Q1: $[0 - 165]$	-	reference	reference
	Q2: $(165 - 215]$	-	$0.97\ (0.81, 1.17)$	$0.98\ (0.84, 1.13)$
	Q3: $(215 - 356]$	-	$1.02\ (0.88, 1.18)$	$1.05\ (0.89, 1.24$
	Q4: $(356 - \infty)$	-	$1.18\ (1.03, 1.34)$	$1.14\ (0.99, 1.32)$
Daily ICS Intake	No	-	reference	reference
	Yes	-	$1.18\ (1.03, 1.35)$	$1.14\ (1.00, 1.29)$

5.2 Fitting the Univariate Distributions

In this section, we provide details on the fitting of the univariate distributions needed for our proposed model. Our data consists of repeated measurements taken at different points in time for each participant in the pilot and bridge studies. We are interested in identifying the trend over time for each participant and also the variation between the participants. We have two response variables; the change in vitamin D between successive visits, $Y^{(1)}$, and the number of asthma attacks that occurred between successive visits and require the use of OCS, $Y^{(2)}$. In addition to the covariates discussed in Section 5.1, we also have a time covariate, t, and an indicator for each participant, ID. Table 5.5 provides details on the variables. Since we are considering data that happened between two time points, namely $[0,3]$ and $[3,6]$, we have to perform some data manipulation as explained below:

- Include only the participants that continued the study until the end,
- Use the average values between the two time points for Age and Z-score of BMI, and
- Use the midpoint of the dates between the visits, and accordingly specify the season

Note that most of the variables were regarding information over the past 3 months (Asthma Severity, SPF usage, Dietary and Supplementary vitamin D, Daily intake of ICS, minutes spent in the sun, and Sunny vacation), and hence for the intervals $[0,3]$ and $[3,6]$, we used the data from the second and third visits, respectively.

Table 5.5: Patients Characteristics for the longitudinal study. Numerical variables are associated with the mean and the 95% confidence interval.

Variable	Categories	$t = [0-3]$ $(n = 58)$	$t = [3-6]$ $(n = 58)$
Study	Pilot	18	18
	Bridge	40	40
Group	Placebo	32	32
	Treatment	26	26
Δ Vit. D in blood	-	7.04 $(-28.6, 42.7)$	1.73 $(-18.5, 22.0)$
# of Asthma Attacks	-	0.5 $(-0.93, 1.93)$	0.41 $(-0.81, 1.63)$
Age	-	2.96 $(0.88, 5.0)$	3.23 $(1.18, 5.28)$
Gender	Male	30	30
	Female	28	28
Z-score of BMI	-	0.56 $(-1.25, 2.36)$	0.54 $(-1.35, 2.43)$
Fitzpatrick Scale	$[1,2]$	35	35
	$[3,4]$	18	18

	$[5, 6]$	5	5
Season	Fall	9	-
	Winter	44	51
	Spring	4	3
	Summer	1	4
Asthma Severity	Persistent	37	31
	Episodic	21	27
Daily Dietary and Supplementary Vit. D	Q1	11	11
	Q2	8	14
	Q3	17	11
	Q4	22	22
Daily inhaled corticosteroids (ICS) intake	No	9	10
	Yes	49	48
Minutes spent in the sun per day in past 3 months	$\leq$ median (60 mins)	46	23
	$>$ median (60 mins)	12	35
SPF usage in past 3 months	< 30	49	10
	≥ 30	9	48
Sunny Vacation in past 3 months	No	50	58
	Yes	8	0

5.2.1 Fitting the Change in Vitamin D

Since we are working with 58 participants, observing the individual plot for each of them would be cumbersome, and hence we plot the trend of the mean of $Y^{(1)}$ across studies and groups over time, as shown in Figure 5.1. For the pilot participants, we can observe that the change in vitamin D is higher in the interval $[0, 3]$ than the interval $[3, 6]$, which is because at baseline, the vitamin D in the blood is tested prior to any bolus or daily vitamin D intake, however at $t = 3$ and $t = 6$, the participants have been exposed to daily vitamin D supplementary intake and the bolus that was taken at the prior visit. In addition, the participants from the treatment group have a significantly higher increase, as expected. On the other hand, for the bridge study, we observe that the participants from the placebo group start with a slightly negative change (or almost no change in the average amount of vitamin D), which can be attributed to less exposure to the sun in the winter season. It is then followed by a slightly positive change (or almost no change) for the second interval. Regarding the treatment group of the bridge study, they experience a positive change in vitamin D, due to the bolus intake, followed by no change.

