Revista Colombiana de Estadística
Junio 2009, volumen 32, no. 1, pp. 33 a 58

Modelado de parejas aleatorias usando cópulas

Modelling Random Couples Using Copulas

Gabriel Escarela[1,a], Angélica Hernández[1,b]

[1]Departamento de Matemáticas, Universidad Autónoma Metropolitana Unidad Iztapalapa, Ciudad de México, México

Resumen

Las cópulas se han convertido en una herramienta útil para el modelado multivariado tanto estocástico como estadístico. En este artículo se revisan propiedades fundamentales de las cópulas que permitan caracterizar la estructura de dependencia de familias de distribución bivariadas definidas por la cópula. También se describen algunas clases de cópulas, enfatizando en la importancia de la cópula Gaussiana y la familia Arquimediana. Se resalta la utilidad de las cópulas para el modelado de parejas de variables aleatorias continuas y el de las discretas. La aplicación de la cópula se ilustra con la construcción de modelos de regresión de Markov de primer orden para respuestas no Gaussianas.

***Palabras clave*:** dependencia, cópula, medida de asociación, [estadística aplicada], τ de Kendall, ρ de Spearman, correlación serial.

Abstract

Copulas have become a useful tool for the multivariate modelling in both stochastics and statistics. In this article, fundamental properties that allow the characterization of the dependence structure of families of the bivariate distributions defined by the copula are reviewed. Also, the importance of both the Gaussian copula and the Archimedean family is emphasized while some classes of copulas are described. The usefulness for modelling either discrete or continuous random couples is highlighted. The construction of first-order Markov regression models for non-Gaussian responses illustrates the application of the copula.

***Key words*:** Dependence, Copula, Measure of association, [Applied statistics], Kendall τ, Spearman ρ, Serial correlation.

[a]Profesor investigador. E-mail: ge@xanum.uam.mx
[b]Estudiante de doctorado. E-mail: cbi206280113@xanum.uam.mx

1. Introducción

Las cópulas bidimensionales son funciones bivariadas que juntan o bien "copulan" dos funciones de distribución univariadas para construir funciones de distribución bivariadas continuas. La cópula representa una forma paramétrica conveniente para modelar la estructura de dependencia en distribuciones conjuntas de variables aleatorias, en particular para parejas de variables aleatorias. Varias cópulas con diversas formas están disponibles para representar a familias de distribuciones bivariadas.

El uso de la cópula es atractivo, pues permite una gran flexibilidad para modelar la distribución conjunta de una pareja aleatoria que pueda surgir de prácticamente cualquier disciplina, y lo hace de forma sencilla ya que solo se necesita especificar la función que copula y las marginales. Las cópulas pueden extraer la estructura de dependencia de la función de distribución conjunta de un vector de variables aleatorias y, al mismo tiempo, permiten separar la estructura de dependencia del comportamiento marginal. Al igual que en el caso univariado, es posible usar transformaciones que permitan crear funciones de distribución bivariadas discretas a partir de las distribuciones continuas; de esta forma, puede aprovecharse la cópula cuando el objetivo es modelar parejas aleatorias discretas.

Las cópulas fueron presentadas originalmente por Sklar (1959), quien resolvió algunos problemas formulados por M. Fréchet sobre la relación entre una función de distribución de probabilidad multidimensional y sus marginales de menor dimensión. En la actualidad, las cópulas se han convertido en una poderosa herramienta de modelado multivariado en muchos campos de la investigación donde la dependencia entre varias variables aleatorias, continuas o discretas, es de gran interés, y para las cuales la suposición de normalidad multivariada puede ser cuestionable.

Algunos ejemplos del modelado de parejas aleatorias continuas se pueden encontrar en aplicaciones biomédicas donde el interés puede centrarse en los tiempos de ocurrencia de una enfermedad en órganos pares (*e.g.* Wang & Wells 2000), o en los tiempos de ocurrencia de un evento cuando este se clasifica en dos tipos de causas mutuamente excluyentes, *i.e.* datos de riesgos concurrentes (*e.g.* Carriere 1995, Escarela & Carriere 2003). En ambos casos, cuando se construye la función de distribución bivariada correspondiente, es importante asignar funciones de distribución marginales del tipo de supervivencia como la Exponencial, la Weibull o la Burr, que permitan hacer comparaciones entre sus ajustes; además, es preponderante entender el mecanismo de dependencia de las parejas aleatorias.

Otras disciplinas que se favorecen de la utilización de las cópulas son la hidrología y los cálculos actuariales. La primera porque generalmente los fenómenos hidrológicos son multidimensionales; por tanto, se requiere modelar conjuntamente diferentes procesos (*e.g.* Genest & Favre 2007), mientras que la segunda lo hace en el análisis de portafolios de seguros -por mencionar un ejemplo-, donde el interés puede centrarse en la estimación de la distribución conjunta de los montos correspondientes a dos tipos de indemnización (*e.g.* Klugman & Parsa 1999); a este tipo de datos se les conoce en la literatura anglosajona como *loss data*.

Una ilustración para parejas aleatorias discretas, la cual también se puede beneficiar del modelo de cópula, proviene del análisis de series de tiempo discretas, como el número de casos mensuales de una enfermedad casi erradicada: la poliomielitis (*e.g.* Escarela et al. 2006). Cuando se tienen variables aleatorias discretas correlacionadas en serie de la forma AR(1), las funciones de distribución marginales en la distribución conjunta de las variables aleatorias adyacentes deben de tener la misma forma y -al igual que su contraparte continua- es crucial usar una estructura de dependencia. En la literatura existen muy pocas distribuciones bivariadas para parejas aleatorias discretas con estas propiedades; las distribuciones construidas con la cópula discretizada son unas de ellas.

Los lectores que buscan más aplicaciones encontrarán en los artículos de Frees & Valdez (1998) y de Clemen & Reilly (1999) una revisión más detallada. En cuanto al asunto de la implementación, el trabajo de Jan (2007) y las rutinas en el lenguaje de distribución gratuita R expuestas ahí pueden facilitar la programación de los modelos.

El propósito de este artículo es exponer algunas propiedades importantes de las cópulas y algunos detalles para su aplicación. En la segunda sección se define la cópula y se revisan propiedades fundamentales de las cópulas, las cuales permiten caracterizar la estructura de dependencia de familias de distribución bivariadas definidas por la cópula. En la tercera sección se revisan los conceptos de correlación, concordancia y dependencia. En la cuarta sección se describen tres familias de cópula relevantes en la literatura y se dan algunas consideraciones sobre la selección de la cópula; además, se muestran algunos contornos de funciones de densidad bivariadas generadas a través de varias clases de cópulas. La quinta sección muestra dos ejemplos, los cuales se enfocan en el modelado de series de tiempo AR(1) para respuestas diferentes a las Gaussianas. El primero consiste en comparar el ajuste de varios modelos de cópula a respuestas de valor extremo, mientras que el segundo presenta la representación de una matriz de transición para series de tiempo binarias en presencia de información concomitante; en ambas ilustraciones se trata de emular al modelo AR(1) para respuestas Gaussianas.

2. La cópula

2.1. Definición estadística de cópula

Una cópula bidimensional es una función de distribución bivariada de un vector aleatorio $\mathbf{V} = (V_1, V_2)$ cuyas marginales V_1 y V_2 son uniformes en el intervalo $\mathbf{I} = (0, 1)$. Es decir, una cópula es una función $C : \mathbf{I}^2 \to \mathbf{I}$ que satisface las siguientes condiciones:

1) de acotamiento

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

$$\lim_{v_j \to 0} C(v_1, v_2) = 0 \tag{2}$$

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

2) de incremento

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

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

La importancia de las cópulas en estadística matemática se describe en el siguiente teorema (*e.g.* Nelsen 1999).

Teorema 1. *(Teorema Sklar) Sean Y_1, Y_2 variables aleatorias con función de distribución conjunta F, con marginales F_1 y F_2 respectivamente. Entonces existe una cópula C tal que satisface*

$$F(y_1, y_2) = C\big[F_1(y_1), F_2(y_2)\big] \tag{3}$$

para toda $y_1, y_2 \in \mathbb{R}$. Si F_1 y F_2 son continuas, entonces C es única; de otra forma C está determinada en forma única sobre el rango $F_1 \times$ rango F_2. Inversamente, si C es una cópula y F_1, F_2 son funciones de distribución, entonces F definida en la ecuación (3) *es una función de distribución conjunta con marginales F_1 y F_2.*

Este teorema establece que, en el contexto de parejas aleatorias continuas, es posible construir una función de distribución bivariada en términos de dos funciones de distribución continuas univariadas y una cópula que permite relaciones de dependencia entre dos variables aleatorias individuales. Una demostración del teorema de Sklar puede encontrarse en Schweizer & Sklar (1983). Las cópulas pueden emplearse para definir distribuciones bivariadas con marginales discretas, de manera que satisfagan la ecuación (3); sin embargo, en contraste con el caso continuo, no hay una forma única para expresar la distribución conjunta de dos variables aleatorias discretas como una función de sus distribuciones marginales (ver Denuit & Lambert 2005). Cuando las variables son discretas, la unicidad solo se encuentra en rango $F_1 \times$ rango F_2.

Corolario 1. *Dada una función de distribución conjunta F con marginales continuas F_1 y F_2, como está indicado en el teorema de Sklar, es fácil construir la cópula correspondiente como se muestra a continuación:*

$$C(v_1, v_2) = F\Big(F_1^{(-1)}(v_1), F_2^{(-1)}(v_2)\Big)$$

donde $F_j^{(-1)}$ es la función cuasi-inversa de F_j, dada por $F_j\Big[F_j^{(-1)}(u)\Big] = u$ si $u \in$ rango F_j, o por $F_j^{(-1)}(u) = \sup\{z \mid F_j(z) \leq u\}$ si $u \notin$ rango F_j, para $j = 1, 2$; aquí las cuasi-inversas se usan para funciones de distribución no estrictamente crecientes. Nótese que si Y_1 y Y_2 son variables aleatorias continuas con funciones de distribución F_1 y F_2, respectivamente, entonces C es la función de distribución conjunta de $V_1 = F_1(Y_1)$ y $V_2 = F_2(Y_2)$ ya que $F_1(Y_1)$ y $F_2(Y_2)$ se distribuyen uniformemente en $\mathbf{I}$.

