AJUSTE DE UNA CÓPULA A UN CONJUNTO DE DATOS LONGITUDINALES COMO ALTERNATIVA A LA CONSTRUCCIÓN DE LA ESTRUCTURA DE COVARIANZA DE UN MODELO DE EFECTOS FIJOS

DIEGO ALEXIS VILLADA CANTOR

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
FACULTAD DE INGENIERIA Y CIENCIAS BASICAS
ESPECIALIZACION EN ESTADÍSTICA APLICADA
BOGOTÁ D.C.
2017

AJUSTE DE UNA CÓPULA A UN CONJUNTO DE DATOS LONGITUDINALES COMO ALTERNATIVA A LA CONSTRUCCIÓN DE LA ESTRUCTURA DE COVARIANZA DE UN MODELO DE EFECTOS FIJOS

DIEGO ALEXIS VILLADA CANTOR

Código: 201710025144

Trabajo de grado presentado para optar por el título de

ESPECIALISTA EN ESTADÍSTICA APLICADA

Trabajo dirigido por:

Mg. JUAN CAMILO SANTANA CONTRERAS

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES

FACULTAD DE INGENIERIA Y CIENCIAS BASICAS

ESPECIALIZACION EN ESTADÍSTICA APLICADA

BOGOTÁ D.C.

2017

Nota de Aceptación

El trabajo de grado titulado AJUSTE DE UNA CÓPULA A UN CONJUNTO DE DATOS LONGITUDINALES COMO ALTERNATIVA A LA CONSTRUCCIÓN DE LA ESTRUCTURA DE COVARIANZA DE UN MODELO DE EFECTOS FIJOS , realizado por el estudiante DIEGO ALEXIS VILLADA CANTOR, cumple con los requisitos exigidos por la FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES para optar al título de ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

Firma del Presidente del Jurado

Firma del jurado

Firma del Jurado

Dedicado a...

... a mi esposa Diana y mis padres Jaime y Magdalena por hacerme soñar y no dejarme desfallecer, y a mis hijas Liz Ariana y Gabriela como parte de mi humilde legado

Agradecimientos

A mi querido amigo Oscar Beltrán por su persistencia en la labor de animador y consejero, a mis hijas y mi esposa por brindarme no solo un mótivo más para culminar este proceso, sino todo su amor y comprensión en los momentos que demandó este trabajo. A mis amigos Andrés Cruz, Diana Ferrucho y Adriana Pineda por su apoyo y por supuesto, mis compañeros de estudio Ricardo Macías, Mónica Rodriguez, Daniel Contreras y Elkin Ortiz que me acompañaron durante esta etapa de mi vida. Al grupo de Docentes de la Fundación Universitaria Los Libertadores, Profesores Juan Camilo Santana, Jorge Nisperuza y Jhon Forigua, quienes me guiaron en la concepción de este trabajo.

Índice general

Índice de cuadros

Índice de figuras

Índice de ecuaciones

Resumen

Algunos de los métodos de análisis de datos longitudinales implementados hasta la fecha, consideran que las datos deben cumplir con supuestos distribucionales necesarios para el adecuado ajuste de los modelos. Si estos supuestos distribucionales no se cumplen, se presentan los modelos de estructura de covarianza como una solución para su modelamiento. Por otra parte, las cópulas son aplicaciones que se usan para construir funciones multivariadas adaptables a las datos. Se expone, por medio de este trabajo, las procedimientos mínimos y necesarios para el ajuste de un modelo para datos longitudinales mediante funciones cópula, aclarando suposiciones y condiciones tanto del conjunto de observaciones como de las funciones cópula aplicables a este modelamiento.

Por último se abordará un problema de interés particular que pueda servir de ejemplo para el modelamiento de los datos longitudinales usando una cópula, comparándolo con el modelamiento tradicional por Modelos Lineales Generalizados y Ecuaciones de Estimación

Palabras claves: Copula, Datos Longitudinales, Modelamiento, Normalidad, Baja correlación lineal, Modelos de estructura de covarianza, matriz de correlación de trabajo

Capítulo 1

Introducción

El principal problema del análisis estadístico sigue siendo encontrar una explicación adecuada al por qué y al cómo de los eventos que se presentan en cualquier tipo de situación. La estadística descriptiva se utiliza para dar evidencia de la situación observada, pero se hace necesario que con esta información se pueda inferir resultados futuros o más generales. Este problema es parcialmente cubierto mediante la construcción de modelos matemáticos. Los estudios de corte transversal se enfocan en un instante temporal o circunstancial definido, pero cuando la información se produce con el paso del tiempo, y este representa una variable a tener en cuenta como afectadora de las observaciones, se hace necesario un tipo de análisis distinto. En el análisis de datos longitudinales se busca una metodología para la identificación de patrones de cambio temporales en las características medibles propias de un individuo. Esto disminuye el número de individuos tenidos en cuenta para un estudio, lo que los hace más económicos (Davis, 2002). El modelamiento en datos longitudinales tiene, actualmente, varios frentes de trabajo, enfoques con modelos lineales, no lineales, no paramétricos, funcionales, entre otros, que sumados a un gran número de aplicaciones en diversas ramas del conocimiento, hace que el análisis longitudinal y sus herramientas cobren importancia en la estadística actual.

Desde hace algunos años se trabajan con bastante recurrencia las funciones *cópula*, con variadas bondades que resultan interesantes por su capacidad para develar la estructura de dependencia de la función de distribución conjunta de un vector de variables aleatorias y, simultáneamente, separar esta estructura de dependencia del comportamiento marginal (Escarela, 2009).

Las cópulas permitirían mejorar el análisis longitudinal, y es este estudio se busca la forma de usarlas como alternativa a las ya conocidas y bien documentadas metodologías para el modelamiento de este tipo de datos.

Lo que busca este trabajo es presentar una metodología complementaria a estos ya conocidos tipos de modelos. Las funciones cópulas presentan una diversa cantidad de cualidades que permitirían realizar el modelamiento de este tipo de datos con ajustes muy precisos, teniendo en cuenta que la principal idea del modelamiento con cópulas es construir la distribución de probabilidad multivariada que rige los errores aleatorios. Al conocer la distribución de probabilidad, la matriz de varianzas y covarianzas también podría ser modelada, de manera que realizar un ajuste lineal

sería muy posible también

La pregunta que guía este trabajo es entonces

Frente a las metodologías de modelamiento usuales en datos longitudinales, ¿Es posible especificar la matriz de covarianzas de un modelo lineal de efectos fijos, mediante el uso de cópulas y así mejorar el ajuste del modelo?

Objetivo general

Diseñar un procedimiento para el ajuste de un modelo de cópula a un conjunto de datos longitudinales

Objetivos específicos

1. Describir los pasos necesarios para el ajuste de un modelo de copula a un conjunto de datos longitudinales
2. Diseñar un algoritmo o rutina en R para el modelamiento de datos longitudinales mediante cópulas.
3. Implementar una aplicación de este algoritmo a un conjunto de datos longitudinales

Justificación

El modelamiento de los datos longitudinales tiene como principal proceso de generación a los modelos lineales generalizados y las ecuaciones de estimación generalizadas. Ambos procesos aparecen con bastante recurrencia en la literatura, con resultados muy bien respaldados.

Hedeker y Gibbons (2006), Davis (2002), Diaz (2015), resaltan en sus libros los principales modelos de datos longitudinales, haciendo un repaso por los modelos ANOVA, en donde se analiza el aporte de cada una de los factores a la variable respuesta. En el mismo sentido, se analiza en estos libros y en capítulos independientes lo referente a la formulación de modelos bajo un supuesto de normalidad en los datos y en los residuales. Asimismo, si este supuesto no se cumple, se hacen extensiones a otros tipos de modelos que no requieren de este supuesto para formular una ecuación. Davis (2002), dedica los capítulos 7, 8 , 9 y 10 de su libro al modelamiento de los datos longitudinales usando metodologías cercanas, análogas a las tratadas inicialmente, pero que carecen de la importante condición de la normalidad. Es así como desarrolla los temas relacionados con los modelos lineales generalizados y el ajuste de un modelo mediante métodos no paramétricos.

Por otro lado en Hedeker y Gibbons (2006), la profundización de los temas toma rumbo hacia los modelos con patrones de covarianza y el modelamiento de estas matrices es lo principal en capítulos ulteriores del libro. Mientras que Diaz (2015), examina el modelamiento con cópulas de un conjunto bivariado de datos, pero no toca el tema directamente en los datos longitudinales.

Pero es de resaltar, que es el enfoque presentado por Hedeker y Gibbons (2006) llama la atención del investigador y por tanto, es donde se han centrado los esfuerzos de este trabajo.

La especificación de la matriz de covarianza para la distribución de probabilidad de los residuales de los modelos lineales, más concretamente de los modelos de efectos fijos, puede marcar la diferencia entre el éxito o el fracaso en la consecución de un modelo mejor ajustado. Dependiendo de la situación, la matriz de varianzas-covarianzas asociada a la distribución de probabilidad de los residuales puede ser escogida por el investigador según le parezca adecuada. Esta elección está regida por el tipo de correlación establecida entre las variables participantes, así, es posible establecer un único parámetro de correlación para todas las parejas conformables o calcularlo para cada una de estas. Las cópulas representan, una herramienta para determinar correlación entre variables con mayor más exactitud, tomando como puente el vínculo existente entre estas funciones y los coeficientes τ de Kendall y ρ de Spearman, los coeficientes serán calculados haciendo uso de alguna cópula para luego ser introducidos en la matriz de varianzas y covarianzas de estos modelos.

Esta propuesta procedimental pretende cubrir posibles errores de estimación de los coeficientes de correlación hasta ahora utilizados para la especificación de los parámetros distribucionales de los residuales. En este sentido, lograr el modelamiento adecuado de datos longitudinales aprovechando las bondades de las cópulas, no solo dejaría en evidencia la relación entre las variables y los tiempos de medición, sino que además la forma relacional que pueda existir entre estas.

En otras palabras, permitiría obtener modelos mejor especificados y más concluyentes en estudios de medidas repetidas.

Capítulo 2

Marco teórico o marco conceptual

2.1. Datos Longitudinales

El análisis de medidas repetidas es un término referido a los datos obtenidos de la observación de una característica medible de un individuo en diferentes instancias temporales o circunstanciales. La diferenciación con los datos longitudinales, aún no es clara. Davis (2002), discute al respecto y concluye que las medidas repetidas, pueden ser entendidas como un caso particular de los datos longitudinales.

En el marco de los datos longitudinales se encuentran una serie de características que categorizan este tipo de estudios en diversas formas de presentación y análisis, es el caso de las medidas repetidas, los datos multinivel, los datos de sobrevida y los datos panel, entre otros. La estructura de *Datos Longitudinales* asume que las observaciones realizadas a cada individuo corresponden a un proceso estocástico de t tiempos en el que los datos son obtenidos bajo condiciones controladas de experimentación (Sosa, 2010). (46)

Algunas de las características más sobresalientes de los datos longitudinales son:

- Los tiempos de transición son una sucesión creciente.
- Se asume independencia sobre los n individuos.
- El número de tiempos puede ser aleatorio debido al individuo.
- Algunos modelos tienen un número máximo posible de eventos.

2.1.1. Objetivos del análisis de datos longitudinales

En Sosa (2010), se incluyen algunos objetivos buscados con el análisis de datos longitudinales, entre ellos:

1. Determinar cómo los estados de medición y las covariables afectan la variable respuesta. ¿Las diferencias se incrementan o no con los estados de medición?
2. Identificar y describir:
   a) El comportamiento medio de la variable respuesta a través de los estados de medición.
   b) Patrones individuales de cambio de la variable respuesta a través de los estados de medición.
   c) Diferencias en la tendencia media de la variable respuesta entre grupos. ¿La respuesta media es igual en todos los grupos en todos los estados? ¿El patron de la variable repuesta es el mismo en todos grupos?
3. Predecir los valores de la variable respuesta dada información a priori de la misma en estados de medición anteriores.

2.1.2. Estructura de los datos

En la estructura de datos longitudinales, el proceso del individuo i es seguido durante el intervalo $(0, t_i)$ donde t_i es un tiempo fijo o aleatorio y los tiempos de los eventos son de la forma t_{ij} con $i = 1, ..., n$ y $j = 1, ..., s$. Es decir, los espacios de medición corresponden a los instantes de medición (números reales) o a los momentos de medición (números enteros) (Sosa, 2010).

