Copula-based joint regression models for longitudinal data

By

Aydin Hibbert

A thesis submitted to Macquarie University

for the degree of Master of Research

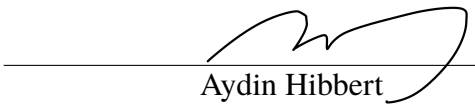
Department of Mathematics and Statistics

April 2019

Examiner's Copy

Typeset in LaTeX 2_ε.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.



Aydin Hibbert

Abstract

A popular approach to modelling longitudinal observations is using mixed models with random effects for subjects. Recent developments in joint regression modelling presents an alternative approach to longitudinal analysis which utilises copulas to flexibly model dependence structures for correlated data. The performance of copula-based regression models has, to date, not been quantified in comparison to random effect based models and other popular methods.

This thesis provides a preliminary analysis of some of the situations in which copula-based joint regression models may be more appropriate than mixed models for longitudinal regression analysis. The models are compared across a range of simulated longitudinal datasets generated from a flexible bivariate distribution, and applied to three real-world datasets. The results of the analysis indicate that in cases where the outcome variables marginal distributions are skewed and there is rank correlation between regression outcome variables, measured by Kendall's tau, mixed models provide biased parameter estimates while copula-based joint regression models provide unbiased estimates with generally lower standard errors than other alternative methods such as generalised linear models or generalised estimating equations.

Contents

1 Introduction

1.1 Background

Dependence between random variables often develops when repeated observations are taken from a sampling unit, for example, cholesterol levels over time for a single patient, or where observations are made from similar sampling units, for example, grades from students in the same classroom. If independence is assumed between these correlated variables, observations will exhibit lower than expected variance which causes both bias and inefficiency in coefficient estimation (Laird and Ware, 1982).

To understand the true effect of covariates, it is of interest to understand both the structure of the dependence, i.e. which observations are more or less dependent, and how this dependence interacts with covariates, i.e. what factors drive different dependencies for different observations. A number of approaches have been developed to date for accounting for dependence between observations in univariate regression, the most popular of which being random effect models (Laird and Ware, 1982) and generalised estimating equations (GEE) (Liang and Zeger, 1986). These methods introduce an adjustment to univariate methods which account for the dependence structure between observations. Both frameworks have been extended substantially since their early inception to apply to a much broader range of response distributions and covariate shapes. More recently, approaches for joint modelling of multivariate dependent data in a regression framework have emerged through the use of copulas, which provide an alternative to the former models.

Models for Correlated Data

Initially developed for only normal response distributions, random effects for covariate intercepts and slopes have been incorporated in multiple generalised regression frameworks including generalised additive models (GAMs) (Hastie and Tibshirani, 1990) and generalised additive models for location scale and shape (GAMLSS) (Stasinopoulos et al., 2017), making them a very popular tool for modelling correlated data.

Generalised additive models for location scale and shape (GAMLSS) (Stasinopoulos et al., 2017) provide a flexible regression framework which extends generalised additive modelling (Hastie and Tibshirani, 1990) to allow for multiple parameters of a target distribution to be modelled simultaneously. In addition, the range of potential distributions that can be fit is extended beyond the exponential family to any distribution with computable derivatives.

Generalised estimating equations (GEEs) have also been extended to apply to a much broader range of potential use cases. Vector generalised additive models (Yee and Wild, 1996, Wild and Yee, 1996) have extended generalised estimating equations to incorporate additive covariates and allow for multivariate fitting of dependent variables while focussing on the marginal distributional fits.

Adjustments for dependence incorporated in GEEs and models with random effect terms are an approximate approach for accounting for multivariate data features in a univariate regression framework. A copula-based joint regression approach directly models the full multivariate distribution, allowing for an extremely high level of flexibility in describing the underlying data. While copula approaches are available for multivariate distributions of any dimension, this thesis will focus on the bivariate case.

Copula Regression

Copulas provide a convenient method for deconstructing a bivariate distribution into a combination of two independent marginal distributions and a copula function (Nelsen, 2007, Trivedi and Zimmer, 2007). If the marginal distributions are continuous, the copula function will be unique and if either of the margins are discrete, the copula can be uniquely determined. By Sklar's theorem (Sklar, 1973) any continuous bivariate distribution can be represented by two independent marginal distribution functions and a copula function defining the dependence structure.

Copulas have been used broadly in economics and finance time series analysis, particularly in portfolio risk (Palaro and Hotta, 2006, Pitt et al., 2006) and insurance loss estimation (Krämer et al., 2013). However, much of the analysis has focussed on Gaussian and unconditional copulas (Kolev and Paiva, 2009), which have limited applications due to the highly restrictive assumptions.

One of the key developments in copula regression was Patton's (2006) introduction of a framework for conditioning a copula on a variable (covariate). This allows for dependence structures to be modelled after the effect of covariates are removed, leaving the true residual dependence structure and significantly expanding the applications of copulas in a regression context. This method is referred to as conditional copula regression. A large amount of literature has focussed on the development of likelihood-based tests for the existence and significance of covariates in a conditional copula regression framework (Acar et al., 2011, Gijbels et al., 2011, Acar et al., 2013, Craiu and Sabeti, 2012, Sabeti et al.).

Generalised Additive Copula Regression

Marra and Radice (2017) and Vatter and Chavez-Demoulin (2015) introduce the first approaches to flexible copula-based joint regression. Both approaches combine univariate models for marginal distributions, generalised additive models for location, scale and shape (gamlss) (Stasinopoulos et al., 2017) and generalised additive models (GAM) (Hastie and Tibshirani, 1990) respectively, with a fit for a copula to capture the dependence structure between the variables of interest.

Marra and Radice (2017) and Vatter and Chavez-Demoulin (2015) differ in their approach to optimisation of the joint likelihood function, with Vatter and Chavez-Demoulin (2015) opting to maximise marginal and copula likelihoods separately while Marra and Radice (2017) introduce a simultaneous estimation approach for the joint likelihood of the copula and marginal distributions.

Klein et al. (2015) have also introduced an alternative approach to Marra and Radice (2017) and Vatter and Chavez-Demoulin (2015) for multivariate regression which incorporates copula methods into generalised estimating equations, building on vector generalised additive models developed by Wild and Yee (1996).

Purpose of this thesis

This thesis focuses on the applications of copula-based joint regression models for correlated data, with direct comparisons to current popular methods for adjusting for variable dependence in univariate regression, in particular, incorporating random effect terms in mixed models and the use of generalised estimating equations.

There is a potential for copula-based joint regression approaches to be extended to the multivariate case, but analysis of the efficiency of copula based models in higher than two dimensions falls outside the scope of this investigation.

1.2 Models for correlated data

1.2.1 Random effect models

Random effect models have significantly evolved from their initial inception by Laird and Ware (1982), now being incorporated in the vast majority of generalised regression toolsets. The core component of these models is the introduction of a random variable which accounts for within-subject variation and is conditioned on in the model definition.

Linear Mixed Models

Laird and Ware (1982) introduced the first random effect models: a general family of two-stage models for adjusting for the correlation of repeated observations in linear models using random effect terms. The model explicitly captures individual and population characteristics separately.

For a set of m sampling units, $i = 1, ..., m$, which are measured at n_i time points, $j = 1, ..., n_i$, we consider the outcome variable $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})$. Then for each sampling unit we have the model equation:

$$y_i = x_i\alpha + z_i b_i + e_i$$

where α is a $p \times 1$ vector of population parameters, $b_i \sim N(0, D)$ is a $m \times 1$ vector of independent individual random effects, x_i is the $n_i \times p$ design matrix of covariates linked to α, z_i is the $n_i \times m$ design matrix of covariates linked to b_i, and $e_i \sim N(0, R_i)$ and are independent error terms.

In this model, each of the y_i are assumed to be marginally independent normal with mean vector $x_i\alpha$ and covariance matrix $\theta_i = R_i + z_i D z_i^T$.

Model estimation uses a two-stage method and is fit using expectation maximisation, where

individual characteristics are considered the missing data. In the first stage of the model, population parameters, individual effects and within-subject variation are fit, followed by the second stage where between-subject variation is fit.

Stage 1 - b_i and α are fit as fixed effects with e_i treated as an independent normal random variable. This allows for an initial model fit assuming independent observations.

Stage 2 - α is treated as fixed and b_i and e_i are allowed to vary to account for correlation between observations within a sampling unit. b_i is fit as a normally distributed random variable centred at zero with covariance matrix D, which is size $m \times m$ and positive definite. e_i is fit as a normally distributed random variable with a covariance matrix R_i which is size $n_i \times n_i$ and positive definite.

Laird and Ware (1982) demonstrate that a Bayesian approach to estimation leads to parameter estimates and variances which are identical to those which are achieved through direct sampling of the expected distributions.

Generalised Linear Mixed Models

Breslow and Clayton (1993) expanded upon Laird and Ware (1982)'s linear mixed models by introducing an approach for inference using random effects for any exponential family linear model. In line with a generalised linear modelling framework, generalised linear mixed models (GLMMs) introduce a link function and variance function to the formulation of linear mixed models.

For a given random effect b_i and response variable y_i for groups $j = 1, .., n_i$ related to explanatory vectors x_i and z_i, the model has

$$E(y_i|b) = \mu_i = h(x_i\alpha + z_i b_i), \quad var(y_i|b) = \phi V(\mu_i), \quad g(\mu_i) = \eta_i, \quad \eta_i = x_i\alpha + z_i b_i,$$

where $V(\cdot)$ is the variance function which depends on μ_i, ϕ is the dispersion parameter, and $g(\cdot)$ is a link function which relates the linear predictor η_i to the conditional mean μ_i and $h(\cdot) = g^{-1}(\cdot)$ is the inverse link function,.

As in linear mixed models, b has a $m-$dimensional multivariate normal distribution with mean 0 and $m \times m$ covariance matrix D.

A key contribution of Breslow and Clayton (1993) is in the introduction of the use of penalised quasi-likelihood as an approximate procedure for inference in GLMMs. A quasi-likelihood function is required when fitting distributions in the presence of overdispersion which presents in the full exponential family of models. The difficulty in attempting to optimise the full likelihood directly lies in the multiple integrations required within the full likelihood function and its derivatives.

Breslow and Clayton (1993) describe an approximation to the quasi-likelihood function which provides a pseudo quasi-likelihood approach for optimising the likelihood function more efficiently, resulting in score equations which can be solved as an iterated weighted least squares problem using Fisher scoring.

Random effects have also been incorporated into recent generalised modelling frameworks, including generalised additive models (Hastie and Tibshirani, 1990) which allows for smooth terms of explanatory variables to be incorporated as fixed effects alongside random effects; and generalized additive models for location shape and scale (Stasinopoulos et al., 2017) which provide the ability to model multiple parameters of a target distribution beyond the mean, and vastly increase the range of potential outcome distributions which can be modelled.

1.2.2 Marginal models

Marginal models are an alternative approach to the use of random effect terms for adjusting for dependence in a univariate regression framework. The key difference between marginal models and random effect models is that marginal models do not make any assumption surrounding the distributional properties of the dependence. The response variable only relies on fixed covariates and does not have a random effect term.

Generalised Estimating Equations

Liang and Zeger (1986) introduced generalised estimating equations (GEEs) for the estimation of covariate effects for longitudinal data. GEEs differ from random effects in that a joint distribution for repeated observations is not introduced, instead weak convergence assumptions for the joint distribution are imposed which give consistent estimates of regression parameters and a model is only fit for the marginal distributions.

The marginal density of y_{ij} is assumed to be exponential family in the following formulation

$$f(y_{ij}) = \exp[\{y_{ij}\theta_{ij} - a(\theta_{ij}) + b(y_{ij})\}\phi]$$

so the first two moments of y_{it} are

$$E(y_{ij}) = a'(\theta_{ij}) \text{ and } var(y_{ij}) = a''(\theta_{ij})/\phi$$

where

y_{ij} is the outcome variable for unit $i = 1, ..., m$ and measurement $j = 1, ..., n_i$,

$y_i = \{y_{i1}, ..., y_{in_i}\}$ and $x_i = \{x_{i1}, ..., x_{in_i}\}$,

$\theta_{ij} = h(\eta_{ij})$ where $\eta_{ij} = x_{ij}\alpha$,

α is the set of population parameters linking x_{ij} to y_{ij},

and ϕ is the dispersion parameter.

The estimators of α are consistent and the estimators of the variance are consistent under a weak assumption that a weighted average of the correlation matrices estimated converge to a single matrix.

The estimating equations take correlation into account to increase the efficiency of the estimators using a working correlation structure R_i which is symmetric and of size $n_i \times n_i$.

Liang and Zeger (1986) define their generalised estimating equations as

$$\sum_{i=1}^{m} D_i^T V_i^{-1} S_i = 0$$

where V_i is equal to the covariance of y_i if R_i is the true structure of the correlation and is defined as

$$V_i = A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}} / \phi$$

where

$$A_i = diag\{a''(\theta_{ij})\}$$

and $D_i = d\{a_i'(\theta)\}\ d\alpha$ and $S_i = y_i - a_i'(\theta)$.

R_i has the freedom to be specified in any number of forms including, for example,:

- Independence: $R_i = I_{n_i}$,
- Compound symmetric / exchangeable: components of R_i are $corr(Y_{it}, Y_{i,t+1}) = \rho$ for $t = 1, ..., n-1$. The compound symmetric GEE provides the same correlation structure as random intercept models where the marginal distributions are normal (Laird and Ware, 1982) with $corr(y_{it}, y_{it'}) = \rho \forall t \neq t'$,
- Auto-regressive: $corr(y_{it}, y_{it'}) = \rho^{|t-t'|}$,
- Totally unspecified: $\frac{1}{2}n(n-1)$ parameters are introduced to the model for the structure of R_i.

Liang and Zeger (1986) demonstrate that misspecification of R_i can lead to significant reductions in efficiency of the estimator, especially when correlation is higher. For this reason, a comprehensive understanding of the underlying correlation structure is required prior to application.

1.3 Copula models

This section provides an overview of the theory of copulas and how they are integrated into regression frameworks from their first incorporation in unconditional linear models to the most recent improvements in flexible copula-based joint regression modelling.

1.3.1 Introduction to copulas

Trivedi and Zimmer (2007) provide a detailed overview of the development of multivariate models and their initial applications in finance. They cite, in particular, Sklar (1973), and the development of the proof for a unique copula, Joe (1997), who provides a detailed survey of copula methods, and Nelsen (2007), who compiles a broad range of applications of copulas.

The joint cumulative distribution of a set of random variables $(Y_1, ...Y_m)$ is defined as:

$$F(y_1, ..., y_m) = Pr[Y_i \leq y_i;\ i = 1, ..., m]$$

Sklar's Theorem (1973) states that an m-dimensional copula is a function, C(.), mapping the unit m-cube $[0, 1]^m$ to the unit interval [0, 1], which satisfies the following conditions:

1. $C(1, ..., 1, a_n, 1, ..., 1) = a_n$ for every $n \leq m$ where a_n is the argument in position n of the function and all a_n in [0,1],
2. $C(a_1, ..., a_m) = 0$ if $a_n = 0$ for any $n \leq m$,
3. C is m-increasing.

