UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament de Matemàtica Aplicada III

Estimación Bayesiana de Cópulas extremales en Procesos de Poisson

por **Maribel Ortego Martínez**

Tesis presentada para obtener el título
de Doctora
por la Universitat Politècnica de Catalunya

Director de Tesis

Prof. Juan José Egozcue Rubí

Programa de Doctorado en Matemática Aplicada

Barcelona, Diciembre 2014

A David
A mis padres, María y Domingo

Al andar se hace camino
y al volver la vista atrás
se ve la senda que nunca
se ha de volver a pisar.
Caminante, no hay camino,
sino estelas en la mar.

Caminante, son tus huellas
el camino, y nada más;
caminante, no hay camino,
se hace camino al andar.

Antonio Machado
Proverbios y cantares (XXIX)
Campos de Castilla (1912)

Agradecimientos

La elaboración de una tesis doctoral es un largo trayecto, o muy largo, como ha sido mi caso. No es nada fácil sintetizar en unas líneas mi gratitud hacia todas las personas que me han ayudado. Galeano (1998) les llama cómplices. Aquí va mi lista.

Sin duda alguna, el principal cómplice es el Dr. Juan José Egozcue, mi director de Tesis. Quiero agradecer especialmente su apoyo y las innumerables e interesantes discusiones y cafés compartidos. Trabajador infatigable, contagia su entusiasmo por las Matemáticas, la Estadística y la Ciencia en general. Durante estos años ha confiado en mí y me ha transmitido ilusión y ánimo en las dosis necesarias. Gracias por la paciencia. Gracias, de corazón, por ser un verdadero Maestro.

Son también cómplices los Drs. José Gibergans y Raimon Tolosana. Con ellos no he trabajado en los temas tratados en la tesis, pero hemos compartido cafés, hemos arreglado el mundo y siempre me han apoyado y animado.

No sólo ellos me han apoyado. Doy también las gracias a todos los profesores del *Club de la Estrella*. Me transmitieron ilusión desde mi entrada en el mundo universitario, y tengo la suerte de disfrutar de su compañía de vez en cuando. Gracias a todos.

En un aspecto más formal, debo agradecer el soporte proporcionado por el Departamento de Matemática Aplicada III de la UPC y por el proyecto Métodos estadísticos en espacios restringidos (METRICS), MTM2012-33236, financiado por el Ministerio de Economía y Competividad. Trabajar con la *Girona Gang* en el contexto de ese proyecto ha sido muy enriquecedor.

Por supuesto, no todos los cómplices son del mundo académico. El agradecimiento a mis padres por cuanto trabajaron para que sus hijos tuvieran educación universitaria no cabe en unas líneas. Siempre lo he valorado mucho, pero ocasiones como esta merecen destacarlo. Gracias por tanto...

Y por supuesto, David. Apoyo, comprensión y mucha, mucha paciencia para aguantar mis días de *malos pelos* peleándome con los conceptos, la programación, el texto... Sorprendiéndome cada día. Y espero que vengan muchos más.

Por último, quiero agradecer a Jan (Hans) Cok y Gema Bustos el diseño de la preciosa portada de la memoria de Tesis. ¡La gente de *Esparta* es genial!.

No están todos los que son, pero sí son todos los que están. Disculpadme si no he mencionado a alguien.

Gracias a mis cómplices por acompañarme en el largo camino. Este, en realidad, es el principio...

Resumen

La estimación de probabilidades de ocurrencia de cantidades extremales es imprescindible en el estudio de la peligrosidad de fenómenos naturales. Las cantidades extremales de interés suelen corresponder a fenómenos caracterizados por dos o más magnitudes, que en muchos casos son dependientes entre sí. Por tanto, para poder caracterizar mejor las situaciones que pudieran resultar peligrosas, se deben estudiar conjuntamente las magnitudes que describen el fenómeno. Se ha establecido un modelo Poisson-GPD que permite describir la ocurrencia de los sucesos extremales y sus tamaños marginales: la ocurrencia de los sucesos extremales se representa mediante un proceso de Poisson y cada suceso se caracteriza por un tamaño modelado según una distribución generalizada de Pareto, GPD. La dependencia entre sucesos se modeliza mediante funciones cópula: se utiliza una familia de cópulas Gumbel, adecuada al tipo de datos, y se introduce un nuevo tipo de cópula, la cópula CrEnC. La cópula CrEnC minimiza la información mutua en situaciones donde se dispone de información parcial en forma de restricciones, tales como los modelos marginales o momentos conjuntos de las variables. La representación de estas cópulas en $\mathbb{R}^2$ permite mejorar tanto su estima como la apreciación de la bondad de ajuste a los datos. Se proporciona un algoritmo de estimación de cópulas CrEnC, que incluye una aproximación de las funciones normalizadoras mediante el método Montecarlo.

En este contexto los datos suelen ser escasos, por lo que la incertidumbre en la estimación del modelo será elevada. Se ha establecido un proceso de estimación bayesiana de los parámetros, la cual permite tener en cuenta esta incertidumbre. La bondad de ajuste de diversos aspectos del modelo (bondad de ajuste GPD, hipótesis GPD-Weibull, bondad

de ajuste global) se ha valorado mediante una selección de p-valores bayesianos, los cuales incorporan la incertidumbre de la estimación de los parámetros. Una vez estimado el modelo, se realiza un post-proceso de la información, donde se obtienen cantidades a posteriori de interés, como probabilidades de excedencia de valores de referencia o periodos de retorno de sucesos de un tamaño determinado.

El modelo propuesto se aplica a tres conjuntos de datos de características diferentes. Se obtienen buenos resultados: las cópulas CrEnC introducidas representan correctamente la dependencia en situaciones en las que sólo se dispone de información parcial y la estimación bayesiana de los parámetros del modelo proporciona valor añadido a los resultados, ya que permite evaluar la incertidumbre de las estimaciones y tenerla en cuenta al obtener las cantidades a posteriori deseadas.

Resum

L'estimació de probabilitats d'ocurrència de quantitats extremals és imprescindible a l'estudi de la perillositat de fenòmens naturals. Les quantitats extremals d'interès acostumen a correspondre a fenòmens caracteritzats per dues magnituds o més, que en molts casos són dependents entre si. Per tant, per a poder caracteritzar millor les situacions que poguessin resultar perilloses, cal estudiar conjuntament les magnituds que descriuen el fenomen. S'ha establert un model Poisson-GPD que permet descriure l'ocurrència dels esdeveniments extremals i les seves mides marginals: l'ocurrència dels esdeveniments extremals es representa mitjançant un procés de Poisson i cada esdeveniment es caracteritza per una mida modelitzada segons una distribució generalitzada de Pareto, GPD. La dependència entre esdeveniments es modelitza mitjançant funcions còpula: s'empra una família de còpules Gumbel, adequada al tipus de dades, i s'introdueix un nou tipus de còpula, la còpula CrEnC. La còpula CrEnC minimitza la informació mútua en situacions on es disposa d'informació parcial en forma de restriccions, com els models marginals o moments conjunts de les variables.

En aquest context, les dades acostumen a ser escasses, pel que la incertesa en l'estimació del model serà elevada. S'ha establert un procés d'estimació bayesiana dels paràmetres, la qual permet tenir en compte aquesta incertesa. La bondat d'ajust de diversos aspectes del model (bondat d'ajust GPD, hipòtesis GPD-Weibull, bondat d'ajust global) s'han valorat mitjançant una selecció de p-valors bayesians, els quals incorporen la incertesa de l'estimació dels paràmetres. Una vegada estimat el model, es realitza un post-procés de la informació, on s'obtenen quantitats a posteriori d'interès, com probabilitats d'excedència de valors de referència o períodes de retorn d'esdeveniments d'una mida determinada.

El model proposat s'aplica a tres conjunts de dades de característiques diferents. S'obtenen bons resultats: les còpules CrEnC introduïdes representen correctament la dependència en situacions en les que només es disposa d'informació parcial, i l'estimació bayesiana dels paràmetres del model proporciona valor afegit als resultats, donat que permet avaluar la incertesa de les estimacions i tenir-la en compte en obtenir les quantitats a posteriori desitjades.

Abstract

The estimation of occurrence probabilities of extremal quantities is essential in the study of hazards associated with natural phenomena. The extremal quantities of interest usually correspond to phenomena characterized by two or more magnitudes, often showing dependence among them. In order to better characterize situations that could be dangerous, the magnitudes that describe the phenomenon should be jointly described.

A Poisson-GPD model, which describes the occurrence of extremal events and their marginal sizes, has been established: the occurrence of the extremal events is represented by means of a Poisson process, and each event is characterized by a size modelled by a Generalized Pareto Distribution, GPD. The dependence between events is modelled through copula functions: a family of Gumbel copulas, suitable for the type of data treated, and a new type of copula that is introduced, the CrEnC copula. The CrEnC copula minimizes the mutual information in situations in which only partial information in the form of restrictions is available, such as marginal models or joint moments of the variables.

In this context, data are often scarce, and the uncertainty in the estimation of the model will be great. A Bayesian estimation process that takes into account this uncertainty has been established. Goodness-of-fit of some aspects of the model (GPD goodness-of-fit, GPD-Weibull hypothesis and global goodness-of-fit) has been checked using a selection of Bayesian p-values, which incorporate the uncertainty of the parameter estimation. Once the model has been estimated, a post-process of information has been performed to obtain a posteriori quantities of interest, such as exceedance probabilities of reference values or return periods of events of a certain size.

The proposed model is applied to three datasets, with different characteristics. The results obtained are good: the introduced CrEnC copulas correctly represent the dependence in situations in which only partial information is available, and the Bayesian estimation of the parameters of the model gives added value to the results, because it allows the uncertainty of the posterior estimates, such as hazard and dependence parameters, to be evaluated.

Índice general

Capítulo 1

Introducción

1.1. Motivación

Uno de los objetivos básicos de la Estadística es hacer de puente entre la realidad y los modelos matemáticos que la describen. En algunos problemas se estudia una única característica de la población, pero en la mayoría de situaciones de interés será imprescindible estudiar dos o más características. Por ello, es básica la comprensión de las relaciones entre sucesos multivariados.

En el contexto de peligrosidad de fenómenos naturales, también llamada *hazard*, se estudian fenómenos que pueden implicar riesgos que afectan directa o indirectamente a la población. Estos fenómenos se pueden describir mediante una o varias características, posiblemente interrelacionadas entre sí. La modelización de estos fenómenos implica un conocimiento de las variables que los describen, tanto individual como conjuntamente, especialmente para aquellos valores potencialmente peligrosos. Por ejemplo, en el estudio de un temporal de lluvia, caracterizado por la precipitación total recogida y las intensidades máximas de lluvia registradas, conviene tratar estas variables tanto individualmente como de manera conjunta, en particular para precipitaciones totales grandes e intensidades elevadas.

La estimación de probabilidades de ocurrencia de los fenómenos que implican riesgos con frecuencia presenta dificultades. Los datos de calidad suelen ser un bien escaso, y en muchos casos, los datos disponibles

pueden resultar insuficientes para poder realizar estimas de probabilidades, periodos de retorno o distribuciones con una cierta calidad. Por ello, interesa ampliar la información de que se dispone utilizando información de fuentes diversas, aunque se trate de datos de características dispares o registrados en intervalos de tiempo diferentes. Por ejemplo, en un contexto de ingeniería civil, se podría ampliar la información sobre los temporales marinos registrados en una boya situada en la costa utilizando los datos registrados en una boya cercana, pese a que estos registros abarcaran periodos de tiempo diferentes. O bien se podría ampliar la información contenida en el registro de precipitaciones recogidas cada 30 minutos en un observatorio meteorológico utilizando el registro de precipitación en tiempos inferiores, por ejemplo cada 5 minutos.

Estas situaciones se deben enfocar globalmente, con el objetivo de realizar estimaciones de calidad, incorporando toda la información al alcance y prestando especial atención a la dependencia entre las diversas variables estudiadas o los diferentes conjuntos de datos utilizados.

En estas circunstancias, se pretende desarrollar un método de estimación que permita reducir la incertidumbre sobre los parámetros del modelo. El modelo contiene la selección de distribuciones marginales y la representación de la dependencia entre esas marginales. El uso de funciones cópula, distribuciones de extremos y métodos bayesianos son de particular importancia para el desarrollo metodológico de la Tesis. A continuación se incluye una introducción al estado del arte de estas materias.

1.2. Estado del arte

1.2.1. Cópulas

Entender relaciones entre sucesos multivariados es un problema básico en Estadística. La correlación ha sido el parámetro más popular utilizado para resumir la dependencia de variables aleatorias. Sin embargo, la correlación por sí misma no es capaz de describir la dependencia completa en contextos donde la distribución normal no juega el papel central. Las cópulas multivariadas surgen entonces (Sklar, 1959;

Nelsen, 1999) como una caracterización de las distribuciones conjuntas, independientemente de las distribuciones marginales.

El término cópula proviene del latín, y se refiere a conectar o unir. Pero las cópulas de las que nos ocupamos son conceptos estadísticos que se refieren al modo en el que variables aleatorias se relacionan entre sí. Por un lado, las cópulas son funciones que unen o acoplan funciones de distribución multivariadas con sus funciones de distribución marginal unidimensionales. Por otro lado, las cópulas son funciones de distribución multivariadas cuyas marginales unidimensionales son uniformes en el intervalo (0,1).

Las cópulas son de interés por dos motivos principales: en primer lugar, como un modo de estudiar medidas de dependencia libres de escala; en segundo lugar, como un punto de partida para construir familias de distribuciones bivariadas.

La palabra cópula fue utilizada por primera vez en un sentido matemático o estadístico por Abe Sklar en 1959, en el teorema que actualmente lleva su nombre, describiendo las funciones que unen funciones de distribución unidimensionales para formar funciones de distribución multivariadas. Sin embargo, sin utilizar explícitamente el término cópula, éstas aparecen en el trabajo de Fréchet, Dall'Aglio, Féron y muchos otros (citados en Nelsen, 1999), en el estudio de distribuciones multivariadas con distribuciones marginales univariadas. De hecho, muchos de los resultados básicos sobre cópulas aparecen en el trabajo de Wassily Hoeffding, quien utiliza distribuciones estandarizadas bivariadas con soporte contenido en el cuadrado $[-1/2, 1/2]^2$ y cuyas marginales son uniformes en el intervalo $[-1/2, 1/2]$ (Hoeffding, 1940, 1941).

Hoeffding también obtuvo las desigualdades básicas de cotas mejores posibles para aquellas funciones, caracterizó las distribuciones (dependencia funcional) correspondiendo a aquellas cotas, y estudió medidas de dependencia que son invariantes por escala, es decir, invariantes bajo transformaciones estrictamente crecientes (Hoeffding, 1940, 1941). Desconociendo el trabajo de Hoeffding, Fréchet (1951) obtuvo independientemente resultados que han llevado a términos como "las cotas de Fréchet", y "clases de Fréchet" (Fréchet, 1951). En reconocimiento a la responsabilidad compartida por esas importantes ideas, se utiliza el término "cotas de Fréchet-Hoeffding " y "clases de Fréchet-Hoeffding".

Tras Hoeffding, Fréchet y Sklar, las funciones ahora conocidas como

cópulas fueron redescubiertas por algunos autores más.

En la época en que Sklar escribió su artículo de 1959 con el término cópula (Sklar, 1959), estaba colaborando con Berthold Schweizer en el desarrollo de la teoría de espacios métricos probabilísticos, o espacios PM. Aquí aparecen las t-normas, que como las cópulas, van de $[0,1]^2$ a $[0,1]$, y relacionan funciones de distribución. Algunas t-normas son cópulas, y recíprocamente, algunas cópulas son t-normas.

Entre los resultados más importantes en espacios PM está la clase de t-normas arquimedianas. Las t-normas arquimedianas que son a su vez cópulas reciben el nombre de cópulas arquimedianas. Debido a sus formas simples, la sencillez con la cual pueden ser construidas, y sus propiedades atractivas, las cópulas arquimedianas aparecen frecuentemente en discusiones sobre distribuciones multivariadas (p.e. Genest and MacKay, 1986b).

Las cópulas aparecen implícitamente en trabajos sobre dependencia de muchos autores. Entre otros, Hoeffding además de estudiar las propiedades básicas de distribuciones estandarizadas, es decir, cópulas, las usó para estudiar medidas de asociación no paramétricas como la ρ de Spearman y su índice de dependencia Φ^2 (Hoeffding, 1940, 1941). Por su parte, Deheuvels definió las funciones de dependencia empíricas, es decir, cópulas empíricas, el análogo muestral de las cópulas, para estimar la cópula poblacional y construir varios tests de independencia no paramétricos.

Las cópulas están siendo explotadas en diversos campos, jugando un importante papel en probabilidad, estadística y procesos estocásticos.

La cópula captura aspectos no paramétricos de la relación entre las variables, por lo que las medidas de asociación y conceptos de dependencia son propiedades de la cópula. Las cópulas representarán una aproximación muy útil a la comprensión y la modelización de las variables aleatorias cuya dependencia se desea estudiar, dado que nos permiten centrarnos explícitamente en la estructura de dependencia entre ellas.

1.2.2. Origen de algunas familias de cópulas

Una de las primeras familias de funciones de dependencia introducidas fue la propuesta por Plackett (1965). Desde entonces, otras

familias han sido sugeridas por diversos autores, como Clayton (1978), Cook and Johnson (1981), Oakes (1989) o Tawn (1988), entre otros. Estas primeras familias de funciones de dependencia solían corresponder a una clase específica de distribuciones bivariadas indexadas por un parámetro (uni o bidimensional). Con posterioridad se introdujeron las cópulas Arquimedianas (Genest and MacKay, 1986b,a; Genest and Rivest, 1993), una clase de funciones de dependencia amplia y matemáticamente tratable.

La clase de cópulas arquimedianas incluye a las funciones de dependencia obtenidas a partir de muchas funciones de distribución bivariadas ampliamente conocidas, como las de Gumbel, (Gumbel, 1960); Ali-Mikhail-Haq-Thélot, (Ali et al., 1978) ; Clayton, (Clayton, 1978; Cook and Johnson, 1981); Frank, (Genest, 1987; Frank, 1979; Nelsen, 1986; Hougaard, 1984, 1986). Las cópulas arquimedianas también aparecen en el contexto de las funciones de supervivencia, Oakes (1989), o en el contexto de los modelos de mixturas (Marshall and Olkin, 1988).

La clase de cópulas metaelípticas, derivada de la familia de distribuciones elípticas, que constituyen una extensión de la normal multivariante clásica, fue introducida originalmente por Fang et al. (2002). Las propiedades principales de esta familia de cópulas se describen en Fang et al. (2002) y Abdous et al. (2005).

1.2.3. Métodos Bayesianos

La probabilidad ha sido el objeto de estudio durante cientos de años, pero en cambio, la mayoría de técnicas estadísticas son relativamente recientes (regresión lineal, noción de verosimilitud, tests de hipótesis clásicos...). Como excepción, los métodos bayesianos surgieron a partir del teorema de Bayes, a mediados del sigo XVIII. El área generó cierto interés entre matemáticos del siglo XIX, como Laplace o Gauss, pero durante el principio del siglo XX los estadísticos la ignoraron por completo o bien surgieron oposiciones drásticas por cuestiones filosóficas. Fueron personajes no estadísticos, como Jeffreys o Bowley, quienes mantuvieron durante esta época el interés sobre las ideas bayesianas (a las cuales ellos denominaban probabilidad inversa).

Hacia la mitad del siglo XX, algunos investigadores estadísticos como L.J. Savage, Bruno de Finetti y otros, (De Finetti 1974; Savage

1972), empezaron a utilizar los métodos bayesianos y a defenderlos como remedios a algunas deficiencias observadas en los métodos clásicos, por ejemplo en la estimación de parámetros por intervalo o los tests de hipótesis mediante el método de Neyman-Pearson (Lee, 1997). Durante los años 50 y 60 eran comunes las discusiones estadístico-filosóficas entre partidarios y detractores de estos métodos.

Dado que los métodos bayesianos ofrecen soluciones a diversos problemas o contradicciones del enfoque frecuentista, sorprende que la metodología bayesiana no haya hecho mella en la práctica estadística hasta hace relativamente poco. Se pueden considerar diversos motivos que influyeran en esta puesta en práctica tardía: en primer lugar, los primeros defensores de los métodos bayesianos eran subjetivistas, es decir, creían que todos los cálculos estadísticos debían realizarse sólo después de que uno valorara y cuantificara las creencias previas (a priori) sobre la materia estudiada. Esto generaba dudas sobre la objetividad de los resultados obtenidos, sesgados por los conocimientos del investigador. Por otro lado, y quizá con más importancia desde el punto de vista aplicado, la alternativa bayesiana pese a tener una base teórica simple, requería la evaluación de integrales complejas, incluso en problemas relativamente sencillos. A partir de los años 80, con el gran salto cualitativo en los ordenadores, la evaluación de estas integrales resultó un problema menor. A su vez, la aparición de nuevos investigadores partidarios de métodos más objetivos determinaron la irrupción de las técnicas bayesianas en muchos ámbitos.

La diferencia básica entre el esquema bayesiano y el frecuentista consiste en considerar que los parámetros son aleatorios, mientras que el esquema frecuentista los considera fijos. Al ser aleatorios, los parámetros siguen una ley de probabilidad. A partir de los datos observados, se estima la distribución condicionada de estos parámetros a los datos, de manera que se incorpora al modelo la información aportada por éstos.

Los estadísticos estan interesados en hallar métodos que proporcionen un equilibro efectivo entre eficiencia y robustez. Estas propiedades son de gran importancia, sin tener en cuenta si los datos se tratan desde un punto de vista frecuentista o bayesiano. En muchos casos, se pueden utilizar las mejoras del formalismo bayesiano sin perder la robustez de los métodos frecuentistas. El formalismo bayesiano puede ser incluso más efectivo si lo que se pretende es ajustar el análisis sobre la base de

la opinión personal o sobre información objetiva externa al conjunto de datos que se está tratando, de manera que se enriquece el conocimiento sobre la situación estudiada.

Los métodos Bayesianos han demostrado tener propiedades atractivas tanto para los estadísticos bayesianos como los frecuentistas. El enfoque bayesiano ofrece, en muchos casos, mejoras sobre la metodología clásica, y por eso será el enfoque utilizado en el desarrollo de esta Tesis.

1.2.4. Procesos de Poisson

Los Procesos de Poisson constituyen un modelo adecuado para multitud de fenómenos que ocurren a lo largo del tiempo, y en particular para los fenómenos estudiados desde el punto de vista del *hazard*, como temporales de viento o precipitaciones.

Un proceso puntual es un proceso estocástico que describe la ocurrencia de sucesos en el espacio de modelado (p. ej. el eje temporal o el espacio). Intuitivamente, al mencionar un proceso puntual nos referimos a una serie de sucesos que ocurren en el tiempo (o el espacio, o ambos), según una ley de probabilidad. Los ejemplos clásicos son la llegada de llamadas telefónicas a una centralita, desintegraciones radiactivas o bien la distribución de las semillas de una cierta planta en un campo no cultivado. Cuando se asocia una magnitud o intensidad a cada ocurrencia de suceso, entonces el proceso se denomina un proceso puntual marcado. Reiss (1993) muestra la relevancia de la teoría de procesos puntuales en diversas aplicaciones en Estadística y otros campos. Autores como Leadbetter et al. (1983) o Embrechts et al. (1997) los introducen el uso de los procesos puntuales en el tratamiento de sucesos de tipo extremal, como los excesos sobre un umbral.

El proceso puntual más simple en tiempo continuo es el proceso de Poisson, en el que los sucesos ocurren aleatoriamente a lo largo del eje temporal. En un proceso de Poisson, los tiempos entre sucesos son independientes y distribuidos exponencialmente. El proceso de Poisson ha sido utilizado frecuentemente para estudiar fenómenos como la precipitación. Por ejemplo Rodriguez-Iturbe et al. (1984) estudian la precipitación mediante un proceso de Poisson marcado.

Las monografías de Cox and Isham (1980) y Daley and Vere-Jones (2003) contienen excelentes tratamientos de la teoría de procesos estocásticos puntuales. Daley and Vere-Jones (2003) incluye un capítulo inicial donde con un repaso histórico a los avances en diversas áreas (tablas de supervivencia; problemas de conteo; física de partículas y procesos poblacionales o ingeniería de comunicación), que comparten conceptos con lo que los autores denominan la teoría moderna de los procesos puntuales.

1.2.5. Distribuciones de extremos

Ordenar observaciones de acuerdo con su magnitud e identificar sucesos centrales o extremos es una de las actividades humanas más simples. Por tanto, se pueden obtener referencias tempranas a los estadísticos de orden en diversos libros antiguos. Por ejemplo, J. Tiago de Oliveira menciona la edad de Matusalén (Génesis, La Biblia) en el prefacio de *Statistical Extremes and Applications* (Tiago de Oliveira, 1984). En él se explica que Matusalén vivió 969 años. Esto no debe ser considerado meramente una curiosidad, sino como un indicador de las dificultades de la elección adecuada de un modelo, en conexión con la pregunta. Su enfoque puede compararse con el de Gumbel (1933), dónde E.J. Gumbel se plantea si la distribución de mortalidad tiene un soporte acotado utilizando la edad del mismo personaje bíblico.

La teoría probabilística de valores extremos es un área de investigación que fue emprendida inicialmente por probabilistas teóricos, ingenieros e hidrólogos y más recientemente por estadísticos. En ella se combinan resultados matemáticos específicos con aplicaciones en una gran variedad de áreas. Las aplicaciones a fenómenos naturales (precipitaciones, inundaciones, viento, polución) son las más conocidas, pero existen aplicaciones en muchas otras áreas.

En la introducción de Gumbel (1958) se resalta el trabajo de recopilación de L. Harter, quien generó una bibliografía de publicaciones y referencias sobre estadísticos de orden en dos volúmenes: antes de los años 1950 y en el periodo 1950-59. El primer resultado relevante mencionado en esta recopilación es el de Nicolas Bernoulli, en 1709, quien discutió la distancia media mayor desde el origen dados n puntos situados aleatoriamente sobre una línea de longitud fijada t, el cual

podría ser interpretado como el cálculo de la esperanza del máximo de variables aleatorias uniformes. En los primeros trabajos el uso de la media fue de cierta importancia, debido a su propiedad de minimizar la suma de las desviaciones absolutas al cuadrado. Hay que destacar que Laplace en 1818 demostró la normalidad asintótica de esta media muestral.

La teoría estadística en el siglo XIX se puede caracterizar por el papel de de la distribución normal como ley universal. A su vez, fue el inicio de una fase crítica que surgió del hecho que los extremos frecuentemente no se adecuan a este supuesto de normalidad. Los extremos se veían como dudosos, como observaciones periféricas (*outliers*), que deberían rehusarse. La actitud hacia los extremos en aquella época se podría interpretar como un intento de inmunizar la suposición de normalidad contra la experiencia.

La teoría estadística moderna debe mucho a de R.A. Fisher, quien en 1921 discutió el problema de los *outliers*: "...el rechazo de observaciones es demasiado duro para ser defendido, y a no ser que haya otras razones para el rechazo que meras divergencias respecto la mayoría, sería mucho más filosófico aceptar estos valores extremos, no como errores flagrantes, sino como indicaciones de que la distribución de errores no es normal"(Fisher, 1922).

A finales de los años veinte diversos matemáticos estudiaron la materia. Entre ellos R. Von Mises estudió el comportamiento asintótico del máximo muestral de variables aleatorias normales y no normales (Von Mises, 1923). Bajo condiciones de regularidad débiles, las satisfechas por la función de distribución normal por ejemplo, von Mises probó que el valor esperado del máximo muestral es asintóticamente igual a $F^{-1}(1-1/n)$. Un resultado similar fue deducido por E.L. Dodd para diversos tipos de distribuciones.

Este desarrollo culminó con el artículo de Fisher and Tippet, 1928, donde se demostraba que las distribuciones de extremos pueden ser sólo de tres tipos (dominios de atracción) y se discutió el problema de la estabilidad. La distribución límite denominada de Fréchet fue descubierta de manera independiente por Fréchet, 1927. De hecho, el resultado de Fisher y Tippet y el de Fréchet aparecieron casi simultáneamente en 1928.

Gnedenko en 1943 halló condiciones necesarias y suficientes para que

una función de distribución F pertenezca al dominio de atracción débil de una función de densidad de valor extremo (Gnedenko, 1943). Años más tarde De Haan alcanzó una especificación de la función auxiliar en la caracterización de la función de distribución F realizada por Gnedenko para pertenecer al dominio de atracción de la distribución de Gumbel (de Haan, 1976).

Von Mises enunció las condiciones suficientes para que una función de distribución pertenezca al dominio de atracción de las distribuciones de Fréchet y Weibull (Von Mises, 1936). Otro conjunto de condiciones de von Mises para funciones dos veces derivables se puede hallar en obras de von Mises y Pickands (Von Mises, 1936; Pickands III, 1986). Estas nuevas condiciones, en conjunto con los dominios de atracción fuertes, dieron un nuevo impulso a la materia. La convergencia puntual de las densidades de máximos muestrales fueron probadas independientemente en Pickands III (1967) y Reiss (1989), y en los años ochenta, otros autores investigan las equivalencias de las condiciones de Von Mises con diferentes tipos de convergencia.

Con el fin de establecer la distribución límite del k-ésimo estadístico mayor, diversos autores, entre ellos Leadbetter (Leadbetter et al., 1983) estudiaron el número de excesos de n variables independientes e idénticamente distribuidas sobre un umbral u. El argumento clave fue que este número de excesos tiene asintóticamente una distribución Poisson.

Los desarrollos teóricos de los años 1920 y mediados de los 30 fueron seguidos por una serie de artículos, al final de los años 30 y años 40, sobre aplicaciones prácticas de estadísticos de valores extremos en duración de la vida humana, emisiones radioactivas, resistencia de materiales, análisis sísmico y análisis de precipitaciones, por citar algunos ejemplos. Desde la publicación del libro *Statistics of Extremes* (Gumbel, 1958), el cual causó un gran impacto, se han logrado muchos avances en el área de la teoría de valores extremos. Esto se ha debido en gran parte a ingenieros y científicos quienes comenzaron a darse cuenta de los nuevos problemas que surgían en su práctica diaria, y al trabajo de matemáticos y estadísticos quienes incorporaron muchos de estos problemas de valor extremos a sus áreas de trabajo.

Sin pretender ser exhaustivos, podemos mencionar algunos de estos avances:

- Algunos teoremas que han permitido la caracterización de distribuciones límite, dominios de atracción, etc;
- La descripción única de von Mises de la distribución límite para el caso independiente e idénticamente distribuido (i.i.d.), clásicamente descrita en sus tres casos;
- La identificación de condiciones necesarias y suficientes bajo las cuales las distribuciones límite para el caso i.i.d. aún se mantienen en el caso dependiente;
- Resultados para modelos usuales, como el Gaussiano o las medias móviles, por ejemplo;
- El análisis de distribuciones de valor extremos multivariadas;
- Nuevos métodos de estimación basados en estadísticos de orden;
- La identificación de distribuciones límite posibles para estadísticos de orden k-ésimo en el caso i.i.d., con análisis separado para estadísticos de orden grande, pequeño, moderadamente grande o pequeño y estadísticos de orden central;

Diversos autores (Beirlant et al., 2004; Embrechts et al., 1997; Castillo, 1988; Castillo et al., 2004; Galambos, 1987; Kotz and Nadarajah, 2000; Reiss and Thomas, 1997) resumen los avances más significativos en la teoría asintótica de extremos y sus aplicaciones estadísticas, haciendo énfasis en diferentes áreas, como ingeniería o aplicaciones actuariales.

1.2.6. Coeficientes de dependencia

El concepto de *correlación*, introducido por Galton en 1885 (Galton, 1888), dominó la Estadística durante gran parte del siglo XX, prácticamente sirviendo como la *única* medida de dependencia, en ocasiones conllevando conclusiones erróneas (pese a la advertencia de prudencia del propio autor, Galton (1890)). Durante los últimos treinta años del siglo XX y hasta la actualidad, ha habido un resurgimiento de los trabajos sobre propiedades de dependencia tanto desde el punto de vista estadístico como probabilístico.

Dado que la estructura de dependencia entre variables aleatorias se halla representada por la distribución de cópula C (ver Sección 2.1), parece que un modo natural de estudiar y medir la dependencia entre variables aleatorias es mediante cópulas. Las funciones cópula son invariantes bajo transformaciones estrictamente crecientes de las marginales, y por ello ésta es una propiedad deseable de las medidas de dependencia. Por ejemplo, si se desea estudiar la dependencia entre dos variables aleatorias, X e Y, el resultado debiera ser el mismo al cambiar la escala utilizada para tomar las observaciones. En esta situación, el uso del coeficiente de correlación lineal puede llevar a conclusiones erróneas, dado que no es una medida basada en cópulas. Una discusión interesante sobre correlación nula versus dependencia y algunas malinterpretaciones relevantes al respecto se presenta en Drouet-Mari and Kotz (2001). La cuestión de la correlación nula habiendo dependencia se discute en la mayoría de libros de texto estándar de probabilidad y estadística, como Feller (1968a), Feller (1968b), o Mood et al. (1974).