Se acostumbra en los estudios de carácter comparativo, dividir la muestra en grupos para comparar los tratamientos aplicados a los individuos bajo factores adicionales de experimentación. En particular, un conjunto de datos longitudinales que haya sido sometido a un diseño experimental por factores tendrá la apariencia de Cuadro 2.1.

Una discusión al respecto de otras características y tratamientos comúnmente utilizados en los datos longitudinales, como las consideraciones de datos perdidos, tipos alternativos de modelamiento, la relación y diferenciación con otras formas de medidas repetidas que no son del alcance de este trabajo, se pueden encontrar al respecto en Davis (2002).

2.1.3. Modelamiento En Datos Longitudinales

Este tipo de datos pueden ser modelados mediante diversas metodologías dependiendo de las características de las muestras estudiadas; el tamaño, el numero de tiempos medidos, la ausencia de datos y la baja correlación lineal entre tiempos, pueden ser determinantes para la elección de una metodologia de modelamiento adecuada.

Algunos de los modelos más referenciados en la literatura de los datos longitudinales son los siguientes

Análisis de varianza (ANOVA)

En dicho modelo, cada sujeto representa un factor aleatorio, mientras que las ocasiones y los grupos se denominan factores fijos. Esta estrategia estima los promedios de las observaciones de

Grupo	Individuo	puntos temporales				
		1	...	j	...	t
	1	$y_{[1]1}$	$\cdots$	$y_{[1]j}$	$\cdots$	$y_{[1]s}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
1	i	$y_{[i]1}$	$\cdots$	$y_{[i]j}$	$\cdots$	$y_{[i]s}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	
	n_1	$y_{[n]1}$	$\cdots$	$y_{[n]j}$	$\cdots$	$y_{[n]s}$
	1	$y_{[1]1}$	$\cdots$	$y_{[1]j}$	$\cdots$	$y_{[1]s}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	
h	i	$y_{[i]1}$	$\cdots$	$y_{[i]j}$	$\cdots$	$y_{[i]s}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	
	n_h	$y_{[n]1}$	$\cdots$	$y_{[n]j}$	$\cdots$	$y_{[n]s}$
	1	$y_{[1]1}$	$\cdots$	$y_{[1]j}$	$\cdots$	$y_{[1]s}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	
s	i	$y_{[i]1}$	$\cdots$	$y_{[i]j}$	$\cdots$	$y_{[i]s}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	
	n_s	$y_{[n]1}$	$\cdots$	$y_{[n]j}$	$\cdots$	$y_{[n]s}$

Cuadro 2.1: Estructura General de los Datos Longitudinales

los sujetos y los compara a través de los distintos grupos, o a lo sumo, examina si se ajustan a polinomios conocidos (Arnaú, (3)). La metodología ANOVA es comúnmente usada en modelos de respuesta normal y por lo general se asume que las observaciones son independientes lo que puede representar un impedimento de uso bajo el contexto de los datos longitudinales.

Así, para un individuo i en el tiempo j, se cumple que la base general para el modelo de medidas repetidas en datos longitudinales es :

$$y_{ij} = \mu_{ij} + \pi_{ij} + e_{ij}$$

Ecuación 2.1.1: Modelo ANOVA datos longitudinales

donde μ_{ij} es la media de las observaciones en el tiempo j para los sujetos aleatoria mente seleccionados. este componente tambien es conocido como efecto fijo del modelo. π_{ij} se conoce como el efecto aleatorio del modelo debido a las características propias del individuo, mientras que e_{ij} es error puro de cada observación. Un amplio compendio del tratamiento de estos datos se encuentra en el libro de Davis 2002, en el capítulo 5

Análisis de varianza multivariado (MANOVA)

Corresponde con la versión del ANOVA para múltiples muestras. Su formulación es la siguiente:

$$y_{hij} = \mu + \gamma_h + \tau_j + (\gamma\tau)_{hj} + \pi_{i(h)} + e_{hij}$$

Ecuación 2.1.2: Modelo MANOVA en datos longitudinales

En este modelo μ es la media general de las observaciones y γ_h es el efecto fijo del grupo h, con $\sum_{h=1}^{s} \gamma_h = 0$, además, se considera que τ_j es el efecto fijo del tiempo j con $\sum_{j=1}^{t} \tau_j = 0$, y $(\gamma\tau)_{ht}$ es el efecto fijo de interacción entre el grupo h y el tiempo j, de nuevo, existen limitaciones para el parametro de interacción, de la forma $\sum_{h=1}^{s}(\gamma\tau)_{hj} = \sum_{j=1}^{t}(\gamma\tau)_{hj} = 0$. Los demás parámetros corresponden a efectos aleatorios debidos al grupo del individuo y al error puro (Davis, (13)).

Modelos Lineales Generalizados (MLG)

Es una alternativa que utiliza familias de funciones para vincular los parámetros β en un modelo, con el vector de covariables $\mathbf{x}$ por medio de una función μ, a la cual se le conoce como *función vínculo*.

Este tipo de funciones son obtenidas de la *familia exponencial de distribuciones*, que son las funciones que se pueden escribir de la forma:

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\theta)\right]$$

Ecuación 2.1.3: Forma funcional de la familia exponencial

para algunas funciones específicas $a(\cdot)$,$b(\cdot)$ y $c(\cdot)$. Si ϕ es conocida entonces se tiene un modelo de familia exponencial de un parámetro canónico mientras que si no lo es, la expresión arriba referida sera una función de la familia exponencial biparamétrica (Pinheiro, 2009).

El vínculo entre los componentes aleatorio y sistemático se especifica cómo $\mu = E(y)$ el cual relaciona a las variables explicativas en el predictor lineal. Se puede modelar la media μ directamente o modelar una función de la media $g(\mu)$ que sea monótona diferenciable. La fórmula del modelo se especifica que

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Ecuación 2.1.4: Modelo Lineal General para datos longitudinales

La función $g(\cdot)$ se llama la función de enlace o función de vínculo.

Modelo General Lineal Mixto (o de efectos fijos)

El modelo conocido como *modelo lineal mixto (MLM)*, considera un modelado en dos instancias, una para efectos fijos (como tiempos, grupos o tratamientos) y otra para efectos no controlables llamados aleatorios. Para cualquier individuo incluido en un estudio de naturaleza longitudinal, es posible describir el modelo mediante la ecuación

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\gamma + \epsilon_i$$

Ecuación 2.1.5: Modelo lineal de efectos fijo para el individuo i

en este caso, se definen en la ecuación a $\mathbf{y}_i$ como el vector de respuestas del i-ésimo individuo, mientras que $\mathbf{X_i}$ es una matriz de diseño de tamaño $t_i \times p$, β es el vector $p \times 1$ de parámetros para los efectos fijos del modelo. Para el componente aleatorio del modelo se define $\mathbf{Z}_i$ como una matriz de diseño de tamaño $t_i \times q$ para los efectos aleatorios, γ un vector $q \times 1$ de parámetros para los mismos efectos y ϵ_i un vector $t_i \times 1$ de errores aleatorios.

El modelo dado por la ecuación 2.1.5, se puede expresar en forma matricial para el individuo i de la siguiente manera:

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & x_{i11} & \cdots & x_{i1p} \\ 1 & x_{i21} & \cdots & x_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{ij1} & \cdots & x_{ijp} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{it_i1} & \cdots & x_{it_ip} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} 1 & z_{i11} & \cdots & z_{i1q} \\ 1 & z_{i21} & \cdots & z_{i2q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{ij1} & \cdots & z_{ijq} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{it_i1} & \cdots & z_{it_iq} \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_j \\ \vdots \\ \gamma_{(q-1)} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{ij} \\ \vdots \\ \epsilon_{it_i} \end{pmatrix}$$

Ecuación 2.1.6: Forma Matricial del Modelo Lineal Mixto

Con $i = 1, ..., \mathbf{n}$ donde $\mathbf{n}$ es el número de individuos, la ecuación 2.1.6, es la conformación de las variables en el individuo i-ésimo, el cual presenta t_i observaciones.

Una reunión de los n vectores de respuestas y sus respectivas matrices de diseño, en donde $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)'$, $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)'$ y $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_n)'$, y los vectores de parámetros β y γ se mantengan como hasta ahora están definidos, dará como resultado el modelo que contiene a todos los individuos del estudio y sus respectivas matrices de diseño, el cual se puede expresar mediante la siguiente ecuación matricial

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

Ecuación 2.1.7: Modelo lineal mixto completo

El modelamiento de los datos longitudinales mediante el MLM requiere de la adecuada escogencia de la matriz de varianzas y covarianzas. Esta matriz es fundamental para describir las relaciones existentes entre las variables de los efectos aleatorios. Usando paquetes de software de la plataforma `R`, es posible hacer una estimación de los parámetros de correlación. Estas estimaciones a su vez son usadas para calcular el posible valor de una medida de dependencia que pueda existir entre las variables consideradas en el estudio.

2.2. Cópulas

En colaboración con Schweizer y Fréchet, Abe Sklar (1959) desarrolló el concepto de función *cópula* (19). Hay que decir, que las funciones definidas por Sklar, no necesariamente describen funciones de distribución de probabilidad conjuntas. Por ejemplo, sea (X, Y) un vector aleatorio con función de distribución conjunta $H(x, y)$, entonces las funciones de distribución marginal de X e Y, están dadas por $F(x) := H(x, \infty)$ y $G(y) := H(\infty, y)$, respectivamente. Sklar (1959) demostró que existe una función C, a la cual denominó *cópula*, que establece la relación funcional entre la distribución conjunta y sus marginales unidimensionales (19)

$$H(x, y) = C\left(F(x), G(y)\right)$$

Ecuación 2.2.1: Definición de cópula

Básicamente se trata de encontrar una función multivariada que tenga marginales conocidas que la puedan componer, en realidad, Sklar encontró que dicha función se puede construir con base en las inversas de las marginales. El teorema originalmente se presenta así:

Teorema 2.2.1. ***(Teorema de Sklar)***

Sea H una función de distribución n-dimensional con marginales F_1, ..., F_n. Entonces existe una cópula n-dimensional C tal que para todo $x \in \mathcal{R}^n$

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n))$$

Si F_1, ..., F_n son todas continuas, entonces C es única, por tanto, está unívocamente determinada sobre

$$RanF_1 \times ... \times RanF_n$$

Inversamente, si C es una cópula n-dimensional y F_1, ..., F_n son funciones de distribución, entonces la función H definida anteriormente es una función de distribución n-dimensional con marginales $F_1, ..., F_n$.

Junto con la información suministrada en la primera sección del capítulo, se tiene entonces, en resumen, que una cópula se define como una función bivariada con dominio en el plano $\mathbb{R}^2$, específicamente en la región limitada por el producto cartesiano $\mathbf{I}^2 = [0, 1] \times [0, 1]$, y que tiene su rango en el intervalo $\mathbf{I} = [0, 1] \subset \mathbb{R}$.

2.2.1. Funciones *Cópula*

En términos bidimensionales, una copula definida en el plano $\Re^2$, específicamente en la región limitada por el producto cartesiano $\mathbf{I}^2 = [0, 1] \times [0, 1]$, que tiene su rango en el intervalo $\mathbf{I} = [0, 1] \in \Re$. Esta puede representar una función de distribución acumulada de probabilidad conjunta de un vector aleatorio (v_1, v_2), en el que v_1 y v_2 son valores de las variables aleatorias V_1 y V_2, las cuales

siguen una distribución uniforme con parámetros $[0,1]$, para las cuales se define $v_1 = F_1(x_1)$ y $v_2 = F_2(x_2)$, para algún $x_1 \in DomF_1$ y algún $x_2 \in DomF_2$, con F_1 y F_2 funciones de probabilidad acumulada.