La desigualdad de las *cotas Fréchet-Hoeffding* indica que si F es una función de distribución bivariada con marginales F_1 y F_2, entonces

$$\text{máx}\{F_1(y_1) + F_2(y_2) - 1, 0\} \leq F(y_1, y_2) \leq \text{mín}\{F_1(y_1), F_2(y_2)\}$$

Este resultado es consecuencia del teorema de Sklar. En términos de cópulas, la desigualdad puede expresarse como (Fréchet 1951)

$$W(v_1, v_2) = \text{máx}\{v_1 + v_2 - 1, 0\} \leq C(v_1, v_2) \leq \text{mín}\{v_1, v_2\} = M(v_1, v_2)$$

la cual se conoce como desigualdad de las cotas de Fréchet e indica que cualquier cópula C representa un modelo de dependencia que se encuentra entre los extremos W y M. Las funciones W y M son conocidas como las cotas inferior y superior, respectivamente; de hecho, Fréchet (1951) demuestra que W y M son también cópulas.

2.2. La función de densidad de la cópula y la cópula de supervivencia

Si $F_1(y_1)$, $F_2(y_2)$ y la cópula $C(v_1, v_2)$ son diferenciables, la densidad conjunta de (Y_1, Y_2), correspondiente a la función de distribución conjunta en la ecuación (3), puede expresarse como

$$f(y_1, y_2) = f_1(y_1) f_2(y_2) \times c\big[F_1(y_1), F_2(y_2)\big]$$

donde $f_1(y_1)$ y $f_2(y_2)$ son las funciones de densidad marginales correspondientes, y

$$c(v_1, v_2) = \frac{\partial^2 C(v_1, v_2)}{\partial v_1 \partial v_2} \qquad (v_1, v_2)^T \in (0, 1)^2 \tag{4}$$

es la función de densidad de la cópula. Como consecuencia, se tiene que la función de densidad condicional de Y_2 dada Y_1 puede expresarse convenientemente de la siguiente forma:

$$f_{2|1}(y_2 \mid y_1) = f_2(y_2) \times c\big[F_1(y_1), F_2(y_2)\big] \tag{5}$$

En varias situaciones donde el objetivo es el modelado, es más conveniente hablar de la función de supervivencia conjunta, la cual se define como $S(y_1, y_2) = \Pr\{Y_1 > y_1, Y_2 > y_2\}$, en lugar de la función de distribución conjunta $F(y_1, y_2)$; esta representación es particularmente relevante cuando se tienen parejas de variables aleatorias positivas. En forma análoga a como se construye la función de distribución conjunta, si se dan dos funciones de supervivencia marginales $S_j(y_j) = \Pr\big\{Y_j > y_j\big\}$, con $j = 1, 2$, estas pueden ser "copuladas" para formar una función de supervivencia conjunta, como se muestra a continuación (*e.g.* Wang & Wells 2000)

$$S(y_1, y_2) \equiv C\big[S_1(y_1), S_2(y_2)\big] \tag{6}$$

La función de densidad conjunta correspondiente a la función de supervivencia definida en la ecuación (6) es:

$$f(y_1, y_2) = f_1(y_1) f_2(y_2) \times c\big[S_1(y_1), S_2(y_2)\big]$$

donde $f_1(y_1)$ y $f_2(y_2)$ son las funciones de densidad marginales correspondientes a $S_1(y_1)$ y $S_2(y_2)$, respectivamente, y $c(\cdot, \cdot)$ es la cópula de densidad definida en la ecuación (4).

2.3. El caso discreto

El potencial de las cópulas no se reduce al caso continuo. Cuando el modelado es para parejas de variables aleatorias discretas, se puede obtener una función de probabilidad bivariada al tomar la derivada Radon-Nikodym de $F(y_1, y_2)$ en la ecuación (3) con respecto a la medida contable. De esta forma, la función de probabilidad conjunta de una pareja de variables aleatorias discretas (Y_1, Y_2) puede representarse en términos de la versión discretizada de la cópula y de las funciones de distribución marginales, las cuales tienen la forma $F_j(y_j) = \sum_{z \leq y_j} f_j(z)$, donde $f_j(y_j) = \Pr\{Y_j = y_j\}$ representa la función de probabilidad marginal de Y_j para $j = 1, 2$, como se muestra a continuación (Song 2000):

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

aquí $u_j = F_j(y_j)$ y $v_j = F_i(y_j - 1)$ para $j = 1, 2$.

Cuando se trata de representar la familia de distribuciones condicionales de $Y_2 \mid Y_1$, esta toma la forma

$$F_{2|1}(y_2 \mid y_1) = \Big\{C\big[F_1(y_1), F_2(y_2)\big] - C\big[F_1(y_1 - 1), F_2(y_2)\big]\Big\} \Big/ f_1(y_1)$$

3. Medidas de correlación, concordancia y dependencia

Cuando los modelos de cópula son usados para construir una distribución conjunta de una pareja aleatoria continua, pueden verse como versiones de funciones de distribución conjuntas libres de marginales que tienen la poderosa habilidad de capturar propiedades de dependencia invariante al reescalamiento de las parejas aleatorias (ver *e.g.* Rodríguez-Lallena & Úbeda Flores 2004); aquí, invariante al reescalamiento significa que las propiedades y las medidas se quedan sin cambiar cuando se realizan transformaciones estrictamente crecientes a 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.

Además de la concordancia, existen varios conceptos de correlación, asociación y dependencia, importantes para entender el modelado de parejas aleatorias usando las cópulas (ver *e.g.* Lehmann 1966, Barlow & Proschan 1975, Nelsen 1991). En esta sección se describen algunos de los conceptos y medidas correspondientes, los cuales son relevantes en la literatura.

3.1. El coeficiente de correlación

El coeficiente de correlación es la forma más tradicional para cuantificar la relación de dos variables aleatorias. Este mide la fuerza y dirección de una relación lineal entre dos variables aleatorias.

Axioma 1. *Sean Y_1, Y_2 variables aleatorias con varianzas finitas. Entonces el coeficiente de correlación de Pearson, se define como*

$$\begin{aligned}\mathrm{Cor}(Y_1, Y_2) &= \frac{\mathrm{Cov}(Y_1, Y_2)}{\sqrt{\mathrm{Var}[Y_1]}\,\sqrt{\mathrm{Var}[Y_2]}} \\ &= \frac{\mathrm{E}\big\{\big(Y_1 - \mathrm{E}[Y_1]\big)\big(Y_2 - \mathrm{E}[Y_2]\big)\big\}}{\Big\{\mathrm{E}\big(Y_1 - \mathrm{E}[Y_1]\big)^2\Big\}^{1/2}\Big\{\mathrm{E}\big(Y_2 - \mathrm{E}[Y_2]\big)^2\Big\}^{1/2}}\end{aligned}$$

Entre las propiedades del coeficiente de correlación de Pearson se encuentra que su rango está en el intervalo $[-1, 1]$, con valores ± 1 si y solo si $Y_1 = a + bY_2$. Este coeficiente es simétrico, es decir, $\mathrm{Cor}(Y_1, Y_2) = \mathrm{Cor}(Y_2, Y_1)$, y no cambia bajo transformaciones lineales; esto es, $\mathrm{Cor}\,(Y_1, f(Y_2)) = \mathrm{Cor}(Y_1, Y_2)$ cuando $f(y) = a + b\,y$, donde $b > 0$.

Como las cópulas permiten un camino fácil en el estudio de la dependencia entre variables aleatorias y son invariantes al reescalamiento, entonces el coeficiente de correlación de Pearson se usa con más frecuencia como medida de dependencia, pues es más fácil calcular y es un parámetro importante en distribuciones elípticas; de hecho, a menudo se emplea en la familia normal multivariada y en la distribución t-Student multivariada. Sin embargo, para algunos casos de parejas aleatorias continuas cuyas distribuciones no son elípticas, como es el caso de distribuciones construidas con ciertas cópulas. La utilidad de este coeficiente es poca, pues su valor depende no solo de la cópula, sino también de las marginales, ya que $\mathrm{Cor}(f(Y_1), f(Y_2)) \neq \mathrm{Cor}(Y_1, Y_2)$ cuando $f(y)$ no es lineal; es decir, esta medida no siempre es invariante al reescalamiento.

3.2. Medidas de concordancia

Cuando se considera una pareja de variables aleatorias, es útil saber qué tanto tienden a estar asociados valores grandes de una de las variables aleatorias con valores grandes de la otra, y que tanto están asociados valores pequeños de una con valores pequeños de la otra. Una formalización de la idea intuitiva de este grado de asociación fue propuesta por Yanagimoto & Okamoto (1969), quienes proponen el uso del orden de concordancia de distribuciones bivariadas con marginales univariadas dadas de acuerdo con la fuerza de su asociación positiva, el cual se denota por $\prec$. Este orden estocástico se define a continuación.

Axioma 2. *Dadas dos parejas aleatorias (X_1, Y_1) y (X_2, Y_2) con marginales idénticas, se dice que (X_2, Y_2) es más concordante que (X_1, Y_1), y se denota $(X_1, Y_1) \prec (X_2, Y_2)$, si*

$$\Pr\{X_1 \leq s, Y_1 \leq t\} \leq \Pr\{X_2 \leq s, Y_2 \leq t\}$$

para toda $s, t \in \mathbb{R}$.

Es importante señalar que el uso de las medidas de concordancia permiten construir estimaciones fiables cuando se asume que la cópula pertenece a una familia paramétrica específica.

El término medida de asociación se refiere a una medida de concordancia, un concepto desarrollado por Scarsini (1984) y presentado por Nelsen (1999) como se define a continuación.