Desde los trabajos iniciales de Hoeffding (1940), Kruskal (1958) y Lehmann (1966), se han propuesto múltiples medidas de dependencia entre variables aleatorias y/o muestras (ver Dickinson Gibbons, 1993; Joe, 1997; Nelsen, 1999; Coles et al., 1999). En Drouet-Mari and Kotz (2001) también se puede encontrar una buena cobertura de varias medidas de asociación, así como cuestiones relativas a ordenaciones de dependencia y el origen histórico de estos conceptos de dependencia. Kruskal (1958) presenta un completo resumen de las medidas *ordinalmente invariantes* aparecidas hasta la fecha, y en particular presenta una excelente reseña histórica sobre el origen de los coeficientes más conocidos, como la τ de Kendall (Kendall, 1938) y la ρ de Spearman (Spearman, 1904), entre otros, a finales del siglo XIX y principios del XX, pocos años después de que el coeficiente de correlación lineal se extendiera como herramienta de análisis estadístico. Schweizer and Wolff (1981) y Scarsini (1984) muestran que los coeficientes de dependencia monótona se pueden determinar a partir de la función cópula. En consecuencia, medidas de concordancia como la τ de Kendall o la ρ de Spearman se pueden expresar de manera sencilla en términos de la correspondiente cópula. El estudio realizado por Embrechts et al. (2002) o Genest and Plante (2003) proporcionan una discusión sobre las ventajas de las medidas basadas en cópulas sobre la correlación

lineal.

1.2.7. Densidades de mínima entropía cruzada dadas restricciones

En numerosas situaciones sólo se dispone de información parcial sobre una distribución (o de densidad) multivariada, en forma de propiedades marginales y/o cierto grado de dependencia entre las variables aleatorias. Se ha abordado la descripción o construcción de las correspondiente distribuciones conjuntas utilizando diversas técnicas. En particular, ha sido frecuente el uso de conceptos de teoría de la información. En este tipo de problemas se consideran básicamente dos tipos de restricciones sobre las distribuciones conjuntas: restricciones sobre las marginales y restricciones sobre los momentos, aunque no todos los trabajos consideran estas restricciones simultáneamente.

La distribución bivariada de máxima entropía con marginales dadas ya fue considerada en Fréchet (1951). En Nelsen (1991) se consideran distribuciones bivariadas con marginales y correlación dadas. Otros autores, como Pasha and Mansoury (2008), Gokhale (1999), Arnold et al. (1999) han abordado la determinación de distribuciones de probabilidad de máxima entropía (de Shannon) bajo restricciones.

La entropía cruzada o divergencia de Kullback-Leibler es otro de los coeficientes usados frecuentemente en la construcción de densidades conjuntas sujetas a restricciones. Uno de los primeros trabajos donde se relaciona la minimización de la entropía cruzada y las distribuciones con marginales dadas en el caso multivariado es Kullback (1968). Más recientemente, Miller and Liu (2002), entre otros, abordan la determinación de las distribuciones de probabilidad de mínima entropía relativa bajo restricciones en forma de momentos. Cabe destacar que algunos de los primeros trabajos en el área parecen haber pasado desapercibidos por autores posteriores. Es el caso del trabajo de Rumsey Jr and Posner (1965), quienes consideraban las restricciones en forma de marginales y de momentos (simultáneamente). Otros autores, como Ebrahimi et al. (2008), realizan enfoques similares, llegando a soluciones que presentan algunas carencias.

Meeuwissen and Bedford (1997), quienes estudian la distribuciones bivariadas de mínima información con marginales uniformes y corre-

lación por rangos dada, introducen las funciones cópula y las medidas de correlación *marginal-invariant* en el tratamiento del problema. Más recientemente, Calsaverini and Vicente (2009), profundizan en la necesidad de utilizar correlaciones invariantes por transformaciones y explicitar la dependencia conjunta mediante las funciones cópula.

1.3. Objetivos

En el estudio de peligrosidad de fenómenos naturales (*hazard*) es necesario evaluar con qué probabilidad pueden aparecer sucesos que impliquen riesgos. El objetivo principal es evaluar estas probabilidades en situaciones donde existen dos o más variables implicadas. En muchos casos el objeto de estudio son fenómenos que ocurren en el tiempo y que presentan tamaños diferentes en cada ocurrencia. Por ello, utilizaremos los llamados Procesos de Poisson evaluados o compuestos (Embrechts et al., 1997), que describen tanto esta ocurrencia en el tiempo como el tamaño de la variable de interés en cada instante.

Centramos la atención en dos posibles escenarios :

Escenario 1

Consideramos los sucesos de viento registrados en una estación meteorológica A. Se dice que ha ocurrido un suceso de viento cuando las velocidades registradas en la estación superan un umbral de referencia previamente fijado. El suceso persiste hasta que se detectan velocidades por debajo del umbral definido, y éstas se mantienen por debajo del umbral durante un intervalo de tiempo determinado, usualmente 3 días. Esta definición procura la independencia entre sucesos. El suceso de viento se caracteriza por un tamaño, X_1, y un momento de ocurrencia, t : por ejemplo, se puede considerar que X_1 corresponde al máximo de las velocidades de viento registradas durante el suceso y t al instante de ocurrencia de este valor máximo, aunque existen definiciones alternativas.

El objetivo principal es evaluar la probabilidad de ocurrencia de sucesos con unas características determinadas o bien funciones de estas probabilidades, como p.ej. los periodos de retorno. Es común estimar la probabilidad de que tenga lugar un suceso de viento cuyo tamaño supere un nivel prefijado durante la vida útil de una infraestructura para determinar, por ejemplo, parámetros del diseño de ésta. Otra situación usual, también de cara al diseño de infraestucturas, es el cálculo del periodo de retorno de sucesos con tamaños determinados.

La estimación de estas probabilidades o periodos de retorno no siempre es sencilla. Generalmente los periodos de retorno de interés están entre 50 y 500 años, relacionados con tamaños grandes de la variable, y en cambio, la mayoría de registros disponibles abarcan solamente los últimos 10 ó 20 años. Esta diferencia entre el tiempo de registro y el tiempo de predicción implicará una gran incertidumbre en las estimas de las probabilidades.

Para solventar estas deficiencias, consideraremos el registro de viento de una estación meteorológica, B, cercana a la original, cuyo registro abarque un periodo de tiempo más largo que el de la estación de referencia A. Al poder utilizar un conjunto de datos que puede complementar la información de que disponemos, nos planteamos como objetivo secundario la reducción de la incertidumbre de las estimaciones en la estación A.

Los sucesos de viento registrados en la estación B están caracterizados de la misma manera que los observados en la estación A, es decir, están caracterizados por el máximo de las velocidades observadas durante el transcurso del suceso, que denotaremos X_2, y por el momento de ocurrencia de este máximo, t. Para incorporar la información de la estación B a la estación A, utilizaremos la distribución condicionada: condicionaremos el tamaño de la estación original, X_1, por un valor observado de la estación complementaria, $X_2 = x_2$. Para obtener esta distribución condicionada, y una visión de conjunto de ambas estaciones, necesitaremos calcular previamente la función de distribución conjunta de ambos tamaños.

Dado un suceso de viento, podemos considerar diferentes valores de referencia, que llamaremos umbrales, que caracterizan diferentes niveles de peligrosidad. Utilizaremos la notación h_i para los umbrales correspondientes al tamaño X_i, $i = 1, 2$.

El objetivo es estimar la probabilidad de ocurrencia de sucesos de viento con un tamaño superior a un valor dado, o modelar los excesos sobre un umbral h. Para conseguirlo, aplicaremos métodos específicos, como el de excesos sobre umbral (Peak-Over-Threshold, POT), (Davison and Smith, 1990; Embrechts et al., 1997). Por otro lado, para estimar los riesgos que pueden implicar

los episodios de viento fuerte, se pretende estudiar el tamaño máximo en un periodo de tiempo t_0. Para modelizar la ocurrencia de tamaños máximos se aplicarán métodos de extremos (Castillo, 1988; Castillo et al., 2004; Coles, 2001).

Escenario 2

Consideramos los sucesos de lluvia registrados en una estación meteorológica A. Se produce un suceso de lluvia cuando la precipitación recogida en 24 horas en la estación supera un nivel prefijado, generalmente entre 5 y 25mm. Este suceso persiste hasta que la precipitación recogida es inferior a este nivel y se mantiene por debajo del umbral durante un tiempo determinado, generalmente 3 días. De este modo, los sucesos de lluvia registrados se supondrán independientes entre sí.

El suceso de lluvia se caracteriza por un tamaño, X_1, y un momento de ocurrencia, t, donde X_1 corresponde al total de precipitación recogida durante el suceso.

Describir la lluvia en una determinada región únicamente mediante la precipitación total puede dar una idea incompleta de la realidad. Si se quieren estimar probabilidades de ocurrencia de fenómenos que implican un riesgo elevado para la población, o funciones de estas probabilidades, será útil estudiar otras variables que caractericen el fenómeno. En el estudio de riesgos relacionados con precipitaciones extremas, los sucesos con lluvias de intensidad elevada pueden provocar problemas, por ejemplo en la red de alcantarillado: aunque el total de lluvia recogido al final del episodio no sea elevado, episodios de lluvia intensos y de una cierta duración pueden provocar inundaciones y riesgos a la población.

Por tanto, caracterizaremos los sucesos de lluvia por una segunda variable, X_2, y el tiempo de ocurrencia del suceso, t. Se define intensidad de lluvia como la cantidad de precipitación recogida en un tiempo corto, de 5 a 30 min. En este caso, la variable X_2 corresponderá a la intensidad máxima registrada durante el episodio de lluvia, y t al momento en que este máximo fue registrado.

Dado un suceso de lluvia, podemos definir diferentes niveles o umbrales para cada una de las variables, correspondientes a los diferentes niveles de riesgo. Por ejemplo, conocido el caudal máximo que puede absorber la red de alcantarillado en un momento dado, en función de este valor podemos definir una intensidad umbral, h. El objetivo será evaluar la probabilidad de que la variable intensidad, X_2, supere este umbral h. Por otro lado, se pretende modelar la distribución del máximo de precipitación recogida, X_1, con el objetivo estimar las probabilidades de ocurrencia de máximos de un valor determinado. Aplicaremos métodos adecuados a cada uno de los modelos: excesos sobre umbral *(POT)* (Davison and Smith, 1990; Embrechts et al., 1997) en el caso de excesos sobre un umbral y métodos de extremos (Castillo, 1988; Galambos, 1987; Reiss and Thomas, 1997) en el caso de máximos.

En ambos escenarios se han de estimar probabilidades de ocurrencia de sucesos extremales, aunque el enfoque en los dos escenarios no es exactamente el mismo: en la situación descrita en a), se utilizan los datos de una de las estaciones para aumentar la información disponible en la estación de interés; en cambio, en la situación descrita en b) nos interesa la propia distribución conjunta de las dos variables, que puede aportar información sobre las características, por ejemplo, de diferentes regiones pluviométricas.

Se utilizarán métodos de estimación bayesiana para incorporar la variedad de datos disponibles a la estimación de los parámetros del modelo. A su vez, al optimizar el uso de la información disponible se puede reducir la incertidumbre inherente a las estimaciones.

Los objetivos que se desea alcanzar en esta Tesis son:

1. Definir un modelo de ocurrencia, tamaños y dependencia que permita describir situaciones de *hazard* y realizar cálculos de valores de interés de estas situaciones.

2. Introducir un nuevo tipo de cópulas, las cópulas CrEnC, que permitan describir la dependencia entre dos o más variables a partir la información parcial en forma de restricciones (marginales

y coeficientes de dependencia invariantes por transformaciones monótonas), minimizando la información mutua.

3. Describir la transformación de la función cópula entre dos variables X_1, X_2, a un nivel determinado $\mathbf{h_0} = (h_0^1, h_0^2)$, en dos situaciones de interés:

   a) Al aumentar el nivel de referencia o umbral a uno más elevado, $\mathbf{h} = (h^1, h^2)$, y observar los excesos por encima del nuevo umbral, Y_1, Y_2.

   b) Al extraer los máximos de las variables en un tiempo t_0, Z_1, Z_2.

4. Desarrollar un código que permita realizar la estimación bayesiana de todos los parámetros implicados en el modelo (los de ocurrencia, los de distribuciones marginales y los correspondientes a la dependencia), y realizar el cálculo de probabilidades de cantidades de *hazard* de interés.

En el Capítulo 2, se resumen algunos conceptos teóricos contenidos en la literatura que serán utilizados en el desarrollo metodológico. Los procesos de Poisson modelizarán la ocurrencia de los fenómenos en los escenarios establecidos; las distribuciones de máximos y excesos modelizan los tamaños de las características de estos fenómenos y las funciones cópula describen la dependencia entre dos o más de estas características, en el sentido descrito en los escenarios. El Paradigma Bayesiano y el método de Gibbs se introducen en los apartados 2.5 y 2.5.1 respectivamente, dado que serán necesarios en el desarrollo del método de estimación. Se introducen también los coeficientes de dependencia, las densidades sujetas a restricciones y las medidas de información más comunes, la base para la definición de un nuevo tipo de cópula.

Los Capítulos 3, 4 y 5 constituyen el corpus metodológico de la Tesis, donde se introduce el modelo propuesto, las transformaciones extremales de cópulas y la cópulas CrEnC. El contexto bayesiano de estimación de los parámetros del modelo se explicita en los Capítulos 6 y 7 , donde se obtiene la densidad conjunta a posteriori de los

parámetros del modelo. El método de Gibbs facilita el cálculo de este posteriori. Finalmente, se obtienen probabilidades, periodos de retorno, y otros valores de interés en el entorno del *hazard* a partir de la muestra simulada de la densidad conjunta a posteriori de los parámetros. La estimación del modelo se ha implementado en un programa Fortran elaborado íntegramente para la Tesis, donde se realizan las estimaciones de los parámetros del modelo y los cálculos de las cantidades predictivas correspondientes. Este programa se aplica a tres conjuntos de datos de características diferentes: unos datos simulados (Capítulo 10), un conjunto de datos de precipitación diaria (Capítulo 11) y un conjunto de datos de altura de ola significante (Capítulo 12). Finalmente, se proporcionan unas conclusiones (Capítulo 13).

Capítulo 2

Fundamentos

Una vez definidos los objetivos, conviene introducir algunos conceptos básicos. A continuación se presentan conceptos como las funciones cópula, los procesos de Poisson, las distribuciones de máximos y excesos y las distribuciones de máxima información mutua respecto sus marginales. Se incluyen también apartados que introducen el Paradigma Bayesiano y el método de Gibbs, técnicas a utilizar en los capítulos posteriores.

En cada uno de estos apartados se presentan conceptos básicos y se facilitan referencias bibliográficas que permiten completar la información sobre el tema.

2.1. Cópulas

La noción de cópula es importante desde diversos puntos de vista. Este tipo de funciones de distribución se utilizan principalmente por su capacidad de reflejar la dependencia entre dos o más variables aleatorias, independientemente de la distribución marginal de éstas, y en un sentido más amplio de dependencia que el que proporciona la correlación lineal. A lo largo del siglo XX diversos autores han ampliado el conocimiento sobre el tema, aunque el teorema básico es el debido a Sklar (1959). Las definiciones y teoremas que se introducen a continuación pueden encontrarse en diversas fuentes, por ejemplo en Nelsen (1999).

La noción general de cópula proviene de la definición:

Definición 2.1.1. Una *cópula n-dimensional* (una n-cópula) es una función C del n-cubo unidad $[0,1]^n$ al intervalo unidad $[0,1]$, $C : [0,1]^n \to [0,1]$, satisfaciendo las siguientes condiciones:

(1) $C[1,\ldots,1,a_m,1,\ldots,1] = a_m$ para cada $m \leq n$ y todo a_m en $[0,1]$.

(2) $C[a_1,\ldots,a_n] = 0$ si $a_m = 0$ para cualquier $m \leq n$.

(3) C es n-creciente, en el sentido que el c-volumen de cualquier intervalo n-dimensional es no negativo.

En particular, si $C[u,v]$ es una cópula 2-dimensional, entonces,

(4) $C[u,1] = u$, $C[1,v] = v$, para todo u,v en $[0,1]$.

(5) $C[u,0] = C[0,v] = 0$ para todo u,v en $[0,1]$.

(6) $C[a_2,b_2]-C[a_1,b_2]-C[a_2,b_1]+C[a_1,b_1] \geq 0$, cuando $a_1 \leq a_2$, $a_1 \leq a_2 \in [0,1]$.

El resultado básico debido a Sklar (1959) , es el siguiente:

Teorema 2.1.1. *Si H es una distribución de probabilidad n-dimensional con marginales unidimensionales $F_1,\ldots,F_n$, entonces existe una cópula n-dimensional C tal que, para todo $x_1,\ldots,x_n$ en $\mathbb{R}$,*

(7) $H(x_1,\ldots,x_n) = C[F_1(x_1),\ldots,F_n(x_n)]$.

Si H es continua, entonces C es única; en caso contrario, C está unívocamente determinada sobre el producto cartesiano $(\mathrm{Sop}F_1)\times(\mathrm{Sop}F_2)\times\cdots\times(\mathrm{Sop}F_n)$, *donde* Sop *denota el soporte de las distribuciones.*

Recíprocamente, si C es una cópula n-dimensional y $F_1,\ldots,F_n$ son funciones de distribución unidimensionales, entonces la función H definida por la Ec. 7 es una función de distribución de probabilidad n-dimensional cuyas marginales unidimensionales son $F_1,\ldots,F_n$.

Por tanto, si las funciones de distribución marginales de H son continuas, entonces H tiene una única cópula. Pero si existen discontinuidades en una o más marginales, existirá más de una representación de H en forma de cópula.

Schweizer and Wolff (1981) enuncia el Teorema 2.1.1 en su versión más utilizada y comenta los trabajos más destacados donde se pueden hallar demostraciones del mismo.

Si expresamos el teorema de Sklar en términos de variables aleatorias, éste toma la forma:

Teorema 2.1.2. *Sean $X_1, \ldots, X_n$ variables aleatorias reales, definidas en un espacio de probabilidad común, con funciones de distribución individuales $F_{X_1}, \ldots, F_{X_n}$ y función de distribución conjunta $H_{X_1 \ldots X_n}$. Entonces, existe una cópula n-dimensional $C_{X_1 \ldots X_n}$ tal que, para todo $x_1, \ldots, x_n$ en $\mathbb{R}$,*

(8) $H_{X_1 \ldots X_n}(x_1, \ldots, x_n) = C_{X_1, \ldots, X_n}[F_{X_1}(x_1), \ldots, F_{X_n}(x_n)]$.

Si $F_{X_1}, \ldots, F_{X_n}$ son continuas, entonces $C_{X_1, \ldots, X_n}$ es única; en caso contrario, $C_{X_1, \ldots, X_n}$ está unívocamente determinada en $(\mathrm{Sop} F_{X_1}) \times (\mathrm{Sop} F_{X_2}) \times \cdots \times (\mathrm{Sop} F_{X_n})$.

Para simplificar notaciones, de ahora en adelante en el texto nos referiremos principalmente al caso bidimensional.

Proposición 2.1.3. *Si X_1 y X_2 son variables aleatorias con función de distribución F y G, respectivamente, función de distribución conjunta H y cópula C, se cumple que:*

(8) $\mathrm{Max}\{F(x_1) + G(x_2) - 1, 0\} \leq H(x_1, x_2) \leq \mathrm{Min}\{F(x_1), G(x_2)\}$,

y estas cotas son comúnmente denominadas cotas de Fréchet.

En el caso de variables aleatorias independientes, la cópula que les corresponde es $C[u, v] = uv$. Se puede considerar que las tres cópulas más importantes son:

- la cota inferior de Fréchet, $C[u, v] = \text{máx}\{0, u + v - 1\}$, representada en la Fig. 2.1 ,
- la cota superior de Fréchet, $C[u, v] = \text{mín}\{u, v\}$, representada en la Fig. 2.2, y
- la cópula de la independencia, $C[u, v] = uv$, representada en la Fig. 2.3 .

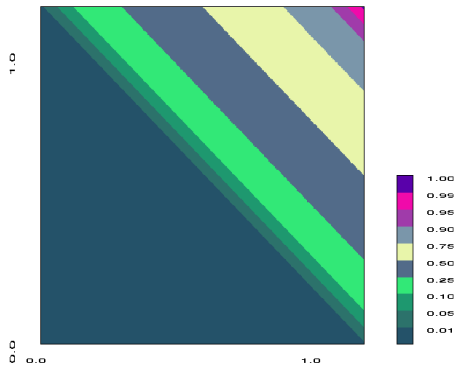

Figura 2.1: Cota inferior de Fréchet. Representación de la función de distribución mediante curvas de isoprobabilidad.

Una de las propiedades que da utilidad a las funciones de cópula es su invariancia bajo transformaciones estrictamente monótonas de las variables aleatorias.

Proposición 2.1.4. *(Invariancia) Sean $X_1, \ldots, X_n$ variables aleatorias reales, con funciones de distribución marginales continuas $F_{X_1}, \ldots, F_{X_n}$ y cópula C. Sean $T_1, \ldots, T_n$ transformaciones estrictamente crecientes. Entonces, el conjunto de variables $T_1(X_1), \ldots, T_n(X_n)$ tiene la misma cópula C que $X_1, \ldots, X_n$.*

Por tanto, se deduce que la forma en que $X_1, \ldots, X_n$ se relacionan es capturada por la cópula, independientemente de la escala en que cada variable sea medida.

2.1.1. Cópulas deducidas de distribuciones

La definición de cópula y la expresión de cópula de variables aleatorias $X_1, \ldots, X_n$ se suelen intercambiar con frecuencia. Esto sugiere que

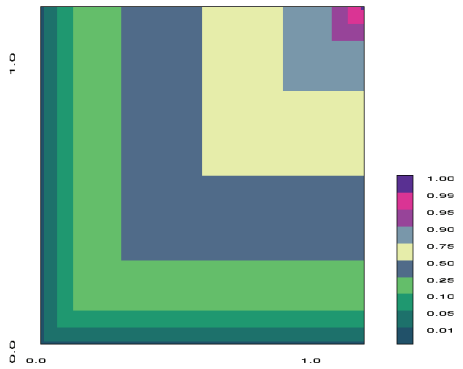

Figura 2.2: Cota superior de Fréchet. Representación de la función de distribución mediante curvas de isoprobabilidad.

distribuciones multivariadas comunes describen estructuras de dependencia importantes. En particular, la distribución normal multivariada conduce a la cópula Gaussiana, y la distribución t de Student conduce a la t-cópula.

Dadas dos variables aleatorias distribuídas normalmente, X_1 y X_2, conjuntamente normales, su correlación

$$\rho(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{Var(X_1) \cdot Var(X_2)}}$$

describe completamente la estructura de dependencia. Esta propiedad se mantiene en la familia de distribuciones elípticas, y suele ser fuente de errores cuando X_1,X_2 tienen distribuciones fuera de esta familia, puesto que este coeficiente no es invariante por transformaciones monótonas (Ver Sec. 2.2)

La cópula Gaussiana bidimensional tiene expresión

$$C_\rho[u_1, u_2] = \Phi_\Sigma(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \ ,$$

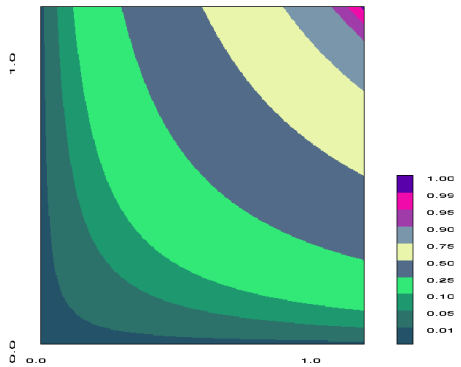

Figura 2.3: Cópula de la independencia. Representación de la función de distribución mediante curvas de isoprobabilidad.

donde Φ denota la función de distribución de una normal estándar y Φ_Σ es la función de distribución de una normal bivariada con media nula y matriz de covarianzas Σ (con 1 en la diagonal y ρ fuera de ella).

En el caso multivariante, la cópula Gaussiana con matriz de correlación Σ tiene expresión:

$$C_\Sigma[u_1, \ldots, u_n] = \Phi_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)) \ .$$

Para distribuciones normales y elípticas, la independencia es equivalente a la correlación nula. Por tanto, para $\rho = 0$, la cópula Gaussiana corresponde a la cópula de la independencia. En el otro extremo, si $\rho = 1$, se obtiene la cópula de completa dependencia, mientras que si $\rho = -1$, se obtiene la dependencia exacta inversa.

Otra distribución conocida de importancia es la distribución t-Student. Una variable η distribuida según una t con ν grados de libertad se puede representar como

$$\eta = \frac{X_1}{\sqrt{\xi/\nu}} = \frac{\sqrt{\nu} \cdot X_1}{\sqrt{Y_1^2 + \cdots + Y_\nu^2}} \ ,$$

donde $X_1, Y_1, \ldots, Y_n$ son variables normal estándard independientes, y ξ tiene una distribución χ^2_ν. La distribución t multivariada, en d dimensiones, con ν grados de libertad se obtiene de manera similar a partir de $\mathbf{X} = (X_1, \ldots, X_d) \sim \mathcal{N}(0, \Sigma)$:

$$\left(\frac{X_1}{\sqrt{\xi/\nu}}, \ldots, \frac{X_d}{\sqrt{\xi/\nu}} \right) ,$$

donde ξ tiene distribución χ^2_ν, independiente de $\mathbf{X}$. En el caso bidimensional, si la correlación entre X_1 y X_2 es 1, se tiene dependencia completa, o completa inversa para la correlación -1. Es importante destacar que correlación nula no implica independencia en este caso, dado que se ha introducido algo de dependencia mediante ξ.

La t-copula o Student copula, obtenida a partir de la distribución t multivariante, viene dada por

$$C^t_{\nu,\Sigma}[u_1, \ldots, u_d] = t_{\nu,\Sigma}(t^{-1}_\nu(u_1), \ldots, t^{-1}_\nu(u_d)) ,$$

donde Σ es una matriz de correlación, t_ν es la función de distribución de una distribución t, y $t_{\nu,\Sigma}$ es la función de distribución de una variable t multivariada.

Si se comparan la cópula Gaussiana y la cópula t, sus densidades son similares en el centro, pero el comportamiento en los extremos difiere. Para la misma correlación, los casos extremos tienen una densidad mucho más elevada en el caso de la t-cópula, indicando diferentes niveles de dependencia en las colas (tail-dependence).

La clase de cópulas metaelípticas incluye a las cópulas normal y t. Esta clase de cópulas fue introducida originalmente por Fang et al. (2002), a partir de la familia de distribuciones elípticas, que a su vez es una extensión de la distribución multivariante normal clásica.

Un vector p-dimensional $\mathbf{X}^*$ tiene una distribución elíptica $\varepsilon_p(\mu, \Sigma, g)$ con vector de medias $\mu \in \mathbb{R}^p$, matriz de covarianzas Σ y generador g , si se puede expresar en la forma

$$\mathbf{X}^* = \mu + RAU ,$$

donde $AA^T = \Sigma$ es la descomposición Cholesky de Σ, y U es un vector aleatorio p-dimensional uniformemente distribuído en la esfera S_p y R

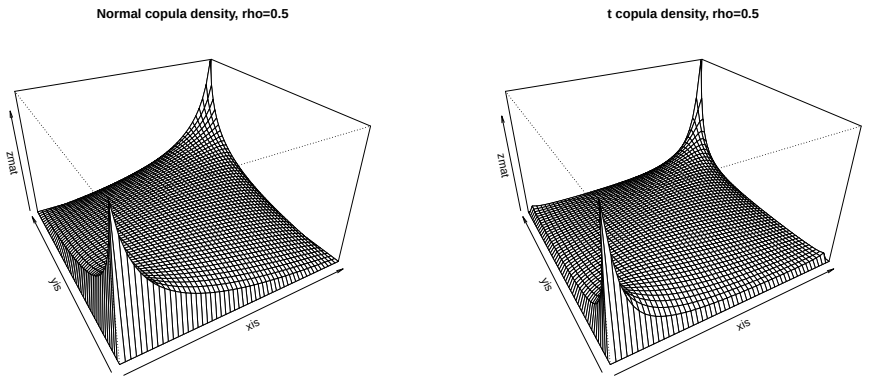

Figura 2.4: Densidad de Cópula Normal ($\rho = 0.5$) (Izq.); Densidad de Cópula t ($\rho = 0.5$) (Dcha.)

es un vector aleatorio no negativo con densidad

$$f_g(r) = \frac{2\pi^{p/2}}{\Gamma(p/2)} r^{p-1} g(r^2),\ r > 0\ ,$$

donde g es una función de escala tal que

$$\int_0^{+\infty} g(t)dt < \infty\ .$$

Cuando $g(t) \propto \exp(-t/2)$, $\mathbf{X}^*$ es una normal multivariante, y R^2 se distribuye según una χ^2_p. Otras elecciones de g dan lugar a distribuciones como la t-Student multivariante o la distribución de Pearson Tipo II.

Formalmente, la cópula metaelíptica $C_{\Sigma,g}$ asociada a $\mathbf{X}^*$ es la función de distribución conjunta del vector $(U_1, \ldots, U_p)$, con $U_i = Q_g(X_i/\sigma_{ii})$, para $i \in \{1, \ldots, p\}$. La función de distribución conjunta de un vector elíptico $\mathbf{X}^*$ no suele estar disponible en forma cerrada. Por tanto, su inversa no es explícita y la expresión de $C_{\Sigma,g}[u_1, \ldots, u_p]$ a partir de ellas no es útil, aunque se pueden simular observaciones mediante un sencillo algoritmo. Las principales propiedades de las cópulas metaelípticas se describen en Fang et al. (2002) y Abdous et al. (2005), entre otros.

De cara a la estimación de estas familias de cópulas, es de particular importancia la conexión entre la correlación entre las variables y el valor teórico de la τ de Kendall entre componentes, establecida en Kruskal (1958), Hult and Lindskog (2002), o Lindskog et al. (2003):

$$\text{Si} \quad \rho(X_1^*, X_2^*) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_{ii}\sigma_{jj}} \text{ , entonces}$$

$$\tau_{ij} = \tau(X_1^*, X_2^*) = \frac{2}{\pi}\arcsin(\rho_{ij}), \ i, j \in \{1, \ldots, p\} \ .$$

Debe tenerse en cuenta que, excepto si g es el generador de la distribución normal multivariante, $\tau_{ij} = \rho_{ij} = 0$ *nunca* corresponde a la independencia entre X_1^*, X_2^*.

Entre las cópulas destacaremos, por su sencillez y propiedades, a las denominadas cópulas arquimedianas. En el siguiente apartado incluimos algunas definiciones básicas, ampliables en Nelsen (1999).

2.1.2. Cópulas arquimedianas

La representación arquimediana de cópulas nos permite reducir el estudio de una cópula multivariada a un única función univariada.

Definición 2.1.2. Sea Φ el conjunto de funciones φ, $\varphi : [0,1] \to [0,+\infty]$, continuas, estrictamente decrecientes, convexas y tales que $\varphi(0) = +\infty$ y $\varphi(1) = 0$. Cada una de estas funciones φ posee una inversa, $\varphi^{-1} : [0,+\infty] \to [0,1]$, decreciente, convexa y tal que $\varphi^{-1}(0) = 1$ y $\varphi^{-1}(+\infty) = 0$.

Dada una función $\varphi \in \Phi$, sea la cópula C generada como

$$C[u,v] = \varphi^{-1}(\varphi(u) + \varphi(v)), \quad 0 \leq u, v \leq 1 \ . \tag{2.1.1}$$

La cópula definida en (ec. 2.1.1) se denomina *arquimediana* y φ se denomina un *generador* de C.

La condición $\varphi(0) = +\infty$ no es imprescindible para obtener una cópula mediante (ec. 2.1.1). Si $\varphi(0) = +\infty$ el generador se denomina estricto. Cuando $\varphi(0)$ es finito se puede obtener una cópula arquimediana mediante la pseudo-inversa de φ:

$$\varphi^{[-1]} = \begin{cases} \varphi^{-1} & \text{si } 0 \leq \text{t} \leq \varphi(0) \\ 0 & \text{si } \varphi(0) \leq \text{t} \leq +\infty \end{cases} \ .$$

En general, se considerarán generadores que cumplan la definición original de Φ.

Ejemplo 2.1.1 (Cópula de la independencia como cópula arquimediana)**.** Sean X_1 y X_2 variables aleatorias independientes, continuas, con funciones de distribución marginales F_1 y F_2 y función de distribución conjunta F_{12}. Dada su independencia, para dos valores reales x_1, x_2,

$$F_{12}(x_1, x_2) = F_1(x_1)F_2(x_2) \quad .$$

Se toma la función $\varphi(t) = -\ln(t)$, $\varphi \in \Phi$. Dado que $\varphi(0) = +\infty$, su inversa viene dada por $\varphi^{-1} = \exp(-t)$. Si generamos una cópula mediante la expresión (2.1.1), obtenemos

$$C[u, v] = \exp(-[(-\ln(u)) + (-\ln(v))]) = uv = \Pi(u, v) \ .$$

y por tanto, $F(x_1, x_2) = \Pi(F_1(x_1), F_2(x_2))$. Así, la cópula producto (o de la independencia), Π, es una cópula arquimediana.

Las familias de cópulas arquimedianas tienen como ventaja su facilidad de construcción. Simplemente es necesario encontrar funciones adecuadas que sirvan como generadores, las cuales definiran la correspondiente cópula. Para cada elección de generadores, obtendremos diferentes familias de cópulas arquimedianas:

Ejemplo 2.1.2. Algunas familias arquimedianas de uso frecuente:

a) Cópula Gumbel: Dado el generador

$$\varphi_\theta(t) = (-\ln t)^\theta \quad ,$$

se obtiene la cópula

$$C[u, v] = \exp(-[(-\ln(u))^\theta + (-\ln(v))^\theta]^{1/\theta}) \quad , \qquad (2.1.2)$$

una de las familias de cópulas de Gumbel, (Gumbel, 1960).

b) Cópula Ali-Mikhail-Haq: Si se toma el generador

$$\varphi_\theta(t) = \ln\left(\frac{1-\theta(1-t)}{t}\right) \quad ,$$

se obtiene la cópula

$$C[u,v|\theta] = \frac{uv}{1-\theta(1-u)(1-v)} \quad , \tag{2.1.3}$$

llamada de Ali-Mikhail-Haq (Ali et al., 1978);

c) Cópula Frank: El generador

$$\varphi_\theta(t) = \ln(e^{-\theta}-1) - \ln(e^{-\theta t}-1) \quad ,$$

lleva a la cópula

$$C[u,v] = \frac{1}{\theta}\ln\left(1 + \frac{\ln(e^{-\theta u}-1)\cdot\ln(e^{-\theta v}-1)}{\ln(e^{-\theta}-1)}\right) \quad , \tag{2.1.4}$$

para $\theta \in \mathbb{R}\setminus\{0\}$, denominada cópula de Frank (Frank, 1979; Genest, 1987).