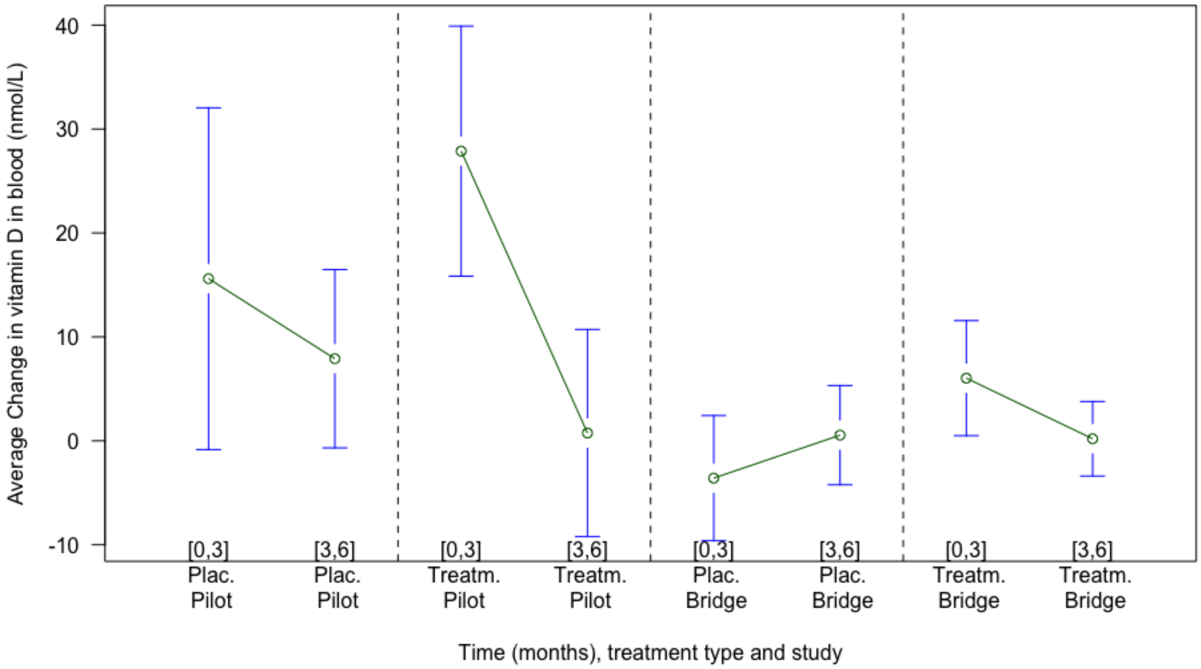

Figure 5.1: The mean of $Y^{(1)}$ over time split by treatment type and study.

We are interested in identifying the trend of the change in vitamin D in the blood, namely $Y^{(1)}$ over time within each participant and to compare this trend with other participants. This requires the use of a mixed-effects model. Since $Y^{(1)}$ is continuous and not strictly positive, we fit a GLMM model with Normal family. In our fitted model, we start with all the available covariates, in addition to two random effects for each participant. The random effects for a certain participant represent the deviation from the population values for the intercept and the slope of that participant's time trend. In other words, the model calculates the value of the intercept and slope of the time variable for the population and then each participant has two additional parameters that explain how he deviates from that population. The additional parameters per participant are one for the intercept, and another for the slope. Further details on random effects are available in Section 1.3.

We then perform variable selection procedures to obtain a parsimonious model. We confirm that the new model, obtained from stepwise variable selection, is not significantly different from the original model, with all predictors, by LRTs, where we obtain a p-value of 0.95. In addition, we obtain an estimate of 0 for the variance of the random effects. This does not imply a lack of variation between the participants, but rather that level of variability between participants is not sufficient to require adding random effects to the model. In other words, fitting a model with only fixed effects through OLS is equivalent to this model. This is confirmed by performing LRTs, in which we obtain a p-value of 1 when we drop the random effects terms from the model.

Table 5.7 provides the coefficients of the fixed effects for the final model. Note that since we used a Normal distribution to model the data, the link function used is the identity link, which results in additive coefficients. The overall average decrease in vitamin D levels that we visually observed in Figure 5.3 is confirmed by the negative intercept value. Additionally, the treatment participants have a higher change in their vitamin D levels, as expected. We also notice that older patients have a positive increase in the change in their vitamin D levels and that participants who are diagnosed with episodic asthma condition experience lower changes in their vitamin D. Additionally, higher amounts of dietary and supplementary vitamin D results in higher positive change in the amount of vitamin D in the blood. Finally, over time, the change in vitamin D becomes smaller, which was already deduced from the visual representation. In addition, we observe the Q-Q plot of the residuals and we obtain a visual confirmation that our model provides a good fit of the data, except for 2 observations.