Axioma 3. *Una medida numérica κ de asociación entre dos variables aleatorias continuas Y_1 y Y_2, cuya cópula es C, es una medida de concordancia si:*

1. *κ esta definida para cualquier pareja de variables aleatorias continuas*
2. *$\kappa \in [-1, 1]$ con $\kappa(Y, Y) = 1$ y $\kappa(Y, -Y) = -1$*
3. *$\kappa(Y_1, Y_2) = \kappa(Y_2, Y_1)$*
4. *si Y_1 y Y_2 son independientes entonces $\kappa(Y_1, Y_2) = 0$*
5. *$\kappa(-Y_1, Y_2) = \kappa(Y_1, -Y_2) = -\kappa(Y_1, Y_2)$*
6. *si dos parejas aleatorias están representadas por las cópulas C_1 y C_2 de manera tal que $C_1 \prec C_2$, y si κ_i denota la medición de concordancia correspondiente a la cópula C_i, donde $i = 1, 2$, entonces $\kappa_1 \leq \kappa_2$*
7. *si $\{\mathbf{Y}_n\}$ es una sucesión de parejas aleatorias continuas con cópula C_n y medida de concordancia κ_n y si $\{C_n\}$ converge a C cuya medida de concordancia es κ, entonces $\lim_{n\to\infty} \kappa_n = \kappa$.*

A continuación se describen dos medidas de asociación importantes en la literatura estadística, las cuales satisfacen la definición de concordancia.

3.2.1. La τ de Kendall

Axioma 4. *Sean (X_1, Y_1) y (X_2, Y_2) vectores aleatorios independientes e idénticamente distribuidos, tales que $(X_i, Y_i) \sim F$, $i = 1, 2$. Entonces, la τ de Kendall se define como*

$$\begin{aligned}
\tau &= \mathrm{E}\Big[\mathrm{signo}\big\{(X_1 - X_2)\,(Y_1 - Y_2)\big\}\Big] \\
&= \Pr\big\{(X_1 - X_2)(Y_1 - Y_2) > 0\big\} - \Pr\big\{(X_1 - X_2)(Y_1 - Y_2) < 0\big\} \\
&= \Pr\{\textit{concordancia}\} - \Pr\{\textit{discordancia}\} \\
&= 2\Pr\big\{(X_1 - X_2)\,(Y_1 - Y_2) > 0\big\} - 1 \\
&= 2\Pr\{\textit{concordancia}\} - 1
\end{aligned}$$

La versión de la τ de Kendall de las entradas de una pareja aleatoria continua (Y_1, Y_2), dada en términos de la cópula C, puede expresarse como

$$\tau = 4 \int\int_{\mathbf{I}^2} C(v_1, v_2)\, \mathrm{d}C(v_1, v_2) - 1 \tag{7}$$

La ecuación (7) indica que la τ de Kendall está completamente determinada por la cópula y no está relacionada con las distribuciones marginales de (Y_1, Y_2).

3.2.2. La ρ de Spearman

La ρ de Spearman, al igual que la τ de Kendall, es una medida de asociación que satisface la definición de concordancia. Esta medida de asociación puede ser definida como el coeficiente de correlación de Pearson, pero no aplicado a las variables aleatorias Y_1, Y_2, sino a sus rangos $V_1 = F_1(Y_1)$ y $V_2 = F_2(Y_2)$.

Axioma 5. *Sean (X_1, Y_1), (X_2, Y_2) y (X_3, Y_3) vectores aleatorios independientes e idénticamente distribuidos, tales que $(X_i, Y_i) \sim H, i = 1, 2, 3$. La ρ de Spearman se define como*

$$\rho = 3\Big[\Pr\big\{(X_1 - X_2)(Y_1 - Y_3) \geq 0\big\} - \Pr\big\{(X_1 - X_2)(Y_1 - Y_3) < 0\big\}\Big]$$

Las variables aleatorias V_1 y V_2 son uniformes en $\mathbf{I} = [0, 1]$; además $\mathrm{E}(V_1) = \mathrm{E}(V_2) = 1/2$ y $\mathrm{Var}(V_1) = \mathrm{Var}(V_2) = 1/12$. Si C es la función de distribución conjunta de U y V, como se especifica en la cópula, entonces se tiene que

$$\begin{aligned} \rho &= \frac{\mathrm{E}\left[V_1 V_2\right] - \mathrm{E}\left[V_1\right]\mathrm{E}\left[V_2\right]}{\sqrt{\mathrm{Var}[V_1]}\sqrt{\mathrm{Var}[V_2]}} = \frac{\mathrm{E}\left[V_1 V_2\right] - 1/4}{1/12} \\ &= 12\mathrm{E}\left(V_1 V_2\right) - 3 = 12\int\int_{\mathbf{I}^2} v_1 v_2 \, dC - 3 \\ &= 12\int\int_{\mathbf{I}^2} \left[C\left(v_1, v_2\right) - v_1 v_2\right] dv_1 dv_2 \end{aligned} \tag{8}$$

3.3. Medidas de dependencia

Un inconveniente de la definición dada en la sección 3 sobre la concordancia es que la cuarta propiedad indica que si las dos variables aleatorias son independientes, entonces la medida es igual a cero, pero no viceversa. Rényi (1959) estableció un marco axiomático que, además de considerar la situación mencionada, formaliza el concepto de medida de dependencia. A continuación se define dicha medida usando la colección de condiciones relevantes de los axiomas de Rényi.

Axioma 6. *Una medida numérica δ de dos variables aleatorias continuas Y_1 y Y_2 es una* medida de dependencia *si satisface las siguientes condiciones:*

1. *δ está definida para cualquier pareja aleatoria $\mathbf{Y} = (Y_1, Y_2)$*
2. *$\delta(Y_1, Y_2) = \delta(Y_2, Y_1)$*
3. *$\delta \in [0, 1]$*
4. *$\delta = 0$ si y solo si Y_1 y Y_2 son independientes*
5. *$\delta = 1$ si y solo si la variable aleatoria Y_{3-i} es una función estrictamente monótona de Y_i casi seguramente, para $i = 1, 2$*
6. *Si f y g son funciones estrictamente monótonas sobre el* rango Y_1 *y* rango Y_2*, respectivamente, casi seguramente, entonces $\delta\big[f(Y_1), g(Y_2)\big] = \delta(Y_1, Y_2)$*

7. *Si las parejas aleatorias* $\mathbf{Y}$ *y* $\mathbf{Y}_n$, $n = 1, \ldots$, *tienen funciones de distribuciones conjuntas* H *y* H_n *respectivamente, y si la sucesión* $\{H_n\}$ *converge débilmente a* H, *entonces* $\lim_{n\to\infty} \delta(\mathbf{Y}_n) = \delta(\mathbf{Y})$.

Es posible demostrar que el valor absoluto de la medida de concordancia de Speraman, $|\rho|$, satisface las condiciones 1 a 7 mencionadas arriba con la importante excepción del punto 4. El mismo resultado es aplicable para el valor absoluto de la medida de concordancia de Kendall, $|\tau|$. El valor absoluto del coeficiente de correlación de Pearson satisface las condiciones 1, 2 y 3; las condiciones 5 y 6 las satisface si y solo si las funciones f y g son lineales; y no satisface las condiciones 4 y 7. Existen varias medidas de dependencia que satisfacen las condiciones dadas. Estas medidas están principalmente basadas en la distancia de la distribución conjunta de la pareja aleatoria en cuestión y el producto de las distribuciones marginales correspondientes. Una ilustración de una medida de dependencia bivariada basada en dicha distancia se presenta a continuación.

3.3.1. La $\boldsymbol{\sigma}$ de Schweizer y Wolff

El integrando de la ecuación (8), que define a la ρ de Spearman, representa el volumen con signo entre las superficies $v_3 = C(v_1, v_2)$ y $v_3 = v_1 v_2$. Schweizer & Wolff (1981) notaron que las variables aleatorias Y_1 y Y_2 son independientes si y sólo si $C(v_1, v_2) = v_1 v_2$; entonces una medida adecuada de distancia normalizada entre $v_3 = C(v_1, v_2)$ y $v_3 = v_1 v_2$, como la norma L_1, podría resultar una medida de dependencia que satisface las siete condiciones dadas en la definición de medida de dependencia. Esta medida, se conoce como la $\boldsymbol{\sigma}$ de Schweizer y Wolff, y se define como

$$\boldsymbol{\sigma} = 12 \int\int_{\mathbf{I}^2} |C(v_1, v_2) - v_1 v_2| \, dv_1 dv_2 \tag{9}$$

3.4. Dependencia en las colas

Diversas cópulas pueden caracterizar de manera diferente la dependencia en el centro de una distribución. De igual forma, hay muchas situaciones en las que se requiere cuantificar la estructura de la dependencia asintótica de datos bivariados para estimar las probabilidades de eventos raros. La siguiente definición es un ejemplo de estas medidas (ver *e.g.* Embrechts et al. 2002, Charpentier & Juri 2006), y las referencias citadas allí.

Axioma 7. *Sean* Y_1 *y* Y_2 *variables aleatorias continuas con función de distribución conjunta* F, *cópula* C *y marginales* F_1 *y* F_2. *El coeficiente de dependencia de cola superior de* Y_1 *y* Y_2 *está dado por*

$$\begin{aligned}
\lambda_u &= \lim_{u\to 1^-} \Pr\big\{Y_2 > F_2^{-1}(u) \mid Y_1 > F_1^{-1}(u)\big\} \\
&= \lim_{u\to 1^-} \Pr\big\{Y_1 > F_1^{-1}(u) \mid Y_2 > F_2^{-1}(u)\big\} \\
&= \lim_{u\to 1^-} \frac{1 - 2u + C(u, u)}{1 - u}
\end{aligned}$$

siempre y cuando el límite λ_u exista.

El coeficiente λ_u mide la dependencia entre valores extremos de Y_1 y Y_2; en particular, la dependencia de cola superior contesta la pregunta: "suponiendo que un valor extremo impacta a Y_1, ¿cuál es la probabilidad de que un valor extremo impacte a Y_2?". Si $\lambda_u = 0$, entonces entre Y_1 y Y_2 no existe dependencia de cola superior; de otra forma, Y_1 y Y_2 son dependientes en la cola superior. En forma similar, el coeficiente de la cola inferior se define como $\lambda_l = \lim_{v\to 0^+} C(v,v)/v$.