Definida como una función $C : \mathbf{I}^2 \to \mathbf{I}$, una cópula cumple con condiciones de acotamiento y de incremento. En las ecuaciones 2.2.2 y 2.2.3 se describen estas condiciones

de acotamiento, lo que significa que

$$\lim_{v_j \to 1^-} C(v_1, v_2) = v_{3-j}$$
$$\lim_{v_j \to 0} C(v_1, v_2) = 0$$

Ecuación 2.2.2: Condiciones de acotamiento de las cópulas

donde $j = 1, 2$ y $(v_1, v_2)^t \in \mathbf{I}^2$, y

$$C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_2) + C(u_2, v_1) \geq 0$$

Ecuación 2.2.3: Condición de incremento de las cópulas

para todo u_1, u_2, v_1 y $v_2 \in \mathbf{I}$ tal que $u_1 \leq u_2$ y $v_1 \leq v_2$.

Por otra parte, una cópula C_θ con θ parámetro denominado *"de dependencia"*, que represente una función de distribución acumulada de probabilidad, siempre que sea absolutamente continua, tendrá su respectiva función de densidad determinada por la expresión

$$\rceil\ c_\theta(v_1, v_2) = \frac{\partial C_\theta(v_1, v_2)}{\partial v_1 \partial v_2},\ (v_1, v_2) \in [0,1]^2$$

Ecuación 2.2.4: Obtención de densidad de la cópula

Luego, si $v_1 = F_1$ y $v_2 = F_2$ son funciones de probabilidad acumulada absolutamente continuas con densidades respectivas $f_1 = F_1'$ y $f_2 = F_2'$, y para C_θ existen las derivadas mixtas de orden 2, entonces la densidad de la cópula es

$$c_\theta(v_1, v_2) = c_\theta(F_1(x_1), F_2(x_2)) = C_\theta'\left[F_1(x_1), F_2(x_2)\right] f_1(x_1) f_2(x_2)$$

Ecuación 2.2.5: Función de densidad de la cópula

donde θ representa un parámetro de dependencia entre las variables F_1 y F_2.

2.3. Medidas de Dependencia

Las cópulas representan, en la mayoría de los casos, distribuciones conjuntas de parejas aleatorias continuas. una característica de estas funciones es poder capturar la dependencia entre variables aleatorias de forma invariante al reescalamiento (20); esto significa que las propiedades y las medidas no cambian al realizar transformaciones estrictamente crecientes sobre las variables aleatorias. De esta forma, las medidas de asociación invariantes bajo reescalamiento, como las de concordancia, pueden estudiarse sin necesidad de especificar las distribuciones marginales.

Drouet y Kotz (2001), hacen en su libro un recorrido por la construcción de las medidas de dependencia, para las cuales Kimeldorf y Sampson (1987) definen las propiedades de dicha medida, basados en los axiomas de Rènyi (1959), quien estableció un marco axiomático que formaliza el concepto de medida global. Estas propiedades también son deseables para las medidas locales, como las de dependencia. Entre los índices usados para caracterizar la dependencia existen algunos, que no son medidas (como el coeficiente de correlación lineal), ya que pueden ser negativos.

Una medida numérica δ de asociación entre dos variables aleatorias continuas X_1 y X_2 cuya cópula es C, es una *medida de dependencia* si esta satisface las siguientes propiedades.

Definición 2.3.1. *Una medida nume?rica δ establecidad entre dos variables aleatorias continuas X_1 y X_2 es una medida de dependencia si satisface las siguientes condiciones:*

1. *δ está definida para cualquier pareja aleatoria $X = (X_1, X_2)$*
2. *$\delta(X_1, X_2) = \delta(X_2, X_1)$*
3. *$\delta \in [0, 1]$*
4. *$\delta = 0$, si y solo si, X_1 y X_2 son independientes*
5. *$\delta = 1$, si y solo si, la variable aleatoria X_{3-i} es una función estrictamente monótona de X_i casi seguramente, para $i = 1, 2$*
6. *Si f y g son funciones estrictamente monótonas sobre el rango X_1 y rango X_2, respectivamente, entonces $\delta[f(X_1), g(X_2)] = \delta(X_1, X_2)$ casi seguramente*
7. *Si las parejas aleatorias (X_1, X_2) y (X_{1n}, X_{2n}), $n = 1, 2 \dots$, tienen funciones de distribuciones conjuntas H y H_n respectivamente, y si la sucesión $\{H_n\}$ converge en distribución a H, entonces $\lim_{n\to\infty} \delta(X_n) = \delta(X)$.*

La cuarta condición mencionada en las medidas de dependencia en realidad resulta siendo la más deseada: que determine independencia.

2.3.1. Medidas de dependencia a través de las cópulas

Entre las mas familiares de estas medidas están los coeficientes de correlación r de Pearson, ρ de Spearman, y τ de Kendall, dados respectivamente por

$$r(X_1, X_2) = \frac{1}{D(X_1)D(X_2)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [H(x_1, x_2) - F_1(x_1)F_2(x_2)] dx_1 dx_2$$

Ecuación 2.3.1: Coeficiente de correlación de pearson

$$\rho(X_1, X_2) = 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [H(x_1, x_2) - F_1(x_1)F_2(x_2)] dF_1(x_1) dF_2(x_2)$$

Ecuación 2.3.2: Coeficiente de Correlación ρ de spearman

$$\tau(X_1, X_2) = 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x_1, x_2) dH(x_1, x_2) - 1$$

Ecuación 2.3.3: Coeficiente de Correlación τ de Kendall

Donde F_1 y F_2 son las funciones marginales de las variables X_1 y X_2, y D representa la desviación estándar (Hoeffding (1948) , Kruskall (1958) y Lehmann (1966)). Es aquí donde se involucran las cópulas como soporte para la construcción de medidas de dependencia. Por ejemplo, el coeficiente τ de Kendall puede ser representado a través del generador $\phi(t)$ de una cópula arquimediana.

Teorema 2.3.1. *Sean X_1 y X_2 dos v.a. con una cópula arquimediana C con generador ϕ. El coeficiente τ de Kendall se puede escribir por medio de la expresión*

$$\tau = 1 + 4 \int_0^1 \frac{\phi(t)}{\phi\prime(t)} dt$$

Ecuación 2.3.4: *τ de Kendall. Versión de cópulas arquimedianas*

La prueba puede ser encontrada en el libro de Nelsen (1999, p. 163)

no solo el coeficiente τ de Kendall puede ser llevado a una expresión con cópulas arquimedianas, también el coeficiente ρ de Spearman, así:

$$\rho = 12 \int \int_{\mathbb{I}^2} [C(v_1, v_2) - v_1 v_2] \, dv_1 dv_2$$

Ecuación 2.3.5: ρ de Spearman. Versión de cópulas arquimedianas

Siendo entonces ambos, excelentes herramientas para la estimación de los parámetros de dependencia para las cópulas buscadas. Los métodos de estimación que se usan son conocidos como inversiones tanto de τ como de ρ, por tanto, se requiere obtener expresiones para el parámetro de dependenci θ en términos de estas medidas. Otras fórmulas para la estimación de la medida τ de Kendall a partir de otras cópulas son mostradas en la tabla 2.2

Familia	Espacio de Parámetros	Generador $\phi(t)$	τ de Kendall (En función de θ)
Clayton (1978)	$\theta \geq 0$	$\theta^{-1}(t^{-\theta} - 1)$	$\dfrac{\theta}{\theta + 2}$
Frank (1979)	$\theta \geq 0$	$-\ln \dfrac{e^{-\theta t} - 1}{e^{-\theta} - 1}$	$1 + \dfrac{4}{\theta}\left[D_1(\theta) - 1\right]$
Gumbel (1960)	$\theta \geq 1$	$(-\ln t)^{\theta}$	$1 - \dfrac{1}{\theta}$

$D_1(\theta) = \theta^{-1} \int_0^{\theta} x/(\exp(x) - 1)dx$ esta función se conoce como *Debye*

Cuadro 2.2: Forma del coeficiente τ de Kendall en términos de algunas familias de cópulas y sus generadores

Capítulo 3

Marco metodológico

3.1. Ajuste de cópulas

El ajuste de los datos mediante cópulas, se propone desde una ecuación que dé cuenta de la relación de las variables respuesta a través del tiempo. Se busca una distribución de probabilidad que explique esta relación. En general, para un vector aleatorio $\mathbf{Y} = (Y_1, ..., Y_t)$ con t el número de observaciones, se puede encontrar su función de probabilidad acumulada mediante

$$F(\mathbf{Y}) = C_\theta(F_1(Y_1), ..., F_t(Y_t))$$

Ecuación 3.1.1: Modelo Cópula

en donde F es una función a estimar, t-variada sobre el vector $\mathbf{Y}$ y con funciones marginales univariadas continuas F_i para $i = 1, ..., t$, θ representa el parámetro de dependencia establecido para la cópula C asociada a la función F.

En un arreglo de datos longitudinales, se considera cada $Y_i = (y_{i1}, ..., y_{ij}, ..., y_{it})$, donde y_{ij} es la respuesta del sujeto i en el tiempo j, con $i = 1, ..., n$ y $j = 1, ..., t$.

Para empezar, se quiere ver si las respuestas en el tiempo j_r están relacionadas con los tiempos j_s por medio de la prueba de independencia ilustrada en la sección 3.2.3.

De obtener en la prueba un resultado que indique dependencia, se busca establecer cómo es la supuesta relación entre estos tiempos, y si ésta puede ser expresada por una cópula $C_\theta(u_i, u_j)$; donde $C_\theta(\cdot)$ representa una cópula con parámetro de dependencia θ, mientras que $u_i = F_i(\mathbf{y_i})$ y $u_j = F_j(\mathbf{y_j})$ representan las probabilidades acumuladas correspondientes a los valores de $\mathbf{y_i}$ e $\mathbf{y_j}$ de acuerdo con sus respectivas distribuciones de probabilidad marginales F_i y F_j, para $i, j = 1, ..., t$.

Para ajustar un modelo de cópulas se debe trabajar con variables uniformes estándar, esto es, $U \sim \mathcal{U}(0, 1)$, por tanto, u_i como u_j son representaciones uniformes de las respuestas en los tiempos i y j.

En caso de desconocer las distribuciones de probabilidad marginales, se usa un procedimiento no paramétrico para determinar las probabilidades empíricas de los valores de las variables, éstas

probabilidades son conocidas como *pseudo-observaciones*, cuya construcción se explica en detalle en la sección 3.2.5.

3.2. Ajuste de cópulas a conjuntos de datos y elección del modelo

3.2.1. Estimación de los parámetros de una cópula

Determinar la cópula adecuada para el ajuste de un grupo de datos, suele ser una tarea muy complicada, más aún, si se tiene en cuenta que el método de estimación de los parámetros de dependencia de la cópula depende, en la mayoría de las veces, de las características de la correlación entre ellos.

Existen tres métodos de estimación de los parámetros de dependencia asociados a las cópulas. Roberto De Matteis (16) discute brevemente estos tres métodos, presentando sus respectivas ventajas y desventajas.

3.2.2. Procedimientos de estimación

En el capítulo anterior se han descrito dos metodologías para la estimación de los parámetros de las cópulas, uno por máxima verosimilitud exacta, y el segundo, en dos etapas, estimando primero las marginales. Xu (1996) (1996) mediante estudios de simulación Monte Carlo verificó que la eficiencia relativa de los dos estimadores es bastante próxima a 1. Estos resultados sugieren que el método de estimación por las marginales es eficiente comparado con el método de verosimilitud exacta. Anjos (2006) justifica que el método de estimación en dos etapas produce estimadores consistentes.

Tras usar uno de estos procedimientos y obtener las estimaciones de los parámetros mencionados, se procede a verificar las cualidades de los diferentes modelos por medio de las pruebas de bondad de ajuste presentadas en la siguiente sección.

3.2.3. Pruebas de bondad de ajuste

Prueba de bondad de ajuste basada en bootstrap

Genest & Rémillard (Genest and Rémillard) describen el siguiente procedimiento para esta prueba

1. A partir de las pseudo-observaciones $\hat{\mathbf{p}}_1, ..., \hat{\mathbf{p}}_n$ calcule el estimador C_n y estime θ a partir de éstas por medio de un estimador basado en rangos θ_n,
2. Calcule el estadístico S_n,
3. Para algún entero grande N, repita los siguientes pasos para cada $k \in 1, ..., N$,

a) genere una muestra aleatoria $X_1^{(k)}, ..., X_n^{(k)}$ a partir de la cópula C_{θ_n} y calcule las pseudo-observaciones asociadas $\mathbf{p}_1^{(k)}, ..., \mathbf{p}_n^{(k)}$,

b) sea

$$C_n^{(k)}(p) = \frac{1}{n} \sum_{X_i=1}^{n} \mathbf{I}(\mathbf{p}_i^{(k)} \leq p)$$

Con $\mathbf{I}(\cdot)$ función indicadora, a partir de $\mathbf{p}_1^{(k)}, ..., \mathbf{p}_n^{(k)}$ calcule una estimación $\theta_n^{(k)}$ de θ, usando el mismo estimador del paso 1 basado en rangos.

c) Calcule una realización independiente de S_n bajo H_0 mediante

$$S_n^{(k)} = \sum_{X_i=1}^{n} \{C_n^{(k)}(\mathbf{p}_i^{(k)}) - C_{\theta_n}^{(k)}(\mathbf{p}_i^{(k)})\}^2$$

.