These same rules can also be described as below:

1. Given all other realisations are known to be marginal probability one, the uncertain outcome is entirely dependent on the marginal probability of the last, unknown variable,
2. The probability of all outcomes is zero if any outcome's marginal probability is zero,
3. The value of C at any point is non-negative and increasing in each argument.

Using a copula approach we can describe a multivariate distribution in terms of its marginal distributions. This can be achieved by using the cumulative distribution function of each marginal distribution, $F_i(y_i)$, where $y_i = F_i^{-1}(u_i)$, which will be distributed as uniform on interval $[0, 1]$, $U(0, 1)$, and expressed as a copula function. That is, we describe the multivariate cdf $F(y_1, ...y_m)$ as:

$$\begin{aligned} F(y_1, ..., y_m) &= F(F_1^{-1}(u_1), ..., F_m^{-1}(u_m)) \\ &= Pr[U_1 \leq u_1, ..., U_m \leq u_m] \\ &= C(u_1, ..., u_m). \end{aligned}$$

If $y \sim F$, and F is continuous, then

$$(F_1(y_1), ..., F_m(y_m)) \sim C,$$

and if $U \sim C$, then

$$(F_1^{-1}(u_1), ..., F_m^{-1}(u_m)) \sim F.$$

To enable adjustment of the shape of the copula, a copula parameter, θ, referred to as the dependence parameter, is included in this formulation:

$$F(y_1, ..., y_m) = C(F_1(y_1), ..., F_m(y_m); \theta).$$

If all the marginal distributions, F_i are continuous, then the copula function C is unique. If any one of the marginal distributions is discrete then the copula function will not be unique to that multivariate distribution but can be uniquely determined, i.e. there is only one copula function which can exist that is defined by those margins but that copula is not necessarily unique to those margins.

Copula distributions

Copula functions are able to capture a broad range of dependence distributions. There are a number of commonly used copulas which utilise only one parameter but are able to describe a varying set of shapes. These include the Gaussian, Clayton, Frank, Gumbel, Joe, AMH, Frank, Plackett and Hougaard copulas (Trivedi and Zimmer, 2007, Joe, 1997) among others. The t-copula is also commonly used as an extension to the Gaussian copula with an additional parameter for degrees of freedom for capturing heavier tail dependence. Figure 1.1 illustrates a number of examples of the flexible shapes which these copulas can capture.

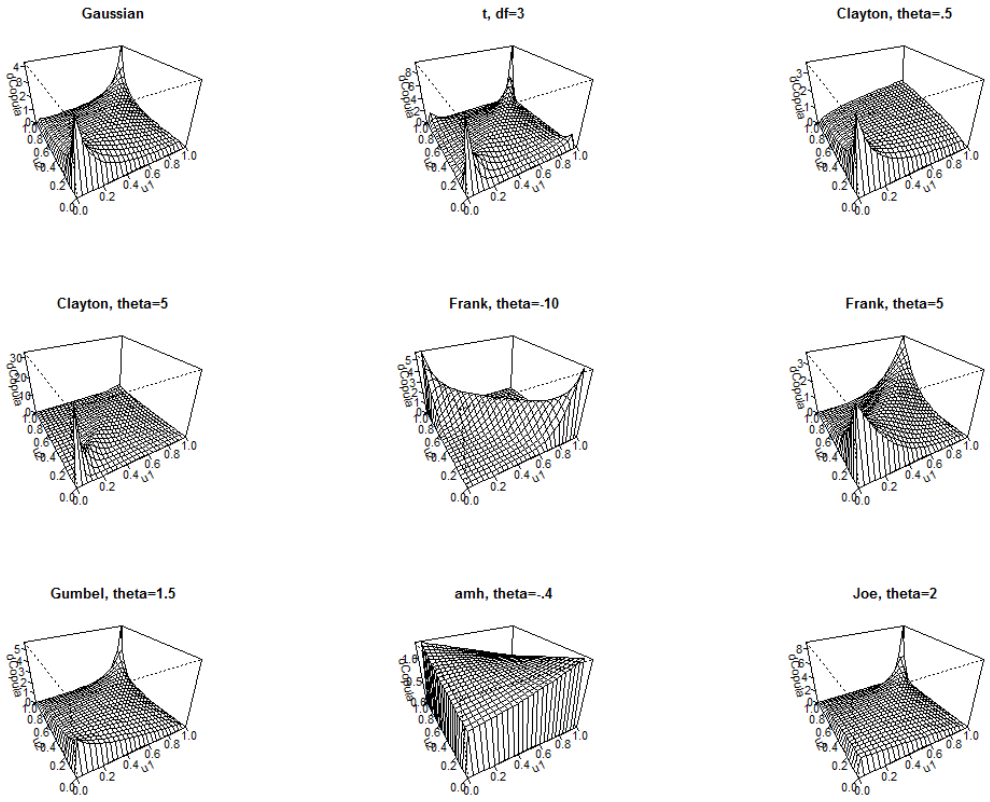


Figure 1.1: Nine copula functions

There are four key methods for generating copulas (Trivedi and Zimmer, 2007):

1. **Method of inversion:** Given a joint distribution, $F(y_1, y_2)$, derive the marginal distributions $F(y_1)$ and $F(y_2)$, then derive the resulting copula. Unfortunately this requires the full description of the joint distribution as a starting point.

2. **Algebraic methods:** Begin with a relationship between marginals based on independence, then introduce a dependence parameter i.e. introduce a parameter for description of the joint distribution as a function of the independent marginals.

3. **Mixtures and convex sums:** Given a copula, C, its lower and upper bounds, C_L and C_U, and the independence copula, C_P, a new copula can be constructed using a convex sum , e.g. the convex sum of the upper bound and independence copulas is also a copula, $C^M = \pi_1 C_P + (1 - \pi_1) C_U$, where $\pi_i \in [0, 1]$.

4. **Archimedean copulas:** class of functions φ which are convex and decreasing of

the form $C(u_1, u_2; \theta) = \varphi(\varphi^{-1}(u_1) + \varphi^{-1}(u_2))$ with conditional density $\frac{\partial}{\partial u_2} C(u_1, u_2) = \frac{\varphi'(u_2)}{\varphi'(C(u_1,u_2))}$.

Initial copula estimation in regression

Pitt et al. (2006) was the first to introduce a computational approach for estimating multivariate copula regression models with any discrete or continuous marginals, using a general Bayesian approach. Prior to Pitt's approach, copula modelling was severely limited by requiring Gaussian marginal distributions. A Markov chain Monte Carlo simulation is used to sample from the resulting distribution for inference. Note that Pitt et al. (2006) only demonstrates this approach for the case of a Gaussian copula.

Pitt et al. (2006) demonstrates the need for non-normal marginal distributions in the case of t-distributed US industrial CAPM returns, where parameter estimation error is significantly reduced by capturing the heavier tails in returns with the t-distribution. Another example used was for multivariate count data of health care utilisation, which requires a zero-inflated geometric distribution, something that couldn't be captured reasonably with a normal approximation.

Consider a sample y_{ij} for $i = 1, ..., n$ and $j = 1, 2$, where y_{ij} depend on explanatory variables x_{ij} with parameters β_j linking x_{ij} to y_{ij}. For $j = 1, 2$ we have:

Denote the marginal density functions as

$$f_j(y_j|x_j;\beta_j) = \frac{\partial F_j(y_j|x_j;\beta_j)}{\partial y_j}$$

and the copula function as

$$C\big\{(F_1|x_1;\beta_j),(F_2|x_2;\beta_j);\theta\big\}$$

then copula density is then written as

$$\begin{aligned} c(F_1(\cdot), F_2(\cdot)) &= \frac{\partial}{\partial y_1 \partial y_2} C\big\{(F_1|x_1;\beta_j),(F_2|x_2;\beta_j);\theta\big\} \\ &= \frac{\partial C\{\cdot\}}{\partial F_1 \partial F_2} f_1(\cdot) f_2(\cdot) \end{aligned}$$

The log likelihood for the sample is

$$\begin{aligned}\ell_n(\beta_1, \beta_2, \theta) &= \log \mathcal{L}_n((y_1|x_1; \beta_1), (y_2|x_2; \beta_2)); \theta) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{2} \log f_{ji}(y_{ji}|x_{ji}; \beta_j) \\ &+ \sum_{i=1}^{n} C_{12}[F_1(y_{1i}|x_{1i}; \beta_1), F_2(y_{2i}|x_{2i}; \beta_2); \theta].\end{aligned}$$

This log likelihood decomposes into two parts which can be estimated using full maximum likelihood by solving the score equations based on:

$$\ell_n(\beta_1, \beta_2, \theta) = \ell_{1,n}(\beta_1, \beta_2) + \ell_{2,n}(\beta_1, \beta_2, \theta).$$

It is also possible to maximise the likelihood using maximisation by parts (two step sequential likelihood maximisation).

1.3.2 Conditional copula regression

Patton (2006) introduces a novel approach for conditioning copulas on a covariate, significantly expanding the types of applications copulas can be used for to any standard regression. Prior to this approach, there was no simple way to exclude the effect of a confounding variable on a copula dependence structure between two variables.

Patton (2006) demonstrates that confounding variables to a dependence structure can introduce high levels of bias in estimates if not accounted for. Patton (2006) uses a clear example of Deutsche Mark-US Dollar and US Dollar-Yen joint distributions which have a step change in dependence following the conversion of the Mark to the Euro in 1999 (see figure 1.2). Without conditioning on the period of time before and after the currency conversion, there is a large confounding effect on estimates of the dependence. Patton describes the structure of a conditional copula by conditioning each component on a variable, i.e.,

Let $F_{Y_1|X}(y_1|x)$ be the conditional distribution of $Y_1|X = x$

Let $F_{Y_2|X}(y_2|x)$ be the conditional distribution of $Y_2|X = x$

Let $F_{Y_1Y_2|X}(y_1y_2|x)$ be the joint conditional distribution of $Y_1Y_2|X = x$

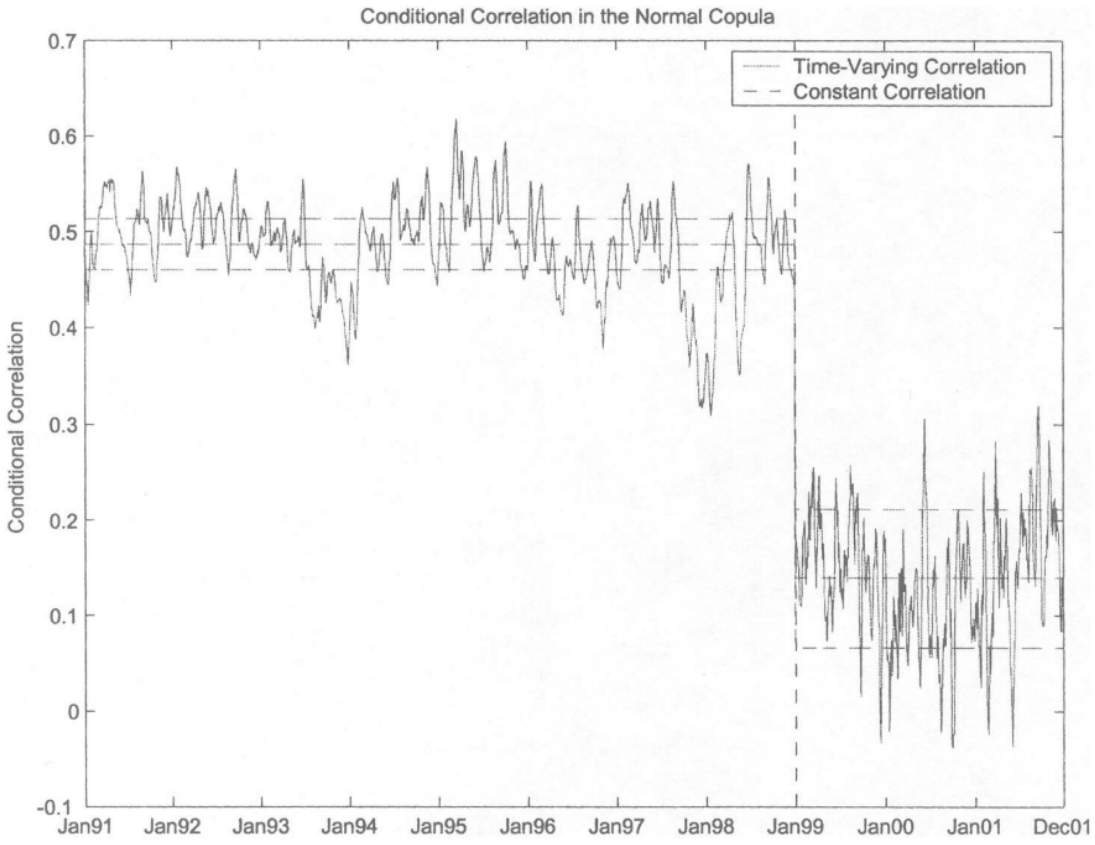


Figure 1.2: Example from 'Modelling asymmetric exchange rate dependence' (Patton, 2006)

Assuming $F_{Y_1|X}(y_1|x)$ and $F_{Y_2|X}(y_2|x)$ are continuous for all x, then there exists a unique conditional copula $C(.|x)$ such that:

$$F_{Y_1Y_2|X}(y_1y_2|x) = C(F_{Y_1|X}(y_1|x), F_{Y_2|X}(y_2|x)|x)$$
$$\forall(y_1, y_2, x) \in \mathbb{R} \times \mathbb{R} \times R_X$$

where R_X is the range of X. Note that the same conditioning must be used for both the marginal distributions and the copula for the joint distribution to hold. The density function is then defined as below, assuming both $F_{Y_1|X}, F_{Y_2|X}$ are once differentiable and $F_{Y_1Y_2|X}, C$ are twice differentiable:

$$f_{Y_1Y_2|X}(Y_1y_2|x) = f_{Y_1|X}(y_1|x) \times f_{Y_2|X}(y_2|x) \times c(u, v|x)$$
$$\forall(y_1, y_2, x) \in \mathbb{R} \times \mathbb{R} \times R_W$$

The MLE can then be calculated using the log of this density function.

Covariate selection

Due to the significant effect of confounding variables on copula models, it is essential to be able to identify the cases in which a conditioning variable is required and what its effect is.

Gijbels et al. (2011) demonstrate how local measures of Kendall's tau and Spearman's rho association can be used to identify whether a covariate needs to be conditioned upon for a copula fit. The approach utilises non-parametric localised estimates of the association measures across the range of potential covariates to motivate the need for conditioning on the variable where there is a significant trend.

Kendall's tau represents the degree of concordance between two random variables realizing values between -1 and 1 and is calculated as

$$\tau = 2P((Y_1 - Y_1')(Y_2 - Y_2') > 0) - 1,$$

where $(Y_1', Y_2')^T$ is an independent realisation of the random variables describing the random vector $(Y_1, Y_2)^T$.

There is a direct relationship between conditional Kendall's tau, $\tau(X) = 4E[H(Y_1, Y_2|X)|X] - 1$ where H is the joint distribution function of $(Y_1, Y_2)^T$, and the copula parameter $\theta(X)$, with parametrisation by either being equivalent (Sabeti et al.).