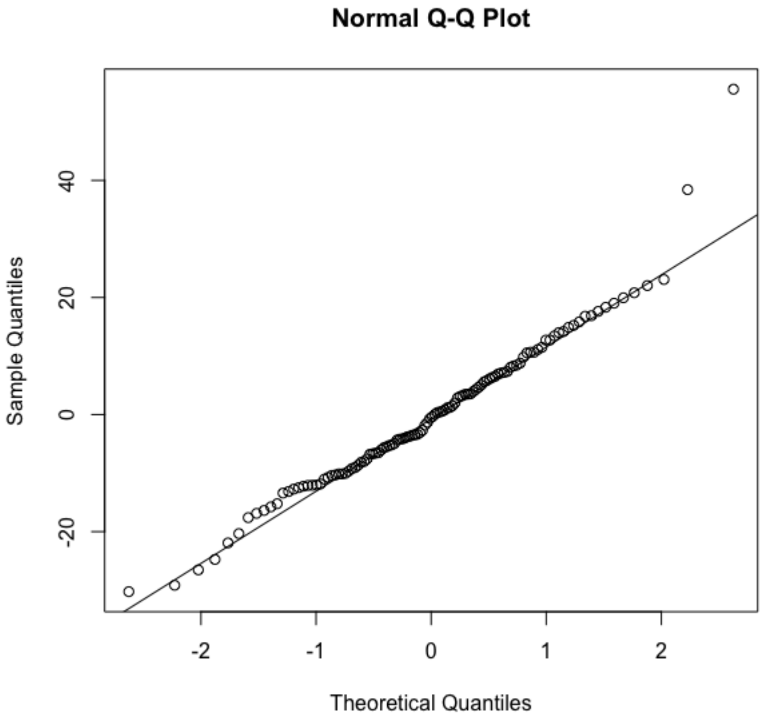

Figure 5.2: Q-Q plot of the residuals of the model with the parameters specified in 5.7

Table 5.7: Parameter estimates and their 95% confidence intervals for the OLS model for the change in the amount of vitamin D between visits.

Variable	Categories	Fixed Effects Coefficients
Intercept	-	$-2.78\ (-15.92, 10.35)$
Group	Placebo	reference
	Treatment	$7.09\ (1.84, 12.33)$
Age		$2.92\ (0.34, 5.50)$
Asthma Severity	Persistent	reference
	Episodic	$-4.59\ (-9.79, 0.60)$
Daily Dietary and Supplementary Vit. D	Q1	reference
	Q2	$-1.64\ (-9.76, 6.49)$
	Q3	$1.87\ (-6.13, 9.88)$
	Q4	$12.12\ (4.96, 19.28)$
Time	-	$-1.75\ (-3.44, -0.06)$

5.2.2 Fitting the Number of Asthma Attacks requiring the use of OCS

We start by visually observing the available data and calculate the mean of $Y^{(2)}$ across the studies and groups, which is shown in Figure 5.3. We observe that in general, there is a slight decrease in the mean of $Y^{(2)}$ over time.

Our variable of interest is a discrete variable that represents the count of events that occur per participant, which means that a Poisson Distribution is an appropriate modeling choice. We are interested in identifying the behavior of $Y^{(2)}$ over time for the population, and identify any variation among patients. Therefore, we fit a GLMM model with Poisson family. Similar to the initial model used to fit $Y^{(1)}$, this model includes two random effects per participant; one for the variation in the intercept from the population values, and the other for the slope of the time trend. We include all the covariates in the model, and then perform methods of variable selection to exclude the variables that do not add significant value to the predictions of the model. This was confirmed by doing LRTs and obtaining a p-value of 0.98.