4. Un valor p aproximado para la prueba está dado por $\frac{1}{N} \sum_{X_k=1}^{N} \mathbf{I}(S_n^{(k)} \geq S_n)$.

Esta prueba esta programada en la librería `copula` (36) del paquete `R` (39), bajo la sentencia `gofCopula` .

Prueba de independencia

Si las variables incluidas en el conjunto de datos son independientes, no tiene mucho sentido elaborar un modelo para el comportamiento de los mismos. Genest & Rémillard (Genest and Rémillard) sugieren una prueba de independencia mutua de $X_1, ..., X_p$ basada en el estadístico

$$I_n = \int_{[0,1]^p} n \left\{ C_n(u) - \prod_{i=1}^{p} u_i \right\}^2 du$$

Ecuación 3.2.1: Estadístico para prueba de independencia en cópulas

Bajo el supuesto de independencia mutua de las componentes de $\mathbf{X}$ el proceso empírico $\sqrt{n}Cn - \Pi_p$, con Π_p la cópula de independencia, puede descomponerse, usando la transformación de Möebius (Rota, 1964), en $2^p - p - 1$ subprocesos $\sqrt{n}\mathcal{M}_A(C_n)$, $A \subseteq 1, ..., p$, y el cardinal de A mayor que uno ($|A| \geq 1$) convergen conjuntamente a procesos Gausianos centrados, mutuamente independientes. Una propiedad fundamental de esta descomposición, cuya forma es dada en Genest & Rémillard (2004) es que la independencia mutua entre los elementos de $X_1, ..., X_p$ es equivalente a tener $\mathcal{M}_A(C)(u) = 0$ para todo $u \in [0, 1]^p$ y todo A tal que $|A| \geq 1$. Esto sugiere que en lugar del estadístico simple I_n se consideren $2^p - p - 1$ estadísticos de prueba de la forma

$$M_{A,n} = \int_{[0,1]^p} n \left\{ M_A(C)(u) \right\}^2 du$$

Ecuación 3.2.2: Esadístico alternativo para independencia entre variables.

donde $A \subseteq 1, ..., p$, $|A| > 1$ que son mutuamente independientes, asintóticamente, bajo la hipótesis nula de independencia de las marginales. Cada estadístico $M_{A,n}$ puede considerarse como enfocado en la dependencia entre las componentes de $\mathbf{X}$ cuyos índices están en A. La prueba descrita en esta sección está implementada en la librería `copula` de `R`

3.2.4. Representaciones gráficas para la elección de la mejor cópula

Después de haber estimado los parámetros de dependencia θ, para algunas de las cópulas consideradas se debe decidir sobre cuál de ellas ajusta mejor a los datos. Para este fin, se usan, por ejemplo, los siguientes métodos gráficos.

QQ-plot sobre la función de densidad condicional de $Y|X$

Como se ha visto en las secciones previas, la función de densidad conjunta de la cópula C está dada por la expresión $h(x, y) = f(x)g(y)C_{1,2}(F(x), G(y))$. Por tanto la función de distribución condicional de $Y|X = x$ con $C_1(u, v) = \frac{\partial}{\partial u}C(u, v)$ es

$$H_{x,y}(x, y) = \frac{C_1(F(x), G(y))}{C_1(F(x), 1)} = C_1(F(x), G(y))$$

Ecuación 3.2.3: Función de distribución condicional

ésto debido a que

$$\frac{\partial}{\partial u}C(u, 1) = \lim_{\Delta u \to 0} \frac{C(u + \Delta u, 1) - C(u, 1)}{\Delta u} = \lim_{\Delta u \to 0} \frac{\Delta u}{\Delta u} = 1$$

Nótese que $H_{Y|X}(x, y)$ es una distribución uniforme estándar, luego, si se usa el estimador $\hat{\theta}$, un QQ-plot de $H_{Y|X}(x, y) = C_1(F(x), G(y))$ aplicado sobre las observaciones x y y enfrentados a los cuantiles de la distribución uniforme estándar debería ilustrar una línea recta. En la fiigura 3.1 se mide el ajuste de los datos a las cópulas Clayton (Family 1) y Gumbel (Family 4) respectivamente, dejando en evidencia, que la cópula clayton puede ser más adecuada

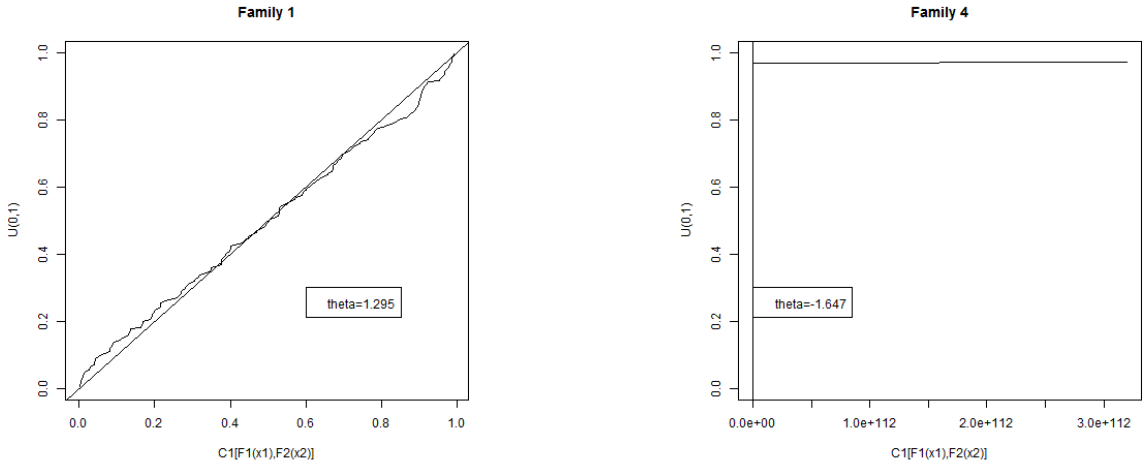


Figura 3.1: Comparación gráfica del ajuste las cópulas mediante el procedimiento gráfico sobre la densidad condicional

QQ-plot sobre sobre la densidad de la cópula

La función de distribución univariada de la cópula está dada por

$$K_C(t) = P[C(U,V) \leq t] = t - \frac{\varphi(t)}{\varphi\prime(t)}$$

Ecuación 3.2.4: Funcion de distribución univariada de una cópula

En consecuencia, aplicando la función $K_C(\cdot)$ a la cópula $C(F(X), G(Y))$ debería ser una distribución estándar uniforme (para mas detalles vea De Matteis(16), Secc. 2 y 3). Usando el estimador $\hat{\theta}$, un QQ-plot de la función $K_{C(F(X),G(Y))}(\cdot)$ contra los cuantiles de la distribución estándar uniforme debería dar una línea recta si la cópula ajusta bien a los datos. En la imagen se comparan los ajustes de las cópulas Clayton (Family 1) y Gumbel (Family 4) para el procedimiento descrito sobre la densidad de cada cópula. En el caso Clayton (Izquierda) resulta tener un ajuste distante de lo ideal, mientras que la cópula Gumbel, se comporta adecuadamente.

3.2.5. Estimador no paramétrico de la función de densidad de la cópula

La idea de Genest y Rivest (24) consiste en determinar dos estimaciones de la función de distribución K de la cópula desconocida C. Una estimación no paramétrica, que es independiente de las marginales y del parámetro de la cópula arquimediana, y una estimación paramétrica, que

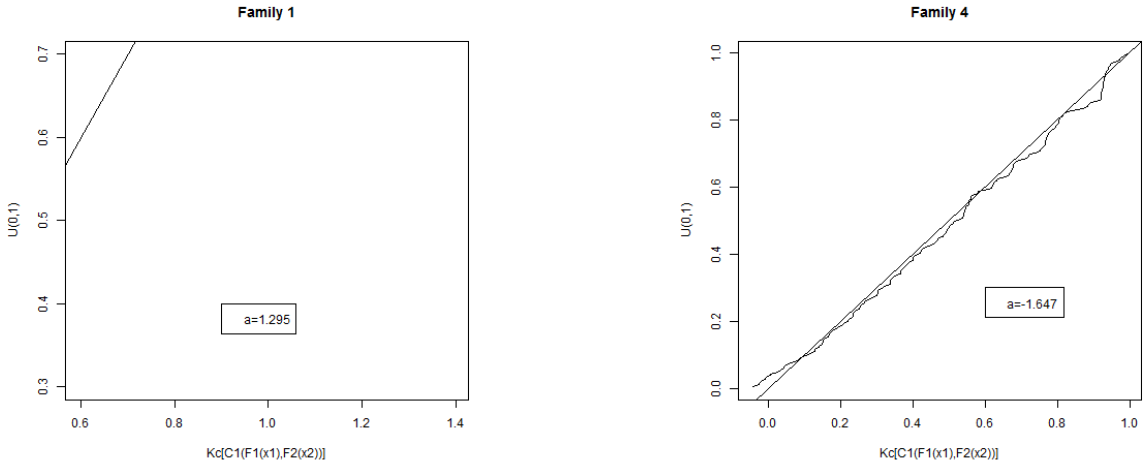


Figura 3.2: Comparación gráfica del ajuste las cópulas mediante el procedimiento gráfico para la densidad de las cópulas

tiene la forma $K_C(t) = t - \frac{\varphi(t)}{\varphi\prime(t)}$. Luego, para la función de densidad paramétrica se necesita el parámetro θ estimado previamente por $\hat{\theta}$

Aunque este tipo de estimación puede ser usado directamente en la construcción de una cópula arquimediana bivariada, es más conveniente hacer varias elecciones de copulas arquimedianas y seleccionar aquella que más se parezca a la estimación No paramétrica (16). La comparación se logra siguiendo el procedimiento propuesto por De Matteis:

i. Determinar la función de densidad no paramétrica Suponga que una muestra aleatoria $(X_1, Y_1), ..., (X_n, Y_n)$ ha sido extraída de una distribución bivariada $H(x, y)$ con marginales continua $F(x)$ y $G(y)$ y una cópula arquimediana $C(x, y)$. Sean $U = F(x)$ y $V = G(y)$ distribuidas uniformemente. Las copulas arquimedianas se caracterizan porque la función densidad de la variable aleatoria $C(x, y)$ puede ser producida por una generadora φ. El procedimiento consiste en estimar primero la función de distribución univariada $K(w) = P\left[C(U, V) \leq w\right] = P[H(X, Y) \leq w]$ sobre el intervalo $(0, 1)$. Al encontrar dicha función se define la variable aleatoria $W := \hat{H}(X, Y)$ donde $\hat{H}$ es el estimado de las distribución bivariada empírica. Se debe encontrar ahora el estimado empírico de la función W. La idea es que $C_n(X_i, Y_i)$ represente la proporción de observaciones en la muestra que son menores o iguales a (X_i, Y_i) para cada $i = \{1, ..., n\}$. Para esto, se calculan las pseudo-observaciones para cada W_i, así:

$$W_i = \hat{H}(X_i, Y_i) = \frac{Card\left\{(X_i, Y_i) : X_j < X_i, Y_j < Y_i\right\}}{(n-1)}$$

Ecuación 3.2.5: Fórmula para el cálculo de las pseudo-observaciones

Luego, $\hat{H}$ es una estimación empírica de una distribución bivariada H. Entonces el estimador no paramétrico de $K(w)$ está dado por:

$$\hat{K(w)} = \frac{Card\{i : 1 \leq i \leq n, W_i \leq w\}}{n} \tag{3.1}$$

ii. Determinar la distribución paramétrica Una vez estimado $K_n(w)$ de la distribución $C(F(x), G(y))$ a través de la ecuación 3.2.5 es posible determinar a partir del K_n, cuál es la cópula arquimediana más cercana a $H(x, y)$. Pero en general es teóricamente más conveniente usar el K_n como una herramienta para ayudar a identificar la familia paramétrica de cópulas arquimedianas que provee el mejor ajuste a los datos. Dado que cada cópula arquimediana tiene su generador específico, se determina a través de $\hat{\theta}$ la estimación paramétrica de la función de densidad la cual toma la forma de $K_C(w) = w - \frac{\varphi(w)}{\varphi\prime(w^+)}$, el cual se conoce como el estimador $K_\varphi(w)$

iii. Comparar las funciones de densidad paramétrica y no paramétrica Si la densidad paramétrica $K_\varphi(w)$ de alguna de las cópulas arquimedianas estimadas es similar a la función de densidad no paramétrica $K_n(\varphi)$, esta debe ser seleccionada como la cópula arquimediana a usar en el ajuste de los datos. Un QQ-plot construido a partir de una muestra y dos funciones de densidad (la paramétrica elegida y la no paramétrica) nos ayudará determinar si la elección del estimador es buena. siendo así, la gráfica debería mostrar una línea recta, por lo que se considera que la cópula ajusta buena los datos.