As described by Nelsen (2007), the copula function can also be used to calculate this quantity:

$$\tau = 4 \int \int C(u_1, u_2) dC(u_1, u_2) - 1.$$

As a measure of dependence, Kendall's tau has the advantage that is is not reliant on the relative scale of observations, as is the case with Pearson correlation. Kendall's tau compares the relative rank of observations between the dependent random variables being considered. In the extreme example, if the matched pairs between two dependent datasets receive the same ranked size within their own datasets but one increases exponentially while the other is linear, Kendall's tau will indicate complete concordance while Pearson correlation may be weak due to the relationship's non-linearity. This shape and scale invariant property of Kendall's tau becomes extremely important when attempting to identify an appropriate dependence structure, as the scale and shape of marginal distributions are eventually stripped

away through the modelling process and the remaining dependence structure is all that is required to be modelled by the copula function.

Gijbels et al. (2011) demonstrate the use of non-parametric localised Kendall's tau on a dataset of life expectancies at birth for males and females in 222 countries. Joint life expectancies for males and females are modelled with a fitted copula for joint dependence and a non-parametric fit of log10(GDP) as the covariate. There is a clear decrease in localised Kendall's tau and Spearman's rho estimators as log10(GDP) increases, which demonstrates the need for conditioning based on this variable to identify the underlying dependence structure and describe its shape. Further recommendations for the practical use of associations in copula regression including review of scatterplots of pairs followed and a review of non-parametric estimators of dependence across the range of each variable are also proposed by Veraverbeke et al. (2011).

Simulation studies indicate that the performance of association measure estimators on an integrated square error basis are in close agreement with those of the copula, but that the estimation of Kendall's tau is very sensitive to the bias of the underlying copula estimator (Gijbels et al., 2011).

Note that there are also cases where a covariate only affects the marginal distributons and is not required to be conditioned on in the copula. In these cases, such a model could hold one less parameter and would require a separate estimation method (Gijbels and Veraverbeke, 2015).

Fitting covariates

Building on the algebraic form for the conditional copula developed by Patton (2006), Acar et al. (2011) introduce an approach for modelling a copula parameter which varies with one covariate. A non-parametric approach based on local likelihood is introduced, which identifies the case where the copula parameter changes with a single covariate.

Let Y_1 and Y_2 be continuous variables of interest and X a continuous covariate which might affect the dependence between Y_1 and Y_2. Assuming the marginals are known, then we can find U_{1i} and U_{2i} being the cumulative distributions of the marginal distributions and the

following conditional model:

$$(U_{1i}, U_{2i})|X_i \sim C(u_{1i}, u_{2i}|\theta(x_i)),$$

where $\theta(x_i) = g^{-1}(\eta(x_i)),\ i = 1, ..., n.$

A local-likelihood framework is used to identify whether there is a significant latent variable affecting the copula shape. The local maximum likelihood estimator can be found via Newton Raphson iteration from the likelihood function, $\mathcal{L}$:

$$\frac{\partial \mathcal{L}(\beta, x, p, h)}{\partial \beta} = 0$$

where h is the bandwidth determining the size of the local neighbourhood and p is the local polynomial fit of degree p. Acar et al. (2011) proposes an optimal bandwidth for local fits based on cross validated maximum likelihood.

Acar et al. (2011) demonstrate the ability of the approach to identify both a linear and quadratic conditioning variable and find that the estimator accuracy quickly deteriorates when the incorrect form for the copula function is selected. Applying the approach to the Twin Births dataset, Acar et al. (2011) demonstrate that there is an age specific dependence structure in the twins dataset. Acar et al. (2013) later introduce a general approach for testing the statistical significance of covariates and their transformations in the form of calibration functions, against alternatives within a conditional copula modelling framework.

We can extend conditioning on a covariate to conditioning on some function of a covariate as below. This function is termed the calibration function as it calibrates the support of the copula function. The form of the conditional copula is written:

$$(U_1, U_2)|X = x \sim C_x\left(u_1, u_2|\theta(x) = g^{-1}(\eta(x))\right)$$

To make this useful for inference, there is a need to test the significance and relative efficiency of different forms for $\eta(x)$.

Assuming the conditional marginal distributions are known, Acar et al. (2013) introduces likelihood tests for comparing three different cases of interest for shapes of $\eta(x)$ for a single covariate: constant, linear and semi-parametric. Below is the likelihood test comparing the

linear and semi-parametric case:

$$\mathcal{L}_n(H_o) = \sum_{i=1}^{n} \ell(a_0 + a_1 X_i, U_{1i}, U_{2i})$$
$$\mathcal{L}_n(H_1) = \sum_{i=1}^{n} \ell(\eta(X_i), U_{1i}, U_{2i})$$

Acar et al. (2013)'s approach assumes the conditional marginal distributions are known. It is worth noting that the choice of the marginal distributions and how they incorporate covariates could have a significant effect on the choice of calibration function.

1.3.3 Generalised copula regression

Sabeti et al. propose the use of additive models for bivariate copula regression and outline potential approaches to, and tools for, model selection and computation.

Building on the development of a mathematical form for conditional copulas (Patton, 2006), and the approach for incorporating and comparing single covariates in fitting a copula function (Acar et al., 2011, 2013), Sabeti et al. introduce an additive model framework to provide a method for analysing the relationship between more than one covariate with a non-parametric form and the copula function.

Consider random variables Y_{i1}, Y_{i2} dependent on covariate $X_i \in \mathbb{R}^p$. Assume each Y_i's marginal distribution is modelled with normal regression. So we assume the marginal distributions are:

$$Y_{ij} \sim N(X_i^T \beta_j, \sigma_j^2),\ j = 1, 2,\ i = 1, .., n$$

and the joint density is

$$f(Y_{i1}, Y_{i2}|X_j) = \prod_{j=1}^{2} \frac{1}{\sigma_j} \phi\left(\frac{Y_{ij} - X_i^T \beta_j}{\sigma_j}\right) \times$$
$$\times\, c^{(1,1)}\left[\Phi\left(\frac{Y_{i1} - X_i^T \beta_1}{\sigma_1}\right) \times \Phi\left(\frac{Y_{i2} - X_i^T \beta_2}{\sigma_2}\right) \middle| \theta(X_i)\right], \qquad i = 1, ..., n,$$

where

ϕ is the probability density function of the standard normal distribution,

Φ is the cumulative density function of the standard normal distribution,

$c^{(1,1)} = \partial^2 C(u, v|\theta)/\partial u \partial v$, and,

θ is the copula parameter which depends on X_i.

Since many copulas require the parameter θ to be restricted to a subset of $\mathbb{R}$, a link function is introduced to the model for θ, $g(\theta) = \eta(X)$.

Utilising the additive model of Hastie and Tibshirani (1990), the form for $\eta(X)$ when $p > 1$ is as below:

$$\eta(X) = \alpha_0 + \sum_{i=1}^{p} \eta_i(X_i).$$

The computational approach utilises a Markov chain Monte Carlo approach to sample from the posterior distribution. Covariates are assumed to be independent random variables with a normal distributon for β_i and inverse gaussian for σ_i^2. Sabeti et al. demonstrate the approach for a given set of covariate splines.

Sabeti et al. proposes model selection based on cross validated, pseudo marginal likelihood (CVML):

$$CVML(\mathcal{M}) = \sum_{j=1}^{n} \log p(Y_{1j}, Y_{2j}|\mathcal{D}_{-j}, \mathcal{M}),$$

where $\mathcal{M}$ is a model that maximises the CVML and D_{-j} is the remaining data after removing covariates and response variables for the jth item.

A Monte Carlo estimator for this value is:

$$\widehat{CVML}(\mathcal{M}) = \sum_{j=1}^{n} -\log\left[\frac{1}{M}\sum_{m=1}^{M} p(Y_{1j}, Y_{2j}|\omega^{(m)}, \mathcal{M})^{-1}\right]$$

where $\omega^{(m)}$, the parameters of the model, are draws from the posterior distribution $\pi(\omega|\mathcal{D}, \mathcal{M})$.

Simulations demonstrate that as the dimension of the covariate vector increases, the efficiency of the estimators decreases and further work is required to understand if computational difficulty becomes too intensive for large covariate vectors. The CVML criteria is applied

to the twin births data. The selected copula agrees with that chosen by the likelihood-based model selection criteria of Acar et al. (2013).

Vatter and Chavez-Demoulin (2015) introduce a computational approach for estimating generalised additive models with dependence structures as an extension to work by Sabeti et al. and Acar et al. (2013).

Vatter and Chavez-Demoulin (2015) describe the model for the dependence structure as a generalised additive model for the conditional concordance measure Kendall's tau because it has a simpler interpretation and it is easier to compare across copula families:

$$\tau(x, \theta) = g\left\{z^T\beta + \sum_{k=1}^{K} h_k(t_k)\right\},$$

where

g is a strictly increasing link function expressing the relationship between the GAM and τ,

z and t are subsets of x,

β is the vector of parameters,

h_k are smooth functions, and

θ is the vector of all parameters.

This can be estimated by maximising the penalised log likelihood,

$$\ell_c(\theta, \gamma) = \ell_c(\theta) - \frac{1}{2}\theta^T p(\gamma)\theta,$$

with $p(\gamma)$ being a $d \times d$ block diagonal matrix with $K + 1$ blocks, and γ is the vector of smoothing parameters. Iteratively re-weighted ridge regression is used as the estimation procedure. The estimator is proven to be $\sqrt{n}$-consistent and asymptotically normal.

Vatter and Chavez-Demoulin (2015) apply the approach to a simulated time-varying Gaussian copula with quadratic, sinusoidal and exponential covariate effects. Interestingly, estimation becomes significantly more difficult for highly correlated covariates, which results in higher error. The method is also applied to foreign exchange pairs for EUR/USD and USD/CHF paired with a seasonality filter, and a GARCH model showing positive results for inference.

Generalised Joint Regression Models

Marra and Radice (2017) have designed and implemented a computational tool for fitting flexible copula-based joint regression models. Copula dependence and marginal distributions are estimated simultaneously and parameters rely on their own set of covariate effects including linear, non-linear, random and spatial effects. The approach differs from that of Vatter and Chavez-Demoulin (2015) who optimise the marginal distribution fits and and conditional copula separately, whereas Marra and Radice (2017) introduce simultaneous estimation of both fits.

For Marra and Radice (2017)'s approach, a joint cumulative distribution function for random variables Y_1, Y_2 conditional on covariates is expressed as

$$F(y_1, y_2|\vartheta) = C(F_1(y_1|\mu_1, \sigma_1, \nu_1), F_2(y_2|\mu_2, \sigma_2, \nu_2); \zeta, \theta)$$

where

$\vartheta = (\mu_1, \sigma_1, \nu_1, \mu_2, \sigma_2, \nu_2, \zeta, \theta)$,

$F_i, j = 1, 2$ are marginal cumulative distribution functions of Y_1, Y_2,

μ_j, σ_j, ν_j are the parameters of the marginal distributions of Y_1, Y_2,

θ and ζ are the copula dependence parameters.

For each parameter of the marginal distributions there is a link function, $g^{-1}(\cdot)$, and linear predictor, η. The log-likelihood for the copula model can then be written as

$$\ell(\delta) = \sum_{i=1}^{n} \log\{C(F_1(y_1|\mu_1, \sigma_1, \nu_1), F_2(y_2|\mu_2, \sigma_2, \nu_2); \zeta, \theta)\} + \sum_{i=1}^{n}\sum_{j=1}^{2} \log(f_j(y_{ji}|\mu_{ji}, \sigma_{ji}, \nu_{ji})$$

where copula density c is

$$c(F_1(y_{1i}), F_2(y_{2i})) = \frac{\partial^2 C\{F_1(y_{1i}), F_2(y_{2i})\}}{\partial F_1(y_{1i}) \partial F_2(y_{2i})}$$

and we can define a parameter vector for estimation, $\delta = (\beta_{\mu_1}^T, \beta_{\mu_2}^T, \beta_{\sigma_1}^T, \beta_{\sigma_2}^T, \beta_{\nu_1}^T, \beta_{\nu_2}^T, \beta_{\zeta}^T, \beta_{\theta}^T)$.

Marra and Radice (2017) propose the use of the penalised log likelihood to achieve greater smoothness in their estimators, so the function to maximise incorporates a penalty to the pure likelihood maximisation:

$$\ell_p(\delta) = \ell(\delta) - \frac{1}{2}\delta^T S \delta,$$

where S depends on the kind of specific quadratic penalty that is chosen, the choice of basis function and the choice of smoothing parameter.

2 Simulation

2.1 Selection of a bivariate distribution

This thesis adopts a simulation approach to compare the performance of flexible copula-based regression with random effect based regression, for modelling longitudinal data. To develop a fair comparison between the copula and random effect approaches, a simulation of a bivariate distribution which is neither strictly a parametric copula dependence structure nor a random effect related dependence structure is required.

A range of bivariate distributions which introduce dependence through multiplicative means were identified (Balakrishnan and Lai, 2009). These bivariate distributions would be of the form $Y_1 = UW$, $Y_2 = VW$ where Y_1 and Y_2 are the random variables representing repeated observations at time 1 and 2 on the same individual and W is a random variable which introduces dependence between Y_1 and Y_2 (Balakrishnan and Lai, 2009). The selected distribution is the bivariate gamma of Nadarajah and Gupta (2006), which induces dependence between two gamma distributions by multiplying by a common beta distribution with some restriction on choice of parameters. The approach is as follows:

1. Generate three independent random variables:

$$W \sim Beta(\alpha, \beta)$$
$$U \sim Gamma(\alpha + \beta, 1/\mu_1)$$
$$V \sim Gamma(\alpha + \beta, 1/\mu_2)$$

2. Let $Y_1 = WU$ and $Y_2 = WV$, inducing a dependence between Y_1 and Y_2 and creating the realisations (y_1, y_2) of the required distribution.

Yeo and Milne (1991) describe the properties of mixed beta-gamma distribution of Y_1 and Y_2.

The resulting marginal distributions are:

$$Y_1 \sim Gamma(\alpha, 1/\mu_1),\ E(Y_1) = \alpha/\mu_1$$

$$Y_2 \sim Gamma(\alpha, 1/\mu_2),\ E(Y_2) = \alpha/\mu_2$$

with a linear correlation coefficient of

$$corr(Y_1, Y_2) = \rho = \frac{\sqrt{\alpha\beta}}{\alpha + \beta + 1}.$$

The bivariate distribution is fully described by its probability density function,

$$f(y_1, y_2) = C\Gamma(\beta)(y_1 y_2)^{\alpha+\beta-1}\left(\frac{y_1}{\mu_1} + \frac{y_2}{\mu_2}\right)^{\frac{\alpha-1}{2}-(\alpha+\beta)} \exp\left[-\frac{1}{2}\left(\frac{y_1}{\mu_1} + \frac{y_2}{\mu_2}\right)\right] W_{\alpha+\frac{1-\alpha}{2},\alpha+\beta-\frac{\alpha}{2}}\left(\frac{y_1}{\mu_1} + \frac{y_2}{\mu_2}\right),$$

for $y_1 > 0$,$y_2 > 0$ and where the constant C is given by

$$\frac{1}{C} = (\mu_1\mu_2)^{\alpha+\beta}\Gamma(\alpha + \beta)\Gamma(\alpha)\Gamma(\beta),$$

and $W_{\lambda,\mu}$ is the Whittaker function (Abramowitz and Stegun, 1972),

$$W_{\lambda,\mu}(p) = \frac{p^{\mu+\frac{1}{2}}\exp(-p/2)}{\Gamma(\mu - \lambda + 1/2)}\int_0^\infty t^{\mu-\lambda-1/2}(1 + t)^{\mu+\lambda-1/2}\exp(-pt)dt.$$

The parameters of the mixing beta distribution fully define the dependence structure between the two marginal gamma distributions.