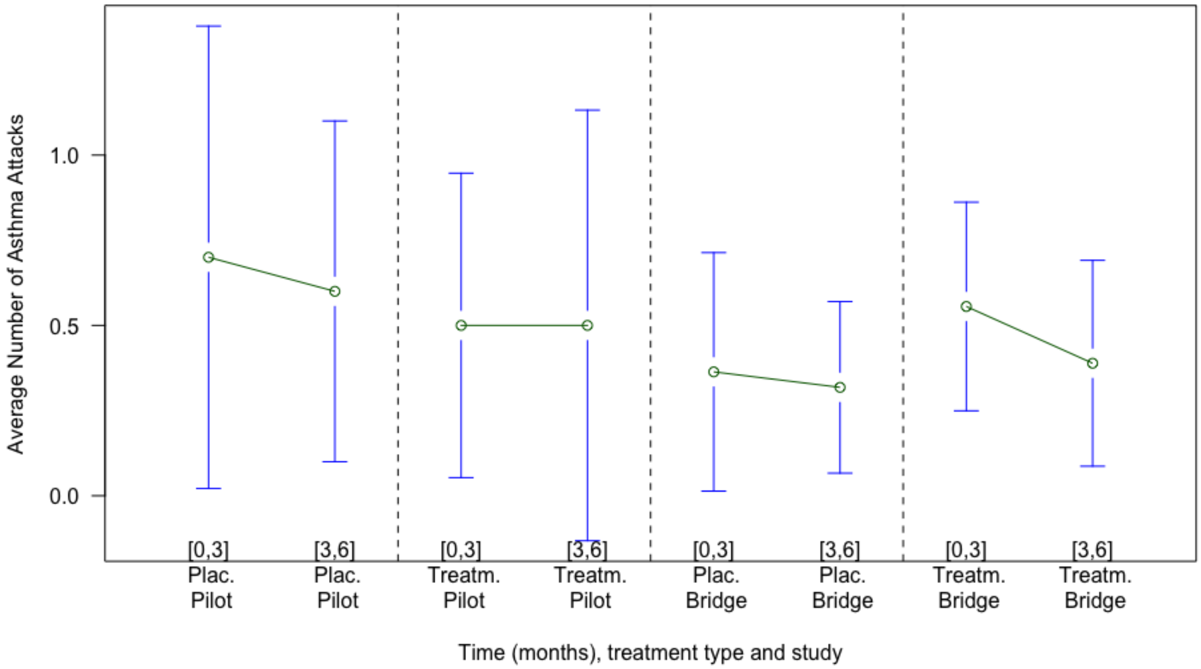

Figure 5.3: The mean of $Y^{(2)}$ over time split by treatment type and study.

Unlike the model for $Y^{(1)}$, we observe variation in the intercept among participants, with a small standard deviation of 0.34. However, the estimate for the random effect for the time variable is 0, which implies that the level of variability over time across the participants is not sufficient to require adding random effects for the slope of the time variable to the model. Therefore, a model with only random effects for the intercept will produce the same results. Figure 5.4 represents the estimates of the random effects for the intercept of each participant in the model versus quantiles of the standard normal distribution. As observed, on average, the variation of the participants lies between $[-0.2, 0.3]$ away from the mean of the population. The 95% confidence interval for each participant is provided in the figure.

Additionally, we perform LRT to compare the model with the random effect, to a model with only fixed effect. A p-value of 0.54 confirms that the fixed effects only model (GLM) does not lose significance value compared to the mixed effects model (GLMM), therefore, the parsimonious model is the GLM model.

Table 5.8 provides the transformed coefficients of the fixed effects for model. Note that those coefficients are multiplicative, as explained in Eq. 1.2.8. We notice that the participants in the bridge group have significantly lower number of attacks, compared to the pilot group, and that females also experience less number of asthma attacks. In addition, as expected, participants who were diagnosed with episodic asthma severity have lower number of attacks compared to patients with persistent asthma severity.

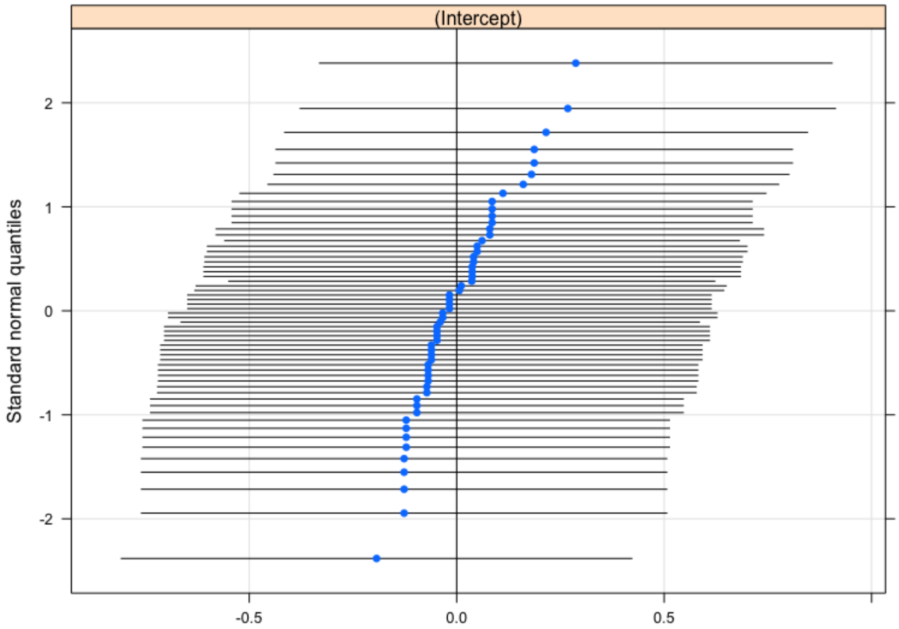

Figure 5.4: 95% confidence intervals of the estimates of random effects for the intercept in the mode of $Y^{(2)}$ versus quantiles of the standard normal distribution.