3.2.6. Pruebas analíticas de ajuste

Se ilustran métodos analíticos que pueden usarse para determinar la cópula de mejor ajuste a los datos. Estas pruebas usan los métodos considerados en la sección anterior para la construcción de las gráficas. Una medida alternativa es enfocarse en la propuesta paramétrica de máxima verosimilitud.

Bondad de ajuste: Kolmogorov-Smirnov

Esta prueba tiene la ventaja de que es un procedimiento libre de distribución. Para muestras pequeñas especialmente. Revela las discrepancias entre una distribución teórica y una empírica. Como prueba estadística, se usa la diferencia maximizada entre la distribución acumulativa empírica y la teórica, luego

$$T = \max_x \left\{ |\hat{F}(x) - F(x)| \right\}$$

Ecuación 3.2.6: Estadístico Kolmogorov-Smirnov para bondad de ajuste

Por otro lado tiene serias deficiencias en la baja potencia que evidencian sus salidas frente a la hipótesis nula.

Bondad de ajuste: prueba χ^2

La prueba usa el estadístico

$$T = \sum_{i=1}^{k} \frac{[f_i - np(x_i)]^2}{np(x_i)}$$

Ecuación 3.2.7: Estadístico chi-cuadrado para bondad de ajuste

donde k representa el número de clases, f_i es la frecuencia absoluta de los datos en la clase i y $np(x_i)$ es la frecuencia teórica de los datos para cada clase i note también que la frecuencia de la prueba crece sin que se incremente el número de clases.

Algunas consideraciones importantes acerca de la conveniencia y la pertinencia sobre el uso de las pruebas anteriormente mencionadas se pueden encontrar en De Matteis(2001)(16).

Criterio de información de Akaike (AIC)

Alternativamente a la comparación de gráficos y valores P para los métodos gráficos en el caso de la propuesta paramétrica de máxima verosimilitud, se pueden comparar los valores negativos de las funciones de log-verosimilitud. Entre 1973 y 1974 Akaike (1) desarrolló una estrategia de toma de decisiones basado en la medida de información de Kullback-Leibler, argumentando que esta medida provee un criterio natural para el ordenamiento alternativo de modelos estadísticos para datos.

Recordemos que el estimador de Máxima verosimilitud está dado por

$$\hat{\theta} = \arg\max_{\theta \in \mathbf{R}} \sum_{i=1}^{n} logL(\theta; U_i, V_i)$$

Ecuación 3.2.8: Estimador de máxima Verosimilitud

y con esto, se puede definir el Criterio de Información de Akaike como sigue

$$AIC = (-2)\log(\text{máxima-verisimilitud}) + 2\rho$$

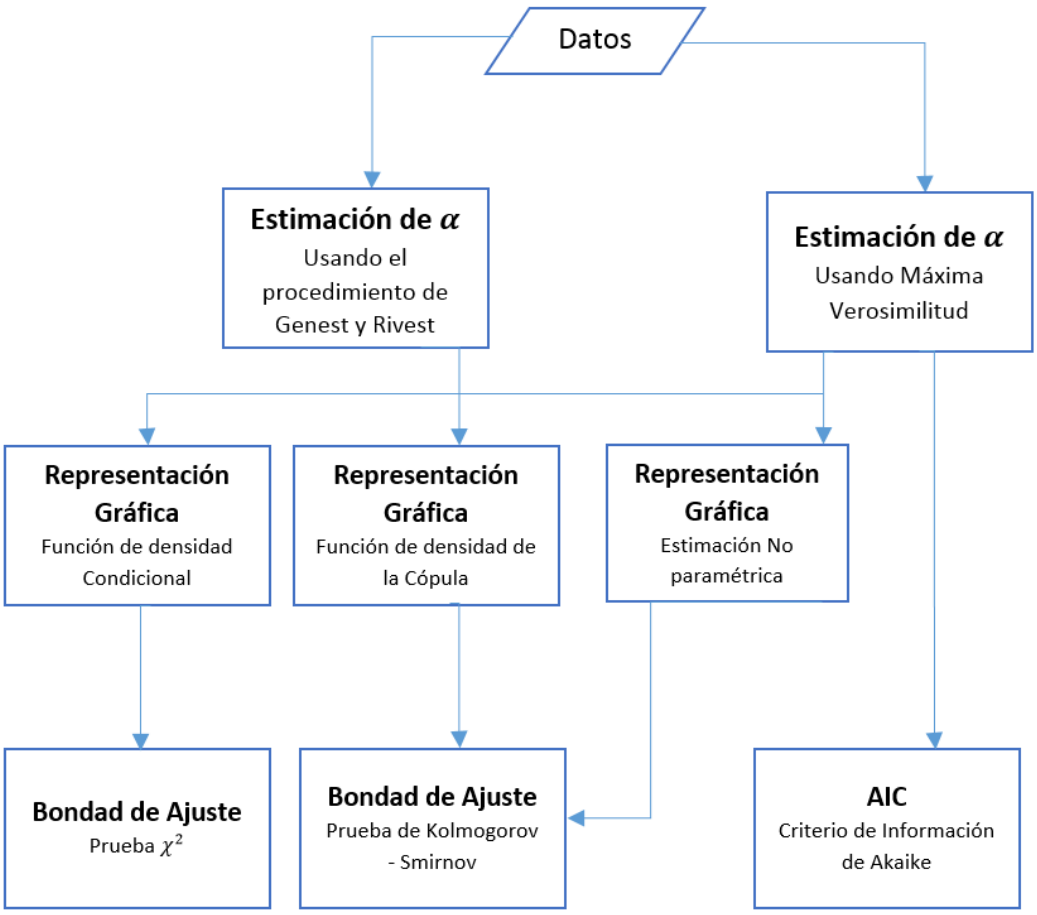
Ecuación 3.2.9: Criterio de información de Akaike

donde ρ es el número de parámetros estimados para el ajuste del modelo

El valor de AIC contiene la información de cual modelo ajusta mejor. El modelo que muestre el menor AIC, se debe considerar como el mejor.

De Matteis (2001), resume el procedimiento de ajuste de una cópula a los datos en unos pocos pasos mediante un diagrama de flujo. este diagrama define los pasos a seguir para tal procedimiento, de acuerdo con la metodología elegida:

El procedimiento a seguir para el ajustar una cópula a un conjunto de datos es ilustrado en la gráfica 3.2.6.



3.3. Ajuste de la cópula a la estructura de covarianza

La gráfica describe los pasos a seguir de acuerdo a

Resumen del procedimiento para el ajuste de la cópula:

1. Calcular las pseudo-observaciones para cada una de las variables incluidas en el estudio. En el caso particular de los datos longitudinales, el tiempo de observación será una de las variables a tomar en cuenta.

2. Estimar el parámetro θ con los procedimiento descritos en la sección 3.2.2, para la mayor cantidad de familias posible. Esto con el fin de obtener un abanico de posibilidades para comparar.

3. Calcular el *AIC* para cada uno de los estimadores de máxima verosimilitud. El *AIC* de menor valor representa el mejor modelo. Usar las tres comparaciones gráficas mencionadas anteriormente.

4. Un método más potente para decidir que tan bien se ajustan los estimadores a los datos es usar las pruebas de bondad de ajuste para los gráficos en el paso 3. Determinar el valor P de ambas pruebas, el mayor de ellos determinará la cópula mejor ajustada.

La adecuada selección de la cópula depende de un análisis cuidadoso de los resultados de los pasos 2, 3 y 4, sin que esto represente un método definitivo para la elección del mejor modelo.

3.3.1. Ajuste de las matrices de covarianzas

Se realiza la especificación de las correlaciones para luego ajustar las matrices de covarianza aprovechando la forma funcional de estas con respecto a las correlaciones, de esta forma las formas funcionales mencionadas en la última columna de la tabla 2.2, son usadas para encontrar el valor de ρ, el cual se incorpora en las matrices de covarianza presentadas a continuación:

Independencia

$$\sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

No estructurada

$$\begin{pmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1n} \\ \theta_{21} & \theta_{22} & \dots & \theta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{n1} & \theta_{n2} & \dots & \theta_{nn} \end{pmatrix}$$

Simetría Compuesta

$$\begin{pmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \dots & \sigma_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^2 & \sigma_1^2 & \dots & \sigma^2+\sigma_1^2 \end{pmatrix}$$

AR(1)

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \dots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-1} & \dots & 1 \end{pmatrix}$$

Ecuación 3.3.1: Formas funcionales de la matriz de covarianzas

Capítulo 4

Resultados

4.1. Una Aplicación

Para hacer efectiva esta comparación se tomaron los datos presentados por Belenky et al. (5) en el que se examinan los datos del tiempo de reacción promedio por día en individuos elegidos para un estudio de privación de sueño. En el día 0 los sujetos contaban con su cantidad normal de sueño. A partir de aquella noche se les limitaba a 3 horas de sueño por noche. Las observaciones representan el tiempo de reacción promedio del sujeto en una serie de pruebas a las que fueron sometidos los primeros diez días del estudio.

4.1.1. Examen de los datos

En las gráficas 4.1 y 4.2 se presentan el comportamiento de los tiempos de respuesta para cada uno de estos individuos

La figura 4.2, presenta las tendencias lineales de las respuestas de los individuos en el tiempo. Al examinar los individuos 308, 350, 331, 351, 370 y 371, se evidencia que las parejas de puntos están lejos de ajustarse a las tendencias lineales, lo que puede indicar baja correlación lineal. además el individuo 335 evidencia un comportamiento inverso, tendencia con pendiente negativa. Para corroborar estas observaciónes, se calcula la matriz de correlaciones .