4. Familias de cópulas

Si se tiene una colección de cópulas, entonces puede emplearse el Teorema de Sklar para construir distribuciones bivariadas con marginales arbitrarias. Una cantidad importante de familias de cópulas puede ser encontrada en la literatura. Estas familias no son equivalentes en términos del tipo de dependencia estocástica que ellas representan o el grado de dependencia que ellas pueden capturar. Como resultado, un problema esencial es la elección de la familia de cópulas para construir una distribución bivariada particular. Aunque se puede usar gran variedad de familias de cópulas para modelar dependencia, en este artículo se adoptan la familia Morgenstern por su simplicidad, la familia Gaussiana por su semejanza con la distribución Gaussiana bivariada, y la familia Arquimediana por ser convenientemente representada a través de una función univariada. A continuación se describen estas tres familias de cópulas. Una revisión más completa de estas y otras familias puede encontrarse en los textos de Joe (1997) y de Nelsen (1999).

4.1. La cópula de Morgenstern

La cópula de Morgenstern es representada por

$$C_\alpha\left(v_1, v_2\right) = v_1 v_2\big[1 + \alpha\left(1 - v_1\right)\left(1 - v_2\right)\big], \qquad \alpha \in [-1, 1]$$

La función de densidad de la cópula correspondiente es

$$c_\alpha(v_1, v_2) = \big[1 + \alpha\left(1 - 2v_1\right)(1 - 2v_2)\big]$$

La τ de Kendall correspondiente está dada por $\tau = 2\alpha/9$, por lo que $\tau \in [-2/9, 2/9]$, lo cual indica que esta cópula tiene un rango de concordancia muy limitado y entonces su aplicabilidad es útil para parejas aleatorias las cuales están asociadas modestamente. De hecho, esta cópula se obtiene a partir de una perturbación de la cópula de independencia dada por $C(v_1, v_2) = v_1\, v_2$ (ver *e.g.* Joe 1997). Esta cópula no exhibe dependencia de cola superior y tampoco inferior, *i.e.* $\lambda_u = \lambda_l = 0$.

4.2. La familia Gaussiana

La función de distribución bivariada que pertenece a la cópula Gaussiana tiene la forma

$$C_r(v_1, v_2) = \Phi_2\big[\Phi^{-1}(v_1), \Phi^{-1}(v_2)\big], \qquad (v_1, v_2)^T \in (0,1)^2 \tag{10}$$

donde $\Phi_2(\cdot,\cdot)$ es la función de distribución conjunta de una Gaussiana bivariada con media $(0,0)^T$ y matriz de covarianza $\mathbf{R}$, igual a una matriz no singular de 2×2 cuyos elementos fuera de la diagonal son cada uno iguales a r, con $r \in (-1,1)$, y los elementos en la diagonal son iguales a uno, y $\Phi^{-1}(\cdot)$ es la función inversa de la distribución acumulada Gaussiana estándar. La función de densidad de la cópula está dada por:

$$c_r(v_1, v_2) = \frac{\phi_2\big[\Phi^{-1}(v_1), \Phi^{-1}(v_2)\big]}{\phi\big[\Phi^{-1}(v_1)\big] \times \phi\big[\Phi^{-1}(v_2)\big]} \tag{11}$$

donde Φ y ϕ denotan, respectivamente, las funciones de distribución y densidad de la Gaussiana estándar univariada, y ϕ_2 denota la densidad bivariada de la Gaussiana definida por:

$$\phi_2(\mathbf{z}) = (2\pi)^{-1}|\mathbf{R}|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{z}^T\mathbf{R}^{-1}\mathbf{z}\right\}, \qquad \mathbf{z}^T \in \mathbb{R}^2$$

Note que si Y_1 y Y_2 están normalmente distribuidas, la función de densidad conjunta resultante generada con la cópula Gaussiana se reduce a la usual función de densidad normal bivariada.

La τ de Kendall y la ρ de Spearman para la cópula Normal están dadas respectivamente por:

$$\tau_r = \frac{2}{\pi}\operatorname{arcsen}(r) \qquad \text{y} \qquad \rho_r = \frac{6}{\pi}\operatorname{arcsen}\left(\frac{r}{2}\right)$$

Conforme a la ecuación (5), se tiene que la función de densidad de $Y_2 \mid Y_1$ correspondiente a la cópula Gaussiana está dada por

$$f_{2|1}(y_2 \mid y_1) = f_2(y_2) \times \frac{\phi_2\left[\Phi^{-1}\left(F_1(y_1)\right), \Phi^{-1}\left(F_2(y_2)\right)\right]}{\phi\big[\Phi^{-1}\left(F_1(y_1)\right)\big] \times \phi\big[\Phi^{-1}\left(F_2(y_2)\right)\big]}$$

Después de algo de álgebra laboriosa, se puede demostrar que

$$f_{2|1}(y_2 \mid y_1) = \frac{f_2(y_2)}{\sqrt{1-r^2}} \times \exp\left\{-\frac{1}{2}\left[\frac{(s_2 - r\, s_1)^2}{1-r^2} - s_2^2\right]\right\}$$

donde $s_i = \Phi^{-1}\big[(F_i(y_i)\big]$ para $i = 1, 2$.

El uso de la cópula Gaussiana bivariada es atractivo ya que codifica la dependencia en la misma forma en que la distribución normal bivariada lo hace usando el parámetro de dependencia r, con la diferencia de que se calcula para variables aleatorias con cualesquiera marginales arbitrarias. Esta cópula tiene la capacidad

de capturar el rango completo de dependencia, ya que incluye tanto las cópulas de cota superior e inferior de Fréchet como el modelo de independencia. Este último caso se obtiene cuando $r = 0$ y define la cópula de independencia. Una desventaja de esta cópula es que para $\rho < 1$, la dependencia en las colas es nula, *i.e.* $\lambda_u = \lambda_l = 0$; cuando $\rho = 1$, entonces $\lambda_u = \lambda_l = 1$. Para mayor información sobre la cópula Gaussiana multivariada, el lector puede referirse a los artículos de Clemen & Reilly (1999) y de Song (2000).

4.3. Cópulas Arquimedianas

Como se ha mencionado, las cópulas proveen una estructura general para modelar distribuciones bivariadas. Una familia de cópulas que permite este modelado a través de una sola función univariada es la Arquimediana. A continuación se enuncia la definición de cópula Arquimediana dada por Genest & Rivest (1993).

Axioma 8. *Una cópula es llamada Arquimediana si esta puede ser expresada en la forma*

$$C(v_1, v_2) = z^{-1}\big[z(v_1) + z(v_2)\big], \qquad (v_1, v_2)^T \in (0,1)^2 \tag{12}$$

para alguna función convexa decreciente z definida en $(0, 1]$ que satisface $z^{-1}(1) = 0$; por convención $z^{-1}(v) = 0$ cuando $v \geq z(0)$.

Las condiciones dadas en la definición de cópula arquimediana son necesarias y suficientes para que la cópula en la ecuación (12) sea una función de distribución bivariada (Schweizer & Sklar 1983). Estas condiciones equivalen a que $1 - z^{-1}(v)$ debe ser una función de distribución unimodal en $[0, \infty)$ con moda en 0. Aquí, la función z es conocida como el generador.

La familia de cópulas Arquimediana permite la definición de modelos generados por la transformada de Laplace $z(s) = \mathrm{E}\big[\exp(-sW)\big] = \int_0^\infty e^{-\theta t}\,\mathrm{d}H(\theta)$, correspondiente a una variable aleatoria W, la tan citada variable *frailty*, que es una variable aleatoria no negativa cuya función de distribución es H. En el contexto de variables aleatorias positivas continuas, Oakes (1989) demostró que para la clase de funciones de distribución conjunta definidas por la cópula en la ecuación (12), las variables aleatorias Y_1 y Y_2 son condicionalmente independientes dada la variable aleatoria W, por lo que $F(y_1, y_2 \mid W = w) = F_1(y_1 \mid W = w)F_2(y_2 \mid W = w)$.

Para encontrar la función de densidad de la cópula $c(v_1, v_2)$ definida en la ecuación (4), defínase $z(C) = z(v_1) + z(v_2)$, y derivando respecto a v_1,

$$z'(C)\,\frac{\partial C}{\partial v_1} = z'(v_1)$$

Derivando esta expresión respecto a v_2, se tiene que:

$$z''(C)\frac{\partial C}{\partial v_1}\frac{\partial C}{\partial v_2} + z'(C)\frac{\partial^2 C}{\partial v_1 \partial v_2} = 0$$

Por tanto

$$c(v_1, v_2) = -\frac{z''(C)z'(v_1)z'(v_2)}{\big[z'(C)\big]^3}$$

Dado (Y_1, Y_2) un par de variables aleatorias con distribución F definida en (12), se puede demostrar que la τ de Kendall correspondiente tiene la siguiente forma (ver *e.g.* Genest & MacKay 1986*b*):

$$\tau = 4 \int_0^1 \frac{z(t)}{z'(t)}\, dt + 1$$

De esta forma, el valor de τ está relacionado en forma lineal con el área bajo la curva de $z(t)/z'(t)$ entre 0 y 1. Note que la cota inferior de Fréchet correspondiente es $z(t)/z'(t) = t - 1$; cuando $z(t)/z'(t)$ tiende a cero, se obtiene la cota superior de Fréchet. De hecho, Genest & MacKay (1986*a*) notan que la convergencia de una sucesión de distribuciones bivariadas construidas con cópulas arquimedianas puede determinarse al observar la gráfica de $z(t)/z'(t)$ de la sucesión.

La familia de cópulas Arquimediana ha recibido la atención de la comunidad estadística debido a que su representación y otras cantidades importantes están dadas en términos de la función $z(s)$. En aplicaciones a finanzas, por mencionar un ejemplo, esta familia es ampliamente usada, pues varios modelos de cópula pueden ser implementados y comparados a través de funciones dadas en términos de $z(s)$ (ver *e.g.* Whelan 2004). A continuación se describen dos cópulas de la familia Arquimediana que sobresalen en la literatura.