Caracterización de las cópulas arquimedianas

Teorema 2.1.5. *Las cópulas arquimedianas son aquellas cópulas $C[u,v]$ tales que satisfacen las siguientes propiedades:*

(1) $C[u,v]$ es asociativa. Es decir, para todo $u,v,w \in [0,1]$, se tiene que

$$C[C[u,v],w] = C[u,C[v,w]] \quad .$$

(2) $C[u,u] < u$ para todo $u \in (0,1)$.

El siguiente criterio, llamado criterio de Abel, permite a su vez dar una caracterización de las cópulas arquimedianas fácil de aplicar.

Teorema 2.1.6. *Una cópula C es arquimediana si es dos veces diferenciable y si existe una función integrable f, $f:(0,1)\to(0,+\infty)$, tal que para valores u,v entre 0 y 1,*

$$f(v)\frac{\partial}{\partial u}C[u,v] = f(u)\frac{\partial}{\partial v}C[u,v] \ .$$

En este caso, el generador φ viene dado por

$$\varphi(t) = \int_t^1 f(s)ds \ , \ 0 \le t \le 1 \ .$$

Ejemplo 2.1.3 (Cotas de Fréchet como cópulas arquimedianas)**.** Aplicando el teorema (2.1.6) a las cotas de Fréchet $M(u,v) = \text{mín}\{u,v\}$ y $W(u,v) = \text{máx}\{0, u+v-1\}$, se puede observar que sólo la cota inferior de Fréchet, $W(u,v)$ es una cópula arquimediana.

Las cópulas permiten enfocar la dependencia entre dos o más variables sin necesidad de considerar la distribución marginal de las variables implicadas. Existen diversos coeficientes que permiten describir esta dependencia a partir de la función cópula (ver Sección 2.2.2). Joe (1997), Nelsen (1999), Rüschendorf et al. (1996) presentan caracterizaciones de diversos tipos de dependencia. Las familias de cópulas arquimedianas o de supervivencia son de utilidad para representar la dependencia en situaciones particulares. Pero en la práctica, hallar cuál es la cópula que representa la dependencia de dos variables aleatorias no es sencillo.

2.1.3. Tests de bondad de ajuste para cópulas

Se desea modelizar la dependencia presente en un conjunto de datos mediante una determinada familia de cópulas C con parámetro θ. ¿Cómo comprobar la calidad global del ajuste con un test formal?. En la literatura se han propuesto diversos procedimientos para comprobar esta bondad de ajuste, que se pueden dividir en tres amplias categorías:

a) aquellos basados en la *probability integral transformation* de Rosenblatt (1952), como Breymann et al. (2003), Berg and Bakken (2006) o Dobrić and Schmid (2007);

b) aquellos que implican procesos tipo kernel, como Fermanian (2005), Panchenko (2005) y Scaillet (2007);

c) aquellos que utilizan el proceso de cópula empírico, como Genest and Rémillard (2008) y Genest et al. (2006);

Ésta es un área de investigación muy activa. Un buen resumen de las técnicas de bondad de ajuste más extendidas se puede encontrar en Genest et al. (2009).

2.1.4. Cautela

Las cópulas son una herramienta general para describir estructuras de dependencia, y han sido aplicadas con éxito en ámbitos diversos. De todos modos, el uso de cópulas presenta algunos inconvenientes de importancia, algunos a nivel teórico y otros a nivel práctico. El artículo Mikosch (2006) es un trabajo muy crítico al respecto, que incluye una interesante discusión. Genest and Rémillard (2006) incluye también comentarios al respecto. En su visión personal, Embrechts (2009) intenta conciliar a partidarios y detractores, aunque apunta los puntos débiles de la técnica.

En particular, cabría destacar los siguientes inconvenientes:

- La representación de la dependencia mediante cópulas se percibe como un recetario.

- No sólo es importante determinar un modelo que puede ajustar los datos, también debe comprobarse que éste ajusta correctamente.

- Existen familias de cópulas que no provienen de distribuciones multivariantes. Por ejemplo esto sucede con algunas de las cópulas arquimedianas, que aparecen principalmente por ser herramientas con un tratamiento matemático sencillo. Su aplicación como modelos naturales para reflejar dependencia debiera ser verificada en cada caso.

- Dependiendo del contexto, puede convenir expresar las marginales de cópula en el soporte $[0,1]^2$ o en otro soporte. Algunos autores proponen expresar las cópulas en soportes diferentes, como por ejemplo el soporte $\mathbb{R}^{+2}$, que correspondería a marginales Fréchet. Elegir uno u otro soporte, dependiendo del contexto, sin argumentar el por qué, da la impresión de discrecionalidad (Klüppelberg and Resnick, 2008).

A pesar de estas (y otras) críticas, las cópulas son una herramienta útil, que permite mejorar el modelado de la dependencia en la práctica, y pone a disposición una gran variedad de posibles estructuras de dependencia.

2.2. Medidas de dependencia

La descripción de la dependencia entre variables aleatorias puede ser un punto clave en aplicaciones de campos tan diversos como ingeniería, medicina, ciencias sociales o política. A continuación, se introducen algunos conceptos relacionados con la descripción de esta dependencia, incluyendo la correlación lineal clásica.

2.2.1. Coeficiente de correlación lineal

El coeficiente de correlación lineal es ampliamente utilizado en aplicaciones, pero constituye una medida de dependencia que no puede capturar relaciones de dependencia no lineales.

Definición 2.2.1. Sean X_1, X_2 dos variables aleatorias con varianza finita, $\mathrm{Var}(X_1) < +\infty$, $\mathrm{Var}(X_2) < +\infty$. Sea $\mathrm{Cov}[X_1, X_2]$ la covarianza entre las variables X_1 y X_2, definida por:

$$\mathrm{Cov}[X_1, X_2] = \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_2 - \mathrm{E}[X_2])] = \mathrm{E}[X_1 X_2] - \mathrm{E}[X_1]\mathrm{E}[X_2] \ .$$

El *coeficiente de correlación lineal* entre X_1 y X_2 viene dado por la expresión:

$$\rho(X_1, X_2) = \frac{\mathrm{Cov}[X_1, X_2]}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}} \ . \tag{2.2.1}$$

Proposición 2.2.1. *Propiedades:*

a) $\rho(X_1, X_2) \in [-1, 1]$.

b) Si X_1,X_2 *son independientes, entonces su coeficiente de correlación lineal es nulo,* $\rho(X_1, X_2) = 0$.

c) Si X_1, X_2 *son dependientes linealmente, es decir* $X_2 = aX_1 + b$, *para dos coeficientes reales* a, b, $a \neq 0$, *entonces* $|\rho(X_1, X_2)| = 1$.

d) El coeficiente de correlación lineal no varía para transformaciones lineales estrictamente crecientes. Es decir, dados dos coeficientes reales no nulos, a, b, $\rho(aX_1 + b, X_2) = \mathrm{sgn}(a)\rho(X_1, X_2)$.

Según Lehmann (1966), la definición 2.2.1 se puede reescribir como:

$$\rho(X_1, X_2) = \frac{1}{\sqrt{Var(X_1)Var(X_2)}} \int\int_{\mathbb{R}^2} [F_{12}(x_1, x_2) - F_1(x_1)F_2(x_2)]\, dx_1 dx_2,$$

donde F_1, F_2 y F_{12}, son las marginales de X_1, X_2 y su distribución conjunta respectivamente. Sustituyendo $u = F_1(x_1)$ y $v = F_2(x_2)$, es decir, empleando la transformada de probabilidad (Whitt, 1976), se obtiene

$$\rho(X_1, X_2) = \frac{1}{\sqrt{Var(X_1)Var(X_2)}} \int\int_{[0,1]^2} [C(u, v) - uv]\, dF_1^{-1}(u) dF_2^{-1}(v)\,,$$

donde $C[u, v] = H(F_1^{-1}(u), F_2^{-1}(v))$. Esta última expresión incluye la cópula C, pero también las varianzas $\mathrm{Var}(X_1), \mathrm{Var}(X_2)$ y por tanto, el coeficiente de correlación lineal no es invariante por transformaciones estrictamente crecientes.

2.2.2. Coeficientes de correlación basados en cópulas

Dado un par de variables X_1, X_2, con función de distribución conjunta F_{12}, se desea resumir su dependencia a partir de algún tipo de medida. Existen numerosos coeficientes que describen y miden dependencia entre variables aleatorias. Muchos de estos coeficientes son, en las palabras de Hoeffding (1940) "invariantes por escala", es decir, no cambian bajo transformaciones estrictamente crecientes de las variables aleatorias. Existen dos clases amplias de medidas de asociación entre variables aleatorias que incluyen a la mayoría de los coeficientes existentes: medidas de dependencia y medidas de concordancia.

La concordancia es una importante noción invariante por escala. De hecho, Schweizer and Wolff (1981) destacan que "es precisamente la cópula quien captura aquellas propiedades de la distribución conjunta que son invariantes bajo transformaciones estrictamente crecientes (c.s.)". Por tanto, las medidas de concordancia y las funciones cópula tienen una estrecha relación. Antes de introducir estas medidas, es necesario realizar algunas definiciones e introducir notación.

Sean X_1, X_2 dos variables aleatorias, y (x_1^i, x_2^i), $i = 1, \ldots, n$ un conjunto de sus observaciones. Denotaremos según (R_i, S_i) los pares de rangos asociados a la muestra, donde R_i corresponde al rango de x_1^i entre $x_1^1, \ldots, x_1^n$ y S_i corresponde al rango de x_2^i entre $x_2^1, \ldots, x_2^n$. En el caso en que las variables X_1, X_2 sean continuas, estos rangos están bien definidos, dado que los empates ocurren con probabilidad nula bajo la hipótesis de continuidad, (Genest and Favre, 2007).

Sea $C[u, v]$ una cópula bivariada y continua, $F_1(x_1)$, $F_2(x_2)$ funciones de distribución marginales de X_1, X_2, absolutamente continuas, y $F_{12}(x_1, x_2)$ la correspondiente función de distribución conjunta. Las funciones de densidad que les corresponden son $c[u, v]$, $f_1(x_1)$, $f_2(x_2)$ y $f_{12}(x_1, x_2)$ respectivamente. El soporte conjunto de las variables aleatorias X_1, X_2 se denota $\text{Supp}(X_1, X_2)$.

Definición 2.2.2. (Hoeffding, 1947) Dadas dos variables aleatorias X_1, X_2, y una muestra (x_1^i, x_2^i), $i = 1, \ldots, n$, se dice que dos pares de observaciones (x_1^i, x_2^i) y (x_1^j, x_2^j) son *concordantes* si ambos valores de un par son mayores que los valores correspondientes del otro par, es decir, $x_1^i < x_1^j$ y $x_2^i < x_2^j$, o si $x_1^i > x_1^j$ y $x_2^i > x_2^j$. Análogamente, (x_1^i, x_2^i) y (x_1^j, x_2^j) son *discordantes* si para un par un valor es mayor y el otro menor que el correspondiente valor del otro par , es decir, $x_1^i < x_1^j$ y $x_2^i > x_2^j$, o si $x_1^i > x_1^j$ y $x_2^i < x_2^j$. De una forma más compacta, podemos decir que (x_1^i, x_2^i) y (x_1^j, x_2^j) son concordantes si $(x_1^i - x_1^j)(x_2^i - x_2^j) > 0$ y , alternativamente, discordantes si $(x_1^i - x_1^j)(x_2^i - x_2^j) < 0$.

Intuitivamente, podemos decir que dos variables son concordantes si crecen de la misma manera y discordantes si cuando una crece, la otra decrece.

Al final de la década de 1950, A. Rényi propuso un conjunto de propiedades para medidas de dependencia de pares de variables aleatorias (ver Schweizer and Wolff, 1981, y las referencias allí incluidas). Treinta años después, Scarsini (Scarsini, 1984) estableció la definición de medida de concordancia de un par de variables aleatorias a través de un conjunto de propiedades:

Definición 2.2.3. (Scarsini, 1984)
Dado $\mathcal{H}$, el espacio de funciones de distribución F_{12} continuas con marginales continuas, y una aplicación J, entre $\mathcal{H}$ y un conjunto totalmente

ordenado A,

$$\begin{array}{rcl} J: \ \mathcal{H} & \rightarrow & A \subseteq \mathbb{R} \\ F_{12} & \rightarrow & l(X_1, X_2) := J(F_{12}) \ , \end{array}$$

J se denomina *medida de concordancia* si satisface los siguientes axiomas:

1. Dominio: $l(X_1, X_2)$ está definida para cualquier (X_1, X_2) con función de distribución continua.

2. Simetría: $l(X_1, X_2) = l(X_2, X_1)$.

3. Coherencia: $l(X_1, X_2)$ is monótona en C_{12}, es decir, si $C_{12} \geq C_{WZ}$ entonces $l(X_1, X_2) \geq l(W, Z)$.

4. Rango: $-1 \leq l(X_1, X_2) \leq 1$.

5. Independencia: $l(X_1, X_2)$=0 si X_1, X_2 son estocásticamente independientes.

6. Cambio de signo: $l(-X_1, X_2) = -l(X_1, X_2)$.

7. Continuidad: Si $(X_1, X_2) \sim H$ y $(X_{1n}, X_{2n}) \sim H_n$, $n \in \mathbb{N}$ y si H_n converge puntualmente a H, (H_n y H continua), entonces $\lim_{n\to\infty} l(X_{1n}, X_{2n}) = l(X_1, X_2)$.

Este conjunto de propiedades se pueden reescribir utilizando la función cópula entre X_1 y X_2:

Definición 2.2.4. (Genest and Plante, 2003) Una medida numérica κ de asociación entre dos variables continuas X_1 y X_2 cuya cópula es C es una medida de concordancia si satisface las siguiente propiedades (escribimos κ_{X_1,X_2} o κ_C según convenga):

1. Dominio: κ_{X_1,X_2} está definida para todo par X_1, X_2 de variables aleatorias continuas.

2. Rango: $-1 \leq \kappa_{X_1,X_2} \leq 1$, $\kappa_{X_1,X_1} = 1$, y $\kappa_{-X_1,X_1} = -1$.

3. Simetría: $\kappa_{X_1,X_2} = \kappa_{X_2,X_1}$.

4. Independencia: Si X_1 y X_2 son independientes, entonces $\kappa_{X_1,X_2} = \kappa_\Pi = 0$.

5. Cambio de signo: $\kappa_{-X_1,X_2} = \kappa_{X_1,-X_2} = -\kappa_{X_1,X_2}$.

6. Coherencia: Si $C1$ y $C2$ son cópulas tales que $C1 \preceq C2$, entonces $\kappa_{C1} \leq \kappa_{C2}$.

7. Continuidad: Si $\{(X_{1n}, X_{2n})\}$ es una sucesión de variables aleatorias continuas con cópulas C_n, y si $\{C_n\}$ converge puntualmente a C, entonces $\lim_{n\to+\infty} \kappa_{C_n} = \kappa_C$.

Las siguientes propiedades se obtienen de la Def. 2.2.4:

8. Si X_2 es una función creciente de X_1 (c.s.), entonces $\kappa_{X_1,X_2} = \kappa_M = 1$; y si X_2 es una función decreciente de X_1 (c.s.), entonces $\kappa_{12} = \kappa_W = -1$, donde M, W son las cotas de Fréchet para cópulas;

9. Si α y β son funciones estrictamente monótonas (c.s.) sobre Rango(X_1) y Rango(X_2), respectivamente, entonces $\kappa_{\alpha(X_1)\beta(X_2)} = \kappa_{X_1,X_2}$.

Definición 2.2.5 (Correlación por rangos de Spearman, Spearman (1904), Kruskal (1958))**.** Dados tres vectores aleatorios independientes e idénticamente distribuidos, (X_1, X_2), (X_1^*, X_2^*), y (X_1^{**}, X_2^{**}) con función de distribución común H y cópula C. El *coeficiente de correlación por rangos de Spearman* es proporcional a la probabilidad de concordancia menos la probabilidad de discordancia para los vectores (X_1, X_2) y (X_1^*, X_2^{**}):

$$\rho_S = 3(\mathrm{P}[(X_1 - X_1^*)(X_2 - X_2^{**}) > 0] - \mathrm{P}[(X_1 - X_1^*)(X_2 - X_2^{**}) < 0]) \ .$$

La correlación por rangos de Kendall está definida en términos de la concordancia introducida anteriormente (Def. 2.2.2):

Definición 2.2.6 (Correlación por rangos de Kendall, Kendall (1938))**.** Dados dos vectores aleatorios independientes e idénticamente distribuidos,

(X_1, X_2), (X_1^*, X_2^*), el *coeficiente de correlación por rangos de Kendall* se define como:

$$\tau(X_1, X_2) = \mathrm{P}[(X_1 - X_1^*)(X_2 - X_2^*) > 0] - \mathrm{P}[(X_1 - X_1^*)(X_2 - X_2^*) < 0] \ .$$

es decir, es una medida de la probabilidad de concordancia menos la de discordancia.

Observación 2.2.1. Existen distribuciones de cola pesada (p. ej. la Cauchy), para las cuales el coeficiente de correlación de Pearson (Def. 2.2.1) no existe. En cambio, las medidas de concordancia están definidas para todo par de variables aleatorias continuas.

Las medidas de concordancia definidas anteriormente se pueden escribir en función de su correspondiente cópula (Schweizer and Wolff, 1981; Dupuis, 2007; Genest and Favre, 2007; Nelsen, 1999):

Proposición 2.2.2. *Dadas X_1, X_2 dos variables aleatorias, con funciones de distribución F_1 y F_2 respectivamente, F_{12} la función de distribución conjunta de ambas variables, C la cópula asociada a ellas, y $\rho_S(X_1, X_2)$, $\tau(X_1, X_2)$ los coeficientes de correlación por rangos de Spearman y Kendall, se cumple que:*

$$\rho_S(X_1, X_2) = -3+12 \int\int_{[0,1]^2} uv dC[u, v] = -3+12 \int\int_{[0,1]^2} C[u, v] du dv \tag{2.2.2}$$

o equivalentemente,

$$\rho_S(X_1, X_2) = 12 \int_0^1 \int_0^1 (C[u, v] - uv) du dv \quad ; \tag{2.2.3}$$

El coeficiente ρ puede interpretarse como la diferencia promedio (normalizada) entre la cópula C y la cópula de la independencia Π.

$$\tau(X_1, X_2) = -1 + 4 \int_0^1 \int_0^1 C[u, v] dC[u, v] \quad . \tag{2.2.4}$$

Considerando C como una función de distribución conjunta de dos variables uniformes en (0,1), se obtiene también

$$\tau(X_1, X_2) = 4\mathrm{E}(C[U, V]) - 1 \quad .$$

Observación 2.2.2. En ocasiones el coeficiente de correlación por rangos de Spearman se define como:

$$\rho_S(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)) \ ,$$

donde ρ denota el coeficiente de correlación lineal de Pearson. Reescribiendo la ecuación 2.2.2,

$$\rho_S(X_1, X_2) = -3 + 12 \ \mathrm{E}[UV] \ ,$$

y dado que U, V son variables aleatorias Unif[0,1], se tiene que $\mathrm{E}[U] = \mathrm{E}[V] = 1/2$, $\mathrm{Var}[U] = \mathrm{Var}[V] = 1/12$. Reescribiendo la expresión anterior,

$$\rho_S(X_1, X_2) = \frac{\mathrm{E}[UV] - 1/4}{1/12} = \frac{\mathrm{E}[UV] - \mathrm{E}[U]\mathrm{E}[V]}{\sqrt{\mathrm{Var}[U] = \mathrm{Var}[V]}} \ ,$$

y por tanto, en efecto, el coeficiente de Spearman puede considerarse como el coeficiente de correlación entre las variables $U = F(X_1)$ y $V = F(X_2)$.

El coeficiente τ de Kendall i el ρ de Spearman son los más utilizados como alternativa al coeficiente de correlación lineal de Pearson. Nešlehová (2007) estudia la relación entre ambos coeficientes.

Para Cópulas Arquimedianas (Def. 2.1.2), el coeficiente de correlación por rangos de Kendall, τ, se puede expresar como una función del generador de la cópula $\varphi(l)$, (Genest and MacKay, 1986b,a),

$$\tau = 1 + 4 \int_0^1 \frac{\varphi(l)}{\varphi'(l)} \, dl \ . \tag{2.2.5}$$

En algunos casos esta expresión se simplifica considerablemente, quedando en función del parámetro de la cópula. Por ejemplo, para la cópula de Gumbel (Ec. 2.1.2), se obtiene la expresión $\tau = \frac{\delta - 1}{\delta}$, que relaciona el parámetro δ y el coeficiente τ. Se obtiene así un posible método de estimación para el parámetro de esta familia de cópulas.

Definición 2.2.7. El coeficiente γ *de Gini* (Dall'Aglio, 1991; Schweizer, 1991) se obtiene como variación de una de las expresiones del coeficiente de Spearman:

$$\rho_S = 3 \int_0^1 \int_0^1 ([u + v - 1]^2 - [u - v]^2) dC[u, v] \ .$$

El coeficiente γ se obtiene considerando valores absolutos en lugar de cuadrados en la expresión anterior:

$$\gamma = 2 \int_0^1 \int_0^1 (|u + v - 1| - |u - v|) dC[u, v] \ .$$

Teorema 2.2.3. *(Nelsen, 1998) Las versiones poblacionales del coeficiente τ de Kendall, ρ de Spearman y γ de Gini son medidas de concordancia. El coeficiente de correlación lineal de Pearson, ρ_{X_1,X_2} no lo es, aunque cumple que ρ_{X_1,X_2}=0 si y sólo si X_1 e X_2 son independientes.*

Existen diferentes familias de generalizaciones multivariantes tanto de la τ de Kendall como de la ρ_S de Spearman (ver Joe, 1990; Nelsen, 1996, entre otras). Asimismo, Taylor (2007) presenta una generalización de la caracterización de medidas de concordancia introducida en Scarsini (1984).

Otra gran clase de medidas de asociación son las *medidas de dependencia* (Schweizer and Wolff, 1981). Estas medidas resumen el grado de relación entre dos variables aleatorias asignando un número entre cero y uno, con extremos en la independencia mutua y la dependencia monótona, pero no dan información sobre el signo de la relación, es decir, si la asociación es positiva o negativa.

Definición 2.2.8. (Desiderata de Rényi modificada, Schweizer and Wolff (1981))

Una medida numérica κ de asociación entre dos variables aleatorias continuas es una *medida no paramétrica, simétrica, de dependencia* si satisface las siguientes propiedades (escribiremos κ_{X_1,X_2} o κ_C según convenga):

1. Dominio: κ_{X_1,X_2} está definido para todo par X_1, X_2 de variables aleatorias continuas;
2. Simetría: $\kappa_{X_1,X_2} = \kappa_{X_2,X_1}$;
3. Rango: $0 \leq \kappa_{X_1,X_2} \leq 1$;
4. Independencia: $\kappa_{X_1,X_2} = 0$ si y sólo si X_1 y X_2 son independientes;
5. Monotonía: Si ϕ y ψ son monótonas c.s. sobre Rango(X_1), Rango(X_2), respectivamente, entonces $\kappa_{\phi(X_1),\psi(X_2)} = \kappa_{X_1,X_2}$;

6. Caso Normal: si la distribución conjunta de X_1 y X_2 es normal bivariante, con coeficiente de correlación r, entonces κ_{X_1,X_2} es una función estricta (ϕ) de $|r|$;

7. Continuidad: Si (X_1, X_2) y $\{(X_{1n}, X_{2n})\}$ son pares de variables aleatorias con funciones de distribución F_{12} y F_{1n2n} respectivamente y si la sucesión F_{1n2n} converge débilmente a F_{12}, entonces $\lim_{n\to+\infty} \kappa_{X_1nX_2n} = \kappa_{X_1,X_2}$.

Estas medidas de dependencia monótona no son medidas de concordancia (Def. 2.2.3), pero satisfacen el converso de la propiedad de independencia. Schweizer and Wolff (1981) propone algunos ejemplos de estas medidas, como

$$\sigma_{X_1,X_2} = 12 \int_{[0,1]^2} |C_{UV}(u,v) - uv| dudv \ , o$$

$$\gamma_{X_1,X_2} = \left(90 \int_{[0,1]^2} (C_{UV}(u,v) - uv)^2 dudv \right)^{1/2} ,$$

y sugiere que cualquier medida de distancia, convenientemente normalizada, entre la superfície $z = C_{UV}$ y $z = uv$, p.ej. cualquier distancia L_p, debería proporcionar una medida de dependencia no paramétrica y simétrica.

Existen numerosas medidas de asociación no dependientes de las marginales, es decir, cuya expresión puede reescribirse mediante funciones cópula. Algunas de estas medidas de asociación son medidas de concordancia; otras medidas de dependencia, aunque algunas de ellas no satisfacen las propiedades ni de unas ni de otras. La Tabla 2.1 presenta algunas medidas de asociación conocidas, así como sus expresiones mediante cópulas. Se muestra la expresión del coeficiente en el caso bidimensional. Trabajos como (Schmid and Schmidt, 2007; Schmid et al., 2010) muestran los equivalentes multidimensionales de algunos coeficientes como τ de Kendall (ec. 2.2.6), ρ de Spearman (ec. 2.2.5), β de Blomqvist o γ de Gini. Los dos primeros coeficientes son los más utilizados como alternativa al coeficiente de Pearson, pero otros menos conocidos, como β de Blomqvist (Blomqvist, 1950) se han presentado como alternativas a la Ec. 2.2.5 para la estimación de

parámetros de cópulas (Genest et al., 2013). En particular, para cópulas arquimedianas, β está relacionado con el generador de la cópula:

$$\beta = -1 + 4\varphi'\left(2\varphi\left(\frac{1}{2}\right)\right) \quad , \tag{2.2.6}$$

Este coeficiente β, también denominado coeficiente de correlación medial, puede interpretarse como una diferencia normalizada entre la cópula C y la cópula de la independencia Π en $(1/2, 1/2)$, y tiene expresión cerrada para muchas de las familias de cópula más populares.

Definición 2.2.9. (Joe, 1989) El *coeficiente* δ^*,

$$\delta^* = [1 - \exp(-2\delta)]^{(1/2)} \ ,$$

es una transformación de la entropía cruzada (def. 2.6.2),

$$\delta = \int_{\mathbb{R}^2} \log\left(\frac{f_{12}}{f_1 f_2}\right) dF_{12}$$

.

Proposición 2.2.4. *(Joe, 1989) El coeficiente δ^* es un coeficiente de dependencia, pero no de concordancia.*

Este coeficiente satisface una generalización de la desiderata de condiciones para conceptos de dependencia bivariada enunciado por Rényi:

Proposición 2.2.5. *(Joe, 1989) Una medida de dependencia debe cumplir que:*

a) Es nula si las variables son independientes o condicionalmente independientes. En caso contrario la medida tendrá valores entre 0 y 1.

b) Es invariante bajo transformaciones uno-a-uno separadas de las variables.

c) Está definida tanto para variables categóricas como continuas (...)

d) Es equivalente al coeficiente de correlación para una variable normal bivariada .

Observación 2.2.3. La entropía relativa (def. 2.6.2) satisface sólo las propiedades b) y c).

Esta es una generalización de los postulados de Rényi para dependencia bivariada (Schweizer and Wolff, 1981), pero no se requiere que las medidas estén definidas para todo vector aleatorio. Esta simplificación de las condiciones para medidas de dependencia facilita la extensión de estas medidas a variables no continuas y/o el caso multivariante. Es más, Joe (1989) insiste en que esta nueva medida de dependencia debiera proporcionar valores útiles para resumir la dependencia no monótona o no lineal (si la dependencia es monótona, la que miden las medidas de concordancia, el valor del coeficiente será similar al de otros coeficientes como la τ de Kendall, pero si la dependencia no es monótona, los valores de ambos coeficientes seran diferentes).

Proposición 2.2.6. *Los coeficientes Type I y Type II son generalizaciones de las medidas de asociación más conocidas. Permiten representar tipos de dependencia que usualmente son difíciles de reproducir.*

Otra noción de asociación entre variables aleatorias, pero apenas utilizada, son las *medidas de concordancia copulares*, definidas en Edwards et al. (2004):

Definición 2.2.10. (Edwards et al., 2004) Una *medida de concordancia copular* es una medida de la forma $\kappa_C = \alpha \int_{[0,1]^2} C dA - \beta$, donde A es una cópula fijada y $\alpha, \beta \in \mathbb{R}$.

Ejemplo 2.2.1. El coeficiente ρ de Spearman $\left(12 \int_{[0,1]^2} C \; d\Pi - 3\right)$ y la γ de Gini
$\left(12 \int_{[0,1]^2} C \; d((M+W)/2) - 2\right)$ son medidas de concordancia copulares, mientras que τ de Kendall $\left(4 \int_{[0,1]^2} C dC - 1\right)$ no es una de estas medidas.

Versiones muestrales de algunas medidas

Los rangos son estadísticos maximalmente invariantes para las funciones cópula (Genest and Plante, 2003). Por tanto, las versiones muestrales de medidas de asociación expresables mediante cópulas, es decir, no dependientes de marginales, suelen depender de rangos.

Definición 2.2.11. Un estimador muestral de la correlación por rangos de Spearman $\rho_S(X_1, X_2)$ (Def. 2.2.5) viene dado por:

$$\widehat{\rho_S}(X_1, X_2) = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \left(R_i - \frac{n+1}{2} \right) \left(S_i - \frac{n+1}{2} \right) \ ,$$

o equivalentemente (Genest and Favre, 2007),

$$\widehat{\rho_S}(X_1, X_2) = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} R_i S_i - 3\frac{n+1}{n-1} \ .$$

Definición 2.2.12. (Kruskal, 1958) Un estimador muestral de la correlación por rangos de Kendall (Def. 2.2.6) es

$$\widehat{\tau} = \frac{c_n - d_n}{\binom{n}{2}} = \frac{4}{n(n+1)} c_n - 1 \ ,$$

donde c_n y d_n denotan el número de pares concordantes y discordantes en la muestra, respectivamente. En caso de variables contínuas, los empates aparecerán con probabilidad nula. Es necesario observar que $\widehat{\tau}$ es función de los rangos de las observaciones, dado que $(x_1^i - x_1^{*j})(x_2^i - x_2^{*j}) > 0$ si y sólo si $(R_i - R_j^*)(S_i - S_j^*) > 0$.

La Tabla 2.2 incluye las expresiones muestrales de algunas medidas de asociación usuales.

En este apartado se han introducido coeficientes que permiten describir diferentes tipos de dependencia entre variables aleatorias. Se ha mostrado la relación entre cópulas y coeficientes de dependencia, reafirmando la utilidad de estas funciones como transmisoras de la estructura de dependencia y de sus matices. En los siguientes apartados se introducen los procesos de Poisson y las distribuciones de máximos y excesos, útiles para modelizar el contexto en el que las cópulas se utilizarán más adelante.

Tabla 2.1: Algunas medidas de asociación usuales y sus expresiones mediante cópula.

Coeficiente	Expresión mediante cópulas
Spearman's ρ (Spearman, 1904)	$3\int_{[0,1]^2}[(1-u-v)^2-(u-v)^2]dC_{UV}$
Spearman's ρ	$12\int_{[0,1]^2}(C_{UV}-uv)dudv$
Kendall's τ (Kendall, 1938)	$-1+4\int_{[0,1]^2}C[u,v]dC[u,v]$
Blest (I) (Blest, 2000)	$2-12\int_{[0,1]^2}(1-u)^2v\ dC_{UV}$
Blest simetrizada (Genest and Plante, 2003)	$-4-6\int_{[0,1]^2}uv(4-u-v)\ dC_{UV}$
Gini's γ (Nelsen, 1998)	$2\int_{[0,1]^2}(\lvert 1-u-v\rvert-\lvert u-v\rvert)\ dC_{UV}$
Blomqvist β (Blomqvist, 1950)	$-1+4C[1/2,1/2]$
Spearman's footrule φ (Nelsen, 1998)	$1-3\int_{[0,1]^2}\lvert u-v\rvert dC_{UV}$
K-L Cross Entropy (Kullback and Leibler, 1951)	$\int_{[0,1]^2}\log(C[u,v])dC[u,v]$
Type I (Prop.2.2.6):	$\int_{[0,1]^2}C_{UV}\cdot c_{UV}\ dC_{UV}$
Type II (Prop. 2.2.6):	$\int_{[0,1]^2}uv\cdot c_{UV}\ dC_{UV}$

Tabla 2.2: Algunas medidas de asociación usuales y sus expresiones muestrales

Coeficiente	Expresión muestral
Spearman's ρ	$r_S = 1 - \frac{6}{n(n^2-1)} \sum (R_i - Q_i)^2$
Spearman's ρ	$r_S = \frac{3}{n(n^2-1)} \left[\sum (n+1-R_i-S_i)^2 - \sum (R_i - Q_i)^2 \right]$
Spearman's ρ	$r_S = \frac{12}{n(n^2-1)} \sum (R_i \cdot S_i) - 3\frac{n+1}{n-1}$
Kendall's τ	$\tau_n = \frac{2}{n^2-n} \sum_{1 \le i < j \le n} sign(R_i - R_j) sign(S_i - S_j)$
Blest (I)	$\nu_n = \frac{2n+1}{n-1} - \frac{12}{n^2-n} \sum_{i=1}^{n} \left(1 - \frac{R_i}{n+1}\right)^2 S_i$
Blest simetrizada	$\xi_n = -\frac{4n+5}{n-1} + \frac{6}{n(n^2-1)} \sum_{i=1}^{n} R_i S_i \left(4 - \frac{R_i+S_i}{n+1}\right)$
Gini's γ	$g = \frac{1}{[n^2/2]} \left[\sum |n+1-R_i-S_i| - \sum |R_i - S_i| \right]$
Blomqvist β	$-1 + 2\frac{n_1}{n1+n2}$, n_1 # points both components greater or lower than medians, $n_2 = 1 - n_1$
Spearman's footrule φ	$f_S = 1 - \frac{3}{n^2-1} \sum |R_i - S_i|$
K-L Cross Entropy	$n^{-1} \sum_{j=1}^{n} \log(\widehat{c}(R_{ij}))$

2.3. Procesos de Poisson

Con frecuencia se desea establecer un modelo estocástico que capture las características principales de fenómenos que ocurren a lo largo del tiempo (lluvias, viento, etc) y que permita el cálculo de cantidades de interés. Los procesos estocásticos en general, y en particular los procesos de Poisson (evaluados o no) son modelos aplicables en ese contexto. En este apartado se introducirán algunas definiciones básicas y propiedades de los procesos de Poisson. Estas son necesarias para el desarrollo metodológico posterior, pero no se pretende ser exhaustivos. Se puede hallar una descripción más amplia de los procesos puntuales en Daley and Vere-Jones (2003); Grandell (1997); Embrechts et al. (1997), entre otros.