Dependence structures can be visualised by plotting the cumulative distribution functions of the marginal distributions against each other (Trivedi and Zimmer, 2007). In the context of a regression, the shape and parameters of the marginal distribution need to be calculated to allow this, however in a simulation, the simulation parameters can be employed for the transform. Some examples of the shapes taken by the dependence structure of this bivariate distribution are shown in figure 2.1.

All the dependence structures from this bivariate gamma incorporate a skew towards lower value dependence with differing strength of overall dependence, and a differing extent to which higher value dependence is also exhibited. The highest rank correlation within the dependence structure is realised when α is low and β is high. This is shown in figure 2.1 and table 2.1.

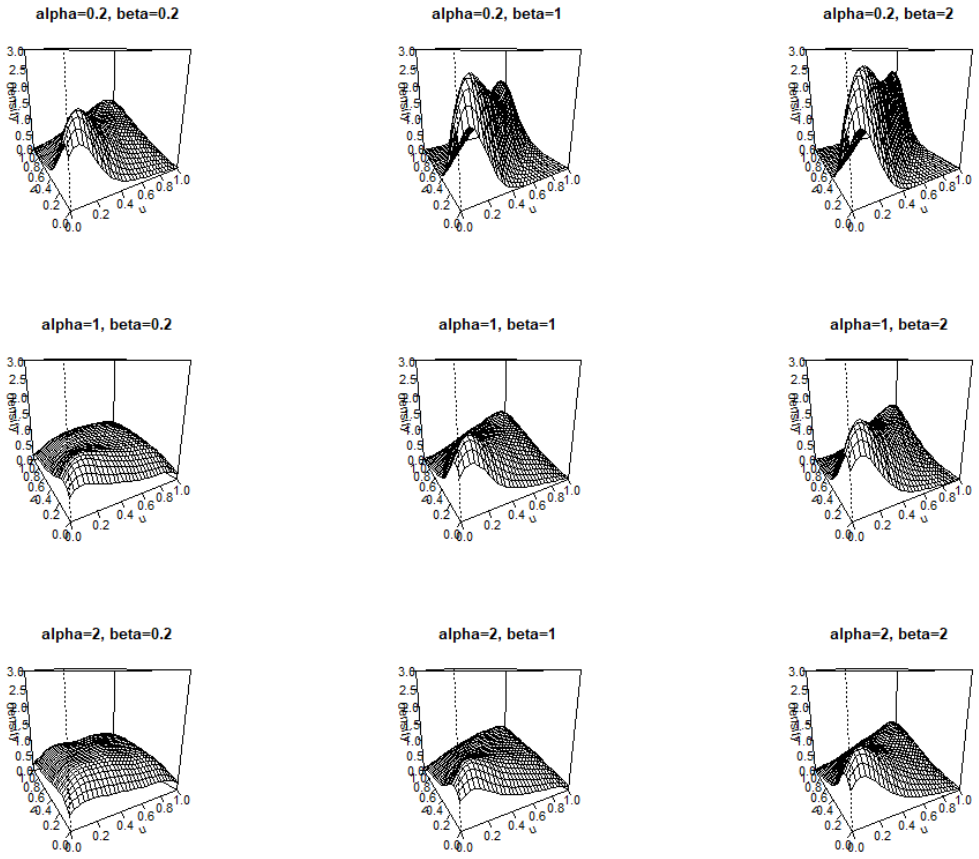


Figure 2.1: Dependence structure for the bivariate gamma of Nadarajah and Gupta (2006) under varying parameters for the multiplicative beta distribution. Cumulative distribution functions of the marginals are plotted against one another as a bivariate density plot.

The second parameter of the multiplicative beta distribution, β, has no effect on the marginal distributions of Y_1 and Y_2 but is an important component in defining the dependence between the two variables. α and β together determine the strength of the rank correlation of the dependence structure. α has the added effect of increasing marginal skewness with lower values. This relationship is shown in table 2.1.

Tau	β									
α	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	1-1.2	1.2-1.4	1.4-1.6	1.6-1.8	1.8-2	2-2.2
0.2-0.4	0.38	0.51	0.60	0.65	0.68	0.71	0.73	0.75	0.77	0.78
0.4-0.6	0.25	0.37	0.45	0.51	0.55	0.58	0.61	0.63	0.65	0.67
0.6-0.8	0.18	0.28	0.36	0.42	0.46	0.49	0.53	0.55	0.58	0.59
0.8-1	0.15	0.23	0.30	0.36	0.40	0.43	0.47	0.49	0.51	0.54
1-1.2	0.12	0.20	0.26	0.31	0.35	0.38	0.42	0.44	0.47	0.48
1.2-1.4	0.10	0.17	0.23	0.27	0.32	0.35	0.38	0.40	0.43	0.45
1.4-1.6	0.09	0.15	0.20	0.25	0.28	0.31	0.35	0.37	0.40	0.41
1.6-1.8	0.08	0.14	0.18	0.22	0.26	0.29	0.32	0.34	0.37	0.38
1.8-2	0.07	0.13	0.17	0.21	0.23	0.27	0.30	0.32	0.34	0.36
2-2.2	0.06	0.12	0.15	0.19	0.23	0.25	0.28	0.30	0.32	0.34

Skew	α									
Margin	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	1-1.2	1.2-1.4	1.4-1.6	1.6-1.8	1.8-2	2-2.2
$\mu = 1$	4.06	3.00	2.49	2.17	1.95	1.79	1.66	1.56	1.47	1.40
$\mu = 2$	4.06	3.00	2.49	2.17	1.95	1.79	1.66	1.56	1.47	1.40

Table 2.1: Average values of marginal skewness and Kendall's tau for differing values of α and β of the generating distribution.

2.2 Model specification

There are five defining characteristics of the bivariate gamma which need to be captured for accurate modelling of the full bivariate distribution:

- The mean and dispersion of the marginal gamma for time 1, μ_1 and σ_1,
- The mean and dispersion of the marginal gamma for time 2, μ_2 and σ_2, and
- The dependence structure between the two marginals which is captured through a dependence parameter θ.

In a copula-based model, each of these parameters is modelled directly. There is a fit for each marginal distribution: μ_1 and σ_1 for the first marginal distribution, and μ_2 and σ_2 for the second marginal distribution, and a copula fit to the dependence structure with parameter θ.

In a generalised linear mixed modelling framework, the same five parameters are captured but slightly differently. The most comparable model is the generalised linear mixed model with a random effect term and time-dependent σ. The model has:

- a parameter for the intercept at time 1, $\beta_{\mu 1}$, which estimates μ_1,
- a parameter for time, $\beta_{\mu t}$, which combines with $\beta_{\mu 1}$ to estimate μ_2,
- an estimate for the intercept of the dispersion parameter, $\beta_{\sigma 1}$, which estimates σ_1,
- a parameter for the effect of time on the dispersion of the distribution, $\beta_{\sigma t}$ which combined with $\beta_{\sigma 1}$ estimates σ_2, and
- a parameter for the variance of the random effect, θ.

This analysis compares both the full generalised linear mixed model and the copula model described above as well as three other models as a comparison point:

- a generalised linear model which only has a single dispersion parameter σ so does not capture differing dispersion between time points, and no parameter for adjusting for dependence between observations (3 parameters total),
- a generalised estimating equations approach which includes a parameter for the dependence but does not directly capture the differing dispersion between time points (4 parameters total),
- a generalised linear mixed model with a random effect but without a time dependent sigma (4 parameters total).

Overview of each model

Table 2.2 summarises the models run for each simulation and their associated parameters.

Model	**No. Parameters**	**Components estimated**
Generalised linear model (GLM)	3	μ_1, μ_2, σ
Generalised estimating equation (GEE)	4	$\mu_1, \mu_2, \sigma, \theta$
Generalised linear mixed model (GLMM no sig)	4	$\mu_1, \mu_2, \sigma, \theta$
Generalised linear mixed model (GLMM)	5	$\mu_1, \mu_2, \sigma_1, \sigma_2, \theta$
Generalised joint regression model (GJRM)	5	$\mu_1, \mu_2, \sigma_1, \sigma_2, \theta$

Table 2.2: Summary of specified models and associated parameters

As the response variable is gamma for each marginal, log link functions for μ_j and σ_j are appropriate.

The generalised linear model can be used as a baseline comparison point which does not adjust for dependence or differing dispersion between time points. The model structure is:

$$\log(\mu_i) = \beta_{\mu 1} + \beta_{\mu t} t$$

$$t = \begin{cases} 0, & \text{if time} = 1 \\ 1, & \text{if time} = 2 \end{cases}$$

where

$\beta_{\mu 1}$ is the intercept,

$\beta_{\mu t}$ is the coefficient for time,

σ^2 is the estimate for the dispersion parameter.

Generalised estimating equations take into account the dependence between marginals. The model requires one additional parameter for modelling the dependence between the time points:

$$\log(\mu_i) = \beta_{\mu 1} + \beta_{\mu t} t$$

$$t = \begin{cases} 0, & \text{if time} = 1 \\ 1, & \text{if time} = 2 \end{cases}$$

where

$\beta_{\mu 1}$ is the intercept,

$\beta_{\mu t}$ is the coefficient for time,

σ^2 is the estimate for the dispersion parameter, and

an additional parameter, θ, for covariance is incorporated.

The generalised linear mixed model introduces a dependence parameter θ for the random effect and adjusts the structure of the model.

$$\log(\mu_{ij}) = \beta_{\mu 1} + \beta_{\mu t} t + b_{ij},\ j = 1, 2,\ i = 1, \ldots, m$$
$$\log(\sigma_j) = \beta_{\sigma 1} + \beta_{\sigma t} t,$$

$$t = \begin{cases} 0, & \text{if time} = 1 \\ 1, & \text{if time} = 2 \end{cases}$$

where

$\beta_{\mu 1}$ is the intercept,

$\beta_{\mu t}$ is the coefficient for time,

$\beta_{\sigma 1}$ is the estimate for the dispersion parameter at time 1, and

$\beta_{\sigma t}$ is the estimate for the effect of time on the dispersion, and

an additional parameter, θ, is estimated for the random effect, $b_j \sim N(0, \theta)$.

For the generalised joint regression model, the following structure is used:

$$\log(\mu_1) = \beta_{\mu 1},$$
$$\log(\mu_2) = \beta_{\mu 2},$$

where

$\beta_{\mu 1}$ is the intercept for the mean at time 1,

$\beta_{\mu 2}$ is the intercept for the mean at time 2,

σ_1, σ_2 are estimated as the dispersion parameters for the marginal distributions at time 1 and 2, and

θ is the estimate for the copula parameter.

2.2.1 Simulations

Simulations of Nadarajah and Gupta (2006)'s bivariate gamma have been run across the range of shapes of the distribution for a comparison of mean estimates. The distribution was simulated for mixing beta distribution parameters α in $(.2, .3, ..., 2.0, 2.1)$ and β in $(.2, .3, ..., 2.0, 2.1)$, and marginal pre-mix means of $\mu_1 = 1$ and $\mu_2 = 2$.

Six specific models are fit for the estimate of the mean at time one and two as a comparison:

- Generalised linear model (GLM)
- Generalised estimating equation (GEE)
- Generalised linear mixed model with a random effect for units (GLMM no sig)
- Generalised linear mixed model with a random effect for units and a differing estimate for σ^2 at time one and time two (GLMM)
- Generalised joint regression model with a Clayton copula (GJRM Clayton)
- Generalised joint regression model with a Normal copula (GJRM Normal)

The code used for simulations in this chapter is available in the supplementary materials, section 5.1.3.

2.2.2 Software

This section provides a short overview of the software used to fit the various models to the bivariate distribution in this simulation.

For GLMs, the base R function *glm* in the core R package *stats* (R Core Team, 2018) is used. The package provides the functionality to fit any exponential family GLM and extract parameter estimates and diagnostics of fit.

GEEs are fit using *geepack* (Hojsgaard et al., 2006), a modern alternative to older GEE methods with significantly increased computational speed.

For mixed models with random effects, multiple packages were tested and compared. The three main packages were *glmer* from *lme4* (Bates et al., 2015), *gamm* from *mgcv* (Wood, 2017) and *gamlss* (Stasinopoulos et al., 2017). *glmer* (Bates et al., 2015) allows for fitting any exponential family mixed model in a linear modelling framework. *gamm* (Wood, 2017) provides the functionality for fitting any mixed model and incorporates the ability to fit smooth covariates in an additive model framework. *gamlss* (Stasinopoulos et al., 2017) is a highly flexible regression toolkit which provides the ability to fit any parametric distribution, and incorporates the ability to fit random effects. The package also incorporates a very broad range of tools for model diagnostics and interrogation. Initial testing comparing the three packages did not indicate significant differences in the estimates between them.

Copula-based regression models and copula fit diagnostics was completed using GJRM, the package developed by Marra and Radice (2017). The package provides a comprehensive computational framework for modelling joint random variables with copulas for dependence fits, includes the functionality to test multiple dependence structures for the appropriate fit and incorporates gamlss (Stasinopoulos et al., 2017) for fitting marginal distributions.

2.2.3 Results

The following four parameter estimates and their errors for each of the models have been compared across the range of realisations:

- Mean and standard error of $\hat{\beta}_{\mu 1}$, the estimate for the mean at time 1,
- Mean and standard error of $\hat{\beta}_{\mu 2}$, the estimate for the mean at time 2.

Parameter estimates at time 1

In terms of standard error for the estimate at time 1, the GLM, GEE and Copula models show a consistent trend of increasing parameter standard error with higher values of τ, while the two random effect models exhibit generally decreasing standard error with increasing values of τ of realisations. See figure 2.2 and differences in average mean estimates in table 2.3.