Table 5.8: Parameters estimates, transformed parameters estimates and their 95% confidence intervals for the GLM model for the number of asthma attacks that require the use of OCS between visits.

Variable	Categories	Fixed Effects Coefficients	Transformed Fixed Effects Coefficients
Intercept	-	$0.26\ (-0.39$ - $0.86)$	1.30 (0.68 - 2.36)
Study	Pilot	reference	reference
	Bridge	$-0.79\ (-1.39$ - $-0.17)$	0.45 (0.25 - 0.85)
Gender	Male	reference	reference
	Female	$-0.76\ (-1.38$ - $-0.16)$	0.47 (0.25 - 0.85)
Asthma Severity	Persistent	reference	reference
	Episodic	$-0.62\ (-1.26$ - $-0.04)$	0.54 (0.28 - 0.96)

5.3 Fitting the Joint Distribution

Prior to fitting the joint distribution, we would like to mention that having only two observations per patient may result in inaccurate estimations for the dependence structure. From the initial graphs and the modeling of the marginal distributions, we observe smaller number of asthma attacks over time, which can confirm the positive effect of vitamin D. However, we also observe that the change in vitamin D becomes smaller over time, which is attributed to the big positive change from baseline, and then followed by a smaller change in vitamin D. Therefore, the combined effect shows that bigger amount of change in vitamin D corresponds to more asthma attacks. This counterintuitive result is due to having only 2 observations per patient and a longer study with more time points will produce more accurate measures and confirm the original hypothesis. However, we will proceed with the modeling of the dependence between $Y^{(1)}$ and $Y^{(2)}$, even though the results may be inaccurate due to the shortcomings of the data. Note that the proposed model can have more than 2 responses, however, our application is constrained by the availability of the data, and hence we only need to fit 1 copula, namely $C_1^{(\theta_1)}$.

We have fitted the models for $Y^{(1)}$ and $Y^{(2)}$, which are an OLS and GLM with Poisson distribution and log link function, respectively. Accordingly, we calculate the vectors $\hat{\boldsymbol{\epsilon}}^{(1)}$ and $\hat{\boldsymbol{\epsilon}}^{(2)}$, as per the formulas indicated in Eq. 4.0.2 and 4.0.3, respectively. Let $\hat{U}_{ij}^{(1)} = F_{n1}(\epsilon_{ij}^{(1)})$ and $\hat{U}_{ij}^{(2)} = F_{n2}(\epsilon_{ij}^{(2)})$ represent the empirical marginal cdf of the standardized residuals for the j^{th} time point of the i^{th} participant. Note that $F_{n1}(\epsilon_{ij}^{(1)}) = \frac{R_{ij}}{n+1}$ and $F_{n2}(\epsilon_{ij}^{(2)}) = \frac{S_{ij}}{n+1}$, where R_{ij} and S_{ij} represent the rank of the observation and we divide by $(n+1)$ to ensure that all ranks lie strictly between 0 and 1.

Table 5.9 provides the estimates for the dependence coefficients. Additionally, the null hypothesis of independence is tested by checking if the empirical estimates of the dependence coefficients are significantly different from 0. We notice that independence is rejected under the rank-based tests of independence.

Table 5.9: Estimates for dependence measures between $\hat{\epsilon}_{ij}^{(1)}$ and $\hat{\epsilon}_{ij}^{(2)}$.

Measure	$\rho_P(\hat{\epsilon}_{ij}^{(1)}, \hat{\epsilon}_{ij}^{(2)})$	$\rho_S(\hat{\epsilon}_{ij}^{(1)}, \hat{\epsilon}_{ij}^{(2)})$	$\tau(\hat{\epsilon}_{ij}^{(1)}, \hat{\epsilon}_{ij}^{(2)})$
Estimate	0.137	0.158	0.126
p-value	0.143	0.089	0.088

Table 5.10 provides the estimates of the dependence parameter and its standard deviation for six copula families fitted to the pair $\hat{\epsilon}_{ij}^{(1)}$ and $\hat{\epsilon}_{ij}^{(2)}$. Red p-values indicate rejected copulas at significance level $\alpha = 10\%$ and bold text indicates the best fit copula. Note that the degrees of freedom for the t-copula have been estimated along with the dependence parameter. The Cramér Von Mises goodness of fit test rejects only the Clayton copula at a significance level $\alpha = 10\%$. However, by observing their AIC_c and BIC values, we notice that out of the accepted models, the Gumbel copula provides the best fit for the data. This has also been confirmed by using the *BiCopSelect* function from the *VineCopula* R package, which tests a wide variety of possible copulas.

Table 5.10: Estimates of copula parameters, p-values of goodness of fit test, AIC_c and BIC values.