Definición 2.3.1. Un *proceso estocástico*, formalmente denotado como $\{X(t),\, t \in T\}$, es una sucesión de variables aleatorias $X(t)$, donde el parámetro t, usualmente el tiempo, tiene valores en un conjunto de índices T. El espacio de estados es el conjunto de posibles valores para las variables aleatorias $X(t)$. Si el conjunto T es contable, el proceso estocástico es discreto; si T es continuo, el proceso estocástico es continuo.

Ejemplo 2.3.1. El resultado de n lanzamientos de una moneda es un proceso estocástico discreto con posibles resultados {cara,cruz} y conjunto de índices $T = \{1, 2, \ldots, n\}$. El numero de llegadas de paquetes en un router durante un cierto intervalo de tiempo $[a, b]$ es un proceso estocástico continuo porque $t \in [a, b]$. Otros ejemplos de procesos estocásticos son la medida de la temperatura cada día o el registro del valor de un stock bursátil cada minuto.

Definición 2.3.2. En un proceso estocástico continuo, se definen los incrementos como $X(t) - X(u)$. Un proceso estocástico continuo tiene *incrementos independientes* si cambios en el valor del proceso en diferentes intervalos de tiempo son independientes. Un proceso estocástico continuo tiene *incrementos estacionarios* si $X(t + s) - X(s)$ tiene la misma distribución para cualquier desplazamiento s.

Los procesos estocásticos pueden clasificarse según su espacio de estados, según su conjunto de índices T o según las relaciones de dependencia entre las variables aleatorias $X(t)$.

Ejemplo 2.3.2. Un proceso de Poisson, definido a continuación, es un proceso estocástico $N(t) \geq 0$ con realizaciones no continuas, incrementos estacionarios e independientes y con $N(t)$ distribuida según una Poisson. Una generalización del proceso de Poisson es un proceso contador.

Definición 2.3.3. Un *Proceso de Poisson* con parámetro o tasa $\lambda > 0$ es un proceso estocástico en tiempo continuo y con valores enteros, $\{N(t), t \geq 0\}$, tal que:

a) $N(0) = 0$.

b) para todo $t_0 = 0 < t_1 < \ldots < t_n$, los incrementos $N(t_1) - N(t_0), N(t_2) - N(t_1), \ldots, N(t_n) - N(t_{n-1})$ son variables aleatorias independientes

c) para $t \geq 0, s > 0$ y enteros no negativos k, los incrementos tienen distribución de Poisson de parámetro λ:

$$\mathrm{P}[N(t+s) - N(s) = k] = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \quad k = 0, 1, \ldots$$

Observación 2.3.1. El proceso de Poisson N(t) es un proceso contador especial, donde el número de sucesos en cualquier intervalo de longitud t se puede especificar mediante la condición c). De la condición c) se deduce también que los incrementos son estacionarios.

Teorema 2.3.1. *Sea $\{N(t), t \geq 0\}$ un proceso de Poisson de parámetro $\lambda > 0$ y sean $t_0 = 0 < t_1 < \ldots < t_n$ los tiempos de ocurrencia sucesivos de sucesos. Entonces, los tiempos entre sucesos, $\tau_n = t_n - t_{n-1}$, $n = 1, \ldots$ son idependientes e idénticamente distribuidos, con distribución exponencial de media $1/\lambda$.*

Observación 2.3.2. El converso del Teorema anterior también se cumple. Si los tiempos entre sucesos $\{\tau_n\}$ de un proceso contador $\{N(t), t \geq 0\}$ son variables aleatorias exponenciales i.i.d. con media $1/\lambda$, entonces $\{N(t), t \geq 0\}$ es un procesos de Poisson de parámetro λ.

El proceso de Poisson satisface algunas propiedades de conservación que resultan útiles en aplicaciones:

Proposición 2.3.2. *Si $N(t)$ y $M(t)$ son dos procesos de Poisson independientes con parámetros λ_1 y λ_2 respectivamente, entonces $Z(t) = N(t) + M(t)$ es un proceso de Poisson de parámetro $\lambda_1 + \lambda_2$.*

Proposición 2.3.3. *Dado un proceso de Poisson de parámetro λ, $\{N(t), t \geq 0\}$, supongamos que cada suceso se clasifica en la clase i con probabilidad $p_i \in (0,1)$, $i = 1, \ldots K$. Sea $N_i(t)$ el numero de sucesos de tipo i en el intervalo $(0,t]$. Entonces, $\{N_1(t), t \geq 0\}, \ldots, \{N_K(t), t \geq 0\}$ son procesos de Poisson independientes, con parámetros $\lambda p_1, \ldots, \lambda p_K$, respectivamente.*

A cada uno de los sucesos en el tiempo dados por un proceso de Poisson se le puede asociar un *tamaño*, el valor de una variable aleatoria X en ese punto, dando lugar a un *proceso evaluado* o marcado (por ejemplo, se indica el valor de la precipitación en 24h., tamaño, en un día lluvioso, suceso). A continuación, presentamos algunas definiciones básicas que permiten modelizar conjuntamente la ocurrencia de los fenómenos en el tiempo y el tamaño de cada suceso:

El modelo de Cramér-Lundberg, (Grandell, 1997), engloba al modelo de Poisson:

Definición 2.3.4. (Modelo de Cramér-Lundberg) Un proceso puntual evaluado es un modelo de Cramér-Lundberg si:

a) Los tamaños $(X_k)_{k \in \mathbf{N}}$ son variables aleatorias positivas independientes, idénticamente distribuidas (i.i.d.) con función de distribución común F_X.

b) Los sucesos aparecen en los instantes aleatorios de tiempo $0 < T_1 < T_2 < \ldots$

c) El número de sucesos en el intervalo $[0, t_u]$ se denota por

$$N(t) = \sup\{n \geq 1 : T_n \leq t_u\} \ , \ t_u \geq 0 \ ,$$

usando como convención $\sup\{\emptyset\} = 0$.

d) Los tiempos entre llegadas,

$$Y_1 = T_1 \ , \ Y_k = T_k - T_{k-1} \ , \ k = 2, 3, \ldots ,$$

son i.i.d., exponencialmente distribuidos y con media finita $\mathrm{E}[Y_i] = 1/\lambda$.

e) Las sucesiones (X_k) e (Y_k) son independientes entre sí.

Definición 2.3.5. Se llama *proceso contador o proceso de contaje* a:

$$S(t) = S_{N(t_u)} = \begin{cases} 0 & \text{si } \mathrm{N(t_u)} = 0 \ , \\ X_1 + \cdots + X_{N(t)} & \text{si } \mathrm{N(t_u)} \geq 1 \ , \end{cases} \quad t_u \geq 0 \ ,$$

donde $\{N(t_u), t_u \geq 0\}$ es un proceso estocástico en $[0, +\infty)$ tal que las variables aleatorias $N(t)$ son no negativas y con valores enteros y $(X_k)_{k\in\mathbb{N}}$ son sus tamaños.

Una consecuencia de la definición del modelo de Cramer-Lundberg es que $\{N(t_u)\}$ es un proceso homogéneo de Poisson con intensidad $\lambda > 0$:

Definición 2.3.6. Un *proceso de Poisson homogéneo con parámetro de intensidad* λ es un modelo de Cramér-Lundberg (Def. 2.3.4), considerando (X_n) y $\{N(t_u)\}$ independientes, es decir, es un proceso contador (Def. 2.3.5) tal que $T_n = Y_1 + \cdots + Y_n, \ n \geq 1$, y (Y_n) (los tiempos entre ocurrencias de los sucesos) son variables aleatorias i.i.d., exponenciales, con esperanza $1/\lambda$.

Si $\{N(t_u)\}$ y (X_n) son independientes, entonces el proceso $\{S(t_u), t_u \geq 0\}$ se denomina un *proceso de Poisson evaluado o marcado.*

Definición 2.3.7. El tiempo esperado entre sucesos se llama *periodo de retorno o de recurrencia.* En el caso en que este tiempo entre sucesos tenga una distribución exponencial, el periodo de retorno es $\tau = \lambda^{-1}$.

Propiedades

Consideremos un proceso de Poisson evaluado, es decir, en un tiempo t_u se producen N sucesos, de forma que $N \sim Poisson(\lambda)$, y cada suceso tiene un tamaño X que se supone independiente del proceso de ocurrencia de sucesos y de otros tamaños. Los tamaños están idénticamente distribuidos según $F_X(x)$.

Proposición 2.3.4. *La distribución del máximo de los tamaños X en un tiempo t_u viene dada por la expresión*

$$F_Z(z) =$$
$$\mathrm{P}[Z \leq z|N=0] \cdot \mathrm{P}[N=0|\lambda, t_u] + \sum_{n=1}^{\infty} F_Z(z|N=n) \cdot \mathrm{P}[N=n|\lambda, t_u] =$$
$$\exp[-\lambda t_u + F_X(z)\lambda t_u] = \exp[-\lambda t_u(1-F_X(z))] \quad . \tag{2.3.1}$$

Proposición 2.3.5. *Sean Y los excesos sobre un umbral h, $Y = X - h$, $h < X < +\infty$. Si $N(h)$ es el número de sucesos que exceden h en un tiempo t_u, se tiene que*

$$\mathrm{P}[N(h) = n|\ t_u, \lambda, h] = \frac{[\lambda(1-F_Y(h))]^n \exp(-\lambda(1-F_Y(h)))}{n!} \ ,$$
$$n = 0, 1, 2, \ldots \tag{2.3.2}$$

es decir, los excesos sobre un umbral h son un nuevo proceso de Poisson con parámetro λ_h, $\lambda_h = \lambda(1 - F_Y(h))$, donde la distribución de los excesos, $F_Y(y)$, corresponde a

$$F_Y(y) = \mathrm{P}[Y \leq y|X > h] = \frac{F_X(y+h) - F_X(h))}{1 - F_X(h)} \quad ,$$
$$y > 0,\ 0 < h < y + h \quad .$$

Para un umbral h, el periodo de retorno de los sucesos que superen este umbral vendrá dado por la expresión $\tau(h) = \lambda(h)^{-1}$.

Los procesos de Poisson evaluados proporcionan un modelo conjunto de ocurrencia y tamaño para fenómenos naturales. En el apartado siguiente se introducen conceptos que permiten una descripción más adecuada de los tamaños en el caso en que éstos sean extremales.

2.4. Distribuciones de excesos y máximos

Una vez caracterizada la ocurrencia de los sucesos de interés mediante un proceso de Poisson, debemos centrar la atención en la modelización de los tamaños correspondientes. En este apartado caracterizaremos el tamaño de un suceso de interés en el caso en que éste sea extremal. Es decir, describiremos las distribuciones del máximo y del exceso sobre un umbral de una variable aleatoria. Una caracterización más amplia de estas distribuciones puede encontrarse en Castillo (1988), Castillo et al. (2004), Embrechts et al. (1997), Kotz and Nadarajah (2000) entre otros.

Sean (X_1, X_2) dos variables aleatorias, denominadas tamaños. Estas variables se pueden describir mediante su función de distribución conjunta $F_{X_1X_2}(x_1, x_2)$ y sus marginales, $F_{X_1}(x_1)$, $F_{X_2}(x_2)$. Utilizaremos la notación (Y_1, Y_2) para nombrar a los excesos de las variables (X_1, X_2) sobre un umbral bidimensional (h_1, h_2), con función de distribución conjunta $F_{Y_1Y_2}(y_1, y_2)$ y marginales $F_{Y_1}(y_1)$, $F_{Y_2}(y_2)$.

Basándonos en el Teorema de Pickands, (Castillo, 1988; Davison and Smith, 1990), consideraremos que las funciones de distribución marginales $F_{Y_1}(y_1)$, $F_{Y_2}(y_2)$ son distribuciones generalizadas de Pareto, *GPD*. El teorema de Pickands, (Pickands III, 1975), afirma que para un umbral suficientemente alto, $h_1 > h_0$, los excesos de la variable X sobre h_1, dado que $X > h_1$, con notación $Y = X - h_1$, tienen aproximadamente distribución generalizada de Pareto, $GPD(\xi, \beta)$.

Definición 2.4.1 (Distribución generalizada de Pareto)**.** Sea la distribución G_ξ, definida por

$$G_\xi(x) = \begin{cases} 1 - (1+\xi x)^{-1/\xi} & \text{si } \xi \neq 0 \\ 1 - \exp\{-x\} & \text{si } \xi = 0 \end{cases},$$

para

$$\begin{array}{cc} x \geq 0 & \text{si } \xi \geq 0 \\ 0 \leq x \leq -1/\xi & \text{si } \xi < 0 \end{array}.$$

G_ξ se denomina distribución generalizada de Pareto estándard (GPD). Se puede introducir la familia de escala y posición relacionada, $G_{\xi,\nu,\beta}$, sustituyendo el argumento x en la definición anterior por $(x-\nu)/\beta$, para parámetros $\nu \in \mathbb{R}$ y $\beta > 0$. El dominio ha de ser ajustado

convenientemente. $G_{\xi,\nu,\beta}$ recibe el nombre de GPD, aunque en general, la distribución denominada $GPD(\xi,\beta)$ corresponde al parámetro $\nu = 0$:

$$G_{\xi,\beta}(x) = 1 - \left(1 + \frac{\xi x}{\beta}\right)^{\frac{-1}{\xi}} \text{ , para } \begin{cases} x \geq 0 & \text{si } \xi \geq 0 \\ 0 \leq x \leq -\beta/\xi & \text{si } \xi < 0 \end{cases} .$$

Dependiendo de los valores del parámetro ξ, la distribución corresponderá a uno de los tres dominios de atracción posibles:

- Si $\xi < 0$, la distribución tiene un soporte limitado superiormente $0 < y < -\beta/\xi$. Diremos que la distribución pertenece al dominio de atracción de Weibull.
- Si $\xi = 0$, la distribución pertenece al dominio de atracción de Gumbel, para $0 < y < +\infty$.
- Si $\xi > 0$, la distribución pertenece al dominio de atracción de Fréchet, para $0 < y < +\infty$.

Por tanto, las funciones de distribución marginales de los excesos sobre un umbral h_i, $i = 1, 2$, $F_{Y_1}(y_1)$ y $F_{Y_2}(y_2)$, que según el teorema de Pickands están distribuidos según una GPD, tendrán la expresión:

$$F_{Y_i}(y_i) = 1 - \left(1 + \frac{\xi y_i}{\beta}\right)^{\frac{-1}{\xi}} \text{ , } 0 < y < y_{sup} \text{ ,}$$

donde la cota superior y_{sup} dependerá de las características del tamaño modelizado. Si el soporte del tamaño X_i es limitado, entonces el soporte de la distribución de los excesos sobre umbral también debe serlo, y por tanto consideraremos que la variable Y_i tiene una distribución GPD en el dominio de Weibull. La hipótesis de GPD en el dominio de Weibull es usual al modelizar precipitación diaria, velocidad del viento o altura de ola significante en una boya, entre muchos otros tamaños.

Por otro lado, se desea modelizar también los máximos de las variables (X_1, X_2) registrados en un intervalo de tiempo t_0, (Z_1, Z_2), con función de distribución conjunta $F_{Z_1 Z_2}(z_1, z_2)$ y marginales $F_{Z_1}(z_1)$, $F_{Z_2}(z_2)$.

Consideraremos que las marginales $F_{Z_i}(z_i)$, $i = 1, 2$, son distribuciones generalizadas de extremos, $GEVD$, a partir del teorema de Fisher-Tippet y la representación de von Mises (Castillo, 1988; Embrechts et al., 1997; Kotz and Nadarajah, 2000):

Teorema 2.4.1. *(Teorema de Fisher-Tippet, distribuciones límite para máximos)*

Sea (X_n) una sucesión de variables aleatorias i.i.d. Si existen constantes normalizadoras $c_n > 0$, $d_n \in \mathbb{R}$ y alguna función de distribución no degenerada H tal que

$$c_n^{-1}(M_n - d_n) \to^d H \quad ,$$

entonces H pertenece a alguno de los siguientes tipos de funciones de distribución:

$$\textit{Fréchet:} \quad \Phi_\alpha(x) = \begin{cases} 0 & x \leq 0 \\ & \alpha > 0 \\ \exp\{-x^{-\alpha}\} & x > 0 \end{cases}$$

$$\textit{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp\{-(-x^\alpha)\} & x \leq 0 \\ & \alpha > 0 \\ 1 & x > 0 \end{cases}$$

$$\textit{Gumbel:} \quad \Lambda(x) = \exp\{-\exp\{-x\}\} \ , \quad x \in \mathbb{R}.$$

En adelante será útil utilizar una representación que englobe en una sola familia las tres posibles distribuciones límite: Gumbel, Fréchet y Weibull. Para ello, se introduce un parámetro ξ tal que

$\xi = \alpha^{-1} > 0$ corresponde a la distribución de Fréchet Φ_α ,
$\xi = 0$ corresponde a la distribución de Gumbel Λ ,
$\xi = -\alpha^{-1} < 0$ corresponde a la distribución de Weibull Ψ_α .

Definición 2.4.2 ($GEVD$, representación de Jenkinson- von Mises de las distribuciones de valor extremo)**.** Se define la función de distribución H_ξ como

$$H_\xi(x) = \begin{cases} \exp\{-(1+\xi x)^{-1/\xi}\} & \xi \neq 0 \\ \exp\{\exp\{-x\}\} & \xi = 0 \end{cases} \quad ,$$

donde $1 + \xi x > 0$. Por tanto, el dominio de H_ξ corresponde a

$$x > -\xi^{-1} \ \text{ para } \ \xi > 0 \ ,$$

$$x < -\xi^{-1} \ \text{ para } \ \xi < 0 \ ,$$

$$x \in \mathbb{R} \quad \text{para } \ \xi = 0 \ .$$

La familia equivalente $H_{\xi,u,\psi}$ se puede introducir sustituyendo el argumento x anterior por $(x-u)/\psi$, para parámetros $u \in \mathbb{R}, \psi > 0$. El

parámetro u corresponde al parámetro de localización, $-\infty < u < +\infty$; ξ es el parámetro de forma y ψ es el parámetro de escala, $\psi > 0$. La distribución con la nueva parametrización, $H_{\xi,u,\psi}$, también se denomina $GEVD$:

$$H_{\xi,u,\psi}(x) = \exp\left[-\left(1+\xi\left(\frac{x-u}{\psi}\right)\right)^{\frac{-1}{\xi}}\right] \quad ,$$

para $1 + \xi(x-u)/\psi > 0$.

Las tres distribuciones de máximos existentes se obtienen en función de los valores del parámetro de forma:

- Para $\xi < 0$ se obtiene la distribución de Weibull, siempre que el soporte sea acotado superiormente, $-\infty < z \leq -\frac{\psi}{\xi} + u$;
- Para $\xi = 0$ se obtiene la distribución de Gumbel, para $-\infty < z < +\infty$;
- Para $\xi > 0$ se obtiene la distribución de Fréchet , $-\frac{\psi}{\xi} + u \leq z < +\infty$.

Las distribuciones marginales de los máximos de las variables (X_1, X_2) registrados en un intervalo de tiempo t_0, $F_{Z_i}(z_i)$, $i = 1, 2$, que según el teorema de Fisher-Tippet tienen aproximadamente distribución GEVD, serán de la forma:

$$F_{Zi}(z_i) = \exp\left[-\left(1+\xi\left(\frac{z_i-u}{\psi}\right)\right)^{\frac{-1}{\xi}}\right] \quad ,$$

definida para valores de z que cumplan $1 + \xi(z_i - u)/\psi > 0$. Si el soporte del tamaño X_i es limitado superiormente, entonces el soporte de la distribución del máximo en un intervalo de tiempo también debe serlo. En estos casos se realiza la hipótesis suplementaria de distribución GEVD-Weibull.

La descripción de los extremos de las variables introducida será de mucha utilidad en el contexto de los procesos de Poisson evaluados, presentados anteriormente. A continuación se introduce la relación entre la distribución GPD y la $GEVD$, la cual permite una simplificación del proceso de estimación de estos modelos extremales.

2.4.1. Relación entre GPD y $GEVD$

Para el desarrollo metodológico posterior es necesario relacionar de una manera sencilla la distribución de excesos sobre un umbral, GPD, y la distribución de máximos, GEVD, en un intervalo de tiempo. Dado que esta relación no se ha hallado de manera explícita en la bibliografía consultada, se detalla a continuación el proceso de obtención:

En un proceso evaluado de Poisson de parámetro λ (Def. 2.3.6), suponemos que los tamaños, X, tienen distribución generalizada de Pareto, $GPD(\xi, \beta)$ (Def. 2.4.1), dada por la expresión:

$$F_X(x|\xi,\beta) = 1 - \left(1 + \frac{\xi x}{\beta}\right)^{\frac{-1}{\xi}} \ , \ 0 < y < y_{sup} \ ,$$

para valores superiores a uno de referencia, $x \geq x_0$. Utilizando la expresión 2.3.1, la distribución de máximo de estos tamaños en un tiempo t_0, denotado por Z, será:

- Si $\xi \neq 0$, $F_Z(z|\xi,\beta) = \exp\left[-\lambda t_0 \left(1 + \frac{\xi}{\beta}\right)^{\frac{-1}{\xi}}\right]$, para $0 < z < -\beta/\xi$, o bien
  $0 < z < +\infty$ (dependiendo del dominio de atracción que corresponda);
- Si $\xi = 0$, $F_Z(z|\xi,\beta) = \exp\left[-\lambda t_0 \exp\left(\frac{-z}{\beta}\right)\right]$, para $0 < z < +\infty$.

Proposición 2.4.2. *Sea un un Proceso de Poisson de parámetro λ, evaluado con tamaños $X \sim GPD(\xi, \beta)$. La distribución del máximo tamaño, Z en un tiempo t_0 se puede aproximar por una distribución $GEVD(u^*, \xi^*, \psi^*)$, donde se mantiene el parámetro de forma de la GPD y varían los parámetros de localización y escala, es decir,*

$$\xi = \xi^* \ ,$$

$$\frac{\xi}{\beta(\lambda t_0)^{\xi}} = \frac{\xi^*}{\psi^*} \ \Rightarrow \psi^* = \beta(\lambda t_0)^{\xi} \ ,$$

$$u^* = \frac{\beta((\lambda t_0)^{\xi} - 1)}{\xi} \ .$$

Demostración. Dado que el objetivo es hallar la relación entre la distribución de excesos $GPD(\xi, \beta)$ y la distribución generalizada de extremos, se compara el resultado obtenido con la distribución generalizada de extremos de parámetros u^*, ξ^* y ψ^*, $GEVD(u^*, \xi^*, \psi^*)$, cuya expresión es:

- Si $\xi^* \neq 0$, $F_Z(z|u^*, \xi^*, \psi^*) = \exp\left[-\left(1 + \frac{\xi^*(z-u^*)}{\psi^*}\right)^{\frac{-1}{\xi^*}}\right]$, para valores tales que $1 + \frac{\xi^*(z-u^*)}{\psi^*} > 0$,
- Si $\xi^* = 0$, $F_Z(z|u^*, \xi^* = 0, \psi^*) = \exp\left[-\left(\exp\left(\frac{(z-u^*)}{\psi^*}\right)\right)\right]$, para valores tales que $1 + \frac{\xi^*(z-u^*)}{\psi^*} > 0$.

Operamos con la expresión de $F_Z(z|\xi, \beta)$, de manera que podamos identificar los parámetros de ambas distribuciones:

$$F_Z(z|\xi, \beta) = \exp\left[-\left(1 - 1 + (\lambda t_0)^{-\xi} + \frac{(\lambda t_0)^{-\xi}\xi z}{\beta}\right)^{\frac{-1}{\xi}}\right] =$$

$$\exp\left[-\left(1 + \frac{\xi}{\beta(\lambda t_0)^{\xi}}\left(z - \frac{(\lambda t_0)^{-\xi}\beta}{\xi} + \frac{\beta}{\xi}\right)^{\frac{-1}{\xi}}\right)\right] .$$

Por tanto, la relación entre los parámetros de la distribución de Pareto, $GPD(\xi, \beta)$, y la de máximos, $GEVD(u^*, \xi^*, \psi^*)$, viene dada por las expresiones:

$$\xi = \xi^* ,$$

$$\frac{\xi}{\beta(\lambda t_0)^{\xi}} = \frac{\xi^*}{\psi^*} \Rightarrow \psi^* = \beta(\lambda t_0)^{\xi} ,$$

$$u^* = \frac{\beta((\lambda t_0)^{\xi} - 1)}{\xi} ,$$

es decir, se mantiene el parámetro de forma y varían los parámetros de localización y escala. □

2.5. Paradigma Bayesiano

El propósito principal de la teoría estadística es obtener una inferencia sobre la distribución de probabilidad de las observaciones de un fenómeno aleatorio. Es decir, proporciona un análisis o descripción de un fenómeno pasado y/o algunas predicciones sobre un suceso futuro de una naturaleza similar.

Cuando se observa un fenómeno aleatorio determinado, X, distribuido según un modelo que consideraremos dependiente de un parámetro $\boldsymbol{\theta}$, $F_X(\cdot|\boldsymbol{\theta})$, los métodos estadísticos permiten obtener de las observaciones una inferencia sobre $\boldsymbol{\theta}$, mientras que la modelización probabilista caracteriza el comportamiento de las observaciones futuras condicionadas al parámetro $\boldsymbol{\theta}$. A continuación se presentan algunos conceptos básicos de este enfoque, denominado Bayesiano. Una visión más amplia puede hallarse en Bernardo and Smith (1994); Lee (1997); Robert (1994).

El teorema de Bayes clásico es la base del enfoque bayesiano:

Teorema 2.5.1 (Teorema de Bayes)**.** *Si A y B son sucesos tales que $\mathrm{P}(B) \neq 0$, $\mathrm{P}(A|B)$ y $\mathrm{P}(B|A)$ están relacionadas por*

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B|A)\mathrm{P}(A) + \mathrm{P}(B|A^c)\mathrm{P}(A^c)} \quad .$$

Este teorema, además de una inversión de probabilidades puede verse como un principio de actualización, dado que describe la actualización de la probabilidad de A, $\mathrm{P}(A)$, a $\mathrm{P}(A|B)$, una vez que el suceso B ha sido observado.

La versión continua del teorema de Bayes se puede enunciar como :

Teorema 2.5.2. *Dadas dos variables aleatorias X_1 e X_2, con distribución condicional $f(x_1|x_2)$ y marginal $g(x_2)$, la distribución condicional de x_2 dado x_1 es*

$$g(x_2|x_1) = \frac{f(x_1|x_2)g(x_2)}{\int f(x_1|\eta)g(\eta)d\eta} \quad .$$

En el enfoque bayesiano los parámetros desconocidos del modelo, usualmente los parámetros de la distribución, se consideran aleatorios.

Sea $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ el conjunto de k valores desconocidos de un modelo, donde k puede ser mayor que 1. Los parámetros que determinan fenómenos aleatorios pueden ser percibidos también como variables aleatorias, los conocimientos o creencias que existen *a priori* sobre sus valores pueden ser expresados en términos de la función de probabilidad $\pi(\boldsymbol{\theta})$ sobre el espacio de parámetros Θ. Sea $\mathbf{X} = (X_1, \ldots, X_n)$ un conjunto de n observaciones, con distribución de probabilidad que depende de los k parámetros desconocidos, de tal modo que la función de densidad (continua o discreta) del vector $\mathbf{X}$ depende del vector $\boldsymbol{\theta}$ en una forma conocida, $f(\mathbf{x}|\boldsymbol{\theta})$.

En esta situación, la inferencia se basa en la distribución de $\boldsymbol{\theta}$ condicional a $\mathbf{X} = x$, $\pi(\boldsymbol{\theta}|\mathbf{x})$, denominada *distribución a posteriori*, (Lee, 1997; Robert, 1994), y definida por

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \quad .$$

Es necesario observar que $\pi(\boldsymbol{\theta}|\mathbf{x})$ es de hecho proporcional a la distribución de X condicionada a $\boldsymbol{\theta}$, es decir, la verosimilitud multiplicada por la distribución a priori de $\boldsymbol{\theta}$.

Debido a este teorema, podemos escribir

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) \quad .$$

Se puede considerar $f(\mathbf{x}|\boldsymbol{\theta})$ como una función de $\boldsymbol{\theta}$, y en este caso se denomina *función de verosimilitud*, utilizando la notación $\ell(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})$.

Con esta definición, la definición de $\pi(\boldsymbol{\theta})$ como la densidad a priori de $\boldsymbol{\theta}$, y $\pi(\boldsymbol{\theta}|\mathbf{x})$ como la densidad a posteriori para $\boldsymbol{\theta}$ dada $\mathbf{x}$, podemos pensar el teorema de Bayes como

$$\textit{posteriori} \propto \textit{priori} \times \textit{verosimilitud} \quad .$$

En términos estadísticos, el Teorema de Bayes actualiza la información sobre el parámetro $\boldsymbol{\theta}$ dada la información contenida en $\mathbf{x}$.

Un modelo estadístico Bayesiano está constituido por un modelo estadístico paramétrico $f(\mathbf{x}|\boldsymbol{\theta})$ y una distribución a priori sobre los parámetros, $\pi(\boldsymbol{\theta})$.

Todos los datos son útiles para actualizar la información, tal y como asegura el siguiente principio:

Proposición 2.5.3 (Principio de verosimilitud)**.** *La información aportada por una observación $\boldsymbol{X} = x$ sobre $\boldsymbol{\theta}$ está enteramente contenida en la función de verosimilitud $\ell(\boldsymbol{\theta}|\boldsymbol{x})$. Es más, si x_1 y x_2 son dos observaciones dependientes del mismo parámetro $\boldsymbol{\theta}$, tales que existe una constante c satisfaciendo*

$$\ell_1(\boldsymbol{\theta}|x_1) = c\ell_2(\boldsymbol{\theta}|x_2) \quad ,$$

para todo $\boldsymbol{\theta}$, ambas aportan la misma información sobre $\boldsymbol{\theta}$ y llevarán a inferencias idénticas.

Dado que el enfoque Bayesiano está basado por completo en la distribución a posteriori

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{\ell(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\int \ell(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \quad ,$$

la cual depende de $\mathbf{X}$ solamente a través de $\ell(\boldsymbol{\theta}|\mathbf{x})$, el principio de verosimilitud se satisface automáticamente en un esquema bayesiano.

Si se dispone de un modelo Bayesiano completo, es decir, se dispone de una distribución muestral $f(\mathbf{x}|\boldsymbol{\theta})$, y de una distribución a priori sobre $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, se pueden construir diversas distribuciones de utilidad: (a) La *distribución conjunta* de $(\boldsymbol{\theta}, \mathbf{X})$,

$$\varphi(\boldsymbol{\theta}, \mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \quad ;$$

(b) la *distribución marginal* de $\mathbf{X}$,

$$p(\mathbf{x}) = \int \varphi(\boldsymbol{\theta}, \mathbf{x})d\boldsymbol{\theta} = \int f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta} \quad ;$$

que recibe el nombre de *distribución predictiva* de $\mathbf{X}$, dado que representa las predicciones del valor de X teniendo en cuenta tanto la incertidumbre sobre el valor de $\boldsymbol{\theta}$ como la incertidumbre residual sobre $\mathbf{X}$ cuando $\boldsymbol{\theta}$ es conocido.

(c) la *distribución a posteriori* de $\boldsymbol{\theta}$, obtenida mediante la fórmula de Bayes,

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{f(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{p(\mathbf{x})} \quad .$$

De entre todas estas funciones, la herramienta básica del enfoque bayesiano es la distribución a posteriori.

En la práctica, este enfoque bayesiano permite incorporar al modelo utilizado la información aportada por las observaciones, mejorando las estimaciones e inferencias. De todos modos, los cálculos implicados en este enfoque no son siempre sencillos, y por ello es necesario introducir métodos que permitan solventar algunas de las dificultades que pueden aparecer. En particular, el cálculo de la distribución a posteriori de los parámetros, dadas las observaciones, puede resultar difícil o bien pesado, en detrimento del resultado final. A continuación se introduce el método de Gibbs, el cual permite hallar alternativas de cálculo de este posteriori.