Model / Tau	0-0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8+
GLM	0.0765	0.0794	0.0915	0.1402	0.2099
GEE	0.0765	0.0794	0.0915	0.1378	0.1855
GLMM no sig	0.0710	0.0561	0.0469	0.0512	0.0538
GLMM	0.0720	0.0557	0.0443	0.0000	0.0494
GJRM (Clayton)	0.0779	0.0788	0.0916	0.1375	0.2034
GJRM (Normal)	0.0780	0.0789	0.0929	0.1495	0.2671

Table 2.3: Median standard error for estimates of $\beta_{\mu 1}$

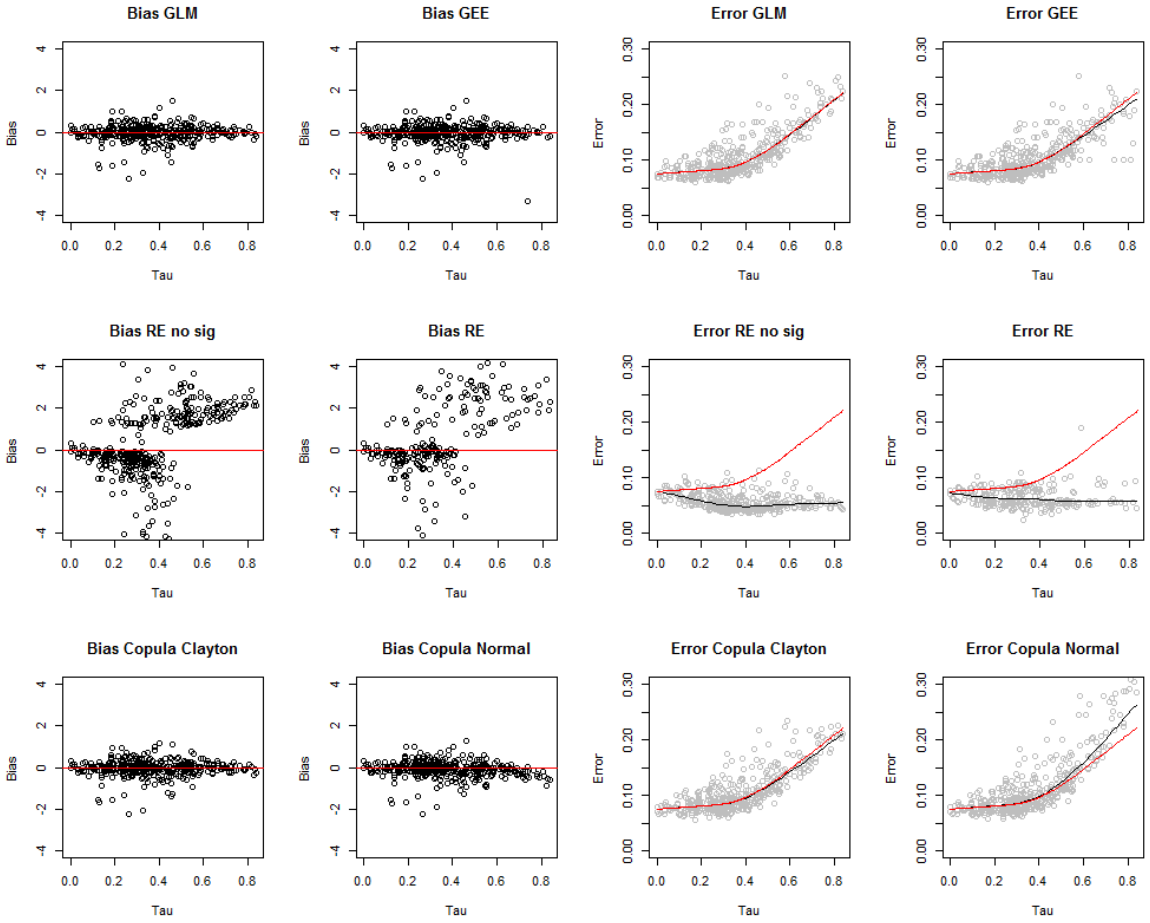


Figure 2.2: Bias and standard error estimates for the mean of the gamma distribution at time 1 across all realisations of the simulated bivariate gamma.

The red line on the left six charts is the line representing zero bias in the model estimates.

The grey line on the right six charts is the smoothed curve fit to the error of the given model.

The red line on the right six charts is the smoothed curve fit to the GLM error as a reference point.

For the same realisations, the GLM, GEE and Copula models exhibit no particular bias in the parameter estimate for time 1 while the GLMMs have a significant bias especially where τ is higher (see Table 2.4). Regardless of the presence of the additional parameter for estimating σ at both time points, this bias is still exhibited in the GLMM estimates.

Model / Tau	0-0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8+
GLM	0.0332	-0.0103	-0.0052	0.0515	-0.0564
GEE	0.0332	-0.0103	-0.0052	0.0398	-0.1039
GLMM no sig	-0.1644	-0.3231	-0.4120	1.7724	2.1497
GLMM	-0.1757	-0.3117	-0.2317	2.0494	2.3190
GJRM (Clayton)	0.0398	-0.0012	-0.0047	0.0437	-0.0663
GJRM (Normal)	0.0329	-0.0045	-0.0156	-0.0851	-0.3531

Table 2.4: Median bias for estimates of $\beta_{\mu 1}$

On the other hand, the Clayton GJRM provides comparable time 1 standard error estimates to the GEE and GLM while maintaining minimal bias.

Note that the normal copula GJRM exhibits slightly higher standard error estimates, above that of the GLM and GEE, and is slightly biased for very high τ. This bias and higher standard error is due to the normal copula not being an appropriate fit for the high skewness and correlation in the joint dependence structure in these cases.

Parameter estimates at time two

Across all realisations of the bivariate distribution, all models remain relatively unbiased in their estimates of the mean at time two across the differing values of τ (rank correlation) except for the time-variant sigma GLMM. See figure 2.3 and table 2.6.

For the estimate of the effect of time 1, the copula models significantly outperform the GLM and GEE in terms of standard error across the range of realisations and values of τ while the random effect continues to present the same trend of decreasing standard errors for parameter estimates when there is higher tau.

Model / Tau	0-0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8+
GLM	0.1082	0.1122	0.1294	0.1983	0.2968
GEE	0.1082	0.1122	0.1294	0.1948	0.2623
GLMM no sig	0.1004	0.0793	0.0663	0.0724	0.0762
GLMM	0.0998	0.0861	0.0872	0.0922	0.0909
GJRM (Copula)	0.0776	0.0796	0.0932	0.1363	0.1993
GJRM (Normal)	0.0776	0.0804	0.0947	0.1497	0.2646

Table 2.5: Median standard error for estimates of $\beta_{\mu 2}$

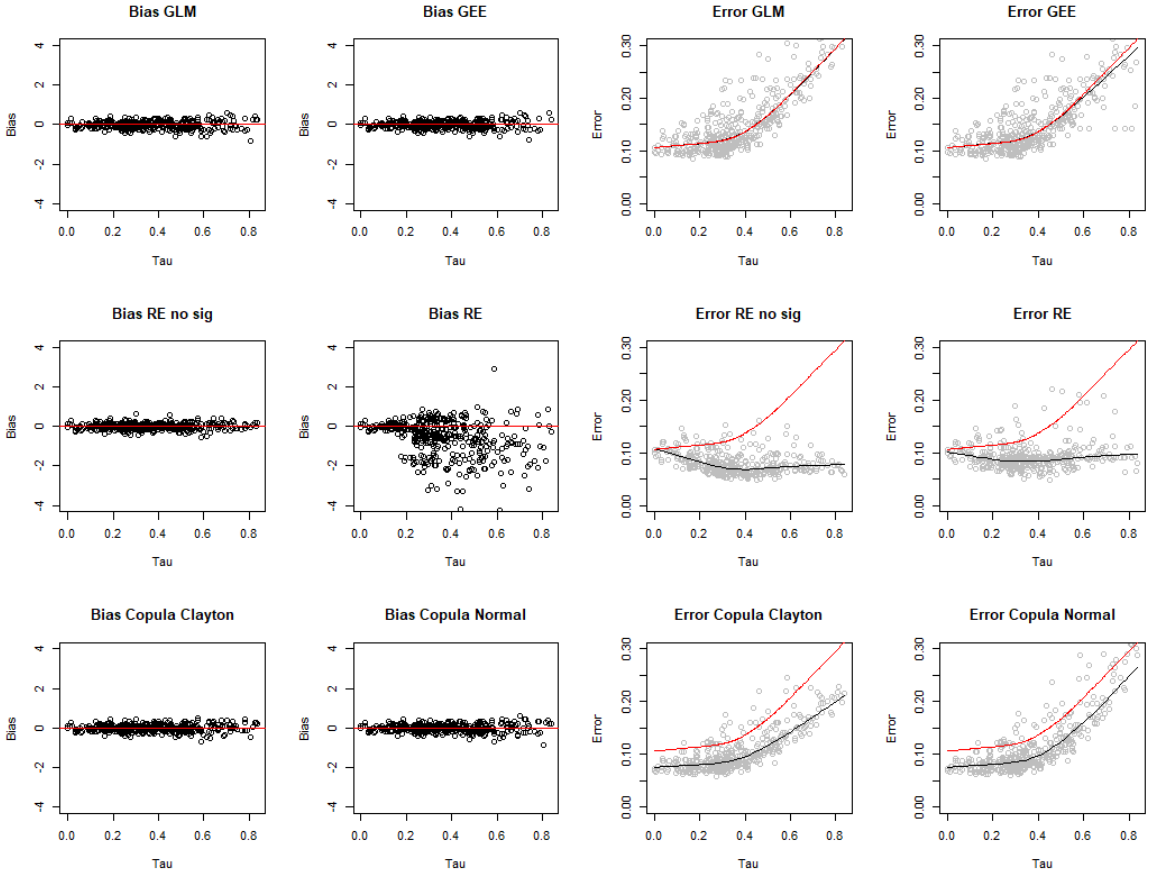


Figure 2.3: Bias and standard error estimates for the mean of the gamma distribution at time 2 across all realisations of the simulated bivariate gamma.

The red line on the left six charts is the line representing zero bias in the model estimates.

The grey line on the right six charts is the smoothed curve fit to the error of the given model.

The red line on the right six charts is the smoothed curve fit to the GLM error as a reference point.

The GLMM with the additional parameter for σ presents a significantly higher variation in observed bias than even the comparable GLMM without the additional parameter. In part, this increased bias for the random effect model with the additional parameter may be explained by the model's increased difficulty in converging, with multiple of these models failing to converge even over an extremely large number of iterations and being excluded from the charts in Figures 2.2 and 2.3.

Model / Tau	0-0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8+
GLM	-0.0164	-0.0215	-0.0196	-0.0461	0.1072
GEE	-0.0164	-0.0215	-0.0196	0.0083	0.1072
GLMM no sig	-0.0256	0.0042	-0.0262	0.0177	0.0280
GLMM	-0.0255	-0.1255	-0.5133	-0.8388	-1.0586
GJRM (Copula)	-0.0196	-0.0192	-0.0174	0.0049	0.0704
GJRM (Normal)	-0.0158	-0.0226	-0.0116	-0.0078	0.1626

Table 2.6: Median bias for estimates of $\beta_{\mu 2}$

2.2.4 The effect of marginal skew on estimate bias

Further analysis of the drivers of bias for the random effect model indicates that marginal skewness also plays a key role alongside rank correlation.

Across the set of realisations it is clear that increasing marginal skewness holding τ constant or increasing τ and holding skewness constant will increase time 1 bias for the random effect models. This is shown in table 2.7 for the time-variant-σ random effect model and table 2.8 for the non-time-variant-σ random effect model with no apparent differences in the trend of the bias between the two models. Interestingly, note that the GLM, GEE and GJRM with Clayton copula remain relatively unbiased across all values of τ and marginal skewness. However, the GJRM with Normal copula becomes biased when marginal skewness and τ are very high, likely because at this point the shape of the dependence is highly skewed and non-normal so the fit is not appropriate. These bias results are shown in tables 2.9, 2.10, 2.11 and 2.12. The values in the tables are calculated on the difference between the transformed results (e^{β_i}) and the true simulated intercept.

	Marginal Skewness						
Tau	1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5
0-0.1	-0.06	-0.05	-0.09				
0.1-0.2	-0.14	-0.18	-0.08				
0.2-0.3	-0.18	-0.28	-0.55	-0.61			
0.3-0.4	-0.10	-0.25	-0.54	-0.67	-0.91	-0.77	
0.4-0.5	-0.28	-0.34	-0.49	-0.79	-0.91	-0.88	-0.95
0.5-0.6		-0.32	-0.58	-0.61	-0.82	-0.88	-0.93
0.6-0.7				-0.79	-0.76	-0.89	-0.98
0.7-0.8					-0.90	-0.90	-0.98
0.8-0.9							-0.98

Table 2.7: Average bias for time-variant sigma random effect model for differing values of τ and marginal skew of the base distribution

	Marginal Skewness						
Tau	1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5
0-0.1	-0.06	-0.05	-0.11				
0.1-0.2	-0.09	-0.15	-0.08				
0.2-0.3	-0.16	-0.22	-0.38	-0.45			
0.3-0.4	-0.18	-0.26	-0.39	-0.49	-0.72	-0.78	
0.4-0.5	-0.26	-0.32	-0.42	-0.59	-0.76	-0.75	-0.91
0.5-0.6		-0.28	-0.52	-0.59	-0.75	-0.84	-0.97
0.6-0.7				-0.73	-0.78	-0.86	-0.97
0.7-0.8					-0.81	-0.90	-0.96
0.8-0.9							-0.98

Table 2.8: Average bias for non-time-variant sigma random effect model for differing values of τ and marginal skew of the base distribution.

	Marginal Skewness						
Tau	1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5
0-0.1	-0.01	0.00	0.03				
0.1-0.2	-0.01	-0.01	0.10				
0.2-0.3	0.00	-0.02	-0.06	-0.03			
0.3-0.4	0.01	-0.01	-0.01	0.05	-0.04	0.00	
0.4-0.5	-0.01	-0.02	0.01	0.05	-0.11	0.13	0.42
0.5-0.6		0.05	-0.03	0.05	-0.04	0.13	-0.03
0.6-0.7				-0.03	0.01	-0.04	-0.13
0.7-0.8					0.24	0.02	0.08
0.8-0.9							0.21

Table 2.9: Average bias for GJRM with Clayton Copula for differing values of τ and marginal skew of the base distribution.

	Marginal Skewness						
Tau	1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5
0-0.1	-0.01	0.00	0.03				
0.1-0.2	-0.01	-0.01	0.10				
0.2-0.3	-0.01	-0.02	-0.05	-0.02			
0.3-0.4	0.00	0.00	0.00	0.07	0.00	0.08	
0.4-0.5	0.00	-0.02	0.01	0.07	-0.01	0.23	0.67
0.5-0.6		0.07	0.00	0.11	0.05	0.26	0.30
0.6-0.7				0.09	0.15	0.18	0.35
0.7-0.8					0.51	0.34	0.75
0.8-0.9							1.52

Table 2.10: Average bias for GJRM with Normal Copula for differing values of τ and marginal skew of the base distribution

	Marginal Skewness						
Tau	1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5
0-0.1	-0.01	0.00	0.03				
0.1-0.2	-0.01	-0.01	0.10				
0.2-0.3	-0.01	-0.02	-0.06	-0.04			
0.3-0.4	0.00	-0.01	-0.01	0.05	-0.04	0.03	
0.4-0.5	0.00	-0.02	-0.01	0.02	-0.08	0.14	0.44
0.5-0.6		0.05	-0.04	0.05	-0.04	0.06	-0.03
0.6-0.7				-0.05	-0.01	-0.04	-0.10
0.7-0.8					0.20	-0.01	0.05
0.8-0.9							0.18

Table 2.11: Average bias for the GLM for differing values of τ and marginal skew of the base distribution

	Marginal Skewness						
Tau	1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5
0-0.1	-0.01	0.00	0.03				
0.1-0.2	-0.01	-0.01	0.10				
0.2-0.3	-0.01	-0.02	-0.06	-0.04			
0.3-0.4	0.00	-0.01	-0.01	0.05	-0.04	0.03	
0.4-0.5	0.00	-0.02	-0.01	0.02	-0.08	0.14	0.44
0.5-0.6		0.05	-0.04	0.05	-0.04	0.06	error
0.6-0.7				-0.05	-0.01	-0.04	error
0.7-0.8					0.20	-0.01	error
0.8-0.9							error

Table 2.12: Average bias for the GEE for differing values of τ and marginal skew of the base distribution. Note the cells denoted as 'error' are cases where the GEE has provided an extreme estimate which results in an extremely large difference which cannot be displayed but seems likely to be a convergence error.