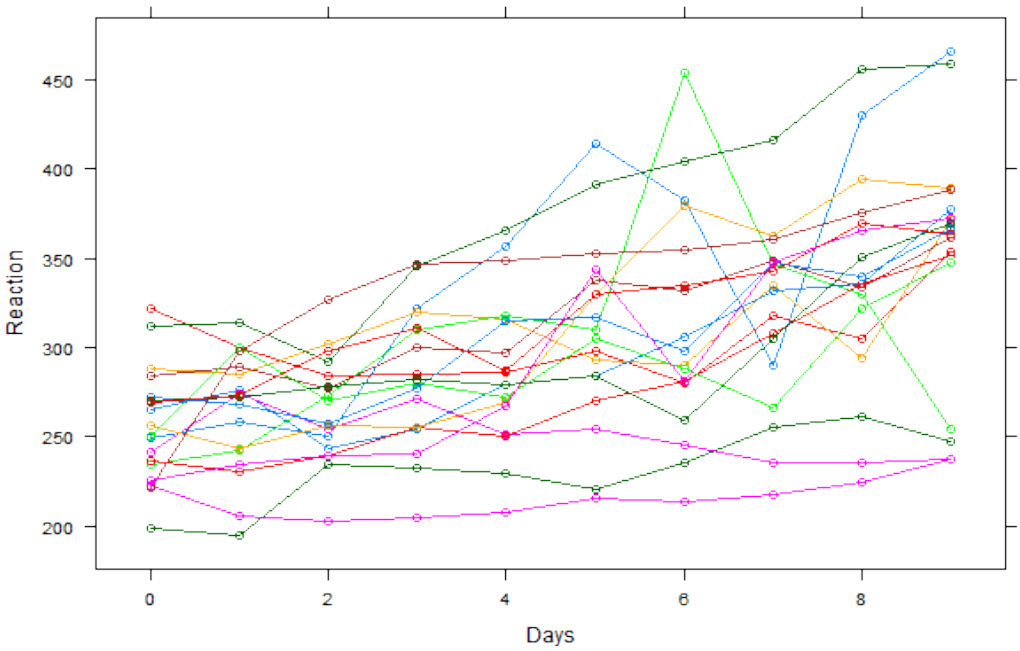


Figura 4.1: Gráfico de perfiles de los tiempos de respuesta de los individuos del estudio en los primeros diez días del estudio

	$t0$	$t1$	$t3$	$t4$	$t5$	$t6$	$t7$	$t8$	$t9$	$t10$
$t1$	$1,0000$	$0,7367$	$0,4698$	$0,4637$	$0,4489$	$0,3720$	$0,2218$	$0,4929$	$0,3291$	$0,5157$
$t2$	$0,7367$	$1,0000$	$0,7704$	$0,7415$	$0,6510$	$0,5290$	$0,3150$	$0,4762$	$0,3952$	$0,5465$
$t3$	$0,4698$	$0,7704$	$1,0000$	$0,8751$	$0,6940$	$0,4921$	$0,4543$	$0,5903$	$0,4064$	$0,4232$
$t4$	$0,4637$	$0,7415$	$0,8751$	$1,0000$	$0,9137$	$0,7221$	$0,6748$	$0,5996$	$0,5958$	$0,5657$
$t5$	$0,4489$	$0,6510$	$0,6940$	$0,9137$	$1,0000$	$0,8541$	$0,7493$	$0,6949$	$0,7446$	$0,7156$
$t6$	$0,3720$	$0,5290$	$0,4921$	$0,7221$	$0,8541$	$1,0000$	$0,7432$	$0,6903$	$0,9010$	$0,8377$
$t7$	$0,2218$	$0,3150$	$0,4543$	$0,6748$	$0,7493$	$0,7432$	$1,0000$	$0,7035$	$0,7287$	$0,4601$
$t8$	$0,4929$	$0,4762$	$0,5903$	$0,5996$	$0,6949$	$0,6903$	$0,7035$	$1,0000$	$0,7617$	$0,6587$
$t9$	$0,3291$	$0,3952$	$0,4064$	$0,5958$	$0,7446$	$0,9010$	$0,7287$	$0,7617$	$1,0000$	$0,8814$
$t10$	$0,5157$	$0,5465$	$0,4232$	$0,5657$	$0,7156$	$0,8377$	$0,4601$	$0,6587$	$0,8814$	$1,0000$

Los resultados arrojados por la matriz evidencian ausencia de correlación entre puntos lejanos en el tiempo. En términos del modelo a elegir, se espera que la construcción de un modelo mixto con estructura de correlación AR(1) sea suficiente para el problema. Para comprobarlo, se construyen los modelos mixtos con tres distintas estructuras de covarianza incluyendo la ya mencionada, autorregresiva de orden 1.

4.1.2. Construcción del modelo

Se realiza el procedimiento descrito en la sección **??**, se ajustan las tres estructuras de covarianza consideradas para comparar y decidir acertadamente el mejor modelo para los datos del estudio del sueño. La tabla 4.1 muestra los resultados de los distintos ajustes.

Estr. Correlación	Modelo	AIC	logLik
Sin estructura	linear	1965.04	-980.524
Simetría Compuesta	fit.cs	1910.32	-952.163
	fit.cs1	1794.46	-893.232
	fit.cs.het	1774.04	-874.024
AR(1)	fit.ar1	1780.33	-887.168
	fit.ar1.1	1747.20	-869.603
	fit.ar1.het	1728.22	-851.113
No estructurada	fit.un	1770.99	-838.499
	fit.un.1	1739.72	-821.860
	fit.un.het	1730.19	-808.097

Cuadro 4.1: Indices de Ajuste de los modelos lineales mixtos considerados

Mediante MLM

En los resultados de la tabla 4.1, se observa que el mejor ajuste está determinado por el modelo con estructura de covarianza AR-1 heterogénea, el cual presenta, los índices más bajos.

Mediante cópulas

En esta etapa de la comparación, se prueban las distintas cópulas de acuerdo a la descripción dada en la sección **??** para determinar cuál es la que mejor se ajusta.

La cópula se construye sin especificar las distribuciones de sus variables marginales. Para esto se halla una estimación del parámetro de dependencia y se calculan las pseudo-observaciones de los datos, con las cuales se ajustan los distintos modelos de cópulas arquimedianas a comparar. El parámetro de dependencia de éstas cópulas se puede estimar, aún, con el desconocimiento de las marginales (Genest y Rivest (24)).

Se realizan estimaciones para cópulas bivariadas; el tiempo es tomado como una de las variables. Se consideran tres cópulas: clayton, frank, y gumbel, de las cuales por problemas en la especificación de los coeficientes de correlación, no se tuvo en cuenta finalmente, la cópula frank. Definidas de dimensión 2, lsus distribuciones de probabilidad correspondientes se muestran en la tabla 4.2, donde $u = F_{y_{ij}}$ y $v = G_{t_j}$

El parámetro de dependencia θ para cada cópula, es estimado de acuerdo con el procedimiento descrito en la sección 3.2.2, en el cual, se calcula inicialmente el coeficiente τ de Kendall, para

Cópula	$C_\theta(u,v)$
Clayton	$\text{máx}\left(\left(u^{-\theta}+v^{-\theta}-1\right),0\right)$
Gumbel	$\exp\left(-\left[(-\ln(u))^{\theta}+(-\ln(v))^{\theta}\right]\right)$

Cuadro 4.2: Fórmulas de las cópulas producidas por los generadores de la tabla 2.2

después invertirlo, de acuerdo con la fórmula correspondiente (ver tabla 2.2). Así, el parámetro de dependencia θ, es estimado a través del coeficiente τ.

Las estimaciones de los parámetros de dependencia, para las cópulas clayton, frank y gumbel, y sus correspondientes pruebas de bondad de ajuste, se presentan a continuación:

```
 "Parametric bootstrap based GOF test with 'method'="Sn", 'estim.method'="itau""
       parameter statistic p.value      data.name
gof.cl 1.294508  0.3546206 0.0004995005 "x"
gof.gb 1.647254  0.2090318 0.0004995005 "x"
        ll.cl     ll.gb
[1,] 8.637997  30.33055
```

En la última linea de la salida, se pueden observar las verosimilutudes de los modelos, ajustados con las cópulas mencionadas . La cópula clayton presenta la menor verosimilitud.

En segunda instancia, se hace la estimación de los parámetros de las cópulas, usando la metodología de máxima verosimilitud (ML). En este caso, aunque computacionalmente más demorado, los modelos son comparables debido a que el indice AIC puede ser calculado. Los resultados de la estimación y ajuste de los modelos son los siguientes:

```
 "Parametric bootstrap based GOF test with 'method'="Sn", 'estim.method'="ml""
       parameter statistic p.value      data.name
gof.cl 0.7135556 0.5553941 0.0004995005 "x"
%gof.fr 3.862553  0.2566275 0.0004995005 "x"
gof.gb 1.584489  0.2419902 0.0004995005 "x"

         Clayton     Gumbel
AIC    -31.94066  -59.07948
LogLik  16.97033   30.53974
```

La estimación por máxima verosimilitud, sugiere que el mejor modelo para los datos en es la cópula clayton. Sin embargo, los anteriores modelos Cópula no tienen en cuenta que los datos tienen una correlación debida al tiempo.

Por esta razón, se construye bajo los mismos parámetros de prueba modelos copula para 10 variables (una por cada tiempo) a fin de encontrar cuál de las copulas consideradas modela mejor los datos.

Desconocidas las marginales, se inicia la estimación de los parámetros de dependencia para las cópulas, siguiendo la metodología de inversión del τ :
Copula Clayton

```
    Parametric bootstrap based GOF test with 'method'="Sn",
    'estim.method'="itau"
data:  x
statistic = 0.0731, parameter = 1.852, p-value = 0.3252
loglik  48.40571
```

Copula Frank Copula Gumbel

```
  Parametric bootstrap based GOF test with 'method'="Sn",
  'estim.method'="itau"
 data:  x
 statistic = 0.1986, parameter = 1.926, p-value = 0.005495
 loglik 42.10838
```

Se compara el p-valor del parámetro de dependencia, y la log-verosimilitud de los tres modelos. El mayor de los p-valores, corresponde al modelo clayton. Pero la log-verosimilitud más baja se observa en el modelo frank.

La estimación por máxima verosimilitud permite comparar los AIC de cada modelo. Los siguientes resultados, dan cuenta de la calidad de los modelos. Respecto a los que consideraban únicamente dos variables, el AIC de los modelos de dimensión 10 es mayor, pero en comparación con los modelos lineales mixtos, los AIC, son mucho menores, incluso que el modelo de mejor desempeño en esa categoría, el modelo con estructura AR-1 heterogénea.
Copula Clayton

```
Parametric bootstrap based GOF test with 'method'="Sn",
'estim.method'="ml"
data:  x
statistic = 0.1441, parameter = 1.32, p-value = 0.05944
```

Copula Gumbel

```
Parametric bootstrap based GOF test with 'method'="Sn",
'estim.method'="ml"
data:  x
statistic = 0.2721, parameter = 1.671, p-value = 0.0004995
```

Indice AIC y log verosimilitud para las cópulas con 10 Variables

```
         Clayton     Gumbel
AIC    -105.1314  -87.63189
LogLik   53.5657   44.81594
```

Entre los dos modelos cópula, y con este ajuste de 10 variables, se elige la cópula *Gumbel* para realizar la especificación de los coeficientes de correlación con los cuales se producirán las matrices de varianzas y covarianzas.

4.1.3. Ajuste de las matrices de covarianzas

Se realiza la especificación de las correlaciones para luego ajustar las matrices de covarianza aprovechando la forma funcional de estas con respecto a las correlaciones, de esta forma las formas funcionales mencionadas en la última columna de la tabla 2.2, son usadas para encontrar el valor de ρ, el cual se incorpora en las matrices de covarianza de la siguiente forma:

se calculan las pseudo-observaciones

```
ps.sleep<-pobs(sleep[,c("Reaction","Days")])
```

luego se calcula el coeficiente de correlación sobre estas pseudo-observaciones y con las formulas de la tabla 2.2, se calculan parámetros iniciales para las cópulas.

```
ktau<-cor(ps.sleep,method = "kendall")[1,2]
#Estimación de los parámetros de las cópulas
# family no. 1  ==============   (Clayton)
alpha.cl<-function(ktau){a.cl=(2*ktau)/(1-ktau)
ifelse (a.cl >= -1 && a.cl!= 0, {isValid<-" is valid!"},{isValid<-" is not valid!"})
cat("alpha.cl=", a.cl , isValid,"\n")}
alpha.cl(ktau)
a.cl=(2*ktau)/(1-ktau)
# family no. 4  ===============   (Gumbel)
alpha.gu<-function(ktau) {a.gu=-1/(ktau-1)
ifelse (a.gu >= 1, {isValid=" is valid"}, {isValid=" is not valid"})
cat("alpha.gu=", a.gu , isValid,"\n")}
alpha.gu(ktau)
a.gu=-1/(ktau-1)
```

El procedimiento `fitCopula` utiliza la primera estimación introducida en la especificación de la cópula para refinar el parámetro de cópula y volver a especificar el coeficiente de correlación con esta medida, ya refinada:

```
#Definición de las cópulas con los parámetros estimados
CopClayton<- archmCopula(family = "clayton", dim = 2, param = a.cl)
```

```
CopGumbel <- archmCopula(family = "gumbel", dim=2, param = a.gu)

#Ajustando la Cópula por metodo de inversión Kendall's tau
fitCopula(copula = CopClayton, ps.sleep, method="itau", estimate.variance=FALSE)
fitCopula(copula = CopGumbel, ps.sleep, method="itau", estimate.variance=FALSE)

# calculo de tau
rho.cl<-a.cl/(a.cl+2)
rho.cl

#cálculo de rho
rho.gu<-1-1/(a.gu)
rho.gu
```

Ahora estos valores se llevan a las estructuras de covarianza y se calculan de nuevo los modelos de efectos fijos para medir así su desempeño frente a los datos, obteniendo los siguientes resultados

Matriz de simetría compuesta
Estimación de los parámetros de las Cópulas por cada tiempo

$$\hat{r} = 0{,}3929291 \ ,(\theta_{Gumbel} = 1{,}647 \ , \ \theta_{clayton} = 1{,}295)$$

Modelo	AIC	logLik
SimComp (S)	1794.465	-893.2325
SimComp (C)	1797.302	-895.6509

**S= Sin Cópulas, C=Con Cópulas*

Matriz Autorregresiva de orden 1

$$\hat{r} = 0{,}3929291 \ ,(\theta_{Gumbel} = 1{,}647 \ , \ \theta_{clayton} = 1{,}295)$$

Modelo	AIC	logLik
AR(1) (S)	1794.465	-893.2325
AR(1) (C)	1797.302	-895.6509
AR(1) Het (C)	1774.049	-874.0246

**S= Sin Cópulas, C=Con Cópulas*

Matriz no estructurada

$$\hat{r} = 0{,}3929291 \ ,(\theta_{Gumbel} = 1{,}647 \ , \ \theta_{clayton} = 1{,}295)$$

Modelo	AIC	logLik
No Estr (S)	1770.999	-838.4997
No Estr (C)	1739.722	-821.8608
No Estr Het (C)	1730.195	-808.0973

**S= Sin Cópulas, C=Con Cópulas*

Se observa una reducción de los valores de los AIC en tanto se utillizan las cópulas para la especificación de los coeficientes de correlación.