2.5.1. Muestreo de Gibbs

El muestreo de Gibbs, o método de Gibbs, es quizá uno de los algoritmos de muestreo MCMC (Markov Chain Monte Carlo) más utilizados. Éste es un método iterativo que proporciona muestras con una determinada distribución conjunta. Resulta una herramienta útil si se aplica a la estimación de los parámetros de un modelo. En un contexto bayesiano (Sec. 2.5), se dispone del priori de los parámetros del modelo y es necesario explicitar el posteriori de éstos, dadas las observaciones. Este posteriori puede resultar difícil de calcular, o bien presentar una expresión complicada, dado que tendrá tantas dimensiones como parámetros intervengan. Sin necesidad de adoptar un modelo demasiado complicado, podemos encontrarnos con posterioris de dimensión 6 o 7, lo que puede implicar dificultades importantes de manejo. El método iterativo de Gibbs (Chen et al., 2000; Robert and Casella, 2000), permite realizar una estimación conjunta de los parámetros implicados en el modelo, utilizando únicamente verosimilitudes condicionales, mucho más simples y fáciles de manejar, dado que trata de densidades univariantes. Este enfoque presenta ventajas fundamentales:

– Se reducen dimensiones, dado que se utiliza la verosimilitud de cada parámetro condicionado a los demás, verosimilitudes unidimensionales. De este modo se simplifica enormemente la notación y el tratamiento de expresiones complejas.

– Se reemplaza la densidad conjunta a posteriori de los parámetros por una muestra simulada, manteniendo sus propiedades. A partir de esta muestra simulada se pueden realizar aproximadamente los mismos cálculos que con la propia densidad a posteriori.

A continuación se introducen algunos conceptos básicos, que pueden encontrarse ampliados en Chen et al. (2000); Robert and Casella (2000); Tanner (1993), entre otros.

Denotaremos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ a un vector p-dimensional de parámetros, y $\pi(\boldsymbol{\theta}|D)$ a su distribución posterior dados los datos D.

Para aplicar el método de Gibbs es preciso determinar las distribuciones a priori condicionadas respecto cada uno de los parámetros, $(\pi(\theta_1|\theta_2, \ldots, \theta_p), \pi(\theta_2|\theta_1, \ldots, \theta_p), \ldots, \pi(\theta_p|\theta_1, \ldots, \theta_{p-1}))$ y las verosimilitudes condicionales. Una vez calculadas, se procede a simular una muestra del posteriori mediante un proceso iterativo. El esquema básico del método de Gibbs es el siguiente (Chen et al., 2000):

Algoritmo de muestreo de Gibbs

- **Paso 0**. Se escoge un punto de partida arbitrario $\boldsymbol{\theta}_0 = (\theta_{1,0}, \ldots, \theta_{p,0})'$, y se coloca el contador $i = 0$.
- **Paso 1**. Dada la iteración $(i + 1)$, el vector de parámetros se denota $\boldsymbol{\theta}_{i+1} = (\theta_{1,i+1}, \ldots, \theta_{p,i+1})'$. En esta iteración se genera $\boldsymbol{\theta}_{i+1}$ de la manera siguiente:

  1. Se genera una realización $\theta_{1,i+1}$ a partir de la densidad condicional correspondiente,

  $$\theta_{1,i+1} \sim \pi(\theta_1|\theta_{2,i}, \ldots, \theta_{p,i}, D) \ ,$$

  2. dada la realización de $\theta_{1,i+1}$ y la densidad condicional de $\theta_{2,i+1}$, se genera una realización de este parámetro:

  $$\theta_{2,i+1} \sim \pi(\theta_2|\theta_{1,i+1}, \theta_{3,i}, \ldots, \theta_{p,i}, D) \ ,$$

  3. dadas las realizaciones de $\theta_{1,i+1}$ y $\theta_{2,i+1}$, se genera una realización de $\theta_{3,i+1}$ a partir de su densidad condicional correspondiente:

  $$\theta_{3,i+1} \sim \pi(\theta_2|\theta_{1,i+1}, \theta_{2,i+1}, \ldots, \theta_{p,i}, D) \ ,$$

4. ...
5. dadas las realizaciones de $\theta_{1,i+1}, \dots \theta_{p-1,i+1}$, se genera una realización de $\theta_{p,i+1}$ a partir de su densidad condicional correspondiente:

$$\theta_{p,i+1} \sim \pi(\theta_2|\theta_{1,i+1}, \theta_{2,i+1}, \dots, \theta_{p-1,i+1}, D) \ .$$

- **Paso 2**. Colocar el contador a $i = i + 1$ y regresar al paso 1.

El proceso iterativo se repite hasta obtener la muestra deseada de los parámetros, $\boldsymbol{\theta}_l = (\theta_{1,l}, \dots, \theta_{p,l})'$. En cada paso, se pasa por cada componente de $\boldsymbol{\theta}$ en el orden natural, y un ciclo en este esquema requiere que se generen p variables aleatorias. Gelfand and Smith (1990) demuestran que bajo ciertas condiciones de regularidad, la sucesión de vectores $\{\boldsymbol{\theta}_i, i = 1, 2, \dots\}$ tiene distribución que converge geométricamente a $\pi(\boldsymbol{\theta}|D)$.

Determinar el número de iteraciones l necesario para obtener muestras independientes entre sí no es sencillo, pero existen métodos (Robert and Casella, 2000) que permiten validarlas. El proceso se repite n veces, para obtener un número suficientemente grande de muestras del posteriori, que sustituirán al posteriori en los cálculos en que éste presenta problemas de cálculo. El criterio de Gelman (Gelman et al., 1995) es uno de los más utilizados para verificar la convergencia de este proceso iterativo.

2.5.2. Contraste bayesiano del modelo

Comprobar la coherencia de un modelo con los datos disponibles es un paso fundamental en cualquier tratamiento estadístico. Un análisis estadístico puede conducir a conclusiones erróneas si el modelo adoptado resulta inadecuado, tanto si se trabaja con un enfoque bayesiano como uno frecuentista. Por tanto, el análisis (bayesiano) de unos datos debería incluir una comprobación del modelo, para decidir si éste proporciona un resumen razonable de los datos disponibles, y, en caso contrario, si debe ser rechazado.

En particular, nos centramos en el problema de estudiar la compatibilidad del modelo con los datos (*model checking*) en el caso en que este

modelo tiene parámetros desconocidos (p.ej. modelo $GPD(\xi, \beta)$, con ξ y β desconocidos). Las medidas de compatibilidad utilizadas usualmente, el enfoque estándar clásico, son los p-valores, también llamados significación de la muestra. Estos p-valores están basados en estadísticos con distribución conocida, evaluados en la muestra. Se estima que realizaciones del estadístico con valores grandes, es decir, probabilidades en la cola pequeñas, indican incompatibilidad entre los datos y el modelo. La esencia del enfoque clásico consiste en comparar el valor observado del estadístico o de la discrepancia con la distribución de referencia obtenida, por ejemplo la distribución muestral, considerando válido el modelo adoptado. La probabilidad de la cola, o p-valor, es un método útil y computacionalmente sencillo para localizar el valor observado en la distribución de referencia, Cuando el modelo adoptado tiene parámetros desconocidos, los p-valores no están definidos de manera única. En esta situación, existen diversas propuestas de cálculo de p-valores, tanto desde el punto de vista frecuentista como desde el bayesiano.

En la literatura se han introducido numerosas definiciones de p-valores alternativos, que pretenden extender la esencia del enfoque clásico al contexto bayesiano, con la intención de proporcionar métodos prácticos de evaluación del ajuste de un modelo, especialmente en situaciones complejas donde no es posible el cálculo de la distribución de referencia del estadístico. Se suele utilizar la nomenclatura *evaluación* en lugar de *comprobación* (*assessing* en lugar de *testing*) para destacar la diferencia fundamental entre evaluar las discrepancias entre un modelo y lo datos y el comprobar la corrección de un modelo.

Dada una variable aleatoria X, con función de densidad $f(x|\theta)$, se pretende evaluar si el modelo escogido es compatible con la muestra de datos obtenida, denotada por x_{obs}. Existen diversas definiciones de p-valores, que se adaptan a la información disponible en cada caso.

Definición 2.5.1. *p-valor frecuentista*:
Dado un estadístico de contraste adecuado, T, el p-valor frecuentista clásico se define como

$$p = \mathrm{P}[t(\mathbf{X}) \geq t(\mathbf{x}_{obs})] \ ,$$

calculando esta probabilidad respecto a $f(\mathbf{x}|\theta)$ si el parámetro θ es conocido.

En el caso en que θ es desconocido, es necesario seleccionar otra distribución de probabilidad respecto a la que hacer el cálculo. Generalmente se utiliza una estimación del parámetro θ, lo que da lugar a un nuevo p-valor:

Definición 2.5.2. *plug-in-p-valor* (p_{plug}):

$$p_{plug} = \mathrm{P}^{f(\cdot|\widehat{\theta})}[t(\mathbf{X}) \geq t(\mathbf{x}_{obs})] \quad . \tag{2.5.1}$$

La distribución $f(\mathbf{x};\theta)$ es sustituida por $f(\cdot;\widehat{\theta})$, donde $\widehat{\theta}$ es una estimación del parámetro desconocido. Generalmente se utiliza el estimador máximo verosímil del parámetro, pero otras estimas proporcionan también buenos resultados.

Una solución alternativa para θ desconocido da lugar a (p_{sim}):

Definición 2.5.3. *p-valor similar* (p_{sim}):

$$p_{sim} = \mathrm{P}^{f(\cdot;u_{obs})}[t(\mathbf{X}) \geq t(\mathbf{x}_{obs})] \quad . \tag{2.5.2}$$

En este caso, el parámetro θ desconocido se elimina condicionando a un estadístico suficiente para θ, U. La distribución $f(\mathbf{x}|u_{obs};\theta)$ no depende del parámetro θ, y el cálculo del p-valor puede realizarse utilizando $f(\mathbf{x};u_{obs})$.

Las definiciones anteriores de p-valores (p_{plug}, p_{sim}, (Ec. 2.5.1-2.5.2) intentan subsanar desde el punto de vista frecuentista el desconocimiento sobre el valor de θ, y en consecuencia sobre la distribución que permite el cálculo de la probabilidad de la cola. Con la definición del p-valor predictivo a priori (Box, 1980), Box introduce el concepto de p-valor Bayesiano. Estos p-valores constituyen una alternativa útil a los p-valores frecuentistas, dado que pueden calcularse en muchos de los casos en que la distribución de referencia es desconocida. Otros autores, como Bayarri and Berger, 2000; Gelman et al., 1995, 1996; Meng, 1994, han propuestos diversos p-valores de tipo Bayesiano. Se presentan aquellos utilizados en el desarrollo posterior. Otros destacados se incluyen en el Anexo A.

La definición del p-valor predictivo a posteriori de Rubin, (Rubin, 1984), se basa en los desarrollos previos de Guttman, (Guttman, 1967):

Definición 2.5.4. *posterior predictive p-value* (p_{post}) (Guttman, 1967; Rubin, 1984):

$$p_{post} = \mathrm{P}^{m_{post}(\cdot|\mathbf{x}_{obs})}[t(\mathbf{X} \geq t(\mathbf{x}_{obs})] \quad ,$$

donde $m_{post}(\mathbf{x}|\mathbf{x}_{obs})$ es la *distribución predictiva a posteriori*,

$$m_{post}(\mathbf{x}|\mathbf{x}_{obs}) = \int f(\mathbf{x};\theta)\pi(\theta|\mathbf{x_{obs}})d\theta \quad ,$$

y $\pi(\theta|\mathbf{x_{obs}})$ es la *distribución a posteriori* para θ.

Definición 2.5.5. *discrepancy p-value* (p_{dis}) (Gelman et al., 1995, 1996; Meng, 1994)

El estadístico de contraste $t(\mathbf{X})$ se sustituye por una discrepancia $t(\mathbf{X}, \theta)$,

$$p_{dis} = \mathrm{P}^{m_{dis}(\cdot)}[t(\mathbf{X}, \theta) \geq t(\mathbf{x}_{obs}, \theta)] \quad ,$$

donde $m_{dis}(\mathbf{x}, \theta|\mathbf{x}_{obs})$ es ,

$$m_{dis}(\mathbf{x}, \theta|\mathbf{x}_{obs}) = f(\mathbf{x};\theta)\pi_{post}(\theta|\mathbf{x_{obs}}) \quad ,$$

y $\pi(\theta|\mathbf{x_{obs}})$ es la *distribución a posteriori* para θ.

Pros y contras:

Dada la variedad de p-valores propuestos en la literatura, la selección del más adecuado para evaluar la bondad de ajuste de un modelo dependerá de la información disponible. En general, se valora como importante la facilidad de cálculo, aunque en muchos casos, las técnicas computacionales existentes permiten el cálculo en un tiempo razonable. Asimismo se considera de utilidad que el estadístico de contraste tenga una distribución de probabilidad conocida, al menos asintóticamente, así como que el resultado sea fácilmente interpretable. Pero la cuestión que podríamos considerar clave es la uniformidad del p-valor. Sería deseable que el p-valor fuera uniforme en el intervalo $[0, 1]$, de manera que su interpretación fuera análoga a la del p-valor frecuentista: valores pequeños del p-valor indican evidencia en contra de la hipótesis primaria, es decir, incompatibilidad del modelo con la muestra.

Entre los p-valores definidos en el apartado anterior, el *plug-in-p-value* es el más fácil de calcular, pero se ignora la incertidumbre

contenida en los datos, dado que únicamente se utiliza una estimación puntual del parámetro desconocido. Las alternativas bayesianas, como el *posterior predictive p-value* o el *discrepancy p-value* no suelen ser difíciles de calcular y consideran la incertidumbre, pero el p-valor obtenido no es uniforme, lo que provoca problemas de interpretación (Robins et al., 2000).

2.6. Densidad de mínima información mutua sujeta a restricciones

Dado un conjunto de variables aleatorias, es frecuente disponer de información parcial sobre ellas: posibles densidades marginales, información vaga sobre la dependencia entre ellas, etc. Se desea determinar la forma de la densidad de probabilidad conjunta de estas variables aleatorias, dadas ciertas restricciones (en particular sus distribuciones marginales y/o momentos conjuntos) de manera que se optimice la información representada. En el caso unidimensional, este problema ya fue abordado por Kullback en 1959 (Kullback, 1968) y en el caso bivariante por Rumsey y Posner (Rumsey Jr and Posner, 1965).

A continuación se introducen definiciones básicas de los conceptos de entropía más utilizados, y se destaca la relación entre éstos y las funciones cópula. Posteriormente se presenta la solución univariada de Kullback y la bivariada de Rumsey y Posner.

Definición 2.6.1. (Shannon, 1948)
La entropía de Shannon de una distribución d-dimensional F es

$$H(F|\mathcal{S}) = -\int_{\mathcal{S}} f(\mathbf{x}) \log(f(\mathbf{x})) d\nu(\mathbf{x}) \ , \tag{2.6.1}$$

donde f es la densidad de F con respecto a la medida ν. Se utiliza el símbolo ν para denotar medidas diferentes de la de Lebesgue y la medida contadora. Se indicará el soporte $\mathcal{S}$ cuando sea necesario enfatizarlo.

Definición 2.6.2. (Kullback and Leibler, 1951)
La divergencia de Kullback-Leibler entre F y G es

$$K(F : G|\mathcal{S}) = \int_{\mathcal{S}} \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) f(\mathbf{x}) d\nu(\mathbf{x}) \geq 0, \tag{2.6.2}$$

donde F es absolutamente continua respecto a G, f y g son densidades respecto a ν, y $K(F : G|S) = 0$ si y sólo si g(x) = f (x) casi para todo.

Este funcional mide la distancia en cuanto a información entre una función de densidad f y una densidad de referencia g. También recibe el

nombre de *información relativa*, *discrepancia de información* o *entropía cruzada*.

En particular, será de interés la entropía cruzada $K(F : G|\mathcal{S})$ entre una distribución d-dimensional F y el producto de sus d marginales univariantes, $G = \Pi_{i=1}^{d} F_i(x_i)$ la cual permite medir el grado de asociación entre estas d variables aleatorias. En este caso se dice que $K(F : \Pi_{i=1}^{d} F_i(x_i)|\mathcal{S})$ es la *información mutua* entre las d variables aleatorias.

En el caso bivariado, considerando la distribución $F(x_1, x_2)$ con soporte $\mathcal{S}$, $G(x_1, x_2) = F_1(x_1)F_2(x_2)$, y la medida de Lebesgue,

$$H(F|\mathcal{S}) = -\int\int_{\mathcal{S}} f(x_1, x_2) \log(f(x_1, x_2)) dx_1 dx_2 \ , \tag{2.6.3}$$

$$K(F : F_1F_2|\mathcal{S}) = \int\int_{\mathcal{S}} \log\left(\frac{f(x_1, x_2)}{f_1(x_1)f_2(x_2)}\right) f(x, x_2)\ dx_1 dx_2 \geq 0 \ . \tag{2.6.4}$$

Tanto la entropía de Shannon como la entropía relativa respecto al producto de marginales (información mútua) son medidas independientes de las marginales utilizadas.

En la Sección 2.1 se introdujo la representación de la función de distribución $F(x_1, x_2)$ mediante su correspondiente función cópula: $F(x_1, x_2) = C[F_1(x_1), F_2(x_2)|\delta]$, donde $C[u, v|\delta]$ es una cópula bivariada parametrizada por el vector δ. Aplicando el cambio de variable $F_1(x_1) = u,\ F_2(x_2) = v$ (*probability integral transform*, Whitt (1976)) se obtiene

$$\begin{aligned} &K(F : F_1F_2|\mathcal{S}) = \\ &\int\int_{\mathcal{S}} \log\left(\frac{c[F_1(x_1), F_2(x_2)]f_1(x_1)f_2(x_2)}{f_1(x_1)f_2(x_2)}\right) c[F_1(x_1), F_2(x_2)]f_X(x_1)f_2(x_2)dx_1dx_2 = \\ &\int\int_{[0,1]^2} \log(c[u, v])c[u, v]dudv = K(C : \Pi|[0, 1]^2) \ \geq 0 \ , \end{aligned} \tag{2.6.5}$$

donde $f_{12}(x_1, x_2) = c[F_1(x_1), F_2(x_2)|\delta]f_1(x_1)f_2(x_2)$ es la función de densidad correspondiente a $F(x_1, x_2)$ y $c(u, v|\delta)$ la función de densidad de la cópula.

Así, la entropía relativa de Kullback-Leibler entre $C[u, v|\delta]$ y la cópula de la independencia $\Pi[u, v] = u \cdot v$ coincide con la entropía de Shannon de la cópula $C[(u, v|\delta]$ sean cuales sean las marginales.

La siguiente observación permite reafirmar esta relación entre conceptos:

Observación 2.6.1. La información relativa de Kullback- Leibler respecto al producto de marginales (información mútua) cumple:

$$K(F : F_1F_2|\mathcal{S}) = H(F_1|\mathcal{S}_\mathcal{X}) + H(F_2|\mathcal{S}_\mathcal{Y}) - H(F|\mathcal{S}) \ .$$

En el caso bivariado, para distribuciones marginales especificadas (como es el caso de las cópulas), $H(F_1|\mathcal{S}_\mathcal{X})$ y $H(F_2|\mathcal{S}_\mathcal{Y})$ están determinadas, y por tanto minimizar la información mutua $K(F : F_1F_2|\mathcal{S})$ es equivalente a maximizar la entropía conjunta $H(F)$. En el caso de las cópulas, esta equivalencia se puede deducir también de la expresión (2.6.5).

Proposición 2.6.1. *La entropía de Shannon y la información de Kullback-Leibler no dependen de las marginales, sino sólo de la relación de dependencia entre ellas, representada por la función cópula correspondiente.*

2.6.1. Distribución de mínima información mutua dados momentos. Caso univariante

Definición 2.6.3. Un conjunto M de medidas en $\mathcal{S}$ se denomina dominado si existe una medida ν sobre $\mathcal{S}$, ν no necesariamente de M, tal que cada miembro del conjunto M es absolutamente continua respecto a ν,

$$\mu(E) = \int_E f(x)d\nu(x) \ ,$$

para todo $E \in \mathcal{S}$.

En el caso unidimensional se desea hallar una densidad generalizada $f(x)$, $f(x) \in M$, tal que sea la más cercana en el sentido de la menor divergencia cruzada a una densidad $g(x)$ dada. Es decir,

$$K(F : G|\mathcal{S}) = \int_\mathcal{S} f(x) \log\left(\frac{f(x)}{g(x)}\right) d\nu(x)$$

mínima. Dado que la entropía relativa es positiva, $K(F : G|\mathcal{S}) \geq 0$, y la igualdad se alcanza si y sólo si $f(x) = g(x)[\nu]$, se deben imponer

algunas restricciones adicionales sobre $f(x)$ si se desea que la medida de probabilidad "más cercana" sea diferente de la propia medida μ_2, $\mu_2(E) = \int_E g(x)d\nu(x)$.

Teorema 2.6.2. *(Kullback, 1959)*

Sean $f(x), g(x)$ densidades generalizadas de un conjunto acotado de medidas de probabilidad M, $g(x)$ dadas. Sea $Y = T(x)$ un estadístico medible tal que

$$\theta = \int T(x)f(x)d\nu(x) existe \ ,$$

(siendo θ un parámetro multidimensional de la población u otra característica de ésta), y sea

$$M_2(\lambda) = \int g(x)e^{-\lambda T(x)}d\nu(x) \ ,$$

que existe para λ en algún intervalo.

Entonces

$$K(F : G|\mathcal{S}) \geq -\lambda\theta + \log(M_2(\lambda) = K(* : G) \ ,$$

con $\theta = \frac{d}{d\lambda}\log M_2(\lambda)$. La igualdad se alcanza si y sólo si

$$f(x) = f^*(x) = \frac{g(x)e^{-\lambda T(x)}}{M_2(\lambda)} \ [\nu] \ .$$

Observación 2.6.2. Se dice que $f^*(x) = g(x)e^{-\lambda T(x)}/M_2(\lambda)$ genera una familia exponencial de distribuciones, la familia de tipo exponencial determinada por $g(x)$, conforme λ varía en su intervalo. Se puede encontrar información detallada al respecto en Kullback (1968).

2.6.2. Densidades de mínima información mutua dados sus momentos

En la Sección 2.6.1 se ha presentado la forma de la densidad unidimensional de probabilidad de mínima entropía relativa respecto a una densidad dada, $g(x)$, y dados sus momentos, $T(x)$. Se desea generalizar el problema al caso multidimensional. El objetivo es hallar la densidad

$f(\mathbf{x})$ multivariante de mínima entropía relativa respecto al producto de sus marginales, $g(x) = \prod f_i(x_i)$, (es decir, su información mútua) sujeta a restricciones en forma de momentos. Para simplificar la notación nos centraremos en el caso bidimensional.

Dadas dos variables X_1 y X_2, se desea determinar su densidad conjunta $f(x_1, x_2)$. Se dispone de información parcial sobre esta densidad en forma de restricciones: se conocen las funciones de densidad marginales de X e Y,

$$f_1(x_1) = \int_{-\infty}^{+\infty} f(x_1, x_2)dy \ ; \quad f_2(x_2) = \int_{-\infty}^{+\infty} f(x_1, x_2)dx \ ,$$

donde ambas ecuaciones se cumplen casi para todo; además se puede disponer de un número finito, J, de otras restricciones en forma de momentos que la función de densidad $f(x_1, x_2)$ debe cumplir. Estas restricciones se pueden expresar en la forma

$$\begin{aligned} &\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} T_j(x_1, x_2) f(x_1, x_2) dx_1 dx_2 = \\ &\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} T_j(x_1, x_2) dF(x_1, x_2) = \theta_j, \quad j = 1, \ldots, J \ , \end{aligned}$$

donde las T_j son funciones dadas y las θ_j son constantes dadas. Debe observarse que pueden no existir restricciones de este tipo, es decir, J puede ser cero.

El problema consiste en hallar la densidad $f(x_1, x_2)$ tal que satisface las anteriores restricciones y tal que además tiene máxima entropía de Shannon $H(f)$ (Def. 2.6.1):

$$\begin{aligned} &\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x_1, x_2) \cdot \log f^{-1}(x_1, x_2) dx_1 dx_2 = \\ &\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \log f^{-1}(x_1, x_2) dF(x_1, x_2) \ , \end{aligned}$$

o equivalentemente (Obs. 2.6.1) hallar la densidad $f(x_1, x_2)$ tal que satisface las anteriores restricciones y tal que además tiene mínima

entropía cruzada respecto las marginales dadas (Def. 2.6.2) :

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x_1,x_2)\cdot\log\left(\frac{f(x_1,x_2)}{f_1(x_1)f_2(x_2)}\right)dx_1dx_2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\log\left(\frac{f(x_1,x_2)}{f_1(x_1)f_2(x_2)}\right)dF(x_1,x_2)\ .$$

Denotaremos Ω_F al conjunto de distribuciones que satisfacen el conjunto de restricciones:

$$\Omega_F = \{F : E_F[T_j(x_1,x_2)] = \theta_j, j = 1,...,J; f_1(x_1); f_2(x_2)\}\ , \quad (2.6.6)$$

donde $T_j(x_1,x_2)$ son funciones reales, integrables respecto a $dF(x_1,x_2)$, y $\theta_j,\ j = 1,\ldots,J$ son momentos conocidos (y adecuados). Se consideran en particular la distribuciones $F(x_1,x_2)$ de Ω_F absolutamente continuas con respecto a $G = F_1F_2$, donde $F_j,\ j = 1,2$, son las funciones de distribución marginales de $F(x_1,x_2)$.

Definición 2.6.4. Sean $f_1(x)$ y $f_2(x_2)$ dos funciones de densidad fijadas, definidas sobre $(-\infty,+\infty)$. Sean $T_1(x_1,x_2),\ldots,T_J(x_1,x_2)$ funciones fijadas, reales, definidas en el plano (x_1,x_2). Supondremos que ninguna combinación lineal no nula de los T_j es igual casi para todo a una función de x_1 más una función de x_2, dado que los valores esperados de esas sumas se pueden determinar a partir de las distribuciones marginales y por tanto no tienen influencia en el problema de optimización. Sean $\theta_1,\ldots,\theta_k$ constantes dadas. Sea

$$\Omega_F = \{F : E_F[T_j(x_1,x_2)] = \theta_j, j = 1,...,J; f_1(x_1)\ ;\ \ f_2(x_2)\}\ ,$$

el conjunto de distribuciones que satisfacen las siguientes condiciones:

a) $\int_{-\infty}^{+\infty} f(x_1,x_2)dy = f_1(x_1)$,

b) $\int_{-\infty}^{+\infty} f(x_1,x_2)dx = f_2(x_2)$, y

c) $(f(x_1,x_2)T_j(x_1,x_2)) \in L_1$ para $j = 1,\ldots,J$, con

$$E_F[T_j(x_1,x_2)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\log\left(\frac{f(x_1,x_2)}{f_1(x_1)f_2(x_2)}\right)dF(x_1,x_2) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x_1,x_2)\cdot\log\left(\frac{f(x_1,x_2)}{f_1(x_1)f_2(x_2)}\right)dx_1dx_2 = \theta_j\ .$$

Las condiciones $a), b)$ se cumplen para todo. L_1 denota el conjunto de funciones integrables Lebesgue.

Diremos que Ω_F es un *conjunto admisible de funciones* si además se cumplen las siguientes condiciones:

d) Ω_F es no vacío, y al menos existe una función $f(x_1, x_2) \in \Omega_F$ tal que $f(x_1, x_2) \log f(x_1, x_2) \in L_1$,

e) $f_1(x_1) \log f_1(x_1) \in L_1$, $f_2(x_2) \log f_2(x_2) \in L_1$,

f) Si $f(x_1, x_2) \in \Omega_F$, entonces $\int \int_{f<1} f(x_1, x_2) \log^2 f(x_1, x_2) \leq A$, donde A es constante (dependiendo de Ω_F),

g) Ω_F es cerrado en L_1.

Observación 2.6.3. A partir de la definición no es sencillo determinar si un conjunto Ω_F es admisible o no. Por ejemplo, no es sencillo determinar si Ω_F es no vacío. Este problema se ha tratado en la literatura, p.ej. en Nataf (1962).

El siguiente teorema proporciona una generalización del Teorema (2.6.2):

Teorema 2.6.3. *(Rumsey Jr and Posner, 1965)*

Sean las funciones $f_1(x_1), f_2(x_2), T_1(x_1, x_2) \ldots, T_J(x_1, x_2)$ y las constantes $\theta_1, \ldots, \theta_J$ tales que Ω_F es un conjunto admisible. Entonces el conjunto de ecuaciones integrales simultáneas

$$a(x_1) \int_{-\infty}^{+\infty} b(x_2) \exp[\lambda_j T_j(x_1, x_2)] dy = f_1(x_1) \ ,$$

$$b(x_2) \int_{-\infty}^{+\infty} a(x_1) \exp[\lambda_j T_j(x_1, x_2)] dx = f_2(x_2) \ ,$$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} a(x_1) b(x_2) \exp[\alpha_j T_j(x_1, x_2)] dx_1 dx_2 = \theta_i \ , \ i = 1, \ldots, J$$

tiene una solución única en las funciones $a(x), b(x_2)$ y las constantes $\alpha_1, \ldots \alpha_J$. Además, la función

$$f(x_1, x_2) = a(x_1) b(x_2) \exp \left[\sum_{j=1}^{J} \alpha_j T_j(x_1, x_2) \right] \ , \ (x_1, x_2) \in \mathbb{R}^2 \ . \quad (2.6.7)$$

es el único elemento de Ω_F con entropía de Shannon máxima.

Observación 2.6.4. Se debe observar que cuando $J = 0$ (no existen restricciones), entonces

$$\sum_{i=1}^{J} \alpha_i T_i(x_1, x_2) = 0 \ ,$$

y $f(x_1, x_2) = a(x_1)b(x_2)$. La solución única, salvo constante, es $a(x_1) = f_1(x_1), b(x_2) = f_2(x_2)$. Este resultado ya fue presentado por Shannon (Shannon, 1948), mostrando que la función de densidad de máxima entropía con marginales dadas es el producto de las densidades marginales, obtenida cuando X_1 y X_2 son variables aleatorias independientes.

Capítulo 3

Un modelo para los valores extremales en procesos de Poisson evaluados

Se desea establecer un modelo general que permita describir la ocurrencia de sucesos extremales a lo largo del tiempo. Se consideran sucesos que ocurren en el tiempo como un proceso de Poisson, el denominado proceso subyacente. A cada suceso del proceso subyacente se le puede añadir un tamaño, medido de algún modo. Así, cada suceso se describe por varias variables aleatorias, dos o más, que pueden ser estadísticamente dependientes (diferentes variables medidas en el mismo momento, la misma variable medida en dos ubicaciones, etc.). De un modo genérico denominaremos estas variables tamaños del modelo y, por simplicidad, nos restringiremos al caso bivariado. Para cada suceso, denotaremos los tamaños aleatorios por X_1, X_2, sin indicar a qué suceso concreto han sido asociados. Los tamaños X_1, X_2 se consideran idénticamente distribuidos de suceso a suceso y su función de distribución conjunta es $F_{X_1X_2}(x_1, x_2)$. Por tanto, las marginales $F_{X_i}(x_i)$, $i = 1, 2$, son las mismas para todos los sucesos. Además, los tamaños se consideran también independientes de suceso a suceso y respecto a la ocurrencia temporal.

Para los tamaños estableceremos un modelo general: consideramos que los valores observados de X_i, $i = 1, 2$, tienen un soporte limitado inferiormente; sean x_{0i}, i=1,2, sus cotas inferiores. Sin embargo, se asig-

narán valores por defecto x^*_{0i}, $i = 1, 2$ a aquellos sucesos no detectados o no observables. Se considera que los valores por defecto son a lo sumo iguales a la cota inferior de los valores observables: $x^*_{0i} \leq x_{0i}$ (ver Fig. 3.1). Por ejemplo, supongamos que se dispone de una base de datos que contiene las precipitaciones diarias registradas en un conjunto de estaciones meteorológicas situadas en ubicaciones cercanas. En cada estación se registra la precipitación diaria superior a 1mm. Si X_i es la precipitación diaria en una ubicación determinada, su cota inferior es $x_{0i} = 1$mm. El valor por defecto en caso de lluvia no registrada, x^*_{0i}, podría ser cero, pero dado que la precipitación diaria es una magnitud con una escala que puede considerarse relativa y tomaremos logaritmos, el valor por defecto escogido en esta escala será $x^*_{0i} = \log(0.05)$.

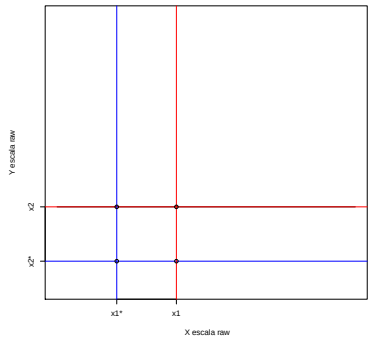

Figura 3.1: Asignación de valores por defecto $x = x^*_{0i}$, $i = 1, 2$ en las marginales, para aquellos valores no registrados. $x = x_{0i}$ indica la cota inferior del soporte.

Las funciones de distribución marginales de X_1, X_2 son de la forma

$$F_{X_i}(x_i) = \mathrm{P}[X_i \leq x_i] = \begin{cases} 0 & \text{si} \quad x_i < x^*_{0i} \ , \\ p_i & \text{si} \quad x^*_{0i} \leq x_i < x_{0i} \ , \\ p_i + \mathrm{P}[x_{0i} \leq X_i \leq x_i] & \text{si} \quad x_{0i} \leq x_i \ , \end{cases} \tag{3.0.1}$$

donde p_i es la probabilidad de asignar el valor por defecto x^*_{0i} a X_i. En la práctica, las cotas inferiores x_{0i} son con frecuencia iguales a los valores

por defecto x^*_{0i} y por tanto, las funciones de distribución descritas en la ecuación 3.0.1 se pueden simplificar.