4.3.1. La cópula de Frank

La cópula de Frank, cuyo generador es $z(t) = -\log\big[(\mathrm{e}^{-\theta t} - 1)/(\mathrm{e}^{-\theta} - 1)\big]$, está definida por

$$C_\theta(v_1, v_2) = -\frac{1}{\theta} \log\left(1 + \frac{\left(e^{-\theta v_1} - 1\right)\left(e^{-\theta v_2} - 1\right)}{(e^{-\theta} - 1)}\right), \qquad \theta \in \mathbb{R} - \{0\}$$

El uso de la cópula de Frank es atractivo ya que puede capturar el rango completo de dependencia; esto es, al igual que la cópula Gaussiana, la cópula de Frank incluye las cópulas de cota superior de Fréchet cuando $\theta \to -\infty$, de cota inferior de Fréchet cuando $\theta \to \infty$, y de independencia cuando $\theta \to 0$. De hecho, cuando se trata de inferencia, algunos estadísticos prefieren usar la cópula de Frank a la Gaussiana, ya que mientras ambas cópulas tienen propiedades similares (ver *e.g.* Carriere 1995), la cópula de Frank proporciona cantidades cerradas y, por tanto, más fáciles de programar (ver *e.g.* Escarela & Carriere 2003). El uso de la cópula de Frank no es recomendable para modelar dependencia de eventos extremos pues no es dependiente en las colas superior ni inferior.

Para evaluar el grado de asociación entre las marginales en el modelo generado por la cópula de Frank, la τ de Kendall correspondiente está dada por:

$$\tau_\theta = 1 - \frac{4}{\theta}\left(1 - \frac{1}{\theta}\int_0^\theta \frac{t}{e^t - 1}\, dt\right)$$

La integral en esta expresión no tiene solución analítica; sin embargo, es posible usar métodos numéricos, como el de cuadraturas Gauss-Kronrod, que pueden dar

buenas aproximaciones. La τ de Kendall de la cópula de Frank toma valores en el rango completo de concordancia. Observando los casos especiales de la cópula de Frank, se puede comprobar que $\lim_{\theta\to-\infty}\tau_\theta = -1$, $\lim_{\theta\to\infty}\tau_\theta = 1$ y $\lim_{\theta\to 0}\tau_\theta = 0$.

4.3.2. La cópula positiva estable

La función de distribución bivariada presentada por Hougaard (1986) toma la forma

$$C_\theta(v_1, v_2) = \exp\left\{-\left[(-\log v_1)^{1/\theta} + (-\log v_2)^{1/\theta}\right]^\theta\right\}, \qquad \theta \in (0,1) \qquad (13)$$

El generador de esta cópula es $z(t) = (-\log t)^{1/\theta}$ y la densidad de cópula correspondiente está dada por

$$c_\theta(v_1, v_2) = \frac{1}{C_\theta}\,\frac{\partial C_\theta}{\partial v_1}\,\frac{\partial C_\theta}{\partial v_2}\left[(\theta^{-1}-1)(-\log C_\theta)^{-1}+1\right]$$

donde

$$\frac{\partial C_\theta}{\partial v_j} = \left(\frac{\log v_j}{\log C_\theta}\right)^{\frac{1}{\theta}-1}\frac{C_\theta}{v_j}, \qquad j = 1, 2$$

La cópula positiva estable exhibe una asimetría porque hay un cluster de valores hacia la cola derecha pero con colas poco pesadas. Esta cópula es útil para modelar variables aleatorias asociadas en forma positiva (ver *e.g.* Nelsen 1999, Joe 1997); valores pequeños de θ proveen dependencia positiva alta entre Y_1 y Y_2, i.e. $\lim_{\theta\to 1} C_\theta(v_1, v_2) = M(v_1, v_2)$, mientras que valores grandes de θ proveen asociaciones cercanas a independencia, *i.e.* $\lim_{\theta\to 1} C_\theta(v_1, v_2) = v_1 v_2$.

En lo que respecta a las medidas de asociación para esta familia de cópulas, la τ de Kendall es definida por $\tau_\theta = 1 - \theta$, mientras que la ρ de Sperman no tiene forma cerrada. Nótese que el rango de la τ de Kendall se encuentra en el intervalo $(0, 1)$, por lo que la cópula positiva estable solo considera concordancias positivas. Esta cópula exhibe dependencia asimétrica en las colas con dependencia nula en la cola inferior $\lambda_l = 0$, y dependencia en la cola superior $\lambda_u = 2 - 2^\theta$. A la cópula positiva estable también se le conoce como *la cópula de valor extremo*, ya que provee modelos flexibles para parejas aleatorias cuyas entradas representan máximos de series de tiempo estacionarias (ver *e.g.* Dupuis 2005).

4.4. Selección y comparación de cópulas

4.4.1. Sobre la selección de cópulas

Escoger una cópula para ajustar un conjunto de datos dado es un problema importante pero difícil; de hecho, no hay un método que la comunidad estadística use rutinariamente. En los últimos años se han propuesto varios métodos para seleccionar una cópula particular. Ané & Kharoubi (2003) muestran un método de selección basado en comparaciones paramétricas y no paramétricas a través de un estimador de distancia. Huard et al. (2006) proponen un método bayesiano basado

en la τ de Kendall para seleccionar la cópula más probable de las dadas en un conjunto. Dobric & Schmid (2007) presentan una prueba de bondad de ajuste basada en la transformación de Rosenblatt; cuando las marginales son especificadas, la prueba funciona bien, pero cuando éstas son estimadas empíricamente, la prueba no es tan útil. Genest & Rivest (1993) propusieron un procedimiento no paramétrico que sin tomar en cuenta a las marginales estima la función que determina a una cópula arquimediana. Este procedimiento representa una estrategia para seleccionar la familia paramétrica de cópulas arquimedianas que provee el mejor ajuste posible para un conjunto de datos apareados. Una técnica de selección de cópulas arquimedianas similar, pero para modelar funciones de supervivencia en presencia de datos con censura, fue propuesta por Wang & Wells (2000).

En general, la selección de una cópula y de las marginales, en particular, depende de la aplicación que se quiera dar. Un juicio informado que incluye rangos de asociación, distribuciones marginales útiles para el problema en cuestión y grados de dependencia en colas puede mejorar el modelado. Otra forma de evaluar si conviene un modelo particular es usar un análisis de residuales. Por ejemplo, para verificar las suposiciones de los modelos AR(1) de la ilustración en la siguiente sección, donde se usa un marco de máxima verosimilitud, se pueden usar los residuales propuestos por Dunn & Smyth (1996), los cuales pueden ser graficados de varias formas, y decidir visualmente si el ajuste es adecuado. La selección de la cópula y la evaluación de la bondad de ajuste son temas actuales de investigación.

4.4.2. Representación gráfica de las cópulas

Las densidades de las cópulas pueden ser graficadas con diagramas de superficie para ver la diferencia entre las diversas cópulas; sin embargo, para tener una mejor visión de lo que hacen las cópulas, en la práctica es útil observar las funciones de densidad correspondientes a una función de distribución o función de supervivencia construida por una cópula con un grado de asociación predeterminado y dos marginales dadas en la literatura.

La figura 1 muestra los contornos de las funciones de densidad correspondientes a la función de supervivencia bivariada definida por la ecuación (6), usando las cópulas Gaussiana, de Frank y positiva estable para asociaciones débil ($\tau = 0.1$), moderada ($\tau = 0.4$) y fuerte ($\tau = 0.7$), con funciones de supervivencia marginales Weibull parametrizadas como

$$S_j(y_j) = \exp\Big\{-\big(b_j y_j\big)^{a_j}\Big\}, \qquad j = 1, 2$$

donde $b_1 = 0.028$, $a_1 = 2$, $b_2 = 0.039$ y $a_2 = 1.5$.

Observando los contornos en la figura 1, es evidente que los parámetros a_j y b_j solo controlan las transformaciones de escala y potencia. El efecto del parámetro de dependencia es más influyente y puede resumirse de igual forma para las tres familias de cópulas. Cuando el grado de asociación es bajo, los tres contornos tienen formas similares. En general, cuando el grado de asociación se incrementa, los contornos son atraídos al origen y concentrados alrededor de la línea $y_1 - y_2 =$ constante. La diferencia principal entre los contornos de las distintas cópulas

graficadas en la figura 1 es la forma del achatamiento de la densidad cuando el valor de τ se incrementa. En particular, es posible notar que el contorno de la cópula positiva estable para dependencia alta tiende a ser más angosto para valores grandes de y_1 y y_2; esta característica puede atribuirse a la propiedad de dependencia en la cola superior que posee esta cópula.

5. Ilustración: modelado de cadenas de Markov

5.1. El contexto de una cadena de Markov de orden 1

Un proceso de Markov estacionario de primer orden a tiempo discreto cuyo espacio de estados es continuo puede construirse a partir de una distribución bivariada dada $F(y_1, y_2) = \Pr\{Y_1 \leq y_1, Y_2 \leq y_2\}$, la cual corresponde a un vector aleatorio conjuntamente continuo (Y_1, Y_2) con ambas distribuciones marginales univariadas iguales a la distribución estacionaria. La distribución de transición, definida como $F_{2|1}(y_2 \mid y_1) = \Pr\{Y_2 \leq y_2 \mid Y_1 = y_1\}$, puede calcularse como (ver *e.g.* Schäbe 1997):

$$F_{2|1}(y_2 \mid y_1) = \frac{\partial F(y_1, y_2)}{\partial y_1} \Big/ \frac{\partial F(y_1, \infty)}{\partial y_1}$$

donde $F(y_1, \infty)$ denota la función de distribución marginal de Y_1. Si el espacio de estados del proceso es finito o un conjunto contable, la distribución de transición está dada por (ver *e.g.* Joe 1997):

$$F_{2|1}(y_2 \mid y_1) = \sum_{z \leq y_2} \frac{f(y_1, z)}{f_1(y_1)}$$

donde $f(y_1, y_2)$ es una función de probabilidad conjunta con ambas funciones de probabilidad marginales igual a $f_1(\cdot)$.

Un ejemplo importante es el caso de una sucesión de variables aleatorias normales cuya respuesta actual depende del valor de la variable aleatoria inmediata anterior, como se describe a continuación. Considere la serie de tiempo estacionaria $\{Y_t, t = 1, 2, \ldots\}$ con respuestas marginales $Y_t \sim \mathrm{N}\Big(\boldsymbol{\beta}^T \mathbf{x}_t, \sigma^2\Big)$, para $t = 1, 2, \ldots$, entonces $\boldsymbol{\beta}^T \mathbf{x}_t$ es la esperanza marginal de Y_t, $\mathbf{x}_t$ es un vector de variables explicativas en el tiempo t, $\boldsymbol{\beta}$ es el vector de coeficientes de regresión y σ^2 es la varianza marginal de las respuestas. Si la correlación entre las respuestas adyacentes Y_{t-1} y Y_t es r, el modelo de transición tiene la siguiente especificación:

$$Y_t \mid Y_{t-1} \sim \mathrm{N}\Big(\boldsymbol{\beta}^T \mathbf{x}_t + r\big[Y_{t-1} - \boldsymbol{\beta}^T \mathbf{x}_{t-1}\big], \nu^2\Big) \tag{14}$$

donde $\nu^2 = \sigma^2\big(1 - r^2\big)$, y $|r| < 1$.