2.2.5 Exploration of an extreme case

It is informative to refer to an extreme case of the bivariate distribution with pre-mixing means as $\mu_1 = 1$, $\mu_2 = 2$, beta parameters as $\alpha = .25$ and $\beta = 1.75$, giving $E(Y_1) = 0.25$ and $E(Y_2) = 0.125$. The shape and dependence structure of the simulated dataset is shown in figure 2.4.

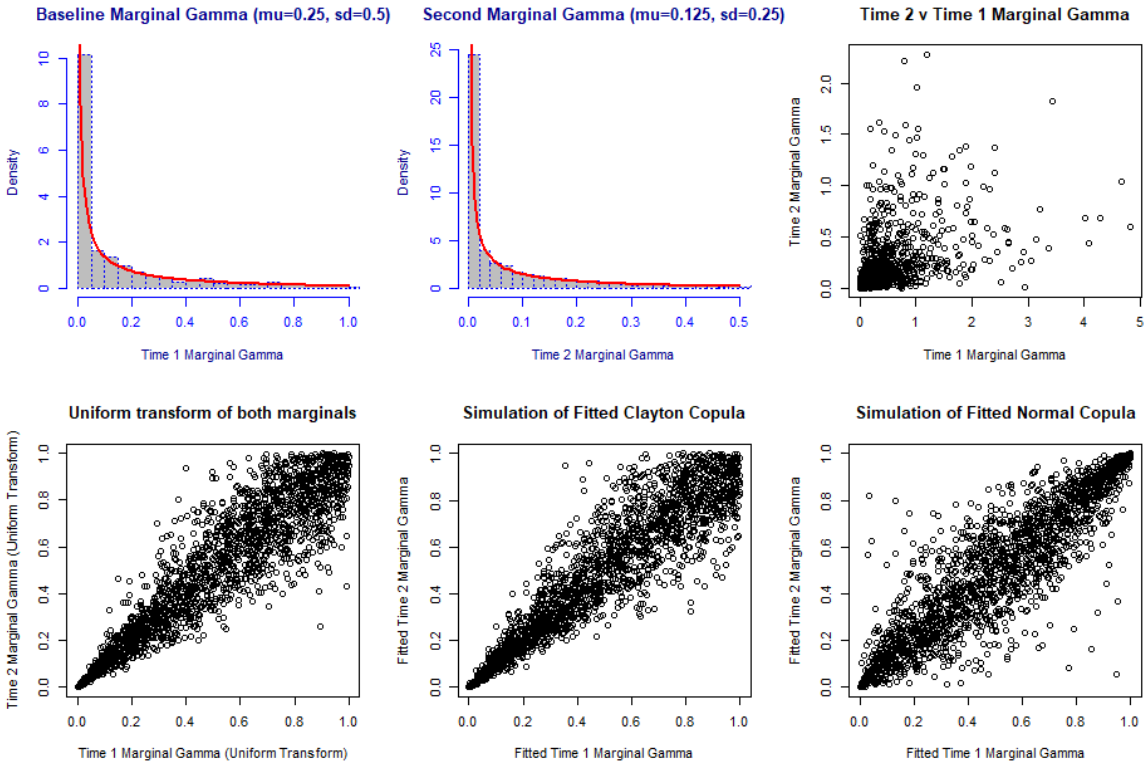


Figure 2.4: From top left to bottom right: 1. Distribution of time one marginal gamma, 2. Distribution of the second marginal gamma for time two, 3. Scatterplot of the time two against time one marginal gamma, 4. Scatterplot of the uniform transform of the time two marginal gamma against the uniform transform of the time one marginal gamma, 5. Scatterplot of a simulation of a fitted Clayton copula, 6. Scatterplot of a simulation of a fitted Normal copula

It's apparent from the plot of the two uniform transforms of the marginals against one another (plot 4) that there is a symmetrically decreasing dependence for larger values of both variables. As can be seen in the fifth chart, the Clayton copula clearly captures this decreasing dependence while the normal copula maintains the same level of dependence throughout. The rank correlation is very high with a Kendall's tau of 0.75 and positive skewness is high

at 4.0 and 3.6 for margins for time one and two respectively.

Model	β_{μ_1}	β_{μ_2}	**SE**(β_{μ_1})	**SE**(β_{μ_2})
Generalised linear model	-1.4882	-0.6414	0.1743	0.2466
Generalised estimating equations	-1.4882	-0.6414	0.1737	0.2453
Generalised linear mixed model (4)	-4.4295	-0.7473	0.0523	0.0739
Generalised linear mixed model (5)	-3.9410	-1.3765	0.1042	0.1042
Generalised joint regression model (Clayton)	-1.4480	-0.7324	0.1922	0.1915
Generalised joint regression model (Normal)	-1.1967	-0.6519	0.2284	0.2300
Simulated true intercepts	**-1.3862**	**-0.6931**		

Table 2.13: Extreme case simulation: parameter estimates and standard errors from the six models against the simulation parameters.

As can be seen in table 2.13, the bias for the generalised linear mixed models is immediately apparent at over three times the true intercept for time one and almost twice the true value for the effect of time two over time one for the time-variant-σ GLMM. As was seen during simulations, the GLMMs exhibit extremely low error alongside the high levels of bias.

The Clayton copula is the most appropriate copula fit in this case and exhibits the lowest error for the effect of time but has a slightly higher error than the GEE and GLM for the estimate at time one. However, in general, across the range of simulations, errors for estimates at time one were similar between the GEE, GLM and copula models. Note the symmetric nature of the error estimates for the copula model compared to the GLM, GEE and GLMMs.

Part of the issue with the random effect models may be that they seem to be assigning an extremely large proportion of model variance to the random effect component. To ensure no single optimisation package or method was at fault, multiple packages which incorporate random effects have been tested including gamlss, lme4 and mgcv. These methods provided similar estimates of the time 1 and time 2 parameters provided above.

The code used for this example case simulation and analysis is available in the supplementary materials, section 5.1.1.

Observations on Computational Complexity

During the development of these simulations, the random effect models, compared to each of the other models, required significantly longer amounts of time to run and this difference widened with increasing sample size. To demonstrate this at a high level, the extreme case in the section above has been rerun ten times for different sample sizes of $n = 100,\ 500,\ 1,000,\ 5,000,$ and $10,000$ with the same randomisation seed and hardware.

In table 2.14, average runtimes across the 10 models at each sample size are provided alongside a scaled calculation of runtime per 1,000 observations. Note that across the range of sample sizes, the runtime for the random effect models is approximately 5-10 times higher than the two copula models which include the same number of parameters. While these total runtime values are not particularly impactful in this case due to the small sample sizes, as model sample sizes increase in the era of big data, it's likely that these computational differences will become increasingly more noticeable.

	Average model runtime (s)					Runtime per 1,000 obs (s)				
Sample	100	500	1,000	5,000	10,000	100	500	1,000	5,000	10,000
GLM	0.01	0.01	0.02	0.10	0.20	0.06	0.02	0.02	0.02	0.02
GEE	0.01	0.04	0.07	0.37	0.65	0.12	0.07	0.07	0.07	0.06
RE (4)	0.11	0.31	9.90	4.29	20.33	1.05	0.61	9.90	0.86	2.03
RE (5)	0.42	1.71	1.41	7.57	14.73	4.25	3.42	1.41	1.51	1.47
GJRM (N)	0.08	0.18	0.35	1.68	3.81	0.83	0.37	0.35	0.34	0.38
GJRM (C)	0.06	0.19	0.35	1.64	3.71	0.62	0.38	0.35	0.33	0.37

Table 2.14: Average model runtimes across different sample sizes

These operations were performed on a Windows operating system with an i7-8750H CPU at 2.2GHz with 32GB RAM without overclocking.

3 Applications

Three datasets are introduced which have similar distributional properties to those described in simulations. The parameters compared in each table of results are as follows:

- β_1, the estimate for the mean at time 1,
- β_t, the estimate for the incremental effect of time 2 over time 1,
- $SE(\beta_1)$ and $SE(\beta_t)$, the standard errors of both of the estimates.

Note that in this setting, compared to the simulation setting, the marginals are not known in advance so both their parameters and shape must be estimated and fit to be able to create the chart of the uniform transform of the variables against one another and to interrogate the shape of the dependence structure.

3.1 ASX200 share prices

The dataset to be considered is share prices on the first trading day of 1998 and 2018 for ASX200 stocks which are included in the index on both dates, resulting in 143 shares (ASXHistoricalData.com). It is of interest to understand the amount by which these stocks have increased in value and the way in which differently priced shares are correlated between time points.

Share prices at times 1 and 2 are weakly correlated, with Pearson correlation of 0.29 and Kendall's tau of 0.35. The marginal distributions are highly positively skewed with skewness of 3.6 and 4.2 for the margins at time 1 and 2 respectively. The charts in figure 3.1 demonstrate that while the prices at both time points are not strongly correlated overall, there is a larger degree of dependence for lower priced shares. This can be seen in Figure 3.1 in the higher density towards the left of the bivariate density plot of the dependence structure (right plot).

Joe	Gumbel	Hougaard	FGM	AMH	Clayton	Normal	Frank	Plackett
-668	-664	-664	-663	-660	-659	-659	-655	-654

Table 3.1: ASX200 Share Prices: AIC values for each copula fit to the dependence structure

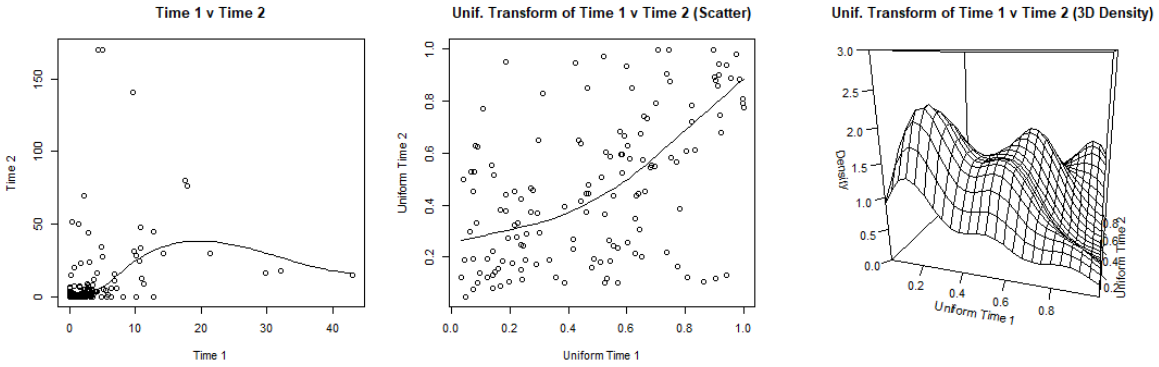


Figure 3.1: ASX200 Share Prices: left to right: scatterplot of raw observations at time 2 against time 1 with a fitted loess curve, scatterplot of the uniform transform of both marginals with a fitted loess curve, plot of bivariate density calculated using the package kde

Each of the models applied in the previous chapter have been applied to this dataset: a GLM, GEE, random effect model, random effect model with additional parameter for sigma at both time points, and nine joint regression models with the same marginal structure but a different copula function. Results are shown in table 3.2.

AIC can be used to select the most appropriate copula to fit to the distribution and is calculated in table 3.1. In this case the Joe copula appears to be the most appropriate.

The random effect models provide vastly different estimates for time 1 and time 2 compared to the copula, GLM and GEE, with significantly lower errors. This is consistent with the simulations in the previous chapter, in that the random effect model estimates are extremely different to all the other models with lower errors, but potentially have quite a high bias due to the high level of marginal skew with some rank correlation in the dataset.

Note that the copula models perform significantly better than the comparable GLM in terms of error for time 1 and 2, and present a similar error to the GLMM with the additional parameter for σ fit for both time points without the difference in estimate.

Model	β_1	β_t	$SE(\beta_1)$	$SE(\beta_t)$	e^{β_1}	e^{β_2}
GLM	1.3260	1.0836	0.1702	0.2408	3.77	11.13
GEE	1.3260	1.0836	0.1345	0.0576	3.77	11.13
Random Effect	0.5989	0.2226	0.0804	0.1137	1.82	2.27
Random Effect w/ Sig	0.2595	2.0866	0.0000	0.1510	1.30	10.44
Clayton	1.3989	1.0504	0.1177	0.1562	4.05	11.58
Normal	1.3733	1.0647	0.1123	0.1544	3.95	11.45
Joe	1.3843	1.0811	0.1142	0.1577	3.99	11.77
Gumbel	1.4203	1.0847	0.1149	0.1580	4.14	12.24
Frank	1.4314	1.0298	0.1084	0.1488	4.18	11.72
AMH	1.3960	1.0581	0.1150	0.1536	4.04	11.64
FGM	1.3859	1.0455	0.1068	0.1469	4.00	11.37
Plackett	1.4085	1.0054	0.1075	0.1469	4.09	11.18
Hougaard	1.4203	1.0847	0.1149	0.1580	4.14	12.24

Table 3.2: ASX200 Share Prices: estimates for the mean and standard error of parameters fit to the ASX200 data

3.2 Avocado prices

The dataset considered is the average price of conventional and organic avocados across 54 regions in the United States at two time points, 1 April 2015 and 25 March 2018 (Hass Avocado Board). The dates were chosen based on being the first and last weeks of observations available.

Prices at times 1 and 2 are reasonably strongly correlated with a Pearson correlation of 0.72 and a Kendall's tau of 0.54. However, note that the marginal distributions are not highly skewed with margin one being only very weakly positively skewed with a skewness of 0.46 and margin two having a skewness close to zero at -0.18. The bivariate dependence is slightly skewed to lower value dependence but still exhibits higher value dependence as can be seen in the right hand chart of bivariate density in figure 3.2.