Copula	Dependence Parameter	Standard Deviation	p-value	AIC_c	BIC
Gauss	0.193	0.107	0.253	-0.213	2.505
t_6	0.187	0.123	0.347	0.874	6.275
Clayton	0.283	0.184	0.087	0.306	3.025
Frank	1.087	0.648	0.177	-0.490	2.228
Gumbel	**1.125**	**0.080**	**0.441**	**-0.539**	**2.179**
Galambos	0.353	0.096	0.242	-0.143	2.575

Therefore, the dependence structure between $Y^{(1)}$ and $Y^{(2)}$ is presented by a Gumbel copula with dependence parameter $\hat{\theta} = 1.125$, which corresponds to $\hat{\tau} = 0.11$.

5.4 Predictive Modeling

As a preventative strategy, one must be able to forecast the expected number of asthma attacks and its standard deviation, given a certain value for the change in vitamin D. The fitted marginals obtained in Sections 5.2.1 and 5.2.2 are represented in terms of the predictors $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ and the regression parameters $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$ as follows:

$$\begin{aligned}\hat{y}_{ij}^{(1)} &= -2.78 + 7.09\left(\mathbf{1}_{\left(x_{ij}^{(1)}=Treatment\right)}\right) + 2.92x_{ij}^{(2)} + -4.59\left(\mathbf{1}_{\left(x_{ij}^{(3)}=Episodic\right)}\right)\\ &\quad - 1.64\left(\mathbf{1}_{\left(x_{ij}^{(4)}=Q2\right)}\right) + 1.87\left(\mathbf{1}_{\left(x_{ij}^{(4)}=Q3\right)}\right) + 12.12\left(\mathbf{1}_{\left(x_{ij}^{(4)}=Q4\right)}\right) - 1.75t,\end{aligned}$$

and

$$\hat{y}_{ij}^{(2)} = \exp\left\{0.26 - 0.62\left(\mathbf{1}_{\left(x_{ij}^{(3)}=Episodic\right)}\right) - 0.79\left(\mathbf{1}_{\left(x_{ij}^{(5)}=Bridge\right)}\right) - 0.76\left(\mathbf{1}_{\left(x_{ij}^{(6)}=Female\right)}\right)\right\}$$

where $x_{ij}^{(p)}$ represents the j^{th} observation of the i^{th} participant for the p^{th} predictor, where $p = 1, \ldots, 6$ represents the group, age, asthma severity, daily dietary and supplementary intake of vitamin D, study and gender, respectively.

In addition,

$$\hat{\epsilon}_{ij}^{(1)} = \frac{y_{ij}^{(1)} - \hat{y}_{ij}^{(1)}}{\hat{\sigma}^{(1)}}, \text{ and } \hat{\epsilon}_{ij}^{(2)} = \frac{y_{ij}^{(1)}}{\hat{\lambda}_{ij}^{(2)}}, \tag{5.4.1}$$

where $\hat{\sigma}^{(1)} = 13.32$.

As shown in Section 5.3, the dependence structure between $\hat{\boldsymbol{\epsilon}}^{(1)}$ and $\hat{\boldsymbol{\epsilon}}^{(2)}$ is represented by a Gumbel copula with dependence parameter $\hat{\theta_1} = 1.125$. Therefore, the proposed model is reduced to

$$\begin{aligned}F\left(\epsilon^{(1)}, \epsilon^{(2)}|\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \hat{\mu}^{(1)}, \hat{\sigma}^{(1)}, \hat{\lambda}^{(2)}\right) &= C_1^{(\hat{\theta}_1)}(U^{(1)}, U^{(2)})\\ &= \exp\left\{-\left[(-\ln U^{(1)})^{\theta_1} + (-\ln U^{(2)})^{\theta_1}\right]^{1/\theta_1}\right\},\end{aligned}$$

where

$$U^{(1)} = F_1\left(\epsilon^{(1)}|\boldsymbol{X}^{(1)}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\mu}^{(1)}, \hat{\sigma}^{(1)}\right), \text{ and } U^{(2)} = F_2\left(\epsilon^{(2)}|\boldsymbol{X}^{(2)}, \hat{\boldsymbol{\beta}}^{(2)}, \hat{\lambda}^{(2)}\right),$$

represent the marginal cdf for $\hat{\boldsymbol{\epsilon}}^{(1)}$ and $\hat{\boldsymbol{\epsilon}}^{(2)}$, respectively, which are approximately

$$\epsilon^{(1)}|\boldsymbol{X}^{(1)}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\mu}^{(1)}, \hat{\sigma}^{(1)} \sim N(0,1)$$

and

$$\epsilon^{(2)}|\boldsymbol{X}^{(2)}, \hat{\boldsymbol{\beta}}^{(2)}, \hat{\lambda}^{(2)} \sim \text{Poisson}(1).$$