La función de distribución conjunta de X_1 y X_2, $F_{X_1X_2}(x_1, x_2)$, tiene características correspondientes a sus distribuciones marginales, las cuales tienen saltos e intervalos de valor constante. De hecho, para $i = 1, 2$, $x_i < x^*_{0i}$, tenemos que $F_{X_1X_2}(x_1, x_2) = 0$. Si $x^*_{01} \leq x_1 < x_{01}$ y $x^*_{02} \leq x_2 < x_{02}$, $F_{X_1X_2}(x_1, x_2) = p_{12}$ donde p_{12} es la probabilidad de asignar ambos valores por defecto. Normalmente se asume $p_{12} = 0$ dado que es la probabilidad de sucesos indetectables en el proceso subyacente. En todo caso, $p_{12} \leq \text{mín}(p_1, p_2) \leq p_1 + p_2$. Dado que los valores, si hubiera alguno, entre x^*_{0i} y x_{0i} no son detectables, entonces, para $x^*_{0i} \leq x_i < x_{0i}$ y$x_j \geq x_{0j}$, $i \neq j$, $F_{X_1X_2}(x_1, x_2)$ es arbitrario siempre que se mantenga la monotonía de la función de distribución conjunta. Finalmente, para $x_i \geq x_{0i}$, $i = 1, 2$, puede ser cualquier función de distribución que se adapte a las funciones de distribución marginales $F_{X_i}(x_i)$, $i = 1, 2$. La Figura 3.2 muestra la forma de este tipo de funciones de distribución.

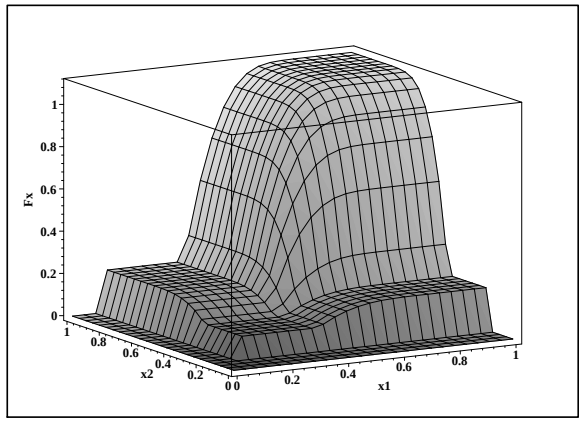

Figura 3.2: Esquema de una función de distribución de $F_{X_1X_2}$, considerando la asignación de valores por defecto. Las discontinuidades de salto se encuentran en $x = x_{0i}$ y $x = x^*_{0i}$, $i = 1, 2$.

El uso de funciones cópula (Sec. 2.1) permite tratar por separado los modelos marginales y la dependencia de las variables. Modelizamos la función de distribución conjunta de X_1, X_2 mediante la cópula

$C_{X_1X_2}[\cdot,\cdot]$, es decir,

$$F_{X_1X_2}(x_1, x_2) = C_{\mathbf{X}}[F_{X_1}(x_1), F_{X_2}(x_2)] \ , \tag{3.0.2}$$

donde $\mathbf{X}$ denota (X_1, X_2).

La elección de los modelos marginales implica algunas particularidades de la cópula. Debido a las posibles discontinuidades en las marginales (Ec. 3.0.1), la cópula no es única (Rüschendorf et al., 1996) y los valores en algunos de sus dominios no juegan ningún papel. En concreto, los valores de $C_{\mathbf{X}}[u, v]$ para $0 \leq u < p_1$ o $0 \leq v < p_2$ se pueden escoger adecuadamente simplemente satisfaciendo condiciones genéricas de las cópulas. Para $p_1 \leq u \leq 1$ y $p_2 \leq v \leq 1$, la cópula debería representar la dependencia verdadera de X_1, X_2. Esto completa un modelo genérico para las variables tamaño de los sucesos.

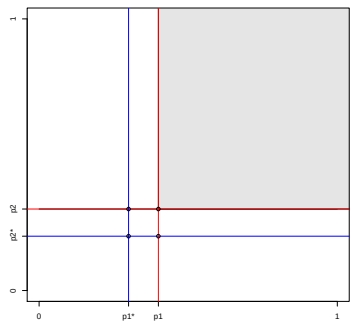

Figura 3.3: Valores destacados en la cópula entre X_1, X_2. La zona sombreada indica el área que representa la dependencia verdadera de las marginales.

En numerosas aplicaciones prácticas se precisa calcular la probabilidad de ocurrencia de un número dado de sucesos en un periodo de tiempo de longitud t_0. En particular, se consideran determinados tipos de sucesos, cuyos tamaños satisfacen una condición. Por ejemplo, sucesos de lluvia cuya precipitación en 5 minutos es mayor de 5mm o cuya precipitación en 30 minutos es mayor de 20mm. El modelo de Poisson de ocurrencia de sucesos en el tiempo (Sec. 2.3) y la función de distribución conjunta de los tamaños permiten determinar de una

manera estándar las probabilidades de ocurrencia de los eventos descritos. Si A es un suceso en el espacio (X_1, X_2) y $N(A)$ es el número de sucesos cuyos tamaños X_1, X_2 están en A, se tiene que (Prop. 2.3.3 o Feller, 1968a)

$$\mathrm{P}[N(A) = n \mid \lambda(A), t] = \frac{[\lambda(A)t]^n \exp[-\lambda(A)t]}{n!}, \quad n = 0, 1, 2, \ldots \tag{3.0.3}$$

donde

$$\lambda(A) = \lambda_0 \ \mathrm{P}[A] \ , \tag{3.0.4}$$

t_0 es el tiempo de observación y λ_0 es la tasa de ocurrencia de sucesos en el proceso de Poisson subyacente.

La probabilidad de sucesos como $A = \{X_1 > x_1, X_2 > x_2\}$ se puede expresar utilizando la función de distribución conjunta de X_1, X_2. La tasa de ocurrencia de los sucesos (Ec. 3.0.4) queda

$$\lambda(A) = \lambda_0 \ \left[\ 1 + F_{X_1X_2}(x_1, x_2) - F_{X_1}(x_1) - F_{X_2}(x_2) \ \right] \ . \tag{3.0.5}$$

Denotamos los excesos de las variables (X_1, X_2) sobre el umbral bivariado (h_1, h_2) como $(Y_1 = X_1 - h_1, Y_2 = X_2 - h_2)$, dado que $X_1 > h_1$ y $X_2 > h_2$.

La función de distribución conjunta de los excesos sobre (h_1, h_2) es

$$F_{Y_1Y_2}(y_1, y_2) = \mathrm{P}[\ Y_1 \leq y_1, Y_2 \leq y_2 \mid X_1 > h_1, X_2 > h_2 \ ] \ ,$$

que es una probabilidad condicional. La relación entre $F_{Y_1Y_2}$ y $F_{X_1X_2}$ es sencilla:

$$\begin{aligned} F_{Y_1Y_2}(y_1, y_2) &= \frac{\mathrm{P}[h_1 < X_1 \leq y_1 + h_1, h_2 < X_2 \leq y_2 + h_2]}{\mathrm{P}[X_1 > h_1, X_2 > h_2]} = \\ &\frac{F_{X_1X_2}(y_1 + h_1, y_2 + h_2) + F_{X_1X_2}(h_1, h_2)}{1 + F_{X_1X_2}(h_1, h_2) - F_{X_1}(h_1) - F_{X_2}(h_2)} - \\ &\frac{F_{X_1X_2}(h_1, y_2 + h_2) + F_{X_1X_2}(y_1 + h_1, h_2)}{1 + F_{X_1X_2}(h_1, h_2) - F_{X_1}(h_1) - F_{X_2}(h_2)} \ . \end{aligned} \tag{3.0.6}$$

Las versiones univariadas de (3.0.6) son

$$1 - \mathrm{P}[Y_i \leq y_i | X_i > h_i] = \frac{1 - F_{X_i}(y_i + h_i)}{1 - F_{X_i}(h_i)} \quad , \quad i = 1, 2 \ , \tag{3.0.7}$$

utilizadas frecuentemente en métodos Peak-Over-Threshold (POT).

En general los valores por defecto establecidos en el modelo no aparecerán en los cálculos de valores de *hazard* de interés relacionados con excesos sobre un umbral, dado que los umbrales $h_i,\ i = 1, 2$ serán superiores a la cota inferior del soporte de X_i, $h_i > x_{0i}$. En el ejemplo de la base de datos de precipitación diaria, puede ser de interés estudiar los excesos de precipitación que superan un umbral $h_i = 20$ mm, un valor superior a la cota inferior del soporte $h_i = 20 > x_{0i} = 1$ mm.

Los umbrales en cambio sí aparecen en los cálculos de valores de *hazard* relacionados con máximos de la magnitud. Si nos centramos en la extracción de los máximos de las observaciones en un tiempo fijado, t_0, denotaremos N al número aleatorio de sucesos en este tiempo t_0, y definiremos Z_i , $i = 1, 2$, el máximo de los tamaños X_i correspondientes a esos sucesos. Si no se han observado sucesos en el tiempo t_0, es decir $N = 0$, asignamos los valores por defecto $Z_1 = x_{01}^*, Z_2 = x_{02}^*$ a los máximos.

La función de distribución conjunta de (Z_1, Z_2), condicionada a N satisface la siguiente relación: si $N = 0$,

$$F_{Z_1 Z_2}(z_1, z_2 \mid N = 0) = \mathrm{I}\{(z_1 \geq x_{01}^*) \cap (z_2 \geq x_{02}^*)\} \ ;$$

mientras que si $N = n \neq 0$,

$$\begin{aligned} &F_{Z_1 Z_2}(z_1, z_2 \mid N = n) = \\ &\mathrm{P}[Z_1 \leq z_1, Z_2 \leq z_2 \mid N = n] \cdot \mathrm{I}\{(z_1 \geq x_{01}^*) \cap (z_2 \geq x_{02}^*)\} = \\ &[F_{X_1 X_2}(z_1, z_2)]^n \cdot \mathrm{I}\{(z_1 \geq x_{01}^*) \cap (z_2 \geq x_{02}^*)\} \ , \end{aligned} \tag{3.0.8}$$

donde $\mathrm{I}\{\cdot\}$ es la función indicadora. La expresión condicional a $N = 0$ se puede incluir en la ecuación 3.0.8 simplemente tomando $n = 0$.

Se debe observar que la expresión 3.0.8 se ha obtenido asumiendo que los vectores (X_1, X_2) son independientes entre sucesos. Aunque X_1 y X_2 se consideran dependientes en el mismo suceso, el anterior supuesto implica que X_1 y X_2 son independientes siempre que estén asociados a diferentes sucesos. Otras hipótesis conducen a expresiones mucho más complicadas.

Eliminando la condición respecto a N, se obtiene la función de

distribución conjunta de (Z_1, Z_2) para un tiempo de observación t_0:

$$F_{Z_1Z_2}(z_1, z_2) =$$
$$\sum_{n=0}^{+\infty} P[N = n \mid \lambda_0, t][F_{X_1X_2}(z_1, z_2)]^n \cdot \mathrm{I}\{(z_1 \geq x_{01}^*) \cap (z_2 \geq x_{02}^*)\} =$$
$$\exp[-\lambda t(1 - F_{X_1X_2}(z_1, z_2))] \cdot \mathrm{I}\{(z_1 \geq x_{01}^*) \cap (z_2 \geq x_{02}^*)\} \ . \qquad (3.0.9)$$

Una vez explicitados los modelos para excesos (Ec. 3.0.6) y máximos (Ec. 3.0.9) en procesos de Poisson, obtendremos las correspondientes cópulas y determinaremos su relación con la cópula de los tamaños originales (X_1, X_2) (Ec. 3.0.2).

Capítulo 4

Transformaciones extremales de cópulas en procesos de Poisson

Dado un proceso de Poisson evaluado, donde la dependencia entre dos tamaños asociados a un suceso se modela mediante su correspondiente función cópula, se desea hallar la transformación de esta cópula en dos situaciones diferentes:

a) extracción de los excesos sobre umbrales superiores al umbral absoluto, uno para cada tamaño;

b) extracción de máximos de cada variable en un tiempo determinado.

Las cópulas transformadas describen la dependencia entre los excesos sobre un umbral bivariado o la dependencia entre los máximos bivariados observados en un proceso de Poisson. Estas transformaciones de cópulas, que denominamos extremales, no son asintóticas, pese a que para umbrales bivariados altos y tiempos de extracción grandes, se aproximan fácilmente a las cópulas asintóticas.

4.1. Transformación de cópulas bajo cambio de umbral

En esta Sección se desea estudiar la relación entre la cópula de los tamaños originales (Ec. 3.0.2) y la cópula de los excesos sobre un

umbral dado, superior al umbral absoluto. Consideremos el modelo especificado en el Capítulo 3. En ese contexto, modelizamos la función de distribución conjunta de los excesos (Y_1, Y_2) mediante la cópula $C_{\mathbf{Y(h)}}$, es decir,

$$F_{Y_1Y_2}(y_1, y_2|X_1 > h_1, X_2 > h_2) = C_{\mathbf{Y(h)}}[F_{Y_1}(y_1), F_{Y_2}(y_2)] \ , \qquad (4.1.1)$$

donde se utiliza el subíndice $\mathbf{Y(h)}$ para indicar que los excesos (Y_1, Y_2) se han tomado sobre el umbral bivariado $\mathbf{h} = (h_1, h_2)$.

El umbral bivariado $\mathbf{h}$ se puede tomar arbitrariamente, pero en general se tomará tal que $x_{0i} << h_i$, $i = 1, 2$, tanto para satisfacer condiciones asintóticas de los excesos como para evitar los valores por defecto de X_i. La cópula $C_{\mathbf{Y(h)}}$ no se halla completamente determinada si los umbrales h_1 o h_2 son inferiores a x_{01} o x_{02} respectivamente. En esos casos, los valores de Y_i, $0 \leq Y_i < x_{0i} - h_i$ con $Y_i \neq x^*_{0i} - h_i$, no son alcanzables, y la cópula no se halla unívocamente definida.

Obtenemos la expresión que relaciona la cópula entre excesos con la cópula de los tamaños originales a partir de la Ec. 3.0.6, la cual relaciona las correspondientes funciones de distribución:

$$C_{\mathbf{Y(h)}}[v_1, v_2] = \frac{C_{\mathbf{X}}[\eta_1, \eta_2] + C_{\mathbf{X}}[u_1, u_2] - C_{\mathbf{X}}[u_1, \eta_2] - C_{\mathbf{X}}[\eta_1, u_2]}{1 + C_{\mathbf{X}}[u_1, u_2] - u_1 - u_2} \ . \qquad (4.1.2)$$

donde

$$F_{Y_i}(y_i) = v_i \quad , \quad F_{X_i}(y_i + h_i) = \eta_i \quad , \quad F_{X_i}(h_i) = u_i \quad ,$$

que serán utilizados como argumentos de las cópulas.

Para expresar η_i en función de u_i y v_i, es útil obtener las probabilidades marginales

$$\mathrm{P}[Y_i \leq y_i|X_1 > h_1, X_2 > h_2] = F_{Y_i}(y_i) = v_i, \quad i = 1, 2 \ ,$$

que proporcionan expresiones implícitas de η_i:

$$\begin{aligned} v_1 &= \frac{\eta_1 + C_{\mathbf{X}}[u_1, u_2] - u_1 - C_{\mathbf{X}}[\eta_1, u_2]}{1 + C_{\mathbf{X}}[u_1, u_2] - u_1 - u_2} \ , \\ v_2 &= \frac{\eta_2 + C_{\mathbf{X}}[u_1, u_2] - u_2 - C_{\mathbf{X}}[u_1, \eta_2]}{1 + C_{\mathbf{X}}[u_1, u_2] - u_1 - u_2} \ , \end{aligned} \qquad (4.1.3)$$

las cuales son no lineales en general, excepto en el caso en que X_1 y X_2 son independientes, es decir, $C_{\mathbf{X}}[\xi_1, \xi_2] = \xi_1\xi_2$, para $0 \leq \xi_i \leq 1$, $i = 1, 2$ (ver Fig. 2.3).

Podemos interpretar la Ec.4.1.2 como una transformación de la cópula $C_{\mathbf{X}}[\cdot, \cdot]$ en $C_{\mathbf{Y(h)}}[\cdot, \cdot]$. La cópula $C_{\mathbf{Y(h)}}[\cdot, \cdot]$ se obtiene como una transformación de $C_{\mathbf{X}}[\cdot, \cdot]$ restringida al rectángulo $(u_1, 1) \times (u_2, 1)$. Los valores u_1 y u_2 dependen del umbral bivariado seleccionado, (h_1, h_2), y por tanto el denominador de la Ec. 4.1.2 es constante para cada umbral.

Una expresión alternativa de la transformación de cópulas se obtiene utilizando cópulas de supervivencia. Siguiendo la notación de Nelsen (1999), una cópula de supervivencia se define como $\widehat{C}[1 - x_1, 1 - x_2] = 1 - C[x_1, x_2]$, donde C es la correspondiente cópula.

La principal dificultad en el uso de la transformación extremal (Ec. 4.1.2) es la definición implícita de η_i dada por Ec.4.1.3. Sin embargo, esta transformación es numéricamente tratable para familias de cópulas $C_{\mathbf{X}}$ cuya función de distribución tiene expresión cerrada. Un buen punto de partida para los procedimientos numéricos es el valor η_i que se obtiene al asumir $C_{\mathbf{X}}(u_1, u_2) = u_1u_2$, que reduce (4.1.3) a la expresión lineal $\eta_i = v_i(1 - u_i) + u_i$. Es más, η_i es monótona respecto a v_i. Resultados previos evidencian que un proceso de bisección proporciona resultados rápidos y precisos.

4.2. Transformación de cópulas bajo extracción de máximos

Centraremos la atención en la relación entre la cópula de los tamaños (Ec. 3.0.2) y la cópula de los máximos de los sucesos observados durante un tiempo t_0. Para cada suceso en el proceso de Poisson subyacente, tomamos los tamaños X_1, X_2. Denotaremos Z_1, Z_2 a los máximos de los tamaños X_1, X_2 respectivamente para sucesos registrados en un tiempo fijado t_0. Modelizamos la función de distribución conjunta de Z_1, Z_2 mediante la copula $C_{\mathbf{Z}}$, es decir,

$$F_{Z_1Z_2}(z_1, z_2) = C_{\mathbf{Z}}[F_{Z_1}(z_1), F_{Z_2}(z_2)] \ ,$$

donde utilizamos el subíndice $\mathbf{Z}$ para indicar (Z_1, Z_2).

La cópula $C_{\mathbf{Z}}[\cdot,\cdot]$ también presenta características heredadas de los modelos marginales (Ec. 3.0.1) y la asignación de valores por defecto (x_{01}^*, x_{02}^*) cuando no se detectan sucesos en el tiempo t_0. Si los valores por defecto de variables tamaño se asignan con probabilidades p_1 y p_2, la probabilidad de asignar valores por defecto a los máximos son $w_{0i} = \exp(-\lambda t(1-p_i))$, $i = 1, 2$.

Para $0 \leq w_1 < w_{01}$ ó $0 \leq w_2 < w_{02}$, los valores de $C_{\mathbf{Z}}[w_1, w_2]$ se pueden escoger adecuadamente, satisfaciendo condiciones sobre las cópulas mientras que, para $w_{01} \leq w_1 \leq 1$ y $w_{02} \leq w_2 \leq 1$, deberían representar la verdadera dependencia de Z_1, Z_2.

En el Capítulo 3, se obtuvo la relación entre las funciones de distribución de los máximos y los tamaños (Ec. 3.0.9). Utilizando cópulas, esta expresión queda

$$C_{\mathbf{Z}}[F_{Z_1}(z_2), F_{Z_2}(z_2)] = \\ \exp[-\lambda t(1 - C_{\mathbf{X}}[F_{X_1}(z_1), F_{X_2}(z_2)])] \cdot \mathrm{I}\{(z_1 \geq x_{01}^*) \cap (z_2 \geq x_{02}^*)\} \ .$$

Como en la Sección anterior, se busca la transformación de $C_{\mathbf{X}}[\cdot,\cdot]$ en $C_{\mathbf{Z}}[\cdot,\cdot]$. Denotamos

$$F_{Z_i}(z_i) = w_i \quad , \quad F_{X_i}(z_i) = u_i \ ,$$

para poder tratar las funciones de distribución marginales como argumentos. La relación entre u_i y w_i se obtiene directamente: $w_i = \exp(-\lambda t(1-u_i))$, o su inversa, $u_i = 1 + \ln w_i/(\lambda t)$, para $i = 1, 2$.

Para $w_1 \geq w_{01}$ y $w_2 \geq w_{02}$ (es decir, para $z_1 \geq x_{01}^*$ and $z_2 \geq x_{02}^*$), se obtienen dos expresiones equivalentes,

$$C_{\mathbf{Z}}[w_1, w_2] = \exp\left[-\lambda t\left(1 - C_{\mathbf{X}}\left[1 + \frac{\ln w_1}{\lambda t}, 1 + \frac{\ln w_2}{\lambda t}\right]\right)\right], \quad (4.2.1)$$

y,

$$C_{\mathbf{Z}}[\exp(-\lambda t(1-u_1))), \exp(-\lambda t(1-u_2)))] = \exp[-\lambda t(1 - C_{\mathbf{X}}[u_1, u_2])], \quad (4.2.2)$$

utilizando alternativamente u_i ó w_i como argumentos. Para $w_1 < w_{01}$ y $w_2 < w_{02}$, la cópula se puede escoger arbitrariamente simplemente preservando la monotonía y las marginales uniformes.

4.3. Un ejemplo de aplicación de las transformaciones extremales

En esta sección ilustraremos el efecto de las transformaciones extremales de cópulas presentadas en los dos apartados anteriores al aplicarlas a la cópula correspondiente a una mixtura de distribuciones normales bivariadas. Una función de distribución normal bivariada se caracteriza por cinco parámetros (dos valores medios, dos varianzas y el coeficiente de correlación de Pearson). Se han mezclado dos de estas distribuciones, F_1 y F_2 de la manera siguiente:

$$F_\delta(x_1, x_2) = \delta F_1(x_1, x_2) + (1 - \delta)F_2(x_1, x_2) \ ,$$

donde δ, $0 \leq \delta \leq 1$, es el parámetro de la mixtura, y μ_{i1}, μ_{i2}, σ_{i1}^2, σ_{i2}^2, ρ_i, son los parámetros de las distribuciones marginales F_i, $i = 1, 2$.

La distribución F_δ puede expresarse en función de su cópula, que depende de los parámetros anteriores:

$$F_\delta(x_1, x_2) = C_\delta[F_1(x_1), F_2(x_2)] \ ,$$

donde $C_\delta[u_1, u_2]$ denota la función cópula.

Las figuras 4.1 y 4.2 muestran respectivamente la función de distribución y de densidad de una cópula de este tipo, $C_\delta[u_1, u_2]$, en $(0, 1)\times(0, 1)$, con los parámetros indicados en la Tabla 4.1.

Tabla 4.1: Parámetros de la cópula de una mixtura de distribuciones normales

μ_{11}	μ_{12}	σ_{11}	σ_{12}	ρ_1	δ
-1	1	0.3	0.3	0.8	0.5
μ_{21}	μ_{22}	σ_{21}	σ_{22}	ρ_2	
1	1	0.2	0.2	-0.6	

La densidad (Fig. 4.2, izquierda) muestra simetrías en ambas diagonales, debido al método de construcción y a los parámetros elegidos. Pequeños cambios en los parámetros producen contornos bastante distintos entre sí. La distribución (Fig 4.1, izquierda) no tiene características destacables. Los cambios en los parámetros producen distribuciones ligeramente diferentes.

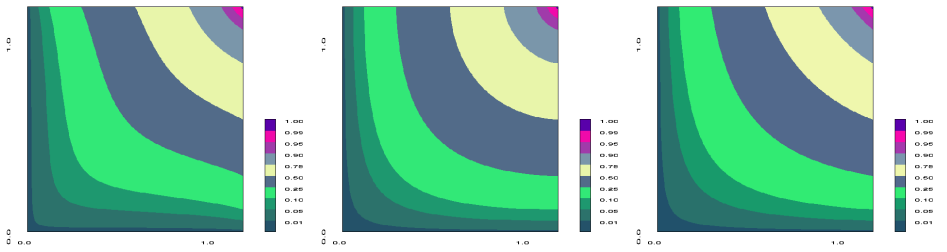

Figura 4.1: Contornos de isoprobabilidad de la función de distribución de la cópula de una mixtura de distribuciones normales de referencia en $[0,1]^2$, con los parámetros mostrados en la Tabla 4.1 (izq.). Contornos de isoprobabilidad de la distribución transformada por extracción de excesos (centro). Contornos de isoprobabilidad de la distribución transformada por extracción de máximos (dcha.).

Se ha aplicado la transformación por excesos (Sec. 4.1) a la cópula de referencia $C_\delta[u_1, u_2]$, con un umbral bivariado $h_1 = 0.5$, $h_2 = 0.7$. Los correspondientes contornos de isoprobabilidad e isodensidad de la cópula y la densidad de cópula transformadas se muestran en las figuras 4.1 y 4.2, centro. Los cambios entre la cópula de referencia y su transformada se pueden apreciar fácilmente y los más destacados se observan en los cuantiles intermedios. Los cambios son mucho más apreciables en los contornos de isodensidad: la simetría en la diagonal secundaria se ha perdido, y se observan valores de densidad mayores en extremo superior derecho, (1,1). Se han transformado los valores de la subcópula de los excesos sobre el umbral escogido, el resto de valores de la cópula no intervienen en la transformación.

La cópula de referencia $C_\delta[u_1, u_2]$ se ha utilizado también para ilustrar la transformación de cópulas bajo extracción de máximos, tal y como se describe en la sección 4.2. Para aplicar esta transformación, se ha asumido que el proceso de Poisson subyacente tiene una tasa de ocurrencia de un suceso por año, y el máximo se extrae de un registro de $t = 500$ años. La cópula transformada por extracción de máximos (Fig. 4.1, derecha), y su densidad (Fig. 4.2, derecha) muestran que la cópula de referencia ha cambiado notablemente. Si se aumenta el tiempo de observación t, la cópula transformada va cambiando, aproximándose

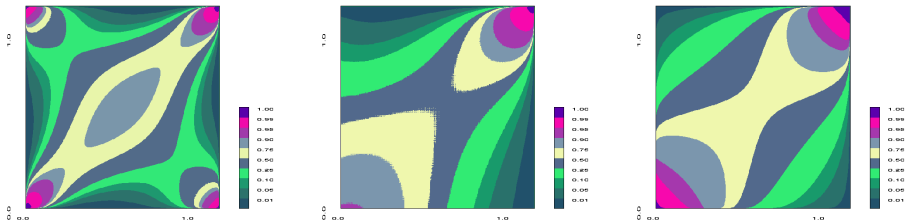

Figura 4.2: Contornos de isodensidad de la función de densidad de la cópula de una mixtura de distribuciones normales de referencia en $[0,1]^2$, con los parámetros mostrados en la Tabla 4.1 (izq.). Contornos de isodensidad de la densidad transformada por extracción de excesos (centro). Contornos de isodensidad de la transformación por extracción de máximos (dcha.).

a la cópula asintótica para máximos. Los cambios son mucho más evidentes, dado que en la transformación de máximos se mezclan sucesos correspondientes a sucesos distintos, mientras que la transformación de excesos mantenía la relación entre excesos correspondientes al mismo suceso. Además, se transforman todos los valores de la cópula, no una subcópula como en el caso de los excesos. Se observa que esta cópula (Fig. 4.1, dcha. ; 4.2, dcha.) no tiene semejanza con las cotas de Fréchet ni con la cópula de la independencia. Esto corresponde a la situación descrita en Coles et al. (1999), donde, en el sentido asintótico, los extremos pueden ser o bien dependientes (con diferentes grados de dependencia), o bien independientes.

Las transformaciones de cópulas presentadas, por extracción de excesos y de máximos, simplifican considerablemente los cálculos de cantidades predictivas de interés. No obstante, ambas transformaciones tienen características distintas: en la transformación de excesos se transforma sólo la subcópula correspondiente, mientras que para máximos se transforma toda la cópula. Y lo más importante, para excesos se transforman excesos correspondientes al mismo suceso, mientras que para máximos se mezclan valores correspondientes a sucesos distintos. Por ello, en el desarrollo posterior se centrará el interés en los excesos sobre un umbral suficientemente alto, y en la dependencia entre ellos.

Capítulo 5

Cópulas de mínima información mutua dados sus momentos (CrEnC)

En la Sección 2.6.2 se ha descrito la forma de la densidad conjunta $f(x_1, x_2)$ de dos variables aleatorias X_1 y X_2 de las que se dispone de información en forma de restricciones: se conocen las funciones de densidad marginales de X_1 y X_2 y se dispone de un número finito de otras restricciones en forma de momentos que la función de densidad $f(x_1, x_2)$ debe cumplir. La densidad $f(x_1, x_2)$ tal que satisface las anteriores restricciones y además tiene mínima entropía cruzada respecto las marginales dadas tiene la expresión (Ec. 2.6.7):

$$f(x_1, x_2) = a(x_1)b(x_2) \exp\left[\sum_{j=1}^{J} \alpha_j T_j(x_1, x_2)\right] \ , \ (x_1, x_2) \in \mathbb{R}^2 \ .$$

En la Proposición 2.6.1 se enunció que la entropía de Shannon y la información de Kullback-Leibler no dependen de la elección de marginales, sino únicamente de la relación de dependencia existente entre ellas, representada por la función cópula correspondiente. Así, se desea determinar la cópula bivariada $C[u, v|\lambda_1^*, \ldots, \lambda_J^*]$ de mínima dependencia (es decir, mínima información mutua o entropía cruzada respecto el producto de sus marginales uniformes) en Ω_C, (ver Def.

2.6.4),

$$\Omega_C = \{C : E_C[\gamma_j(u,v)] = \vartheta_j, j = 1, ..., J; U_{X1}(x_1); U_{X2}(x_2)\} \ , \quad (5.0.1)$$

donde $\gamma_j(u,v)$ son funciones reales, integrables respecto a $dC(u,v)$, y $\vartheta_j, \ j = 1, \ldots, J$ son momentos conocidos (y adecuados) y ambas marginales son Uniformes[0,1]. Es conveniente utilizar momentos basados en cópulas, dada su invariancia por cambio de marginales (ver Sección 2.2.2).

La densidad de la cópula de mínima entropía cruzada, dadas sus marginales uniformes y dados sus momentos, que a partir de ahora denominaremos cópula CrEnC, es de la forma

$$c(u,v) = a(u)b(v) \exp\left[\sum_{j=1}^{J} \alpha_j \vartheta_j(u,v)\right] \ , \ (u,v) \in \mathcal{S} \ . \quad (5.0.2)$$

5.1. Representación de cópulas en $\mathbb{R}^2$

La definición general de función de densidad como derivada de Radon-Nikodým de una probabilidad relativa a una medida es válida en cualquier espacio. Habitualmente trabajamos con variables o vectores aleatorios reales y por lo tanto utilizamos densidades respecto a la medida de Lebesgue del espacio real. Sin embargo, los problemas aparecen cuando trabajamos con espacios soporte donde no es coherente considerar la medida de Lebesgue.

La definición más común de las funciones cópula incluye soporte en un recinto limitado, usualmente $[0,1] \times [0,1]$. Podemos considerar este espacio como un subconjunto de $\mathbb{R}^2$ y consecuentemente podemos utilizar densidades respecto a la medida de Lebesgue. No obstante, Mateu-Figueras et al. (2013) introduce una estructura de espacio vectorial euclidiano diferente a la de $\mathbb{R}^2$ que induce una medida diferente a la habitual del espacio real. En este contexto también se podría definir una densidad de probabilidad respecto a esta medida *adecuada*, correspondiente a la estructura del soporte acotado.

Para evitar ese y otros problemas generados por este carácter acotado del soporte de las cópulas (Schmid et al., 2010; Ortego and Mateu-Figueras, 2006; Pawlowsky-Glahn and Egozcue, 2001), se propone expresar las densidades correspondientes en otro espacio, sacando partido

de las propiedades de las funciones cópula y del uso de momentos basados en éstas, invariantes por cambio de marginales. En la Figura 5.1 se muestra el ajuste de una cópula Gumbel clásica a un conjunto de datos con poca dependencia, tanto en $[0,1] \times [0,1]$ como en $\mathbb{R}^2$. La representación en este soporte no limitado permite apreciar mucho mejor las pequeñas diferencias (incluso visualmente) y el ajuste a una muestra de datos.

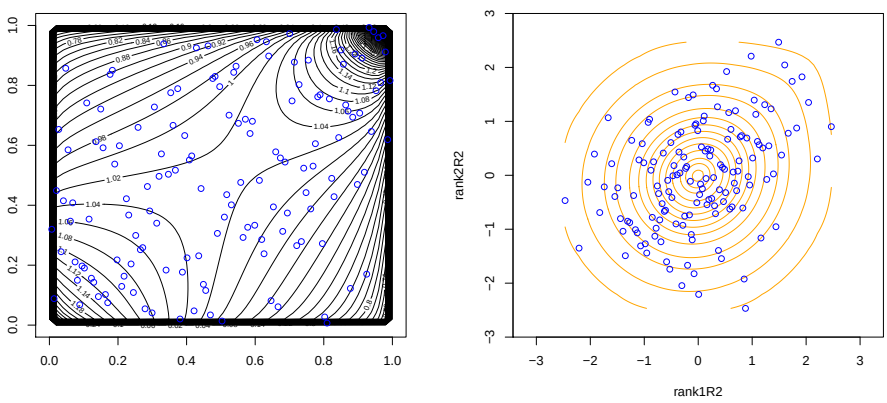

Figura 5.1: Curvas de isodensidad de Cópula Gumbel en $[0,1]^2$ y $\mathbb{R}^2$. En azul, pseudo-observaciones de un conjunto de datos bivariado.