FGM	Joe	Frank	Plackett	AMH	Gumbel	Hougaard	Clayton	Normal
-41	-32	-25	-23	-23	-23	-23	-21	-19

Table 3.3: Avocado Prices: AIC values for each copula fit to the dependence structure

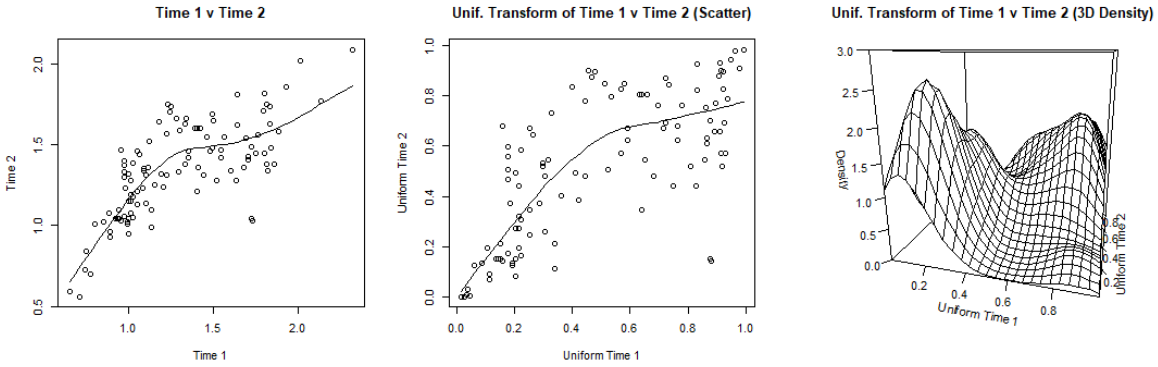


Figure 3.2: Avocado Prices: left to right: scatter-plot of raw observations at time 2 against time 1 with a fitted loess curve, scatter-plot of the uniform transform of both marginals with a fitted loess curve, plot of bivariate density calculated using the package kde

AIC can be used to select the most appropriate copula to fit to the distribution and has been calculated in table 3.2. In this case the FGM copula appears to be the most appropriate.

Each of the models applied in the previous chapter have been applied to this dataset: a GLM, GEE, random effect model, random effect model with additional parameter for sigma at both time points, and nine joint regression models with the same marginal structure but a different copula function. Results are shown in Table 3.4.

For this example, the random effect models provide only slightly differing estimates to the GLM and copula models at approximately 13 percent and 8 percent lower than GLM time 1 estimates and 13 percent higher and 16 percent lower than GLM time 2 estimates.

While in this example, rank correlation, τ, is high, marginal skew is extremely weak which may explain the lower level of difference between the GLMM and other estimates.

Model	β_1	β_t	$SE(\beta_1)$	$SE(\beta_t)$	e^{β_1}	e^{β_2}
GLM	0.2634	0.0344	0.0239	0.0338	1.30	1.35
GEE	0.2634	0.0344	0.0266	0.0011	1.30	1.35
Random Effect	0.2310	0.0454	0.0096	0.0135	1.26	1.32
Random Effect w/ Sig	0.2430	0.0292	0.0181	0.0181	1.28	1.31
Clayton	0.2514	0.0660	0.0243	0.0239	1.29	1.37
Normal	0.2634	0.0346	0.0264	0.0217	1.30	1.35
Joe	0.2646	0.0304	0.0287	0.0219	1.30	1.34
Gumbel	0.2724	0.0301	0.0274	0.0214	1.31	1.35
Frank	0.2264	0.0594	0.0284	0.0214	1.25	1.33
AMH	0.2791	0.0452	0.0221	0.0204	1.32	1.38
FGM	0.2517	0.0421	0.0245	0.0198	1.29	1.34
Plackett	0.2442	0.0488	0.0271	0.0212	1.28	1.34
Hougaard	0.2724	0.0301	0.0274	0.0214	1.31	1.35

Table 3.4: Avocado Prices: Estimates for the mean and standard error of parameters fit to the Avocado data

3.3 Triglyceride levels

The dataset considered is observations of triglyceride levels during a clinical trial of hormone replacement therapy. There were 72 participants at baseline and 24 months after the start of treatment (Nand et al., 1999).

It it of interest to understand how trigliceryde levels change over time for trial participants and whether there is any difference in dependence for individuals with different starting levels of triglycerides.

Observations at time 1 (0 months) and time 2 (24 months) are reasonably correlated with a pearson correlation of 0.62 and a Kendall's tau of 0.48. The marginal distributions for time one and two are slightly positively skewed with skewness of 1.14 and 1.27.

Clayton	FGM	Joe	AMH	Gumbel	Hougaard	Normal	Frank	Plackett
-74	-71	-71	-68	-68	-68	-67	-62	-62

Table 3.5: Tryglyceride Levels: AIC values for each copula fit to the dependence structure

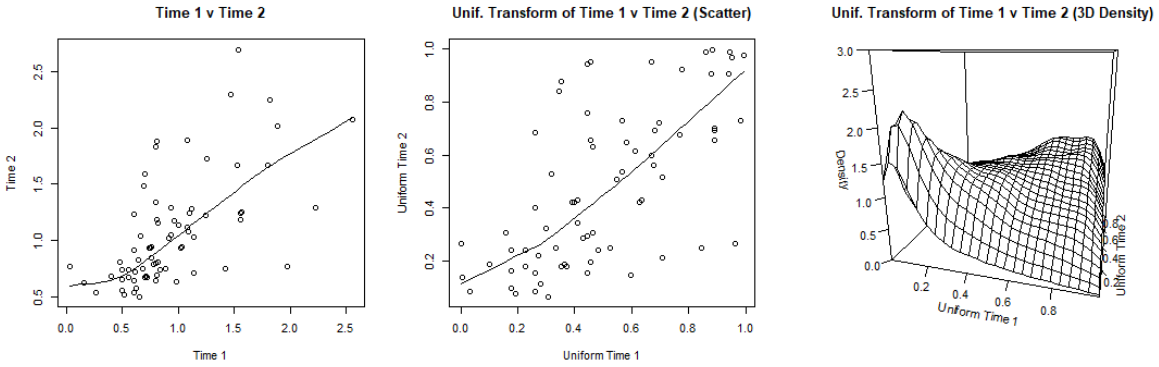


Figure 3.3: Tryglyceride Levels: left to right: scatter-plot of raw observations at time 2 against time 1 with a fitted loess curve, scatter-plot of the uniform transform of both marginals with a fitted loess curve, plot of bivariate density calculated using the package kde

AIC can be used to select the most appropriate copula to fit to the distribution and are shown in table 3.5. In this case the Clayton copula appears to be the most appropriate, as was the case in for the simulated distribution in chapter 2.

Each of the models applied in the previous chapter have been applied to this dataset: a GLM, GEE, random effect model, random effect model with additional parameter for sigma at both time points, and nine joint regression models with the same marginal structure but a different copula function. Results are shown in Table 3.6.

In this example, the GLMM provides an almost three times higher estimate for the intercept at time 1 compared to the GLM, GEE and Clayton copula model but estimates for the effect of time 2 are similar between each of the models. The combination of a moderate marginal skew with a moderate rank correlation, τ, may be the reason for the large differences.

The code used for the analysis of these datasets is available in the supplementary materials, section 5.1.2.

Model	β_1	β_t	$SE(\beta_1)$	$SE(\beta_t)$	e^{β_1}	e^{β_2}
GLM	-0.0569	0.1249	0.0559	0.0791	0.94	1.07
GEE	-0.0569	0.1249	0.0582	0.0062	0.94	1.07
Random Effect	-0.1475	0.1487	0.0307	0.0434	0.86	1.00
Random Effect w/ Sig	-0.0976	0.0792	0.0516	0.0516	0.91	0.98
Clayton	-0.0076	0.0727	0.0681	0.0482	0.99	1.07
Normal	-0.0579	0.1234	0.0609	0.0483	0.94	1.07
Joe	-0.0666	0.1387	0.0601	0.0502	0.94	1.07
Gumbel	-0.0477	0.1258	0.0607	0.0505	0.95	1.08
Frank	-0.0473	0.0978	0.0571	0.0501	0.95	1.05
AMH	-0.0037	0.0887	0.0588	0.0463	1.00	1.09
FGM	-0.0458	0.1124	0.0571	0.0461	0.96	1.07
Plackett	-0.0395	0.0901	0.0568	0.0493	0.96	1.05
Hougaard	-0.0477	0.1258	0.0607	0.0505	0.95	1.08

Table 3.6: Tryglyceride Levels: Estimates for the mean and standard error of parameters fit to the Triglycerides data

4 Conclusion

Simulations were run of a two time point longitudinal dataset generated from a bivariate distribution with gamma marginal distributions and varying non-standard dependence structures. It was identified across these simulations that random effect based regression models provide biased estimates of the mean of the dataset at time 1 and, in some cases, time 2, even though the marginal distributions are correctly specified. Results of the simulations indicate that the higher the skewness of the marginal distributions or the rank correlation between the outcome variables, measured by Kendall's tau, the more pronounced the bias of the random effect based model is likely to be. Concerningly, in the majority of cases analysed, where random effect model estimates are biased, the models also provide lower estimates of standard error compared to other tested models for the fitted biased parameters.

In contrast to the random effect models, copula-based joint regression models, in particular the GJRM implementation (Marra and Radice, 2017), provide a relatively unbiased estimate of the mean of the dataset for time 1 and time 2 across these simulations. Compared to the GLM and GEE, the copula-based joint regression model also has the advantage that it captures the dependence structure between the random variables so provides generally lower standard error estimates for model parameters, especially for the estimate for time two.

The large differences in bias and standard error for parameter estimates between random effect based models and copula-based joint regression models for longitudinal data indicate that a clear approach is required for selecting when a random effect or copula-based model is more appropriate for any given longitudinal dataset. The results of this study indicate that mixed model with random effect terms may not be appropriate where marginal distributions of the joint distribution are skewed and / or rank correlation is high. In these cases, random effect based model estimates should be thoroughly interrogated, compared to alternative methods prior to their use or be replaced with copula-based joint regression models.

5 Discussion

Both the random effect and copula-based joint regression frameworks can capture the same set of parameters for a bivariate distribution, however, the random effect model has two restrictive assumptions that the copula-based joint regression does not: the assumption of normally distributed covariance for the random effect term, and, the same assumed marginal distribution for each time point. In principle it is possible to superimpose a random effect term on two different marginal distributions however we are unaware of software for implementation of such models. For this reason it does not seem surprising that many bivariate distributions fall outside these assumptions and are therefore likely to cause biased or inefficient estimates to be generated from a random effect model.

This thesis provides an initial method for identifying bivariate distributions for which random effect based models are not appropriate, through reviewing whether marginal skewness and rank correlation are present in the bivariate distribution. Further investigation into a broader range of joint distributions may provide a more comprehensive set of rules for identifying which bivariate distributions can reasonably have a random effect model applied to them.

Note that no additional covariates other than time were included in these simulations or applications. Further work is required to understand the extent to which incorporating a covariate impacts the bias and efficiency of copula based joint regression model estimates as compared to random effect based models. An additional consideration in the case of incorporating additional covariates to the copula-based joint regression model will be in parameter comparability whereby a single covariate in random effect model framework may require multiple parameters within the current copula-based joint regression frameworks.

Extensions of copula regression to longitudinal regression estimates in the multivariate case have not been analysed in this thesis and pose an extremely large potential area for further research. The greater flexibility, higher level of transparency and computational simplicity of the copula regression poses significant opportunities for improving the methods used for longitudinal analysis.

References

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Washington, D.C., 1972.

E. F. Acar, R. V. Craiu, and F. Yao. Dependence Calibration in Conditional Copulas: A Nonparametric Approach. *Biometrics*, 67(2):445–453, 2011.

E. F. Acar, R. V. Craiu, and F. Yao. Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics*, 7:2822–2850, 2013.

ASXHistoricalData.com. ASX Market Data. https://www.asxhistoricaldata.com/. Accessed: 2018-03-28.

N. Balakrishnan and C. D. Lai. *Continuous bivariate distributions*. Springer Science & Business Media, 2009.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

A. N. E. Breslow and D. G. Clayton. Approximate Inference in Generalized Linear Mixed Models Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

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

I. Gijbels and N. Veraverbeke. Estimation of a Copula when a Covariate Affects only Marginal Distributions. *Scandinavian Journal of Statistics*, 42:1109–1126, 2015.

I. Gijbels, N. Veraverbeke, and M. Omelka. Conditional copulas, association measures and their applications. *Computational Statistics and Data Analysis*, 55(5):1919–1932, 2011.

Hass Avocado Board. US Hass Avocado Prices. https://www.hassavocadoboard.com/retail/volume-and-price-data. Accessed: 2018-03-28.

T. Hastie and R. Tibshirani. *Generalized additive models*. CRC Press, 1990.

S. Hojsgaard, U. Halekoh, and J. Yan. The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15/2:1–11, 2006.

H. Joe. *Multivariate models and dependence concepts*. Chapman & Hall, 1997.

N. Klein, T. Kneib, S. Lang, and A. Sohn. Bayesian structured additive distributional regression with an application to regional inequality in Germany. *The Annals of Applied Statistics*, 9(2):1024–1052, 2015.

N. Kolev and D. Paiva. Copula-based regression models: A survey. *Journal of Statistical Planning and Inference*, 139:3847–3856, 2009.

N. Krämer, E. C. Brechmann, D. Silvestrini, and C. Czado. Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, 53(3):829–839, 2013.

N. M. Laird and J. H. Ware. Random-Effects Models for Longitudinal Data. *Biometrics*, 38 (4):963–974, 1982.

K. Liang and S. L. Zeger. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73(1):13–22, 1986.

G. Marra and R. Radice. Bivariate copula additive models for location, scale and shape. *Computational Statistics and Data Analysis*, 112:99–113, 2017.

S. Nadarajah and A. K. Gupta. Some bivariate gamma distributions. *Applied Mathematics Letters*, 19(8):767–774, 2006.

S. L. Nand, B. G. Wren, B. A. Gross, G. Z. Heller, et al. Bone density effects of continuous estrone sulfate and varying doses of medroxyprogesterone acetate. *Obstetrics & Gynecology*, 93(6):1009–1013, 1999.

R. B. Nelsen. *An Introduction to Copulas, Second Edition*. 2007.

H. P. Palaro and L. K. Hotta. Using Conditional Copula to Estimate Value at Risk. *Journal of Data Science*, 4:93–115, 2006.

A. Patton. Modeling asymmetric exchange rate dependence. *International Economic Review*, 47(2):527–556, 2006.

M. Pitt, D. Chan, and R. Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554, sep 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

A. Sabeti, M. Wei, and R. V. Craiu. Additive models for conditional copulas. *Stat*, 3(1):300–312. doi: 10.1002/sta4.64. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.64.

A. Sklar. Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6): 449–460, 1973.

D. M. Stasinopoulos, R. Rigby, G. Heller, V. Voudouris, and F. De Bastiani. *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press, 2017.

P. K. Trivedi and D. M. Zimmer. *Copula modeling : an introduction for practitioners*. Now Publishers, Boston, 2007.

T. Vatter and V. Chavez-Demoulin. Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141:147–167, 2015.

N. Veraverbeke, M. Omelka, and I. Gijbels. Estimation of a Conditional Copula and Association Measures. *Scandinavian Journal of Statistics*, 38:766–780, 2011.