The pdf for this copula is given by

$$\begin{aligned} f\left(\epsilon^{(1)}, \epsilon^{(2)}|\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \hat{\mu}^{(1)}, \hat{\sigma}^{(1)}, \hat{\lambda}^{(2)}\right) =& C_1^{(\hat{\theta}_1)}(U^{(1)}, U^{(2)}) \\ &\times f_1\left(\epsilon^{(1)}|\boldsymbol{X}^{(1)}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\mu}^{(1)}, \hat{\sigma}^{(1)}\right) \\ &\times f_2\left(\epsilon^{(2)}|\boldsymbol{X}^{(2)}, \hat{\boldsymbol{\beta}}^{(2)}, \hat{\lambda}^{(2)}\right), \end{aligned}$$

where f_1 and f_2 represent the marginal pdf for $\hat{\boldsymbol{\epsilon}}^{(1)}$ and $\hat{\boldsymbol{\epsilon}}^{(2)}$, respectively.

The conditional cdf for the number of asthma attacks given a fixed value of change in vitamin D is defined as

$$\begin{aligned} C_{U^{(2)}|U^{(1)}}(u^{(2)}|u^{(1)}) &= \frac{\partial}{\partial u^{(1)}} C_1^{(\theta_1)}(u^{(1)}, u^{(2)}) \\ &= C_1^{(\theta_1)}(u^{(1)}, u^{(2)}) \frac{(-\ln u^{(1)})^{\theta_1 - 1}}{u^{(1)}} \left[(-\ln u^{(1)})^{\theta_1} + (-\ln u^{(2)})^{\theta_1}\right]^{\frac{1}{\theta_1} - 1}. \end{aligned}$$

Steps 1-5 in the following simulation procedure are performed to obtain 1 value of the expected number of asthma attacks, for a given value of the change in vitamin D. It is then repeated for several values of the change in vitamin D, as explained below.

1. Specify a given value of $u^{(1)} \sim U(0,1)$ and simulate 5000 observations for $V^{(2)}$, where $V^{(2)} \sim U(0,1)$,

2. Let $U^{(1)} = (u^{(1)}, \ldots, u^{(1)})$ and solve for $U^{(2)}$, such that $C_{U^{(2)}|U^{(1)}}(u^{(2)}|u_j^{(1)}) = v_j^{(2)}, \forall j = 1, \ldots, 5000$. This requires numerical optimization methods,

3. Transform $U^{(1)}$ and $U^{(2)}$ into realizations for the residual values $\hat{\epsilon}^{(1)}$ and $\hat{\epsilon}^{(2)}$ by using the appropriate quantile transformation,

4. Transform the simulated residual values into realizations of $Y^{(1)}$ and $Y^{(2)}$ by using the inverse of Eq. 5.4.1,

5. Calculate the expected value of each variable, where $Y^{(1)}$ will remain constant for each iteration, and

6. Repeat the procedure for a sequence of 100 values of $u_i \in]0,1[$.

Conclusion

In this thesis, we introduced a methodology to estimate the association between responses of longitudinal data. The suggested model incorporates nested copulas in a vine structure, as shown in Figure 4.1. The approach was illustrated by using data from two medical studies performed on preschoolers diagnosed with recurrent, moderate or severe asthma. GLMM models were fit to two responses and covariates, where we consider a random effect for the intercept and the slope of the time component for each participant. Variable and model selection criteria were used to eliminate the variables and models that poorly explained the data, and to choose the best fit model without losing model predictability. LRTs were used to test the significance of the random effects to the models, where the optimal models excluded the random effects, i.e. the optimal models include only fixed effects.

The chosen models were used as the marginals of the copula. The residuals from each model were used to isolate it from the effect of the covariates on the data. The pairwise dependence between the change responses were investigated and modeled by using rank-based procedures. Standard tools for bivariate copula selection, estimation and validation were used. The copula fitting procedure can be repeated iteratively for as many variables as needed such that if we have r responses, we will fit $r-1$ bivariate copulas. Additionally, the dependence model is critical in the estimation of the expected value of a response variable given a fixed value of the other response variable. The most significant limitation of this methodology is that it requires the use of identically distribution responses (or residuals), which might not be the standard case. Additionally, to ensure the accuracy of the model, we need a significant number of repeated observations in the longitudinal data.

Overall, we have established in this thesis that the suggested modeling technique provides great flexibility to model longitudinal data with multiple responses. First, the marginals can be fit by different regression models for each response. Second, bootstrap iterations can be easily performed to minimize uncertainty in the estimation of the parameters. Finally, in addition to predictions for each marginal, predictions can be made for the joint distribution, and more importantly, the conditional distribution of a marginal given the others.

Bibliography

Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198, 2009.