Una alternativa a la representación de las cópulas CrEnC en $[0,1]^2$ sería utilizar el espacio soporte del modelo propuesto, donde las marginales son GPD. Esto trasladaría la expresión de la dependencia de $[0,1]^2$ a $\mathbb{R}^{+2}$, un subconjunto de $\mathbb{R}^2$ donde la escala adecuada también podría ser la relativa. Se ha optado por realizar un cambio de soporte de la cópula CrEnC de $[0,1] \times [0,1]$ a $\mathbb{R}^2$ aplicando la transformación Probit a cada una de las marginales, de manera que la escala en el nuevo espacio es la escala absoluta. El par de variables transformado tiene marginales conocidas ($N(0,1)$) y su dependencia queda inalterada (debido a la selección de momentos y las propiedades de la cópula). Su función de densidad conjunta seguirá siendo la de mínima entropía cruzada respecto a las marginales. Denominaremos a esta función equivalente cópula CrEnC en $\mathbb{R}^2$.

El proceso de estimación de los parámetros de la cópula CrEnC se realiza en $\mathbb{R}^2$ en lugar de en el soporte limitado dado que la dependencia entre las magnitudes queda inalterada. Se propone el siguiente algoritmo de determinación de la cópula de mínima información mutua dadas las marginales y un conjunto de momentos invariantes por transformaciones monónotonas (CrEnC) en Ω_C (Ec. 5.0.1):

Sean U, V, variables aleatorias uniformes en $[0,1]^2$, dependientes.

• Algoritmo:

1. Aplicar la transformación probit ($\Phi^{-1}(u)$) a cada una de las marginales, donde $\Phi(\cdot)$ denota la función de distribución $N(0,1)$. El par de variables transformadas (X_1, X_2) tiene marginales $N(0,1)$, y su dependencia no se ha alterado.

2. Hallar la distribución bivariada $F^*(x_1, x_2)$ de mínima información mutua en Ω_F,

$$\Omega_F = \{F : E_F[T_j(x_1, x_2)] = \theta_j, j = 1, ..., J; \phi_{X1}(x_1); \phi_{X2}(x_2)\} \ ,$$

donde $T_j(x_1, x_2)$ son funciones reales, integrables respecto a $dF(x_1, x_2)$, y $\theta_j,\ j = 1, \ldots, J$ son momentos conocidos. De hecho, $\theta_j,\ j = 1, \ldots, J$ son las transformaciones probit de $\gamma_j(u, v)$ y $\vartheta_j,\ j = 1, \ldots, J$, respectivamente. Asimismo $\phi_{X1}(x_1), \phi_{X2}(x_2)$ denotan las densidades marginales $N(0,1)$.

El Teorema 2.6.3 proporciona una caracterización de la densidad de mínima información mútua respecto al producto de densidades marginales $g(x_1, x_2) = \phi(x_1) \cdot \phi(x_2)$ en Ω_F. Esta densidad conjunta es de la forma (Ec. 2.6.7)

$$f^*(x_1, x_2) = a(x_1)b(x_2) \exp\left[\sum_{j=1}^{J} \alpha_j T_j(x_1, x_2)\right] \ ,$$

y su correspondiente función de distribución es $F^*(x_1, x_2)$. En general, y dependiendo de la elección de los momentos θ_J, esta distribución conjunta no será normal bivariante.

3. El teorema de Sklar permite expresar la distribución conjunta F^* como combinación de sus marginales Normales mediante su

función cópula :

$$F^*(x_1, x_2) = C[\Phi(x_1), \Phi(x_2)|\alpha_1, \ldots, \alpha_J] \ ,$$

donde $C[u, v|\alpha_1, \ldots, \alpha_J]$ denota la correspondiente función cópula, la cópula de mínima información mutua en Ω_C.

5.2. Normalización de la densidad CrEnC

El Teorema 3 de Rumsey y Posner (Rumsey Jr and Posner, 1965, Teorema 2.6.3), permite caracterizar la forma de la densidad conjunta de mínima entropía cruzada respecto a las marginales dadas ($f_{X1}(x_1)$, $f_{X2}(x_2)$) y a un conjunto de momentos dados, θ_i, $i = 1, \ldots, J$. Denotamos esta densidad por $f(x_1, x_2)$ (Ec. 2.6.7).

Según el Teorema 2.6.3, las funciones $a(x_1), b(x_2)$ y los parámetros $\alpha_1, \ldots, \alpha_J$ que conforman esta función de densidad conjunta de mínima entropía cruzada con restricciones son únicas. Estos parámetros y funciones son las soluciones del conjunto de ecuaciones integrales simultáneas:

$$\begin{aligned}
&a(x_1) \int_{-\infty}^{+\infty} b(x_2) \exp\left[\sum_{j=1}^{J} \alpha_j T_j(x_1, x_2)\right] dx_2 = f_{X1}(x_1) \ , \\
&b(x_2) \int_{-\infty}^{+\infty} a(x_1) \exp\left[\sum_{j=1}^{J} \alpha_j T_j(x_1, x_2)\right] dx_1 = f_{X2}(x_2) \ , \\
&\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} a(x_1) b(x_2) \exp\left[\sum_{j=1}^{J} \alpha_j T_j(x_1, x_2)\right] T_j(x_1, x_2) dx_1 dx_2 = \theta_i \ , \\
&i = 1, \ldots, J \ ,
\end{aligned} \tag{5.2.1}$$

donde las funciones $a(x_1), b(x_2)$ normalizan la densidad CrEnC. En el caso particular en que las restricciones en forma de momentos son todas nulas estas funciones corresponden con las densidades marginales. Pero en el caso general estas funciones no tendrán una expresión cerrada y deben aproximarse numéricamente.

A continuación, se propone un algoritmo de aproximación de las funciones $a(x_1), b(x_2)$. Según Rumsey y Posner, la función de densidad conjunta de mínima entropía cruzada dadas las marginales $f_{X1}(x_1), f_{X2}(x_2)$

y los momentos $\theta_1, \dots \theta_J$ es de la forma

$$f(x_1, x_2) = a_\alpha(x_1) b_\alpha(x_2) \exp[\boldsymbol{\alpha}\mathbf{T}(x_1, x_2)] \ , \ x_1, x_2 \in \mathbb{R} \qquad (5.2.2)$$

donde $T_j(x_1, x_2)$ son los estadísticos correspondientes a los momentos θ_j. Los parámetros $\alpha_j, j = 1 \dots, J$ son los correspondientes a estas restricciones.

Sea $(x_1^1, x_2^1) \dots, (x_1^n, x_2^n)$ una muestra aleatoria de las variables X_1,X_2 y $\alpha_j, j = 1 \dots, J$ los parámetros correspondientes a los momentos $\theta_1, \dots \theta_J$, supuestos conocidos. Además, sea $\{\widetilde{x_1}^1, \dots, \widetilde{x_1}^m\}$ una muestra de X_1, generada según $F_{X1}(x_1)$, con m *suficientemente grande*. Análogamente, sea $\{\widetilde{x_2}^1, \dots, \widetilde{x_2}^l\}$ una muestra de X_2, generadas según $F_{X2}(x_2)$, con l suficientemente grande

Dada la función de densidad conjunta, $f(x_1, x_2)$, podemos obtener la marginal de X_1 integrando:

$$f_{X1}(x_1) = a_\alpha(x_1) \int b_\alpha(x_2) \exp(\boldsymbol{\alpha}\mathbf{T}(x_1, x_2)) dx_2 \ , \ x_1 \in \mathbb{R} \ .$$

Esta integral puede aproximarse mediante el método de Montecarlo (Chen et al., 2000). Este método permite aproximar una integral, pensada como valor esperado de una función de una variable aleatoria, por el promedio de las realizaciones de esta función en una muestra de esa variable. Así, la integral de la función de densidad marginal de X_1 puede reescribirse como

$$f_{X1}(x_1) = a_\alpha(x_1) \int b_\alpha(x_2) \frac{\exp(\boldsymbol{\alpha}\mathbf{T}(x_1, x_2))}{f_{X2}(x_2)} f_{X2}(x_2) dx_2 \ ,$$

expresión que corresponde al valor esperado de una función de la variable X_2. Dada una muestra suficientemente grande de X_2, $\{\widetilde{x_2}^1, \dots, \widetilde{x_2}^m\}$, simulada según la distribución $F_{X2}(x_2)$, podemos aproximar esa integral por

$$f_{X1}(x_1) \simeq a_\alpha(x_1) \frac{1}{m} \sum_{j=1}^{m} b_\alpha(\widetilde{x_2}^j) \frac{\exp(\boldsymbol{\alpha}\mathbf{T}(x_1, \widetilde{x_2}^j))}{f_{X2}(\widetilde{x_2}^j)} \ . \qquad (5.2.3)$$

En particular, si el tamaño de la muestra bivariada $(x_1^1, x_2^1) \dots, (x_1^n, x_2^n)$ es grande, podría utilizarse la propia muestra marginal $\{x_2^1, \dots, x_2^n\}$.

La marginal de X_2, $f_{X2}(x_2)$, se puede aproximar de una manera análoga. Dada una muestra suficientemente grande de X_1, $\{\widetilde{x_1}^1, \ldots, \widetilde{x_1}^l\}$, simulada según la distribución $F_{X1}(x_1)$,

$$f_{X2}(x_2) \simeq b_\alpha(x_2)\frac{1}{l}\sum_{i=1}^{l} a_\alpha(\widetilde{x_1}^i)\frac{\exp(\boldsymbol{\alpha}\mathbf{T}(\widetilde{x_1}^i, x_2))}{f_X(\widetilde{x_1}^i)} \ . \tag{5.2.4}$$

Despejando las funciones $a_\alpha(x_1), b_\alpha(x_2)$ de las Ec. 5.2.3 y 5.2.4, se obtiene:

$$a_\alpha(x_1) \simeq \frac{f_{X1}(x_1)}{\frac{1}{m}\sum_{j=1}^{m} b_\alpha(\widetilde{x_2}^j)\frac{\exp(\boldsymbol{\alpha}\mathbf{T}(x_1,\widetilde{x_2}^j))}{f_{X2}(\widetilde{x_2}^j)}} \ , \tag{5.2.5}$$

$$b_\alpha(x_2) \simeq \frac{f_{X2}(x_2)}{\frac{1}{l}\sum_{i=1}^{l} a_\alpha(\widetilde{x_1}^i)\frac{\exp(\boldsymbol{\alpha}\mathbf{T}(\widetilde{x_1}^i,x_2))}{f_{X1}(\widetilde{x_1}^i)}} \ , \tag{5.2.6}$$

donde se observa que las funciones $a_\alpha(x_1), b_\alpha(x_2)$ pueden determinarse mediante un proceso iterativo.

Dadas las variables (X_1, X_2) y los momentos conjuntos $\theta_1, \ldots, \theta_J$, al cual le corresponden los estadísticos $T_1(x_1, x_2), \ldots, T_J(x_1, x_2)$, y la muestra aleatoria $D = \{(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)\}$, sea $f(x_1, x_2)$ la correspondiente función de densidad conjunta de mínima entropía cruzada respecto a las marginales dadas $(f_{X1}(x_1), f_{X2}(x_2))$ y a los momentos $\theta_1, \ldots, \theta_J$. La densidad $f(x_1, x_2)$ tiene expresión $f(x_1, x_2) = a_\alpha(x_1)b_\alpha(x_2)\cdot \exp[\boldsymbol{\alpha}\mathbf{T}(x_1, x_2)]$, $x_1, x_2 \in \mathbb{R}$ (Ec. 5.2.2). Se propone el siguiente algoritmo de estimación para las funciones $a_\alpha(x_1), b_\alpha(x_2)$ de la función de densidad conjunta:

1. Simular una muestra $\{\widetilde{x_1}^1, \ldots, \widetilde{x_1}^m\}$ de X_1, generada a partir de $F_{X1}(x_1)$, con m suficientemente grande.

2. Simular una muestra $\{\widetilde{x_2}^1, \ldots, \widetilde{x_2}^l\}$ de X_2, generada a partir de $F_{X2}(x_2)$, con l suficientemente grande.

3. Evaluar la cantidad $\exp[\boldsymbol{\alpha}\mathbf{T}(x_1, x_2)]$ en la muestra $D = \{(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)\}$: $\exp[\boldsymbol{\alpha}\mathbf{T}(x_1^i, x_2^i)], i = 1, \ldots, n$.

4. Dar valores iniciales de las funciones $a_\alpha(x_1)$,$b_\alpha(x_2)$ en la muestra $D = \{(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)\}$: $a_\alpha^0(x_1^i), b_\alpha^0(x_2^j), i = 1, \ldots, m; j = 1, \ldots, l$.

5. Aproximar de manera iterativa el valor de las funciones $a_\alpha(x_1)$, $b_\alpha(x_2)$ en la muestra D, utilizando las expresiones 5.2.5 y 5.2.6 respectivamente. El proceso se repite como máximo niter veces, donde donde niter es el número máximo de iteraciones establecido previamente.

   - iteración 1
     Dados los valores iniciales $b^0_\alpha(x_2^j), j = 1, \ldots, n$, aproximar $a^1_\alpha(x_1^i)$, $i = 1, \ldots, n$ mediante el método de MonteCarlo (Ec. 5.2.5):

$$a^1_\alpha(x_1^i) \simeq \frac{f_{X1}(x_1^i)}{\frac{1}{m}\sum_{j=1}^{m} b^0_\alpha(\widetilde{x_2}^j)\frac{\exp(\boldsymbol{\alpha}\mathbf{T}(x_1^i,\widetilde{x_2}^j))}{f_{X2}(\widetilde{x_2}^j)}} \ ;$$

     dados los valores $a^1_\alpha(x_1^i)$, aproximar $b^1_\alpha(x_2^j), j = 1, \ldots, n$ mediante el método de MonteCarlo (Ec. 5.2.6):

$$b^1_\alpha(x_2^j) \simeq \frac{f_{X2}(x_2^j)}{\frac{1}{l}\sum_{i=1}^{l} a^1_\alpha(\widetilde{x_1}^i)\frac{\exp(\boldsymbol{\alpha}\mathbf{T}(\widetilde{x_1}^i,x_2^j))}{f_{X1}(\widetilde{x_1}^i)}} \ .$$

   - ...
   - iteración= iter, $1 < iter \leq niter$:
     - Calcular

$$a^{iter}_\alpha(x_1^i) i = 1, \ldots, n \ \ ; \ \ b^{iter}_\alpha(x_2^j), j = 1, \ldots, n \ ;$$

     - si la *diferencia* entre las dos últimas iteraciones es pequeña, parar. Guardar $a^{iter}_\alpha(x_1^i)$; $b^{iter}_\alpha(x_2^j)$.
       Tomamos como medida de la diferencia entre dos iteraciones consecutivas a la variación logarítmica de la verosimilitud estimada en ambas iteraciones (ver Obs. 5.2.1)

$$\sum_{i=1}^{n} \ln(a^{iter}_\alpha(x_1^i)) + \sum_{i=1}^{n} \ln(b^{iter}_\alpha(x_2^i)) - \sum_{i=1}^{n} \ln(a^{iter-1}_\alpha(x_1^i)) - \sum_{i=1}^{n} \ln(b^{iter-1}_\alpha(x_2^i)) \leq 10^{-4} \ . \tag{5.2.7}$$

- ...

6. Si iteración= niter+1, parar.
   Guardar $a_\alpha^{niter}(x_1^i), i = 1, \ldots, n \;\; ; \;\; b_\alpha^{niter}(x_2^j), j = 1, \ldots, n \;\; .$

Observación 5.2.1. [Distancia entre las estimaciones de $a_\alpha(x_1)$, $b_\alpha(x_2)$ en dos iteraciones consecutivas] Dada una muestra $D = \{(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)\}$, con funciones de distribución marginales $F_{X1}(x_1)$, $F_{X2}(x_2)$ respectivamente, la verosimilitud de los parámetros $\alpha_1, \ldots, \alpha_J$ de la función de densidad conjunta de mínima entropía cruzada respecto las marginales y los momentos $\theta_1, \ldots, \theta_J$, respecto a la muestra es

$$L(\alpha_1, \ldots, \alpha_J | D) = \prod_{i=1}^{n} a_\alpha(x_1^i) \prod_{i=1}^{n} b_\alpha(x_2^i) \exp\left(\sum_{j=1}^{J} \alpha_j \sum_{i=1}^{n} T_j(x_1^i, x_2^i) \right) \;\; ,$$

y la correspondiente logverosimilitud

$$l(\alpha_1, \ldots, \alpha_J | D) = \sum_{i=1}^{n} \ln a_\alpha(x_1^i) + \sum_{i=1}^{n} \ln b_\alpha(x_2^i) + \left(\sum_{j=1}^{J} \alpha_j \sum_{i=1}^{n} T_j(x_1^i, x_2^i) \right) \;\; .$$

En el proceso de estimación de $a_\alpha(x_1), b_\alpha(x_2)$, supuesto conocido α, se desea medir la diferencia entre dos estimaciones de esas funciones en la muestra. Sean $a_\alpha^{iter-1}(x_1^i), b_\alpha^{iter-1}(x_2^i)$ y $a_\alpha^{iter}(x_1^i), b_\alpha^{iter}(x_2^i)$, $i = 1, \ldots, n$ las estimaciones de las funciones en la muestra para las iteraciones *iter* e $iter - 1$. La variación relativa de la verosimilitud entre estas dos iteraciones sucesivas, dados $\alpha_1, \ldots, \alpha_J$, es:

$$\frac{\mathrm{L}_{iter}(\alpha_1, \ldots, \alpha_J | (x_1^1, x_2^1) \ldots, (x_1^n, x_2^n))}{\mathrm{L}_{iter-1}(\alpha_1, \ldots, \alpha_J | (x_1^1, x_2^1) \ldots, (x_1^n, x_2^n))} = \frac{\prod_{i=1}^{n} a_\alpha^{iter}(x_1^i) \prod_{i=1}^{n} b_\alpha^{iter}(x_2^i)}{\prod_{i=1}^{n} a_\alpha^{iter-1}(x_1^i) \prod_{i=1}^{n} b_\alpha^{iter-1}(x_2^i)} \;\; .$$

Considerando la escala relativa de estas diferencias, tomaremos la variación logarítimica de la verosimilitud (Ec. 5.2.8) como distancia entre los valores de ambas iteraciones sucesivas:

$$d(iter, iter-1) = \sum_{i=1}^{n} \ln(a_\alpha^{iter}(x_1^i)) + \sum_{i=1}^{n} \ln(b_\alpha^{iter}(x_2^i)) - \sum_{i=1}^{n} \ln(a_\alpha^{iter-1}(x_1^i)) - \sum_{i=1}^{n} \ln(b_\alpha^{iter-1}(x_2^i)) \;\; . \tag{5.2.8}$$

Capítulo 6

Estimación Bayesiana de cópulas paramétricas en procesos de Poisson evaluados

En el Capítulo 3 se ha presentado un modelo apropiado para la modelización de datos de tipo extremal. Se desea establecer un proceso de estimación de los parámetros de este modelo. Se considera el proceso evaluado de Poisson de parámetro λ, donde cada uno de los dos tamaños (X_1, X_2) está distribuido según una $GPD(\xi_i, \beta_i)$, $i = 1, 2$. Además, se describe la dependencia entre los dos tamaños (X_1, X_2) mediante una familia de cópulas. Aquí consideramos una cópula paramétrica $C_{\mathbf{X}}[\cdot, \cdot|\delta]$, dependiente de δ.

Consideramos los tamaños (X_1, X_2), excesos sobre un umbral absoluto $\mathbf{h_0} = (h_1^0, h_2^0)$. Estos umbrales son superiores a las cotas inferiores x_{0i}, $i = 1, 2$ presentadas en el Capítulo 3 (ver Fig. 6.1). En el ejemplo presentado en ese capítulo, una base de datos de precipitación diaria registrada en varias ubicaciones, los tamaños (X_1, X_2) pueden corresponder a los excesos sobre un umbral bivariado $(h_1^0 = 20mm, h_2^0 = 20mm)$, claramente superior a la cota inferior del soporte $(x_{01} = 1mm, x_{02} = 1mm)$.

Los datos disponibles consisten en observaciones de los dos tamaños X_1, X_2 en intervalos de tiempo (t_1, t_2), que pueden ser de diferentes longitudes. En todo caso, en cada intervalo existirá información de

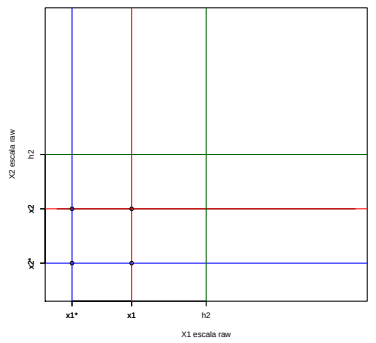

Figura 6.1: Asignación de valores por defecto $x = x^*_{0i}$, $i = 1, 2$ en las marginales, para aquellos valores no registrados. $x = x_{0i}$ indica la cota inferior del soporte y h^0_i los umbrales absolutos sobre los que se definen los excesos.

alguna de las dos variables. Por tanto, la muestra de datos disponible, denotada por D, contendrá observaciones de tres tipos: observaciones sólo de x_1, (x^i_1, na); observaciones sólo de x_2, (na, x^i_2) y observaciones conjuntas de ambas variables (x^i_1, x^i_2). La notación na indica que no existe exceso sobre el umbral h^0_i para esa variable. El comportamiento conjunto de las variables, su dependencia, se estimará mediante las observaciones conjuntas de las variables. El comportamiento marginal se estimará utilizando tanto las observaciones conjuntas como las observaciones de sólo una de las variables. Dado que los sucesos extremales que se desean modelizar son raros, y en general se dispondrá de pocos datos, es importante incorporar al modelo toda la información disponible, aunque ésta sea parcial. Como el modelo propuesto tiene gran cantidad de parámetros y los datos serán escasos, parece adecuado realizar una estimación bayesiana de los parámetros del modelo (Sec. 2.5), de manera que se tenga en cuenta la incertidumbre de la estimación.

Los parámetros del modelo propuesto pueden clasificarse en tres tipologías: parámetros de ocurrencia, de los tamaños y de la dependencia entre los tamaños.

parámetro λ: la ocurrencia de los sucesos en el tiempo se modela mediante un proceso de Poisson de parámetro λ;

parámetros $\zeta_i = (\xi_{Xi}, \beta_{Xi})$**,** $i = 1, 2$**:** los tamaños de los sucesos, (X_1,X_2), considerados excesos sobre un umbral absoluto $\mathbf{h_0}$ establecido a priori, se suponen distribuidos según una distribución $GPD(\xi_{Xi}, \beta_{Xi})$, $i = 1, 2$;

parámetro δ**:** la dependencia entre los tamaños (X_1, X_2) se modela mediante una familia de cópulas paramétrica, $C_{\mathbf{X}}[\cdot, \cdot|\delta]$, donde δ puede ser uni o multidimensional.

Estos parámetros (ocurrencia, tamaño y dependencia) se estimarán conjuntamente utilizando técnicas bayesianas. Mediante el muestreo de Gibbs (Sec. 2.5.1) se obtendrá una muestra del posteriori conjunto de los parámetros, que sustituirá al propio posteriori conjunto. Esta muestra del posteriori se obtiene a partir de las distribuciones condicionales a posteriori, y por tanto solo se precisa determinar las verosimilitudes condicionales y los prioris para cada conjunto de parámetros.

6.1. El priori conjunto de los parámetros

Las hipótesis de independencia establecidas para el modelo (hipótesis de independencia entre la ocurrencia de los sucesos y su tamaño, Definición 2.3.6 p.ej.), se traducen a independencia entre parámetros. Por tanto, el priori conjunto $f_{\delta,\zeta_1,\zeta_2,\lambda}(\delta, \zeta_1, \zeta_2, \lambda)$ (donde ζ_i denota al par de parámetros de la marginal $GPD(\xi_{Xi}, \beta_{Xi}), i = 1, 2$) puede expresarse como producto de los prioris marginales de cada conjunto de parámetros: $f_{\zeta_1}(\zeta_1)$, $f_{\zeta_2}(\zeta_2)$, $f_\delta(\delta)$ y $f_\lambda(\lambda)$. Los prioris establecidos para cada parámetro se describen a continuación.

6.1.1. Parámetros marginales GPD

Sea $f_{Xi}(x_i)$ la densidad de la distribución $GPD(\xi_i, \beta_i)$ marginal de $X_i, i = 1, 2$. El parámetro ξ_i de la distribución tiene soporte en los reales, y el parámetro β_i es real positivo. La distribución generalizada de Pareto (sección 2.4) engloba tres familias de distribuciones (dominios de atracción) según los valores del parámetro ξ_i: dos con soporte no acotado (Gumbel, para $\xi_i = 0$, y Fréchet, para $\xi_i > 0$) y una con soporte acotado superiormente (Weibull, para $\xi_i < 0$).

Se desea construir el priori conjunto de los parámetros marginales $\zeta_i = (\xi_{Xi}, \beta_{Xi})$ de la magnitud $X_i, i = 1, 2$. Se desea que los valores admitidos en el priori conjunto de estos parámetros corresponda a características de la magnitud modelizada. Por ello, el priori se construye a partir de juicio experto sobre el fenómeno. El resultado del juicio experto será un recinto de posibles valores de $\zeta_i = (\xi_{Xi}, \beta_{Xi}), i = 1, 2$ sobre el cual se establece un priori uniforme. Ilustraremos el procedimiento utilizando el ejemplo presentado en el Capítulo 11, donde se modeliza la precipitación diaria en dos ubicaciones cercanas, situadas en la provincia de Alicante.

Inicialmente, se pide al experto que responda algunas cuestiones sobre la magnitud que se modeliza:

1. ¿La magnitud tiene un soporte acotado superiormente?
2. Indicar un umbral de referencia para la magnitud
3. Indicar el periodo de retorno correspondiente al umbral de referencia indicado
4. Indicar un valor de la magnitud que seguro se puede alcanzar
5. Indicar un valor de la magnitud casi imposible (con probabilidad anual 10^{-4} o inferior)

En el ejemplo, el experto consideró que la precipitación diaria en la provincia de Alicante es una magnitud con un soporte acotado superiormente. Dadas las características de los datos disponibles, se tomó un umbral de 20mm como referencia, al cual se le asoció un periodo de retorno de 0.1 años. Se consideró que 200m era un valor de precipitación diaria que se podía alcanzar, mientras que 2000mm era un valor de precipitación con una probabilidad anual inferior a 10^{-4}. La Tabla 6.1 contiene el resumen de estas afirmaciones.

Estas afirmaciones sobre la magnitud permiten determinar un recinto inicial para el priori (Fig. 6.2). Este recinto inicial se puede perfilar añadiendo otras afirmaciones suplementarias sobre la magnitud. Por ejemplo, las afirmaciones pueden ser de los tipos siguientes:

1. Fijar una magnitud de referencia y un intervalo del periodo de retorno correspondiente a esta magnitud, o bien,

Tabla 6.1: Ejemplo de entrada de afirmaciones iniciales para el recinto inicial del priori

1	Finite support of magnitude
20.0	Reference threshold of magnitude for this prior
0.1	Return period for the reference threshold of magnitude
200.0	Magnitude that is surely attainable
2000.0	Almost imposible event (annual probability 10-4 or less)

2. Fijar un periodo de retorno de referencia y un intervalo de magnitudes correspondiente a este periodo de retorno.

En caso de duda sobre el intervalo pedido, éste debe hacerse mayor. En el ejemplo referido, el experto consideró que una magnitud de 50mm de precipitación diaria en la ubicación tiene un periodo de retorno de entre 0.2 y 5.0 años. Además, consideró que en esa ubicación, las precipitaciones diarias que corresponden a un periodo de retorno de 100 años se encuentran entre 150 y 1200mm. La Tabla 6.2 muestra el resumen de esta información.

Tabla 6.2: Ejemplo de entrada de dos afirmaciones complementarias

1	Number of additional prior statements
	.....statement 1.....
1	(1) Magnitude Ref.+Return Per. interval; (2) Return Per. Ref.+ Magnitude interval
50.	Magnitude Reference (1), Return Period Reference (2)
0.2 5.0	Interval for return period (1); Interval of magnitude (2)
	.....statement 2.....
2	(1) Magnitude Ref.+Return Per. interval; (2) Return Per. Ref.+ Magnitude interval
100.	Magnitude Reference (1), Return Period Reference (2)
150. 1200.	Interval for return period (1); Interval of magnitude (2)

A partir de cada afirmación suplementaria se obtiene un recinto diferente para el priori (ver Fig. 6.2). Finalmente, se unifican los recintos obtenidos a partir de las afirmaciones iniciales y de las suplementarias, obteniendo el recinto conjunto (ver Fig. 6.3).

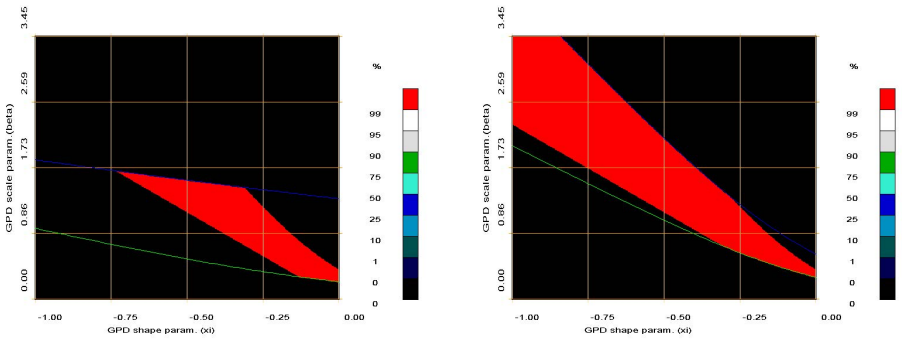

Figura 6.2: Recinto de definición del priori a partir de las afirmaciones iniciales (Izq.). Recinto de definición del priori a partir de la primera afirmación complementaria (Dcha.).

Se define una densidad sobre $\zeta_i = (\xi_{Xi}, \beta_{Xi})$ en el recinto determinado. Esta densidad uniforme se modifica de forma que cerca de los bordes del recinto tome valores cercanos a cero. Esta densidad uniforme suavizada en los borde se toma como priori. La Figura 6.4 muestra el priori utilizado en el ejemplo de precipitación diaria, y la Tabla 6.3 los valores de los límites del recinto utilizado.

Tabla 6.3: Ejemplo de determinación de recinto para priori conjunto.

-0.800	-0.001	min,max values of ξ
0.001	1.800	min,max values of β

6.1.2. Parámetros de la cópula Gumbel

El parámetro δ de la cópula paramétrica de Gumbel toma valores en $\delta \in [1, +\infty]$, y se puede considerar que su escala es relativa. Por tanto, parece conveniente modelizar $\log(\delta)$.

Se ha establecido un priori uniforme para $\log(\delta)$ en un intervalo $(\log(\delta_1), \log(\delta_2))$. Este intervalo se define mediante juicio experto a partir de tres valores: un valor admisible para δ y un intervalo (ρ_1, ρ_2)

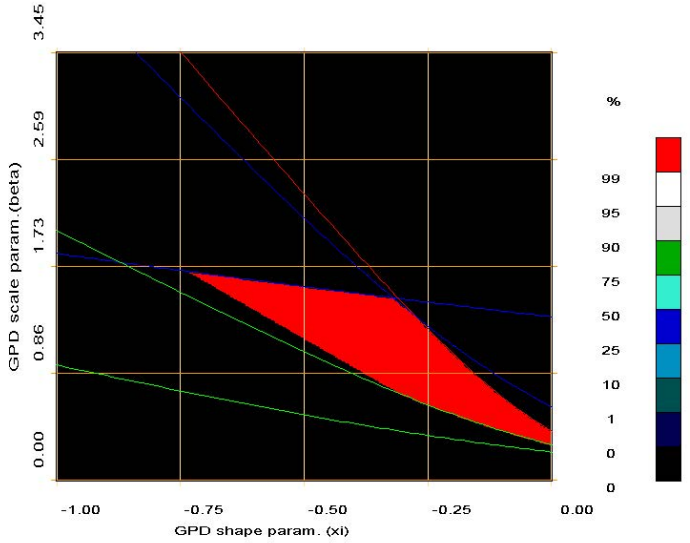

Figura 6.3: Recinto de definición del priori a partir de la combinación de afirmaciones.

para la correlación de Kendall que puede corresponder a las magnitudes modelizadas. La cópula Gumbel es arquimediana, y por tanto existe una relación exacta entre el parámetro de la cópula y el coeficiente de correlación de Kendall (ecuación 2.2.5). El intervalo $(\log(\delta_1), \log(\delta_2))$, donde se define el priori, se obtiene usando la relación entre ambos coeficientes.