Usando la función de densidad condicional de la ecuación (5) en términos de una función de densidad de cópula y una marginal dada, $Y_t \sim f$, se puede construir un modelo de transición para respuestas continuas en una forma similar al

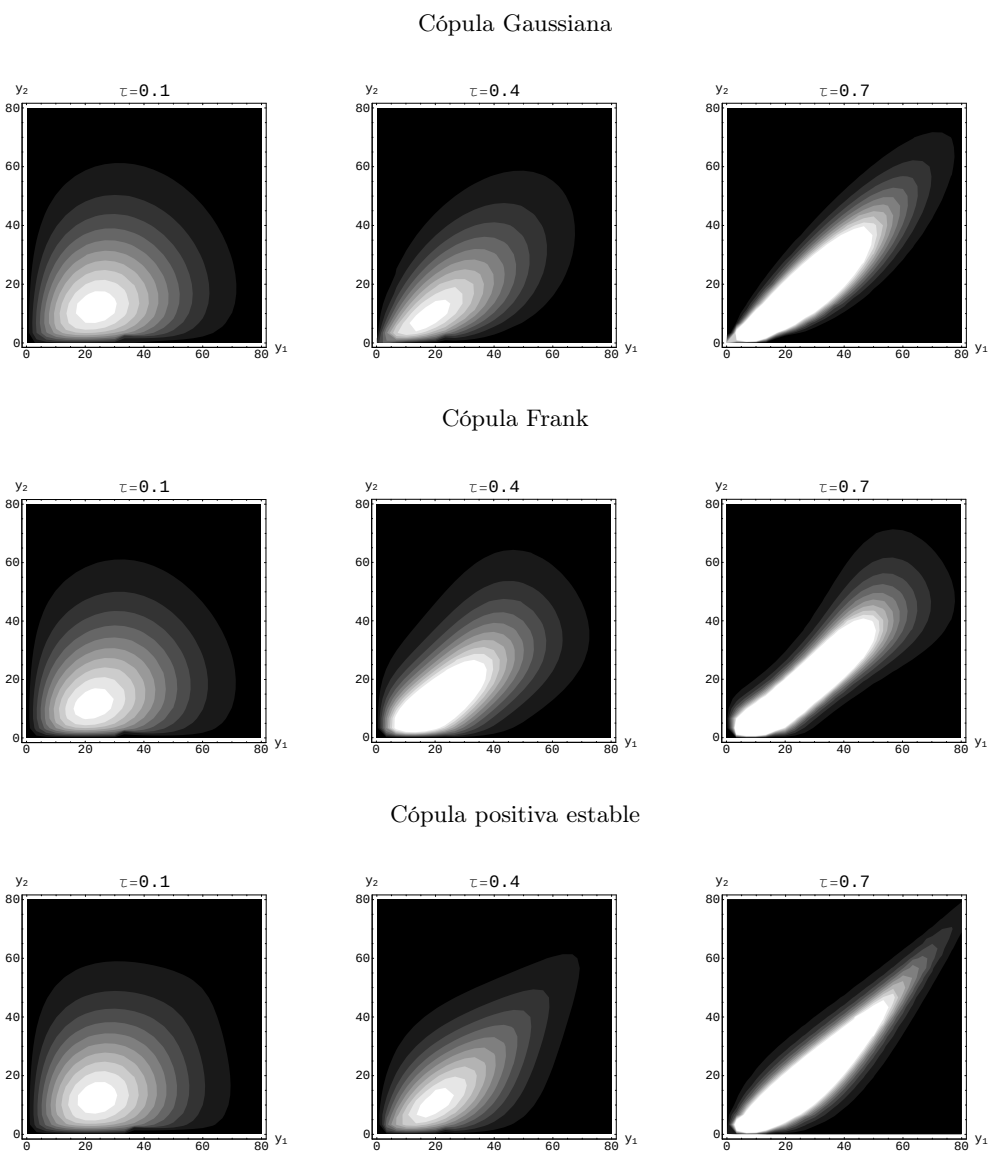

FIGURA 1: Contornos de la función de densidad conjunta resultante cuando se utilizan las cópulas Gaussiana, de Frank y positiva estable para marginales Weibull con parámetros $b_1 = 0.028$, $a_1 = 2$, $b_2 = 0.039$, $a_2 = 1.5$ para asociaciones baja ($\tau = 0.1$), moderada ($\tau = 0.4$) y alta ($\tau = 0.7$).

modelo de transición Gaussiano dado por la ecuación (14). Por ejemplo, si se desea obtener el modelo autoregresivo de primer orden en términos del modelo de transición de cópula, simplemente se necesita proveer la función de densidad

de la cópula Gaussiana $c_\rho(\cdot,\cdot)$ dada en la ecuación (11), la función de distribución marginal estandarizada $F(y_t) = \Phi\Big[\Big(y_t - \boldsymbol{\beta}^T\mathbf{x}_t\Big)\Big/\sigma\Big]$ y la densidad marginal $f(y_t) = \big(2\pi\sigma^2\big)^{-1/2} \times \exp\Big\{-\big(y_t - \boldsymbol{\beta}^T\mathbf{x}_t\big)^2\Big/\big(2\sigma^2\big)\Big\}$.

Cuando los modelos de transición son empleados para respuestas discretas, se puede definir la distribución de transición de $Y_t \mid Y_{t-1}$ al definir una función de probabilidad condicional cuando la distribución de Y_t está dada. Si $f(y_t) = \Pr\{Y_t = y_t\}$ representa la función de probabilidad marginal de Y_t, la familia de distribuciones de transición de $\{Y_t\}$ se puede caracterizar usando la cópula bivariada discretizada y la función de distribución discreta $F(y_t) = \sum_{z\le y_t} f(z)$, como se muestra a continuación:

$$F_{2|1}(y_t \mid y_2) = \Big\{C\big[F(y_2), F(y_t)\big] - C\big[F(y_2-1), F(y_t)\big]\Big\}\Big/ f(y_2) \qquad (15)$$

Se tiene, en consecuencia, que la función de densidad de transición, la cual se define como $f_{2|1}(y_t \mid y_2) = \Pr\{Y_t = y_t \mid Y_2 = y_2\}$, está dada por

$$\begin{aligned} f_{2|1}(y_t \mid y_2) = \Big\{&C\big[F(y_t), F(y_2)\big] - C\big[F(y_t), F(y_2-1)\big] - \\ &C\big[F(y_t-1), F(y_2)\big] + C\big[F(y_t-1), F(y_2-1)\big]\Big\}\Big/ f(y_2) \end{aligned} \qquad (16)$$

5.2. Cadenas de Markov de valor extremo de orden 1

En disciplinas como la ciencia ambiental, el propósito principal de un estudio es analizar datos que corresponden a extremos de algún fenómeno durante varios periodos (*e.g.* Smith 1989). Muchas veces resulta poco verosímil suponer independencia entre las observaciones de la serie de tiempo resultante. Un ejercicio importante en este tipo de aplicaciones es determinar qué tan robusta es la elección de una cópula para modelar cadenas de Markov de primer orden $\{Y_t\}_{t\ge1}$, cuya distribución marginal pertenece a alguna distribución de valor extremo con varios grados de dependencia entre los valores adyacentes Y_{t-1} y Y_t. En esta ilustración se desea realizar una comparación cuantitativa de varios modelos de la distribución condicional de $Y_t \mid Y_{t-1}$ generados con las cópulas, como se estableció en la ecuación (5).

Intuitivamente, la forma más adecuada de modelar una serie de tiempo de valores extremos de primer orden sería escogiendo una cópula cuyos coeficientes de cola inferior o superior sean diferentes a cero, dependiendo si se trata de mínimos o máximos. Si los grados de asociación de las observaciones adyacentes son altos, entonces es imperativo seleccionar una familia de cópulas con coeficientes de cola positivos; sin embargo, si los grados de asociación de las observaciones adyacentes no son muy altos, es posible que la elección de la cópula no sea tan crucial. Para realizar la evaluación correspondiente, se simularon cadenas de Markov estacionarias de primer orden cuya estructura de dependencia entre Y_{t-1} y Y_t está caracterizada por la cópula positiva estable dada en la ecuación (13) y cuya distribución marginal es Gumbel, representada por la siguiente función de

distribución:

$$F(y) = \exp\left[-\exp\left\{-\left(\frac{y-\mu}{\sigma}\right)\right\}\right], \quad y \in \mathbb{R} \tag{17}$$

donde $-\infty < \mu < \infty$ y $\sigma > 0$.

Para la simulación de datos con las características que se acaban de describir se usó la función `evmc` del paquete `evd` (Stephenson 2002) del lenguaje R (Ihaka & Gentleman 1996). Por conveniencia, se fijaron $\mu = 0$ y $\sigma = 1$; además, se seleccionaron tres grados de asociación entre las respuestas adyacentes: baja ($\tau_\theta = 0.1$), moderada ($\tau_\theta = 0.4$) y alta ($\tau_\theta = 0.7$). Se generaron muestras de tamaño $n = 400$ para cada una de las estructuras de dependencia. Las tres series de tiempo resultantes se muestran en la figura 2. Es posible observar cómo se va perdiendo la aleatoriedad cuando la concordancia entre los datos adyacentes va aumentando.