Anas Abdallah, Jean-Philippe Boucher, and Hélène Cossette. Modeling dependence between loss triangles with hierarchical archimedean copulas. *ASTIN Bulletin: The Journal of the IAA*, 45(3):577–599, 2015.

H Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*. Academiai Kiado, 1973.

H Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Michael Patrick Allen. Assumptions of ordinary least-squares estimation. *Understanding Regression Analysis*, pages 181–185, 1997.

Tim Bedford and Roger M Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence*, 32(1): 245–268, 2001.

Tim Bedford and Roger M Cooke. Vines: A new graphical model for dependent random variables. *Annals of Statistics*, pages 1031–1068, 2002.

George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71 (356):791–799, 1976. ISSN 01621459. URL http://www.jstor.org/stable/2286841.

Norman E Breslow. Generalized linear models: checking assumptions and strengthening conclusions. *Statistica Applicata*, 8(1):23–41, 1996.

Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.

Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

Marie-Pier Côté, Christian Genest, and Anas Abdallah. Rank-based methods for modeling dependence between loss triangles. *European Actuarial Journal*, 6(2):377–408, 2016.

Edward W Frees and Ping Wang. Credibility using copulas. *North American Actuarial Journal*, 9(2):31–48, 2005.

Edward W Frees and Ping Wang. Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics*, 38(2):360–373, 2006.

Christian Genest and R Jock Mackay. Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics*, 14(2):145–159, 1986.

Christian Genest, Bruno Rémillard, and David Beaudoin. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and economics*, 44(2):199–213, 2009.

Kilani Ghoudi, Abdelhaq Khoudraji, and Et Louis-Paul Rivest. Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles. *Canadian Journal of Statistics*, 26(1):187–197, 1998.

Gordon Gudendorf and Johan Segers. Extreme-value copulas. *Copula theory and its applications*, pages 127–145, 2010.

Wassily Hoeffding. Massstabinvariante korrelationstheorie. *Schriften des Mathematischen Seminars und des Instituts für Angewandte Mathematik der Universität Berlin*, 5(3):181–233, 1940.

Marius Hofert and David Pham. Densities of nested archimedean copulas. *Journal of Multivariate Analysis*, 118:37–52, 2013.

Marius Hofert, Martin Mächler, et al. Nested archimedean copulas meet r: The nacopula package. *Journal of Statistical Software*, 39(9):1–20, 2011.

Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

Megan E Jensen, Genevieve Mailhot, Nathalie Alos, Elizabeth Rousseau, John H White, Ali Khamessan, and Francine M Ducharme. Vitamin d intervention in preschoolers with viral-induced asthma (diva): a pilot randomised controlled trial. *Trials*, 17(1):353, 2016.

Harry Joe. Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, pages 120–141, 1996.

Harry Joe. *Multivariate models and multivariate dependence concepts.* CRC Press, 1997.

Harry Joe. *Dependence modeling with copulas.* CRC Press, 2014.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Philippe Lambert. Modelling irregularly sampled profiles of non-negative dog triglyceride responses under different distributional assumptions. *Statistics in medicine*, 15(15):1695–1708, 1996.

Philippe Lambert and Francois 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*, 21(21):3197–3217, 2002.

Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3): 285–292, 1984.

Alexander J McNeil. Sampling nested archimedean copulas. *Journal of Statistical Computation and Simulation*, 78(6):567–581, 2008.

Alexander J McNeil and Johanna Nešlehová. Multivariate archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. *The Annals of Statistics*, pages 3059–3097, 2009.

Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: Concepts, techniques and tools.* Princeton university press, 2015.

Steven G Meester and Jock Mackay. A parametric model for cluster correlated categorical data. *Biometrics*, pages 954–963, 1994.

M Mittlbock and Harald Heinzl. Pseudo r-squared measures for generalized linear models. In *Proceedings of the 1st European Workshop on the Assessment of Diagnostic Performance, Milan, Italy*, pages 71–80, 2004.

Roger B Nelsen. An introduction to copulas, volume 139 of lecture notes in statistics, 1999.

Jose Juan Quesada-Molina. A generalization of an identity of hoeffding and some applications. *Statistical Methods & Applications*, 1(3):405–411, 1992.

Marco Scarsini. On measures of concordance. *Stochastica*, 8(3):201–218, 1984.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.

Peng Shi and Edward W Frees. Dependent loss reserving using copulas. *ASTIN Bulletin: The Journal of the IAA*, 41(2):449–486, 2011.

M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

Michael Smith, Aleksey Min, Carlos Almeida, and Claudia Czado. Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association*, 105(492):1467–1479, 2010.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

Stephen M Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, pages 465–474, 1981.

Walter W Stroup. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press, 2012.

Who et al. Who multicentre growth reference study group: Who child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development. *Geneva: WHO*, 2007, 2006.