4.2. Otra Aplicación

Thall and Vail (1990) dan un conjunto de datos para el conteo episodios de 59 epilépticos. El número de episodio fue registrado para una linea base de 8 semanas, y los pacientes fueron asignados aleatoriamente al grupo tratamiento o control. Los conteos fueron registrados para cuatro periodos sucesivos de 2 semanas.

Se enfrentan el número del periodo versus la cantidad de episodios del paciente, y se modela el número de episodios mediante un modelo de efectos fijos:

Para eso se tienen las variables

y número de episodios en un periodo

trt si se asigno al grupo control se denota por `"placebo"`, de lo contrario `"progabide"`

period Un número entero entre uno y cuatro que denota el tiempo de la observación

Los pasos para el modelamiento, se hicieron por separado, primero un modelo para los del grupo control y luego otros para el grupo tratamiento. Se secciona la base de acuerdo con el tipo de tratamiento, y se calculan las medidas de correlación entre los tiempos y las respuestas. Para poder trabajar con cópulas es necesario hallar las pseudo-observaciones y de la misma forma calcular las medidas de correlación para estos tiempos *pseudo observados*. La grafica 4.2, muestra el comportamiento de cada uno de los individuos, a medida que avanza el tratamiento. En la grafica 4.2, se observa el comportamiento de cada individuo en casillas separadas.

se encuentran los siguientes resultados

```
> correl.tiempo(mat)
[1] 1.0000000 0.7823271 0.4940244 0.6745946
> correl.tiempo(mat2)
[1] 1.0000000 0.9070300 0.9124734 0.9713376
```

La gráfica 4.2, muestra como varían las correlaciones de los datos entre tiempos distintos respecto a la primera medición. Así mismo, en colores oscuros se muestran las mediciones de la correlación usando un método no paramétrico sobre las pseudo-observaciones; es clara la diferencia de las primeros cálculos con los segundos.

La estimación de estas correlaciones es la base para el cálculo de los parámetros de las cópulas, las cuales, luego de un procedimiento de máxima verosilmilitud estimarán el parámetro adecuado para cada uno de los grupos y luego un parámetro general para medir el efecto del tratamiento sobre los individuos, y así poder determinar si el número de episodios disminuye con la aplicación continua del tratamiento.

Ajuste de los parámetros de la cópula

Los parámetros de cópula ajustados son los siguientes:

```
>   a.cl;a.gu
[1] a.cl= -0.05469022
[1] a.gu=  0.9726549  invalido!
>   a.cl2;a.gu2
[1] a.cl2= -0.04531925
[1] a.gu2= 0.9773404 invalido!
```

Matriz de simetría compuesta

$$\hat{r} = -0{,}02811389$$
$$\theta_{Gumbel} = No\ calculado$$
$$\theta_{clayton} = -0{,}05469$$

Modelo	AIC	logLik
SimComp (S)	1655.335	-822.6675
SimComp (C)	1873.38	-932.69
SimCompHet(S)	1602.253	-793.1266
SimCompHet(C)	1686.185	-836.0924

**S= Sin Cópulas, C=Con Cópulas*

Matriz Autorregresiva de orden 1

$$\hat{r} = -0{,}02811389$$
$$\theta_{Gumbel} = No\ calculado$$
$$\theta_{clayton} = -0{,}05469$$

Modelo	AIC	logLik
AR(1) (S)	1690.915	-840.4575
AR(1) (C)	1865.757	-929.8785
AR(1) Het (S)	1640.694	-813.3469
AR(1) Het (C)	1855.732 -921.8661	

**S= Sin Cópulas, C=Con Cópulas*

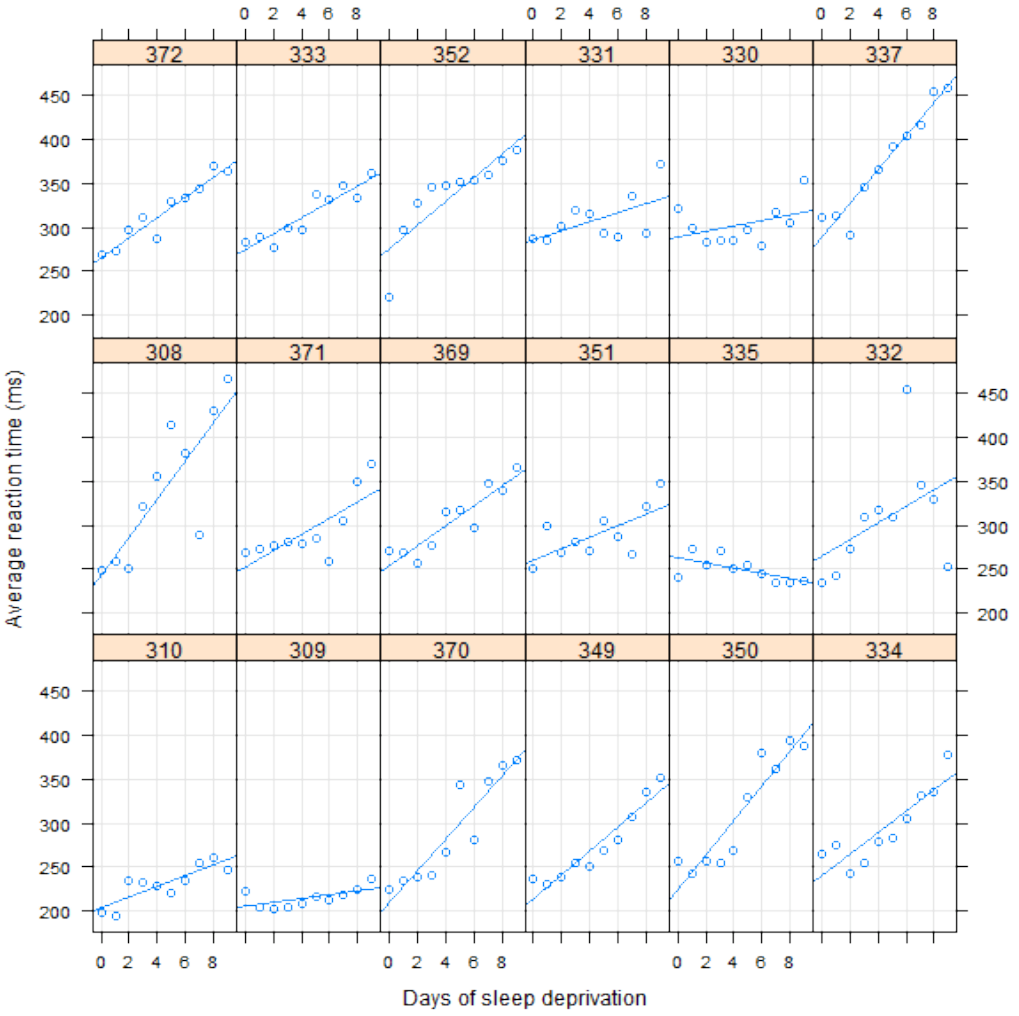


Figura 4.2: Gráfico trellis de los tiempos de respuesta de los individuos del estudio