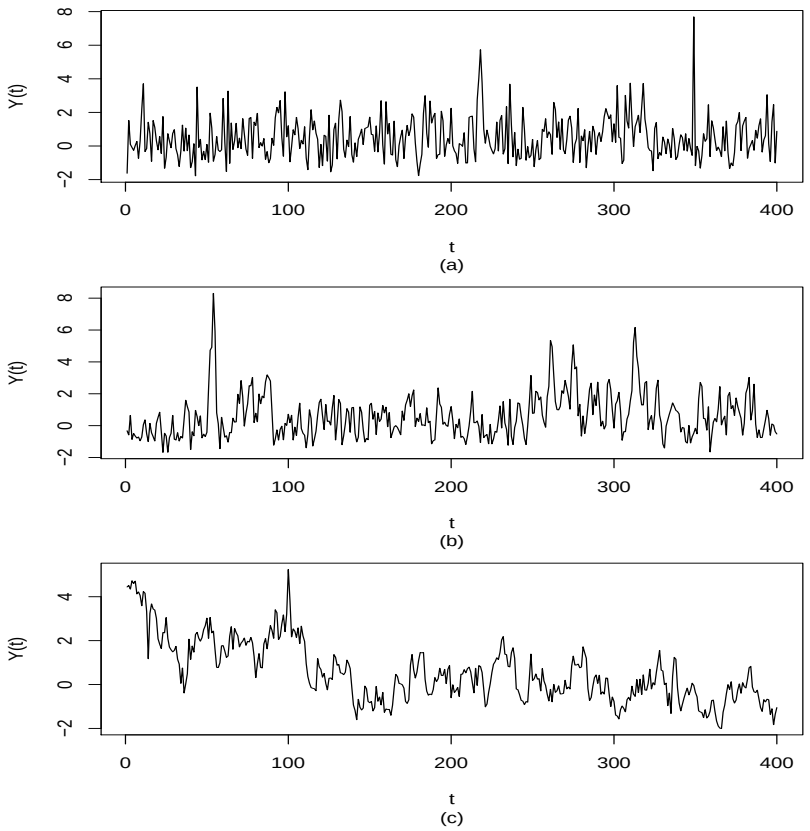

Figura 2: Cadenas de Markov de primer orden simuladas con marginal Gumbel y cópula positiva estable para grados de asociación adyacentes (a) bajo ($\tau_\theta = 0.1$), (b) moderado ($\tau_\theta = 0.4$) y (c) alto ($\tau_\theta = 0.7$).

Para construir los modelos de transición, caracterizados por la función de densidad condicional $f_{2|1}(y_t \mid y_{t-1})$ dada por la ecuación (5), se seleccionaron las cópulas positiva estable, de Frank y Gaussiana, dejando al parámetro de dependencia como desconocido; también se seleccionó la distribución Gumbel caracterizada por la ecuación (17), dejando como desconocidos los parámetros μ y σ. Para ajustar los modelos resultantes, se usó la técnica de máxima verosimilitud; en este caso la función de verosimilitud es $L = f(y_1)\prod_{k=2}^{n} f_{2|1}(y_k \mid y_{k-1})$, donde $f(y)$ es la función de densidad correspondiente a la distribución marginal. La programación correspondiente se realizó en el lenguaje R; para definir cada modelo de transición se usó el paquete `copula` (Jan 2007), y para obtener los estimadores de máxima verosimilitud se usó la función `optim` de R para optimizar la función log verosimilitud.

El ajuste de cada modelo puede ser evaluado mediante la comparación de las funciones de densidad condicionales correspondientes a $Y_t \mid Y_{t-1} = y_p$, donde $y_p = -\log(-\log p)$ es el p-ésimo percentil de la distribución marginal F; *i.e.* y_p satisface $F(y_p) = p$ para $0 \leq p \leq 1$. En este estudio se seleccionaron densidades condicionales correspondientes a p =0.05, p =0.5 y p =0.95. La figura 3 muestra las densidades condicionales resultantes. Es posible observar que cuando los datos muestran un grado de dependencia adyacente modesto, no hay mucha diferencia entre la densidad condicional verdadera y el ajuste de los modelos de las cópulas positiva estable y de Frank, lo cual sugiere que la elección entre estas dos cópulas es robusta. Cuando se trata de datos con dependencia moderada, la cópula positiva estable siempre observa un buen ajuste, mientras que la cópula Gaussiana tiene un ajuste razonable únicamente cuando $p = 0.05$; para valores de p más grandes, las cópulas de Frank y Gaussiana ofrecen un ajuste malo. En presencia de una dependencia alta, solo la cópula positiva estable tiene un ajuste bueno, lo cual sugiere que la elección de la cópula es bastante crucial.

Note que el ajuste de la cópula Gaussiana no aparece para los datos de dependencia baja y alta. Esto se debe a que la función de cópula que se usó presentaba varios problemas numéricos cuando se trataba de optimizar la función log verosimilitud. Estos problemas son formulados por Jan (2007) y no parecen tener una solución trivial.

5.3. Un modelo AR(1) para una serie binaria

Un modelo de transición de orden uno para una cadena de Markov de dos estados puede ser construido fácilmente al predeterminar la marginal F en las ecuaciones (15) y (16) como la distribución Bernoulli con probabilidad de éxito p, y al escoger la función cópula Gaussiana $C_r(u, v)$ con parámetro de dependencia r, como se define en la ecuación (10).

En el planteamiento de regresión, es posible modelar las funciones de probabilidad condicional de Y_t, dada Y_{t-1} con una función de la variable explicativa $\mathbf{x}_t$. Así, un modelo de transición más generalizado supone que la marginal F tiene la siguiente función de distribución Bernoulli, la cual incluye la variable explicativa

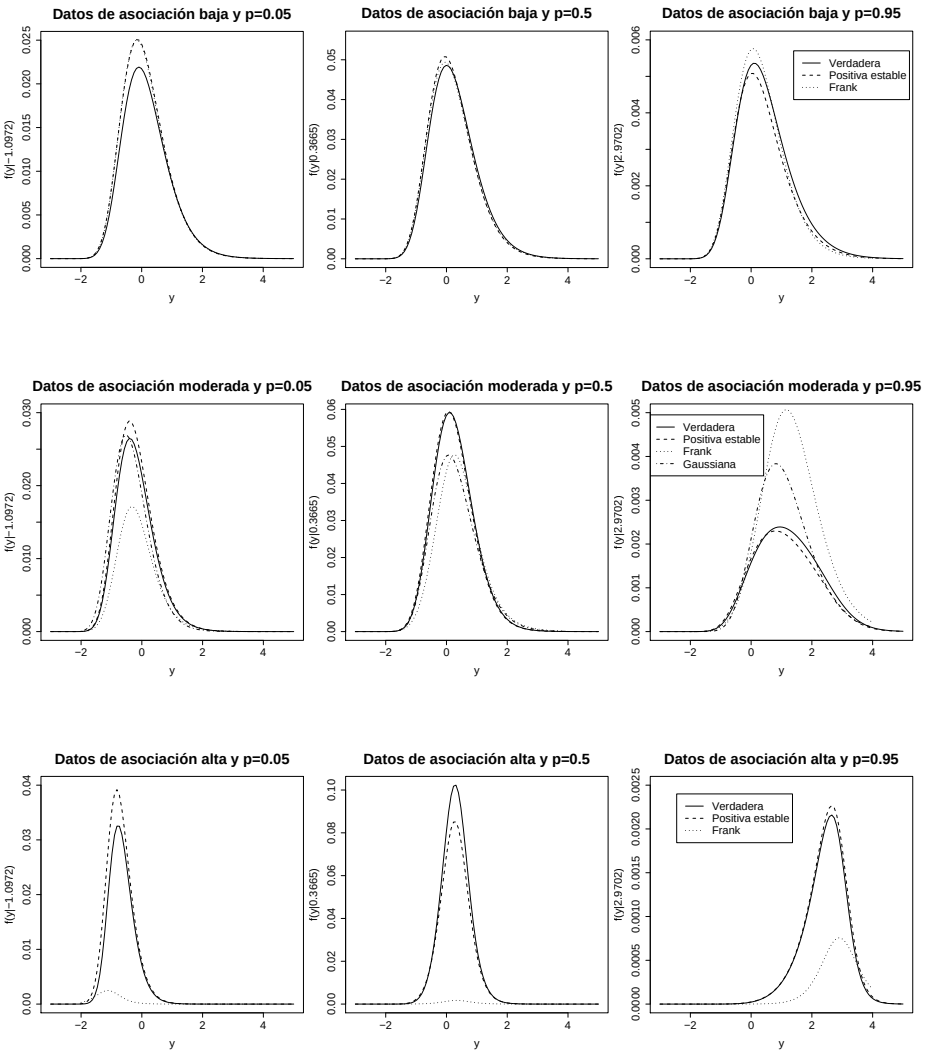

FIGURA 3: Densidades condicionadas a $Y_{t-1} = y_p$ de modelos de transición verdadero y ajustados a datos con dependencia adyacente baja, moderada y alta para distribuciones construidas con marginal Gumbel y cópulas positiva estable, de Frank y Gaussiana con, p =0.05, p =0.5 y p =0.95.