C. J. Wild and T. W. Yee. Additive Extensions to Generalized Estimation Equation Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4):711–725, 1996.

S. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.

T. W. Yee and C. J. Wild. Vector Generalized Additive Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(3):481–493, 1996.

G. F. Yeo and R. K. Milne. On characterizations of beta and gamma distributions. *Statistics and Probability Letters*, 11(3):239–242, mar 1991.

Supplementary Materials

5.1 Code

5.1.1 Example bias case

```r
#Importing required packages
require(gamlss); require(gee); require(lme4);
require(mgcv); require(geepack)

#Setting parameters for extreme case example
set.seed(1)
a=0.25; b=1.75; mu1=1; mu2=2; n=100

#Simulating bivariate gamma of Nadarajah and Gupta
w<-rbeta(n,a,b)
gamma_c_mu1<-w * rgamma(n,shape=a+b,scale=1/mu1)
gamma_c_mu2<-w * rgamma(n,shape=a+b,scale=1/mu2)

#Setting up data in correct format for random effect model
patient<-as.factor(seq(1:n))
dataset<-as.data.frame(rbind(cbind(patient,gamma_c_mu1,0)
                             ,cbind(patient,gamma_c_mu2,1)))
colnames(dataset)<-c("patient","random_variable","time")

#Running GLM, GEE and multiple GLMM packages
model_glm <- glm(random_variable~as.factor(time==1),data=dataset
                 ,family=Gamma(link = "log"),maxit=10000)
model_gee<-geese(random_variable~as.factor(time==1), id=patient
                 , data=dataset, family=Gamma(link="log")
                 , mean.link = "log", corstr = "exchangeable"
                 , control=geese.control(trace=TRUE, maxit=10000))
model_lme4<-glmer(random_variable~as.factor(time==1) + (1 | patient)
                 , data=dataset, family=Gamma(link="log"))
model_gamm<-gamm(random_variable~as.factor(time==1)
                 , random = list(patient=~1)
                 ,data=dataset, family=Gamma(link="log")
                 ,niterPQL=1000)
model_re_nosig <- gamlss(random_variable~as.factor(time==1)
                         + random(as.factor(patient))
                 , data=dataset, family=GA(), method=RS())
model_re <- gamlss(formula=random_variable~as.factor(time==1)
```

```
                    + random(as.factor(patient))
                   , sigma.formula=~as.factor(time==1)
                   , data=dataset, family=GA() , method=RS())

#Running GJRM - note GJRM package is loaded after gamlss
require(GJRM)
eq.mu.1 <- gamma_c_mu1~1
eq.mu.2 <- gamma_c_mu2~1
fl <- list(eq.mu.1, eq.mu.2)
model_copula<-gjrm(fl, margins = c("GA" , "GA"), BivD = "C0"
                    , data=data.frame(gamma_c_mu1,gamma_c_mu2), Model="B")
model_copula_n<-gjrm(fl, margins = c("GA" , "GA"), BivD = "N"
                    , data=data.frame(gamma_c_mu1,gamma_c_mu2), Model="B")
```

5.1.2 Applications

```
#Lipids Data
require(sas7bdat)
lipid <- read.sas7bdat("LipidsData.sas7bdat")
lipids_merged<-(merge(lipid[lipid$MONTH==0,],lipid[lipid$MONTH==24,],by="PATIENT"))
gamma_c_mu1<-lipids_merged$TRG.x
gamma_c_mu2<-lipids_merged$TRG.y

#Stock prices over 10 years
ASX2018<-read.table("20180102.txt", header=FALSE, sep=",")
ASX1998<-read.table("19980102.txt", header=FALSE, sep=",")
ASX98_18<-merge(ASX1998,ASX2018,by="V1")
gamma_c_mu1<-ASX98_18$V6.x
gamma_c_mu2<-ASX98_18$V6.y

#Avocado prices
avo<-read.table("avocado prices.csv", header=T, sep=",")
gamma_c_mu1<-avo[avo$Date=="4/01/2015","AveragePrice"]
gamma_c_mu2<-avo[avo$Date=="25/03/2018","AveragePrice"]

#Setting up data in correct structure for random effect
n=length(gamma_c_mu1)
patient<-as.factor(seq(1:n))
dataset<-as.data.frame(rbind(cbind(patient,gamma_c_mu1,0)
                             ,cbind(patient,gamma_c_mu2,1)))
colnames(dataset)<-c("patient","random_variable","time")

#Loading required pacakges
require(gamlss); require(gee); require(geepack);

#Running each of the models for the application dataset
```

```
model_glm <- glm(random_variable~as.factor(time==1), data=dataset
                 ,family=Gamma(link = "log"), maxit=10000)
model_gee<-geese(random_variable~as.factor(time==1), id=patient, data=dataset
                 , family=Gamma(link="log"), mean.link = "log",corstr = "exchangeable"
                 , control=geese.control(trace=TRUE,maxit=10000))
model_re_nosig <- gamlss(random_variable~as.factor(time==1)+random(as.factor(patient))
                 , data=dataset, family=GA(),method=RS())
model_re <- gamlss(formula=random_variable~as.factor(time==1)+random(as.factor(patient))
                    , sigma.formula=~as.factor(time==1), data=dataset, family=GA()
                    , method=CG(10000))

#Loading GJRM after gamlss to avoid overlapping packages
require(GJRM)

#Setting up GJRM equations
eq.mu.1 <- gamma_c_mu1~1
eq.mu.2 <- gamma_c_mu2~1
fl <- list(eq.mu.1, eq.mu.2)

#Running GJRM for each of the copulas tested
model_copula<-gjrm(fl, margins = c("GA" , "GA") , BivD = "C0"
                 ,data=data.frame(gamma_c_mu1,gamma_c_mu2),Model="B")
model_copula_n<-gjrm(fl, margins = c("GA" , "GA") , BivD = "N"
                 ,data=data.frame(gamma_c_mu1,gamma_c_mu2),Model="B")
model_copula_j<-gjrm(fl, margins = c("GA" , "GA") , BivD = "J0"
                 ,data=data.frame(gamma_c_mu1,gamma_c_mu2),Model="B")
model_copula_g<-gjrm(fl, margins = c("GA" , "GA") , BivD = "G0"
                 ,data=data.frame(gamma_c_mu1,gamma_c_mu2),Model="B")
model_copula_f<-gjrm(fl, margins = c("GA" , "GA") , BivD = "F"
                 ,data=data.frame(gamma_c_mu1,gamma_c_mu2),Model="B")
model_copula_amh<-gjrm(fl, margins = c("GA" , "GA") , BivD = "AMH"
                 ,data=data.frame(gamma_c_mu1,gamma_c_mu2),Model="B")
model_copula_fgm<-gjrm(fl, margins = c("GA" , "GA") , BivD = "FGM"
                 ,data=data.frame(gamma_c_mu1,gamma_c_mu2),Model="B")
model_copula_pl<-gjrm(fl, margins = c("GA" , "GA") , BivD = "PL"
                 ,data=data.frame(gamma_c_mu1,gamma_c_mu2),Model="B")
model_copula_h<-gjrm(fl, margins = c("GA" , "GA") , BivD = "HO"
                 ,data=data.frame(gamma_c_mu1,gamma_c_mu2),Model="B")
```

5.1.3 Full simulations

```
simulateCorrelatedVarNOGJRM <- function(n,a,b,mu1,mu2)  {

  set.seed(100)

  #Loading required packages
```

```
require(gamlss)
require(MASS)
require(gee)
require(VGAM)

#Simulating bivariate random variable according to functional input
w<-rbeta(n,a,b)
gamma_c_mu1<-w*rgamma(n,shape=a+b,scale=1/mu1)
gamma_c_mu2<-w*rgamma(n,shape=a+b,scale=1/mu2)

#Transforming data to format required for random effect models
patient<-as.factor(seq(1:n))
dataset<-as.data.frame(rbind(cbind(patient,gamma_c_mu1,0)
                             ,cbind(patient,gamma_c_mu2,1)))
colnames(dataset)<-c("patient","random_variable","time")

#Running generalised linear model for the realisation
model_glm <- glm(random_variable~as.factor(time==1)
                 , data=dataset
                 , family=Gamma(link = "log")
                 , maxit=1000)

#Running GLMM with no sigma time variable
model_re_nosig <- gamlss(random_variable~as.factor(time==1)+random(as.factor(patient))
                         , data=dataset
                         , family=GA())

#Running GLMM with sigma time variable
model_re <- gamlss(formula=random_variable~as.factor(time==1)+random(as.factor(patient))
                   , sigma.formula=~as.factor(time==1)
                   , data=dataset
                   , family=GA()
                   , start.from = model_re_nosig #Optional
                   , method=CG(1000))


#Running GEE
model_gee<-gee(random_variable~as.factor(time==1)
               , id=patient
               , data=dataset
               , family=Gamma(link = "log")
               , maxiter=25)

#Extracting coefficient estimates from each of the models
summary_glm<-c( summary(model_glm)$coeff[1]
                ,summary(model_glm)$coeff[2]
                ,summary(model_glm)$coeff[3]
                ,summary(model_glm)$coeff[4]
```

```
  )
  summary_gee<-c( summary(model_gee)$coeff[1]
                  ,summary(model_gee)$coeff[2]
                  ,summary(model_gee)$coeff[3]
                  ,summary(model_gee)$coeff[4]
  )

  invisible(capture.output(
    summary_re_nosig<-c( summary(model_re_nosig)[1]
                         ,summary(model_re_nosig)[2]
                         ,summary(model_re_nosig)[4]
                         ,summary(model_re_nosig)[5]
  )
  ))
  invisible(capture.output(
  summary_re<-c( summary(model_re)[1]
                 ,summary(model_re)[2]
                 ,summary(model_re)[5]
                 ,summary(model_re)[6]
  )
  ))

  summary_cop<-c( 0,0,0,0 #Blank as this is combined with GJRM function at a later point
  )
  summary_cop_n<-c( 0,0,0,0 #Blank as this is combined with GJRM function at a later point
  )

  #Calculating true simulated estimates for the distribution based on the parameters
  actuals<-c( log(a*(1/mu1))
              , -log(a*(1/mu1))+log(a*(1/mu2))
              , 0
              , 0
  )

  #Combining simulation estimates into a single table
  output<-rbind(summary_glm
                , summary_gee
                , summary_re_nosig
                , summary_re
                , summary_cop
                , summary_cop_n
                , actuals)

  colnames(output)<-c("Time 1 Intercept","Time 2 Intercept","Time 1 SE","Time 2 SE")

  return(output)
}
```

```
simulateCorrelatedVarGJRM <- function(n,a,b,mu1,mu2)  {

  set.seed(100)

  #Loading required packages
  require(GJRM)
  require(MASS)

  #Simulating bivariate random variable according to functional input
  w<-rbeta(n,a,b)
  gamma_c_mu1<-w*rgamma(n,shape=a+b,scale=1/mu1)
  gamma_c_mu2<-w*rgamma(n,shape=a+b,scale=1/mu2)

  eq.mu.1 <- gamma_c_mu1~1
  eq.mu.2 <- gamma_c_mu2~1
  fl <- list(eq.mu.1, eq.mu.2)
  model_copula <- gjrm(fl
                       , margins = c("GA" , "GA")
                       , BivD = "C0"
                       , data=data.frame(gamma_c_mu1,gamma_c_mu2)
                       , Model="B")
  model_copula_n <- gjrm(fl
                         , margins = c("GA" , "GA")
                         , BivD = "N"
                         , data=data.frame(gamma_c_mu1,gamma_c_mu2)
                         , Model="B")

  #Extracting coefficient estimates from each of the models
  #First four models are set as blank and run in a separate function
  summary_glm<-c( 0,0,0,0
  )
  summary_gee<-c( 0,0,0,0
  )
  summary_re_nosig<-c( 0,0,0,0
  )
  summary_re<-c( 0,0,0,0
  )
  summary_cop<-c( model_copula$coefficients[1]
                  , model_copula$coefficients[2] - model_copula$coefficients[1]
                  , summary(model_copula)$tableP1[2] #SE for time 0
                  , summary(model_copula)$tableP2[2] #SE for time 1
  )
  summary_cop_n<-c( model_copula_n$coefficients[1]
                  , model_copula_n$coefficients[2] - model_copula_n$coefficients[1]
                  , summary(model_copula_n)$tableP1[2] #SE for time 0
                  , summary(model_copula_n)$tableP2[2] #SE for time 1
  )
  actuals<-c( 0
```

```
            , 0
            , model_copula$tau #Capturing tau estimate from copula fits
            , 0
  )

  output<-rbind(summary_glm
                , summary_gee
                , summary_re_nosig
                , summary_re
                , summary_cop
                , summary_cop_n
                , actuals)

  colnames(output)<-c("Time 1 Intercept","Time 2 Intercept","Time 1 SE","Time 2 SE")

  return(output)
}

results<-list()
a=.1+.1*1:20; b=.1+.1*1:20; mu1=1; mu2=2; n=100

#Code to iterate through various shapes of the bivariate distribution and fit the non-GJRM models
i=1; j=1; k=1; l=1; z=1;
start=Sys.time()
for (i in 1:length(a)) {
  for (j in 1:length(b)) {
    for (k in 1:length(mu1)) {
      for (l in 1:length(mu2)) {
        set.seed(z)
        results[[z]] <- rbind(tryCatch({
        simulateCorrelatedVarNOGJRM(n,a[i],b[j],mu1[k],mu2[l])}
        , finally={simulateCorrelatedVarNOGJRM(n,a[i],b[j],mu1[k],mu2[l])})
        , c(a[i],b[j],mu1[k],mu2[l]))
        print(c(z,length(a)*length(b)*length(mu1)*length(mu2)
                ,z/(length(a)*length(b)*length(mu1)*length(mu2))
                , (Sys.time()-start)
                , (Sys.time()-start) / (z/(length(a)*length(b)*length(mu1)*length(mu2)))))
        )
        z = z + 1
      }
    }
  }
}

save(results,file="results_NOGJRM.rds")

results<-list()
a=.1+.1*1:20; b=.1+.1*1:20; mu1=1; mu2=2; n=100
```

```
198
199 #Code to iterate through various shapes of the bivariate distribution and fit the GJRM models
200 i=1; j=1; k=1; l=1; z=1;
201 start=Sys.time()
202 for (i in 1:length(a)) {
203   for (j in 1:length(b)) {
204     for (k in 1:length(mu1)) {
205       for (l in 1:length(mu2)) {
206         set.seed(z)
207         results[[z]] <- rbind(tryCatch({
208         simulateCorrelatedVarGJRM(n,a[i],b[j],mu1[k],mu2[l])}
209         , finally={simulateCorrelatedVarGJRM(n,a[i],b[j],mu1[k],mu2[l])})
210         , c(a[i],b[j],mu1[k],mu2[l]))
211         print(c(z,length(a)*length(b)*length(mu1)*length(mu2)
212         , z/(length(a)*length(b)*length(mu1)*length(mu2))
213                 , (Sys.time()-start)
214                 , (Sys.time()-start) / (z/(length(a)*length(b)*length(mu1)*length(mu2))))
215         )
216         z = z + 1
217       }
218     }
219   }
220 }
```