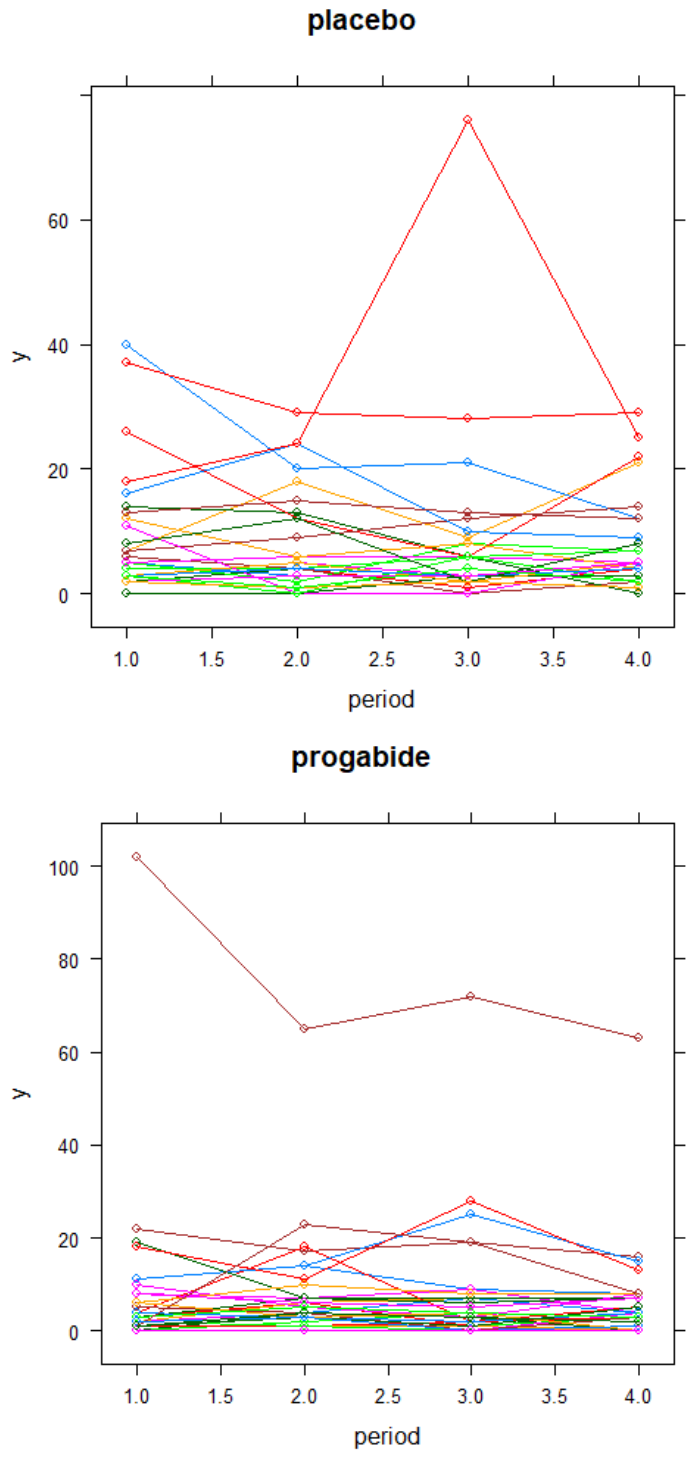


Figura 4.3: Gráfico de perfiles para los individuos en los dos tratamientos

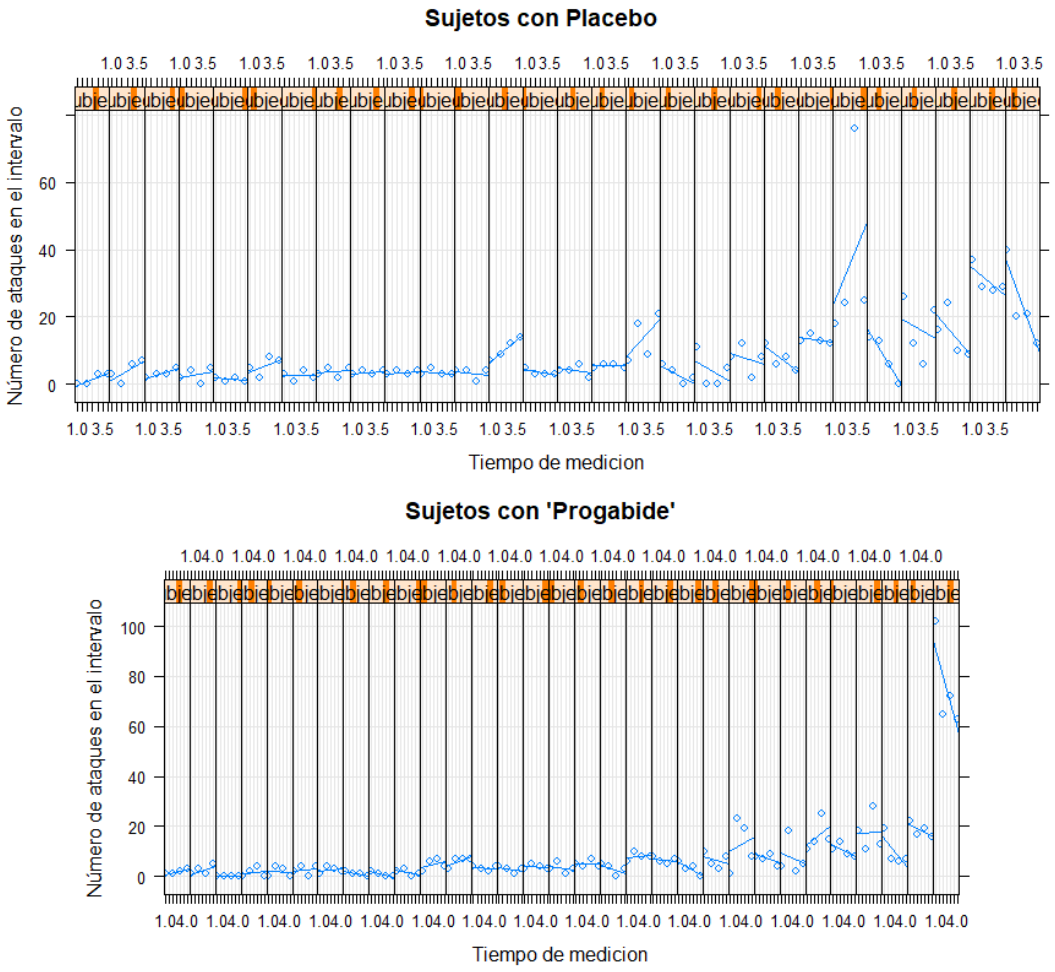


Figura 4.4: Gráfico de perfiles para los individuos en los dos tratamientos

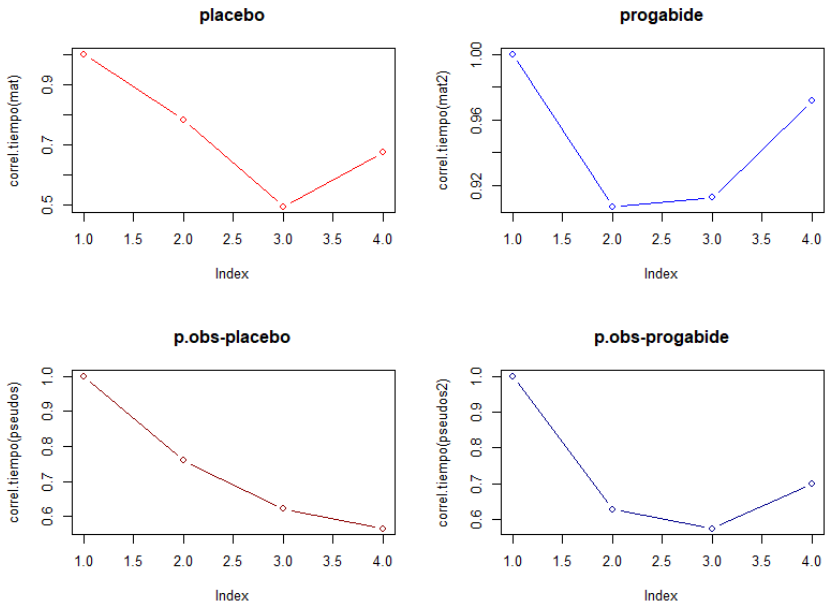


Figura 4.5: Gráfico de perfiles para los individuos en los dos tratamientos

Conclusiones

El trabajo que representa el modelado de datos longitudinales, tiene por no decir mucho más, una gran cantidad de variantes respecto a las metodologías posibles para lograrlo. La construcción de un modelo que separa los errores residuales producidos por los efectos fijos y los errores puros, conlleva a una buena cantidad de consideraciones, como por ejemplo, la consideración de diferentes distribuciones de probabilidad para explicar el comportamiento de cada uno de los tipos de error.

Las cópulas fueron introducidas en este trabajo con el animo de mejorar la estimación de los parámetros incorporados en la matriz de varianzas-covarianzas con la cual se especifica la distribución de probabilidad de los residuales de los efectos fijos.

Se logró evidenciar que las cópulas ayudan a disminuir el error de estimación de los coeficientes de correlación, y en este sentido, obtener mejores estimaciones de los parámetros de los modelos lineales con efectos fijos y así mismo observar que producen en los modelos un desempeño que aunque similar, no deja de ser mejor que los modelos estimados bajo la metodología usual.

En términos generales se lograron las siguientes observaciones:

1. La verosimilitud de los modelos ajustados con las cópulas es más baja.
2. El modelo lineal de efectos fijos mejor comportado se ajustaba con una matriz de correlación No estructurada.
3. En las pruebas hechas no se encontró una diferencia significativa en el ajuste del modelo cuando se usan distintas cópulas.
4. No todas las cópulas cuentan con una forma cerrada de ecuación respecto al coeficiente de correlación, por lo que no todas se pueden usar.
5. Se logró especificar un algoritmo de ajuste de los modelos en el cual se incluyen las cópulas para la especificación de las varianzas y covarianzas.
6. Se encontró que en datos que presenten baja o nula correlación lineal, las cópulas mejoran significativamente el ajuste de los modelos, por lo que se recomienda usar el procedimiento propuesto en este trabajo. Por ejemplo en la segunda aplicación se observó que los índices de los modelos con cópulas no mejoran el desempeño del mismo, sino que por el contrario el AIC se vuelve mayor.

Bibliografía

[1] Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

[2] Alonso, R. y Pardo, M. (2001). *Modelos marginales: nuevos procedimientos de inferencia para datos longitudinales*. PhD thesis, Universidad Complutense de Madrid, madrid.

[3] Arnau, J. y Balluerka, N. (2004). Análisis de datos longitudinales y de curvas de crecimiento. enfoque clásico y propuestas actuales. *Psicothema*, 16(1):156 – 162.

[Ayyad] Ayyad, C. Inferencia y modelización mediante cópulas.

[5] Belenky et al, G. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. (12):1–12.

[6] Blanco, L. (2010). *Probabilidad*. Universidad Nacional de Colombia, 2 edition.

[7] Borges, R. E. (2004). Modelos de análisis de sobreviviencia multivariados. In *Simposio de Estadística: Estadística Multivariada. Cartagena, Colombia.*

[Christian Genest and Rivest] Christian Genest, K. G. and Rivest, L.-P. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions.

[9] Conway, D. (1979). Multivariate distributions with specified marginals. Technical Report 145, Stanford University.

[10] Cook, N.R y Ware, J. (1983). Design and analysis methods for longitudinal research. *Annu Rev Public Health*, 1(1):13–22.

[11] Dahmen, G. y Ziegler, A. (2004). Generalized estimating equations in controlled clinical trials: hypotheses testing. *Biometrika*, 46(2):214–232.

[12] Davidian, M. and Giltinan, D. (1995). *Nonlinear Models for Repeated Measurement Data*. Monographs on Statistics and Applied Probability. Chapman & Hall.

[13] Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer.

[14] Díaz, L. and Morales, M. (2012). *Analisis Estadístico de Datos Multivariados*. Universidad Nacional de Colombia.

[15] Díaz, L. G. and León, L. R. (2015). *Análisis de Medidas Repetidas*.

[16] De Matteis, R. (2001). Fitting copulas to data.

[17] Diggle, P., Liang, K., and Zeger, S. (1994). *Analysis of longitudinal data*. Oxford statistical science series. Clarendon Press.

[18] Drouet, D. and Kotz, S. (2001). *Correlation And Dependence*. Imperial College Press.

[19] Erdely, A. (2009). Copulas y dependencia de variables aleatorias: Una introducción. *Miscelanea Matemática*, 48:7–28.

[20] Escarela, G. and Hernandez, A. (2009). Modelado de parejas aleatorias usando cópulas. *Revista Colombiana de estadística*, 32(1):33 – 58.

[21] Fréchet, M. (1935). Generalizations du theoreme des probabilites totales. *Fund. Math.*, 25:379 – 387.

[22] Frees, E. and Valdez, E. (1998). Understanding relationships using copulas. *North American Actuarial Journal*.

[Genest and Rémillard] Genest, C. and Rémillard, B. Validity of the parametric bootstrap for goodness-of- fit testing in semiparameric models.

[24] Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, 88(423):1034–1043.

[25] Grønneberg, S. and Hort, N. L. (2014). The copula information criteria. *Scandinavian Journal of Statistics*, 41:436 – 459.

[26] Hedeker, D. and Gibbons, R. (2006). *Longitudinal Data Analisys*. John Wiley & Sons.

[27] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325.

[28] Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820.

[29] Kimeldorf, G. and Sampson, A. R. (1987). Positive dependence orderings. *Annals of the Institute of Statistical Mathematics*, 39(1):113–128.

[30] Kojadinovic, I., Yan, J., and Holmes, M. (2011). Fast large-sample goodness-of-fit tests for copulas. *Statistica Sinica*, 21(2):841–871.

[31] Kolev, N., U. d. A. and de Mendes, B. (2006). Copulas: A review and recent developments. *Stochastic models*, 22(4):617–660.

[32] Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861.

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

[34] Lehmann, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.*, 37(5):1137–1153.

[35] Liang and Zeher (1986). Longitudinal data analisys using generalized linear models. *Biometrika*, 73(1):1–24.

[36] Maechler, H. K. and Yan (2015). *copula*. R Foundation for Statistical Computing, Vienna, Austria.

[37] Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer, second edition.

[38] Pinheiro, J. and Bates, D. (2009). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer.

[39] R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[40] Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52(3/4):447–458.

[Rényi] Rényi, A. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3):441–451.

[42] Rothman, K. and Greenland, S. (1998). *Greenland S. Introduction to regresión modeling*. Modern Epidemiology, Filadelfia.

[43] Schweizer, B. and Wolff, E. (1981). On nonparametric measures of dependence for random variables. *The Annals of Stattistics*.

[44] Schweizer B, S. A. (1974). Operations on distribution functions not derivable from operations on random variables. *Studia Math*, 52:43 – 52.

[45] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

[46] Sosa, Berdugo, and Diaz (2010). Análisis de medidas repetidas. Departamento de Estadística. Universidad Nacional de Colombia. Pre-Print.

[47] Trivedi, P. K. and Zimmer, D. M. (2005). *Copula Modeling: An Introduction for Practitioners*. Foundations and Trends® in Econometrics.

[48] V. Gregoire, C. G. . M. G. (2008). Using copulas to model price dependence in energy markets. *Energy Risk*.

[49] Wang, A. (2010). Goodness-of-fit tests for archimedean copula models. *Statistica Sinica*, 20(1):441–453.

[50] Xu, J. (1996). *Statistical Modelling and Inference for Multivariate and Longitudinal Discrete Response Data*.

[51] Zuur, A. F. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer, Los Angeles.