$\mathbf{x}_t$:

$$F(y_t; \mathbf{x}_t) = q(\mathbf{x}_t)\, I_{[0,1)}(y_t) + I_{[1,\infty)}(y_t), \qquad y_t \in \mathbb{R}$$

donde $q(\mathbf{x}_t) = 1 - p(\mathbf{x}_t)$, $p(\mathbf{x}_t)$ es la probabilidad de éxito dada en términos del vector de variables explicativas en el tiempo t, y $I_A(y)$ es la función indicadora de A, la cual es igual a 1 si $y \in A$ y es 0 de otra forma. Empleando las propiedades de la cópula en las ecuaciones (1) y (2), la función de probabilidad de transición

ésta dada por,

$$p_{l|m} = f_{2|1}(l \mid m) = \Pr\{Y_t = l \mid Y_{t-1} = m\}, \qquad l, m \in \{0, 1\}$$

y, usando la parametrización escogida aquí, esta puede representarse como

$$\begin{aligned}
p_{0|0} &= C_r\big[q(\mathbf{x}_t)q(\mathbf{x}_{t-1})\big] \big/ q(\mathbf{x}_{t-1}) \\
p_{0|1} &= \Big\{q(\mathbf{x}_t) - C_r\big[q(\mathbf{x}_t)q(\mathbf{x}_{t-1})\big]\Big\} \Big/ p(\mathbf{x}_{t-1}) \\
p_{1|0} &= \Big\{q(\mathbf{x}_{t-1}) - C_r\big[q(\mathbf{x}_t), q(\mathbf{x}_{t-1})\big]\Big\} \Big/ q(\mathbf{x}_{t-1}) \\
p_{1|1} &= \Big\{1 - q(\mathbf{x}_{t-1}) - q(\mathbf{x}_t) + C_r\big[q(\mathbf{x}_t)q(\mathbf{x}_{t-1})\big]\Big\} \Big/ p(\mathbf{x}_{t-1})
\end{aligned}$$

Una forma conveniente de tomar en cuenta información concomitante en este modelo es usar la liga probit en el modelo de probabilidad marginal. Esta liga implica que $p(\mathbf{x}_t) = \Phi\Big(\boldsymbol{\beta}^T\mathbf{x}_t\Big)$, donde Φ es la función de distribución acumulada normal estándar. Usando la propiedad de simetría de la distribución normal estándar, se tiene que $q(\mathbf{x}_t) = \Phi\Big(-\boldsymbol{\beta}^T\mathbf{x}_t\Big)$, y por tanto, usando la cópula Gaussiana, se obtiene que $C_r\big[q(\mathbf{x}_t), q(\mathbf{x}_{t-1})\big] = \Phi_2\Big(-\boldsymbol{\beta}^T\mathbf{x}_t, -\boldsymbol{\beta}^T\mathbf{x}_{t-1}\Big)$. Aquí, Φ_2 denota la función de distribución correspondiente a ϕ_2 cuyo parámetro de dependencia es igual a r. De esta forma, la elección de la cópula Gaussiana y de la marginal Bernoulli con liga probit provee probabilidades de transición de la serie binaria relativamente fáciles de calcular; para hacerlo, es necesario contar con una rutina que evalúe a la función de distribución Gaussiana bivariada Φ_2. Una aplicación de este modelo de transición puede encontrarse en Escarela et al. (2009).

6. Conclusiones

En la actualidad en muchos campos se ha incrementado el interés en modelar problemas con respuestas multivariadas. En este artículo se ha propuesto el uso de la función cópula para el estudio de respuestas bivariadas el cual puede extenderse al caso multivariado, ya sea para respuestas continuas o discretas. En particular, este trabajo se ha enfocado a la relación de las cópulas y problemas estadísticos aplicados. Debido a que las cópulas son familias paramétricas, el uso de la técnica de máxima verosimilitud resulta bastante atractivo para las inferencias, particularmente cuando se desea incluir variables explicativas. Desde luego, también otras herramientas estadísticas tales como las provenientes de la estadística bayesiana han sido desarrolladas para ajustar a los modelos resultantes.

Como se ha expuesto en este trabajo, las cópulas ofrecen una estructura atractiva para el modelado de parejas aleatorias debido a que permite la investigación del comportamiento marginal y de la estructura de dependencia en forma simultánea. Los autores de este artículo desean que las ideas y conceptos descritos aquí sean de utilidad para los investigadores que cuenten con problemas de dependencia y estén interesados en aplicar la técnica de la cópula, y que los paquetes junto con

las rutinas del lenguaje de acceso gratuito R puedan facilitar significativamente la programación de los modelos que se generen.

Agradecimientos

Los autores agradecen a dos árbitros y al editor por los comentarios tan útiles que ayudaron a mejorar la presentación de este trabajo. Esta investigación ha sido auspiciada por el programa ECOS-ANUIES-CONACYT. Los autores agradecen a CONACYT y PROMEP, México, por su financiamiento.

[Recibido: enero de 2008 — Aceptado: enero de 2009]

Referencias

Ané, T. & Kharoubi, C. (2003), 'Dependence Structure and Risk Measure', *The Journal of Business* **76**, 411–438.

Barlow, R. E. & Proschan, F. (1975), *Statistical Theory of Reliability and Life Testing*, Holt and Rinehart and Winston.

Carriere, J. F. (1995), 'Removing Cancer when it is Correlated with other Causes of Death', *Biometrical Journal* **37**, 339–350.

Charpentier, A. & Juri, A. (2006), 'Limiting Dependence Structures for Tail Events, with Applications to Credit Derivatives', *Journal of Applied Probability* **43**, 563–586.

Clemen, R. T. & Reilly, T. (1999), 'Correlations and Copulas for Decision and Risk Analysis', *Management Science* **45**, 208–224.

Denuit, M. & Lambert, P. (2005), 'Constraints on Concordance Measures in Bivariate Discrete Data', *Journal of Multivariate Analysis* **93**, 40–57.

Dobric, J. & Schmid, F. (2007), 'A goodness of Fit Test for Copulas Based on Rosenblatt's Transformation', *Computational Statistics & Data Analysis* **51**, 4633–4642.

Dunn, P. K. & Smyth, G. K. (1996), 'Randomized Quantile Residuals', *Journal of Computational and Graphical Statistics* **5**, 236–244.

Dupuis, D. J. (2005), 'Ozone Concentrations: A Robust Analysis of Multivariate Extremes', *Technometrics* **47**, 191–201.

Embrechts, P., McNeil, A. J. & Straumann, D. (2002), Correlation and dependence in risk management: Properties and pitfalls, *in* M. A. H. Dempster, ed., 'Risk Management: Value at Risk and Beyond', pp. 176–223.

Escarela, G. & Carriere, J. F. (2003), 'Fitting Competing Risks with an Assumed Copula', *Statistical Methods in Medical Research* **12**, 333–349.

Escarela, G., Mena, R. H. & Castillo-Morales, A. (2006), 'A Flexible Class of Parametric Transition Regression Models Based on Copulas: Application to Poliomyelitis Incidence', *Statistical Methods in Medical Research* **15**, 593–609.

Escarela, G., Pérez-Ruiz, L. C. & Bowater, R. (2009), 'A Copula-Based Markov Chain Model for the Analysis of Binary Longitudinal Data', *Journal of Applied Statistics* . En prensa.

Fréchet, M. (1951), 'Sur les Tableaux de Corrélation dont les Marges sont Donnés', *Annales de l'Université de Lyon* (14), 53–77.

Frees, E. W. & Valdez, E. A. (1998), 'Understanding Relationships Using Copulas', *North American Actuarial Journal* **2**, 1–25.

Genest, C. & Favre, A. C. (2007), 'Everything you always wanted to know about Copula Modeling but were afraid to ask', *Journal of Hydrologic Engineering* **12**, 347–368.

Genest, C. & MacKay, R. J. (1986*a*), 'Copules Archimédiennes et Familles de Lois Bidimensionnelles dont les Marges sont Donnés', *The Canadian Journal of Statistics* **14**, 145–159.

Genest, C. & MacKay, R. J. (1986*b*), 'The Joy of Copulas: Bivariate Distributions with Uniform Marginals', *The American Statistician* **40**, 280–283.

Genest, C. & Rivest, L. (1993), 'Statistical Inference Procedures for Bivariate Archimedian Copulas.', *Journal of the American Statistical Association* **88**, 1034–1043.

Hougaard, P. (1986), 'A Class of Multivariate Failure Time Distributions', *Biometrika* **73**, 671–678.

Huard, D., Évin, G. & Favre, A. C. (2006), 'Bayesian Copula Selection', *Computational Statistics & Data Analysis* **51**, 809–822.

Ihaka, R. & Gentleman, R. (1996), 'R: A Language for Data Analysis and Graphics', *Journal of Computational and Graphical Statistics* **5**, 299–314.

Jan, Y. (2007), 'Enjoy the Joy of Copulas with a Package Copula', *Journal of Statistical Software* **21**, 1–21.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman & Hall, New York, United States.

Klugman, S. & Parsa, R. (1999), 'Fitting Bivariate loss Distributions with Copulas', *Insurance: Mathematics and Economics* **24**, 139–148.

Lehmann, E. L. (1966), 'Some Concepts of Dependence', *Annals of Mathematical Statistics* **37**, 1137–1153.

Nelsen, R. (1991), Copulas and association, *in* Dall'Aglio, ed., 'Advances in Probability Distributions with Given Marginals: Beyond the Copulas', Kluwer, Dordrecht, Netherlands, pp. 51–74.

Nelsen, R. B. (1999), *An Introduction to Copulas*, Springer, New York, United States.

Oakes, D. (1989), 'Bivariate Survival Models Induced by Frailties', *Journal of the American Statistical Association* **84**, 487–493.

Rényi, A. (1959), 'On Measures of Dependence', *Acta Mathematica Academiae Scientiarum Hungaricae* **10**, 441–451.

Rodríguez-Lallena, J. A. & Úbeda Flores, M. (2004), 'A New Class of Bivariate Copulas', *Statistics & Probability Letters* **66**, 315–325.

Scarsini, M. (1984), 'On Measures of Concordance', *Stochastica* **8**, 201–218.

Schäbe, H. (1997), 'Parameter Estimation for a Special Class of Markov Chains', *Statistical Papers* **38**, 303–327.

Schweizer, B. & Sklar, A. (1983), *Probabilistic Metric Spaces*, Dover Publications, New York, United States.

Schweizer, B. & Wolff, E. F. (1981), 'On Nonparametric Measures of Dependence for Random Variables', *The Annals of Statistics* **9**, 879–885.

Sklar, A. (1959), 'Fonctions de Répartition a n Dimensions et Leurs Marges', *Publications de l'Institut Statistique de l'Université de Paris* **8**, 229–231.

Smith, R. L. (1989), 'Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone', *Statistical Science* **4**, 367–377.

Song, P. X. K. (2000), 'Multivariate Dispersion Models Generated from Gaussian Copula', *Scandinavian Journal of Statistics* **27**, 305–320.

Stephenson, A. (2002), 'Evd: Extreme Value Distributions', *R News* **2**(2), 31–32.

Wang, W. & Wells, M. (2000), 'Model Selection and Semiparametric Inference for Bivariate Failure-Time Data', *Journal of the American Statistical Association* **95**, 62–76.

Whelan, N. (2004), 'Sampling from Archimedian Copulas', *Quantitative Finance* **4**, 339–352.

Yanagimoto, T. & Okamoto, M. (1969), 'Partial Orderings of Permutations and Monoticity of a Rank Correlation Statistic', *Annals of the Institute of Statistical Mathematics* **21**, 489–506.