6.1.3. Tasa de ocurrencia de Poisson

Sea τ el periodo de retorno correspondiente a la tasa de ocurrencia de Poisson, λ:

$$\tau = 1/\lambda \ ,$$

donde las unidades de τ son de tiempo, por ejemplo años, y las de λ son sus recíprocas, por ejemplo sucesos (año^{-1}). La escala del periodo de retorno puede considerarse relativa, y por tanto, parece adecuado

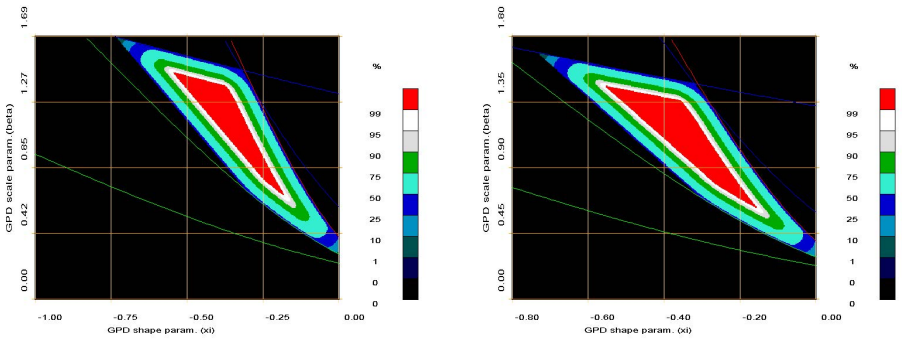

Figura 6.4: Priori sobre el recinto definido y su modificación.

modelizar su logaritmo. Se ha establecido un priori uniforme en un intervalo amplio de $\log(\tau) = -\log(\lambda)$, $(-\log(\lambda)_{min}, -\log(\lambda)_{max})$.

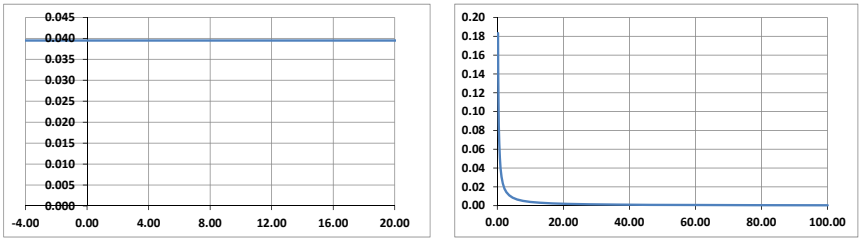

Figura 6.5: Representación del priori utilizado, uniforme en la escala $-\log(\lambda)$ (Izq.). Representación del priori utilizado en la escala de λ (Dcha.).

En el ejemplo presentado en el Capítulo 11 se modeliza la precipitación diaria en dos ubicaciones cercanas. Se ha establecido un priori uniforme en el intervalo (-4.605, 20.723) para el logaritmo del periodo de retorno de los sucesos de precipitación diaria que exceden 20mm , que corresponde a un periodo de retorno en un intervalo amplio entre 0.01 y 10^9 años. La Figura 6.5 representa el priori utilizado en la escala donde es uniforme y en la escala del parámetro λ.

6.2. Verosimilitud de los parámetros

Un punto clave en la estimación bayesiana es la obtención de la función de (log-)verosimilitud conjunta de los parámetros del modelo (ocurrencia, λ; tamaños, $\zeta_1 = (\xi_1, \beta_1), \zeta_2 = (\xi_2, \beta_2)$ y dependencia, δ). El modelo propuesto tiene seis parámetros implicados y la correspondiente verosimilitud conjunta será por tanto poco manejable. Sin embargo, al utilizar el muestreo de Gibbs basta determinar las verosimilitudes condicionales de cada parámetro dados los demás, que tendrán una expresión más sencilla.

Sea D una muestra bivariada de excesos de alguna de las variables (X_1, X_2): $D = \{(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)\}$, donde los pares denotan observaciones sólo de x_1, (x_1^i, na), observaciones sólo de x_2, (na, x_2^i), o bien observaciones conjuntas de ambas variables (x_1^i, x_2^i). Se denota na cuando no se ha registrado exceso sobre el umbral de referencia en una de las variables. Estas observaciones parciales son útiles para mejorar las estimaciones de los parámetros marginales, si las observaciones son escasas.

La hipótesis de independencia entre la ocurrencia de los sucesos y su tamaño (Def. 2.3.6) se traduce a independencia entre el parámetro λ y los parámetros $(\zeta_1, \zeta_2, \delta)$, y por tanto

$$\mathrm{L}(\zeta_1, \zeta_2, \delta, \lambda | D) = \mathrm{L}(\zeta_1, \zeta_2, \delta | D) \cdot \mathrm{L}(\lambda | D) \ ,$$

de manera que podremos determinar por separado las verosimilitudes condicionales de ambos conjuntos de parámetros.

6.2.1. Parámetros marginales. Verosimilitud condicional

Sean $\zeta_i = (\xi_i, \beta_i)$ los parámetros de las distribuciones GPD marginales de $X_i, i = 1, 2$. Denotaremos f_{X_i} las correspondientes densidades $GPD_X(\xi_{Xi},\ \beta_{Xi})$, $i = 1, 2$. Modelizamos la dependencia mediante la cópula $C_{\mathbf{X}}[u, v|\delta]$, cuya densidad es $c_{\mathbf{X}}[u, v|\delta]$.

La verosimilitud condicional de los parámetros marginales $\zeta_1 =$

(ξ_{X1}, β_{X1}) respecto los demás parámetros, se expresa como

$$\mathrm{L}(\zeta_1|D, \zeta_2, \delta) = \prod_{i=1}^{n} c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1), F_{X_2}(x_2^i|\zeta_2)|\delta] \cdot f_{X_1}(x_1^i|\zeta_1) \cdot f_{X_2}(x_2^i|\zeta_2) \ . \quad (6.2.1)$$

Esta expresión puede simplificarse, dado que los parámetros ζ_2 y δ son conocidos:

$$\mathrm{L}(\zeta_1|D, \zeta_2, \delta) \propto \prod_{i=1}^{n} c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1), F_{X_2}(x_2^i|\zeta_2)|\delta] \cdot f_{X_1}(x_1^i|\zeta_1) \ . \quad (6.2.2)$$

De una manera análoga, obtenemos la verosimilitud de los parámetros de la marginal X_2, $\zeta_2 = (\xi_{X2}, \beta_{X2})$ dados los parámetros ζ_1 y δ. Esta verosimilitud, ya simplificada, tiene expresión

$$\mathrm{L}(\zeta_2|D, \zeta_1, \delta) \propto \prod_{i=1}^{n} c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1), F_{X_2}(x_2^i|\zeta_2)|\delta] \cdot f_{X_2}(x_2^i|\zeta_2) \ . \quad (6.2.3)$$

6.2.2. Parámetros de la cópula Gumbel. Verosimilitud condicional

La dependencia entre las marginales X_1, y X_2 se ha modelizado mediante la cópula Gumbel $C_{\mathbf{X}}[\cdot, \cdot|\delta]$ (Ec. 2.1.2),

$$C_{\mathbf{X}}[u, v|\delta] = \exp(-[(-\ln(u))^\delta + (-\ln(v))^\delta]^{1/\delta}) \ , \ \delta \in [1, +\infty] \ .$$

Según el teorema de Sklar (Teorema 2.1.2), la distribución conjunta de X_1 X_2 se puede expresar en función de sus marginales: $F_{12}(x_1, x_2) = C_{\mathbf{X}}[F_{X1}(x_1), F_{X2}(x_2)|\delta]$. La densidad conjunta $f_{X1X2}(x_1, x_2)$ puede por tanto expresarse también en función de la densidad de cópula $c_{\mathbf{X}}[\cdot, \cdot|\delta]$:

$$f_{X_1X_2}(x_1^i, x_2^i|\delta, \zeta_1, \zeta_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{X_1X_2}(x_1^i, x_2^i|\delta, \zeta_1, \zeta_2) = c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1), F_{X_2}(x_2^i|\zeta_2)|\delta] \cdot f_{X_1}(x_1^i|\zeta_1) \cdot f_{X_2}(x_2^i|\zeta_2) \ .$$

La verosimilitud condicional del parámetro δ de la cópula, dados los demás parámetros vendrá dada por la expresión:

$$\mathrm{L}(\delta|D_C,\zeta_1,\zeta_2) = \prod_{i=1}^{n} f_{X_1X_2}(x_1^i, x_2^i|\delta,\zeta_1,\zeta_2) =$$

$$\prod_{i=1}^{n} c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1), F_{X_2}(x_2^i|\zeta_2)|\delta] \cdot f_{X_1}(x_1^i|\zeta_1) \cdot f_{X_2}(x_2^i|\zeta_2) \ , \qquad (6.2.4)$$

donde D_C denota el subconjunto de datos correspondiente a observaciones conjuntas de las variables, $D_C = \{(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)\}$, la densidad f_{X_i} es la densidad $GPD_X(\xi_{Xi}, \beta_{Xi})$, $i = 1, 2$, y $c_{\mathbf{X}}[\cdot,\cdot|\delta]$ denota la densidad de la cópula $C_{\mathbf{X}}[\cdot,\cdot|\delta]$. Estimaremos por tanto la cópula que corresponde a las observaciones conjuntas de ambas variables, una subcópula de la cópula "completa" entre las variables X_1 y X_2. Los parámetros marginales ζ_1, ζ_2 son conocidos.

Debe observarse que, dado que los parámetros marginales $\zeta_1 = (\xi_{X1}, \beta_{X1})$, $\zeta_2 = (\xi_{X2}, \beta_{X2})$ son conocidos, la verosimilitud de la Ecuación 6.2.4 puede simplificarse notablemente:

$$\mathrm{L}(\delta|D_C,\zeta_1,\zeta_2) \propto \prod_{i=1}^{n} c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1), F_{X_2}(x_2^i|\zeta_2)|\delta] \ . \qquad (6.2.5)$$

6.2.3. Tasa de ocurrencia. Verosimilitud condicional

El uso de procesos de Poisson evaluados permite asegurar que la ocurrencia de los sucesos es independiente de su magnitud (Teorema 2.3.4). Por ello, se puede considerar que la tasa de ocurrencia de los sucesos (λ de Poisson) es independiente de los otros parámetros del modelo (marginales y de dependencia). La verosimilitud de λ, se expresa como

$$\mathrm{L}(\lambda|D,\zeta_1,\zeta_2,\delta) = \mathrm{L}(\lambda|D) = \exp(-\lambda \cdot t)\frac{(\lambda \cdot t)^n}{n!} \ , \qquad (6.2.6)$$

donde n es el número de sucesos observados y t es el tiempo de observación.

6.3. Cálculo del posteriori conjunto de los parámetros

Sea D la muestra bivariada de observaciones de X_1 y X_2. Sean ζ_i, $i = 1, 2$, el par de parámetros (ξ_i, β_i) de las distribuciones marginales $GPD_X(\xi_{Xi}, \beta_{Xi})$, $i = 1, 2$ (Ec. 2.4.1) y λ la tasa de ocurrencia del proceso de Poisson. La dependencia entre las marginales X_1 y X_2 se modeliza mediante la cópula $C_{\mathbf{X}}[\cdot, \cdot|\delta]$, escogida de la familia de cópulas de Gumbel(1960), con parámetro δ (Ec. 2.1.2).

La densidad conjunta a posteriori de los parámetros del modelo puede obtenerse a partir de las verosimilitudes condicionales y sus prioris. Su expresión es:

$$f_{\delta,\zeta_1,\zeta_2,\lambda}(\delta, \zeta_1, \zeta_2, \lambda|D) \propto \mathrm{L}(\delta, \zeta_1, \zeta_2|D) f_{\delta,\zeta_1,\zeta_2}(\delta, \zeta_1, \zeta_2) \mathrm{L}(\lambda|D) f_\lambda(\lambda) = f(D|\delta, \zeta_1, \zeta_2) f_{\delta,\zeta_1,\zeta_2}(\delta, \zeta_1, \zeta_2) f(D|\lambda) f_\lambda(\lambda) \quad . \tag{6.3.1}$$

La notación $f(D|\delta, \zeta_1, \zeta_2)$ en la Ec. 6.3.1 indica la densidad conjunta de los datos, dados los parámetros del modelo.

El posteriori conjunto $f_{\delta,\zeta_1,\zeta_2}(\delta, \zeta_1, \zeta_2|D)$ (Ec. 6.3.1) puede resultar complicado de calcular. El método de Gibbs (Sec. 2.5.1) permite simular una muestra del posteriori, que en la práctica lo sustituirá. La muestra del posteriori se obtiene a partir de las densidades condicionales a posteriori de cada uno de los parámetros respecto de los demás. Para cada parámetro del modelo, esta densidad condicional a posteriori se puede expresar en función de su verosimilitud condicional y su priori :

- Para δ,

$$f_{\delta|\zeta_1,\zeta_2}(\delta|\zeta_1, \zeta_2, D) = \frac{\mathrm{L}(\delta, \zeta_1, \zeta_2|D) f_{\delta,\zeta_1,\zeta_2}(\delta, \zeta_1, \zeta_2)}{f_{\zeta_1,\zeta_2}(\zeta_1, \zeta_2|D)} \propto \mathrm{L}(\delta|\zeta_1, \zeta_2, D) f_\delta(\delta|\zeta_1, \zeta_2);$$

- Para $\zeta_1 = (\xi_1, \beta_1)$,

$$f_{\zeta_1|\delta,\zeta_2}(\zeta_1|\delta, \zeta_2, D) \propto \mathrm{L}(\zeta_1|\delta, \zeta_2, D) f_{\zeta_1}(\zeta_1|\delta, \zeta_2);$$

- Para $\zeta_2 = (\xi_2, \beta_2)$,

$$f_{\zeta_2|\delta,\zeta_1}(\zeta_2|\delta, \zeta_1, D) \propto \mathrm{L}(\zeta_2|\delta, \zeta_1, D) f_{\zeta_2}(\zeta_1|\delta, \zeta_1);$$

- Para λ,

$$f(\lambda|D) \propto \mathrm{L}(\lambda|D) f_\lambda(\lambda) \ .$$

El muestreo de Gibbs es un proceso iterativo, mediante el cual se obtiene una muestra del posteriori conjunto de los parámetros. El algoritmo para obtener esta muestra del posteriori, descrito en general en la Sec. 2.5.1, se describe a continuación para los parámetros del modelo propuesto (ocurrencia, tamaño y dependencia):

• *Algoritmo de simulación de una muestra del posteriori: método de Gibbs*

Iteración 0: Elección de un valor inicial de los parámetros:
$\xi^{(0)} = (\delta^{(0)}, \zeta_1^{(0)}, \zeta_2^{(0)}, \lambda^{(0)})$

...

Iteración t, $t \geq 1$**:** Sea $\xi^{(t-1)} = (\delta^{(t-1)}, \zeta_1^{(t-1)}, \zeta_2^{(t-1)}, \lambda^{(t-1)})$ el vector de parámetros del modelo simulado en la iteración $t-1$.

En esta iteración:

1. Se genera una realización de δ, a partir de la densidad condicional a posteriori correspondiente:

$$\delta^{(t)} \sim f_{\delta|\zeta_1,\zeta_2}(\delta|\zeta_1^{(t-1)}, \zeta_2^{(t-1)}, D) \ \ ;$$

2. dada la realización de δ y la densidad condicional a posteriori de ζ_1 se genera una realización de este parámetro:

$$\zeta_1^{(t)} \sim f_{\zeta_1|\delta,\zeta_2}(\zeta_1|\delta^{(t)}, \zeta_2^{(t-1)}, D) \ \ ;$$

3. dadas las realizaciones de δ y ζ_1, se genera una realización de ζ_2 a partir de su densidad condicional a posteriori:

$$\zeta_2^{(t)} \sim f_{\zeta_2|\delta,\zeta_1}(\zeta_2|\delta^{(t)}, \zeta_1^{(t)}, D) \ \ .$$

4. se genera una realización de λ a partir de su densidad condicional a posteriori:

$$\lambda^{(t)} \sim f_\lambda(\lambda|D) \ \ .$$

5. si se cumple el criterio de convergencia, parada y almacenamiento de $(\delta^{(t)}, \zeta_1^{(t)}, \zeta_2^{(t)}, \lambda^{(t)})$
   o bien,
   si no se cumple el criterio de convergencia, pasar a la iteración $t + 1$.

...

El proceso iterativo se repite n veces, hasta obtener la muestra del posteriori de los parámetros deseada, $\xi^{(l)} = (\delta^{(l)}, \zeta_1^{(l)}, \zeta_2^{(l)}, \lambda^{(l)}),\ l = 1, \ldots, n$. Se deben tener en cuenta los distintos criterios de convergencia existentes, como el criterio de Gelman, así como los criterios que permiten determinar el número de iteraciones necesario para obtener muestras independientes entre sí (Robert and Casella, 2000). Para que la muestra pueda sustituir al posteriori, debe tener un tamaño grande, por lo que el muestreo de Gibbs puede resultar computacionalmente intensivo.

A partir de la muestra del posteriori pueden obtenerse las estimaciones de parámetros de *hazard* necesarias (Sec. 8.2): densidades marginales de cada uno de los parámetros; probabilidad de un suceso de un tamaño determinado dado el valor de un parámetro; correlaciones entre parámetros, etc. En particular, en un contexto de *hazard* resultará útil obtener una muestra de periodos de retorno de sucesos de un tamaño determinado, por ejemplo. De este modo, se puede caracterizar este periodo de retorno, y dar la probabilidad, por ejemplo, de que el periodo de retorno esté entre dos valores de interés.

Capítulo 7

Estimación Bayesiana de la cópula de mínima entropía cruzada dados sus momentos, en procesos de Poisson

El modelo propuesto en el Capítulo 3 consiste en un proceso evaluado de Poisson de parámetro λ, donde cada uno de los dos tamaños (X_1, X_2) está distribuido según una $GPD(\xi_i, \beta_i)$, $i = 1, 2$ y la dependencia entre los dos tamaños (X_1, X_2) se modela mediante una función cópula. En el presente capítulo se introduce la modelización de esta dependencia mediante la cópula CrEnC (cópula de mínima entropía cruzada dadas sus marginales uniformes y un conjunto de momentos invariantes por transformaciones monótonas) propuesta en el Capítulo 5. Los parámetros del modelo propuesto son:

parámetro λ: la ocurrencia de los sucesos en el tiempo se modela mediante un proceso de Poisson de parámetro λ;

parámetros $\zeta_i = (\xi_{Xi}, \beta_{Xi})$, $i = 1, 2$: los tamaños de los sucesos, (X_1, X_2), considerados excesos sobre un umbral absoluto $\mathbf{h_0}$ establecido a priori, se suponen distribuidos según una $GPD(\xi_{Xi}, \beta_{Xi})$, $i = 1, 2$;

parámetros $\alpha_1, \ldots, \alpha_J$: la dependencia entre los tamaños X_1 y X_2 se modela mediante una familia de cópulas CrEnC. Los parámetros

$\alpha_1, \ldots, \alpha_J$ corresponden a los coeficientes de los estadísticos de los momentos que conforman la restricción.

Consideramos los tamaños (X_1, X_2), excesos sobre un umbral absoluto $\mathbf{h_0} = (h_1^0, h_2^0)$. Estos umbrales (h_1^0, h_2^0) son superiores a las cotas inferiores presentadas en el Capítulo 3, x_{0i}, $i = 1, 2$ (ver Fig. 7.1). Si consideramos el ejemplo de una base de datos de precipitación diaria registrada en varias ubicaciones, presentado en ese capítulo, los tamaños (X_1, X_2) podrían corresponder a los excesos sobre un umbral bivariado $(h_1^0 = 20\text{mm}, h_2^0 = 20\text{mm})$, claramente superior a la cota inferior del soporte $(x_{01} = 1\text{mm}, x_{02} = 1\text{mm})$.

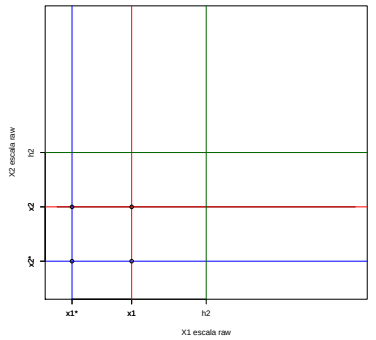

Figura 7.1: Asignación de valores por defecto $x = x_{0i}^*$, $i = 1, 2$ en las marginales, para aquellos valores no registrados. $x = x_{0i}$ indica la cota inferior del soporte y h_i^0 los umbrales absolutos sobre los que se definen los excesos.

Se dispone de una muestra D de observaciones de X_1 y X_2, excesos sobre el umbral absoluto $\mathrm{h}_0 = (h_1^0, h_2^0)$. Las observaciones disponibles son de tres tipos: observaciones de una de las variables (x_1^i, na) o (na, x_2^i) y observaciones conjuntas de ambas variables (x_1^i, x_2^i). La notación na indica que no existe exceso sobre el umbral absoluto en esa variable. Las observaciones conjuntas permitirán estimar tanto los parámetros del modelo de dependencia entre variables como sus marginales; las observaciones de una sola de las variables permitirán mejorar las estimaciones de los parámetros del modelo marginal. Las observaciones

extremales acostumbran a ser escasas, y por eso es importante utilizar toda la información disponible, aunque sea parcial.

Los tamaños X_1,X_2, distribuidos según una $GPD(\xi_{Xi}, \beta_{Xi})$ (sección 2.4) tienen función de densidad $f_{Xi}(x_i),\ i = 1, 2$, de parámetros $\zeta_i = (\xi_{Xi}, \beta_{Xi})$, $i = 1, 2$. La dependencia entre estos tamaños X_1 y X_2 se modeliza mediante una cópula CrEnC, definida en el Capítulo 5, la cual se parametriza según los coeficientes de las restricciones en forma de momentos $(\alpha_1, \ldots, \alpha_k)$. Supongamos que existen k restricciones en forma de momentos, $\theta_1, \ldots, \theta_k$, a las que les corresponden los parámetros $\alpha_1, \ldots, \alpha_k$ en el modelo. Las funciones $T_1(x_1, x_2),\ \ldots, T_k(x_1, x_2)$ son los estadísticos correspondientes a cada uno de los momentos, es decir, son las funciones tales que

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} T_j(x_1, x_2) f_{X1X2}(x_1, x_2) dx_1 dx_2 = \theta_j \ .$$

La función de densidad de la cópula de mínima entropía cruzada dadas sus marginales y estos momentos invariantes por transformaciones monótonas, expresada en $\mathbb{R}^2$, es de la forma

$$f_{X1X2}(x_1, x_2) = a_\alpha(x_1) b_\alpha(x_2) \exp\left[\sum_{s=1}^{k} \alpha_s T_s(x_1, x_2)\right], (x_1, x_2) \in \mathbb{R}^2 \ .$$

Los datos disponibles suelen ser escasos, y la incertidumbre de las estimaciones será elevada. Parece adecuado realizar una estimación bayesiana de los parámetros del modelo. Antes de realizar esta estimación bayesiana, debe determinarse la verosimilitud conjunta de los parámetros, y establecer un priori para éstos. El posteriori conjunto será poco manejable debido al número de parámetros del modelo (parámetros de ocurrencia, marginales y de dependencia). El método de Gibbs permite obtener una muestra extensa del posteriori conjunto a partir de los posterioris condicionales de cada parámetro. Por tanto, será necesaro determinar las verosimilitudes condicionales y los prioris de cada conjunto de parámetros. Los prioris de los tres conjuntos son separables y por tanto pueden establecerse uno a uno.

Las expresiones de las verosimilitudes condicionales y de los prioris para cada uno de los conjuntos de parámetros del modelo se presentan a continuación.

7.1. Restricciones en forma de momentos incluidos en la cópula CrEnC

7.1.1. Coeficientes de las restricciones en forma de momentos. Verosimilitud condicional y priori

La cópula CrEnC, definida en el Capítulo 5, se parametriza según los coeficientes de las restricciones en forma de momentos $(\alpha_1, \ldots, \alpha_k)$. La verosimilitud condicional de estos parámetros depende solo de las observaciones conjuntas, que denotaremos D_C.

La función de verosimilitud condicional de cada uno de los parámetros de la cópula CrEnC dados los demás parámetros del modelo viene dada por la expresión

$$\begin{aligned}&\mathrm{L}(\alpha_j|D_C, \alpha_1, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_k, \zeta_1, \zeta_2, \lambda) = \\ &\prod_{i=1}^{n} a_{\alpha_j}(x_1^i) \prod_{i=1}^{n} b_{\alpha_j}(x_2^i) \exp\left[\sum_{s=1}^{k} \alpha_s \sum_{i=1}^{n} T_s(x_1^i, x_2^i)\right],\ j = 1, \ldots, k,\end{aligned} \tag{7.1.1}$$

y su correspondiente función de log-verosimilitud es

$$\begin{aligned}&l(\alpha_j|D_C, \alpha_1, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_k, \zeta_1, \zeta_2, \lambda) = \\ &\sum_{i=1}^{n} \log(a_{\alpha_j}(x_1^i)) + \sum_{i=1}^{n} \log(b_{\alpha_j}(x_2^i)) + \sum_{s=1}^{k} \alpha_s \sum_{i=1}^{n} T_s(x_1^i, x_2^i), \quad j = 1, \ldots, k\ .\end{aligned}$$

Dado un modelo con k restricciones en forma de momentos, $\theta_1, \ldots, \theta_k$, a las cuales corresponden los parámetros $\alpha_1, \ldots, \alpha_k$, se ha definido un priori normal estándar para cada uno de estos parámetros $\alpha_j,\ j = 1, \ldots, k$.

7.1.2. Coeficientes de las restricciones en forma de momentos. Selección de momentos

Se desea describir la dependencia entre las variables (X_1, X_2) mediante una cópula CrEnC. Supongamos que se dispone de un conjunto

de m posibles restricciones en forma de momentos, $\theta_1, \ldots, \theta_m$, a las que les corresponden los estadísticos $T_1(x_1, x_2), \ldots, T_m(x_1, x_2)$ y los parámetros $\alpha_1, \ldots, \alpha_m$ en el modelo. De entre los momentos disponibles, que habremos escogido invariantes por transformaciones monótonas, se desea seleccionar el subconjunto más adecuado para modelizar la dependencia de los datos disponibles D, $D = \{(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)\}$, mediante una cópula CrEnC. Para obtener el subconjunto de restricciones adecuado a los datos, se ha implementado un algoritmo de inclusión iterativo, *forward*, basado en contrastes de razón de verosimilitudes.

En general, dada una variable aleatoria X cuya densidad depende de un parámetro θ y el contraste de hipótesis

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases},$$

el estadístico de contraste de razón de verosimilitudes tiene expresión

$$\lambda(\vec{x}) = \frac{\sup_{\theta \in \Theta_0} f(x_1, \ldots, x_n|\theta)}{\sup_{\theta \in \Theta} f(x_1, \ldots, x_n|\theta)} = \frac{\mathrm{L}(\widehat{\theta}_0|\vec{x})}{\mathrm{L}(\widehat{\theta}|\vec{x})}, \quad 0 < \lambda(\vec{x}) < 1 \ ,$$

donde $x_1, \ldots, x_n$ es una muestra de X, $\widehat{\theta}_0$ corresponde al estimador máximo verosímil de θ bajo H_0, y $\widehat{\theta}$ corresponde al estimador máximo verosímil de θ bajo $H_0 \cup H_1$. $\mathrm{L}(\cdot|\vec{x})$ denota la verosimilitud del parámetro. Bajo condiciones de regularidad, la distribución asintótica de $-2 \ln \lambda(X_1, \ldots, X_n)$ es conocida. Si $\lambda_n(x_1, \ldots, x_n)$ es el estadístico para cada tamaño de muestra n, se tiene la siguiente convergencia en ley (bajo determinadas hipótesis) (Gómez Villegas, 2005, p. 222):

$$-2 \log \lambda_n(x_1, \ldots, x_n) \rightarrow^{\mathfrak{L}_{\theta_0}} \chi^2_k, \quad \text{donde } k = dim\Theta - dim\Theta_0 \ ,$$

donde el número de grados de libertad de la distribución corresponde a la diferencia entre las dimensiones del espacio paramétrico completo y la del espacio paramétrico obtenido cuando la hipótesis primaria se supone cierta.

El proceso iterativo *hacia delante* de selección de restricciones en el modelo CrEnC que se propone consta de diversas etapas:

1. Se realizan los m contrastes individuales sobre los coeficientes de los momentos:

$$\begin{cases} H_0 : \alpha_s = 0 \\ H_1 : \alpha_s \neq 0 \end{cases}, \quad s = 1, \ldots, m.$$

Si no se puede rechazar la hipótesis primaria para ninguno de los momentos implementados, es decir, ninguno de los momentos implementados modeliza correctamente la dependencia entre las parejas de observaciones, el proceso iterativo se detiene. La cópula utilizada para modelizar la dependencia entre las variables de interés será la de la independencia.

Si sólo se rechaza la hipótesis primaria para uno de los momentos implementados, el k-ésimo, el proceso iterativo también se detiene. La cópula utilizada sólo tendrá una restricción en forma de momento y la densidad conjunta tiene expresión:

$$f(x_1, x_2) = a_{\alpha_k}(x_1) \cdot b_{\alpha_k}(x_2) \exp\left[\alpha_k T_k(x_1, x_2)\right] \ , x_1, x_2 \in \mathbb{R} \ .$$

Si se rechaza la hipótesis primaria para varios de los coeficientes (denotemos j el número de posibles restricciones seleccionadas), el proceso iterativo continúa.

2. Se realizan los j contrastes

$$\begin{cases} H_0 : \alpha_j = 0 \\ H_1 : \alpha_i \cup \alpha_j \neq 0 \end{cases},$$

es decir, se desea comprobar si se puede añadir un nuevo coeficiente α_j al coeficiente α_i. Se elige como coeficiente de referencia al coeficiente con significación más alta en el contraste individual.

Si se rechaza la hipótesis primaria para todos los contrastes, el proceso iterativo se detiene. La dependencia queda bien representada mediante una sola restricción en forma de momento (la correspondiente a α_i). En caso contrario, el proceso iterativo continúa.

3. Si se rechaza la hipótesis primaria para alguno de los contrastes del paso anterior, los contrastes sucesivos son análogos: se parte de l coeficientes de referencia, a los cuales se les desea añadir un nuevo coeficiente de los j seleccionados individualmente. Se elige como conjunto de referencia aquél con mayor significación en los contrastes de la iteración anterior.

4. ...

5. El proceso continúa mientras se pueda añadir al modelo alguna de las j restricciones que resultaron seleccionadas individualmente.

La regla de decisión para los contrastes en cada una de las iteraciones se pueden determinar a partir de las verosimilitudes condicionales de los parámetros (Ec. 7.1.1) y de los estimadores máximo verosímiles de los coeficientes bajo cada una de las hipótesis.

7.2. Parámetros marginales GPD

Una de las hipótesis del modelo propuesto es que las marginales de (X_1, X_2) están distribuídas según una distribución Generalizada de Pareto, GPD, con parámetros marginales $(\xi_{X_i}, \beta_{X_i}), i = 1, 2$.

7.2.1. Parámetros marginales GPD. Verosimilitud condicional y priori

Dada una muestra D de datos de X_1, X_2, denotaremos $f_{X_i}(x_i)$ la función de densidad de la marginal $GPD_X(\xi_{Xi}, \beta_{Xi})$, $i = 1, 2$, y $c_{\mathbf{X}}[\cdot, \cdot|\alpha_1, \ldots, \alpha_k]$ a la densidad de la cópula CrEnC $C_{\mathbf{X}}[\cdot, \cdot|\alpha_1, \ldots, \alpha_k]$, que modeliza la dependencia entre X_1, X_2. Los parámetros $(\alpha_1, \ldots, \alpha_k)$ son conocidos. Para simplificar expresiones, denotaremos ζ_i al par de parámetros marginales (ξ_{Xi}, β_{Xi}) $i = 1, 2$.

La verosimilitud condicional de los parámetros marginales de X_1, $\zeta_1 = (\xi_{X1}, \beta_{X1})$ dados los demás parámetros del modelo viene dada por la expresión

$$\mathrm{L}(\zeta_1|D, \zeta_2, \alpha_1, \ldots, \alpha_k) = \prod_{i=1}^{n} f_{X_1X_2}(x_1^i, x_2^i|\zeta_1, \zeta_2, \alpha_1, \ldots, \alpha_k) =$$
$$\prod_{i=1}^{n} c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1), F_{X_2}(x_2^i|\zeta_2)|\alpha_1, \ldots, \alpha_k] \cdot f_{X_1}(x_1^i|\zeta_1) \cdot f_{X_2}(x_2^i|\zeta_2) \ . \tag{7.2.1}$$

La expresión (7.2.1) puede simplificarse notablemente si se observa que, dados los parámetros conocidos, algunos de los términos son constantes.

Así,

$$\mathrm{L}(\zeta_1|D,\zeta_2,\alpha_1,\ldots,\alpha_k) \propto \prod_{i=1}^{n} c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1),F_{X_2}(x_2^i|\zeta_2)|\alpha_1,\ldots,\alpha_k]\cdot f_{X_1}(x_1^i|\zeta_1)\ . \tag{7.2.2}$$

y su correspondiente log-verosimilitud:

$$l(\zeta_1|D,\zeta_2,\alpha_1,\ldots,\alpha_k) \propto \sum_{i=1}^{n} \log(c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1),F_{X_2}(x_2^i|\zeta_2)|\alpha_1,\ldots,\alpha_k]) + \log(f_{X_1}(x_1^i|\zeta_1))\ . \tag{7.2.3}$$

La expresión de la verosimilitud de los parámetros de la otra marginal, X_2, $\zeta_2 = (\xi_{X2}, \beta_{X2})$, dados los demás parámetros del modelo es análoga. Simplificando los términos constantes queda,

$$\mathrm{L}(\zeta_2|D,\zeta_1,\alpha_1,\ldots,\alpha_k) \propto \prod_{i=1}^{n} c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1),F_{X_2}(x_2^i|\zeta_2)|\alpha_1,\ldots,\alpha_k]\cdot f_{X_2}(x_2^i|\zeta_2)\ , \tag{7.2.4}$$

y su log-verosimilitud,

$$l(\zeta_2|D,\zeta_1,\alpha_1,\ldots,\alpha_k) \propto \sum_{i=1}^{n} \log(c_{\mathbf{X}}[F_{X_1}(x_1^i|\zeta_1),F_{X_2}(x_2^i|\zeta_2)|\alpha_1,\ldots,\alpha_k]) + \log(f_{X_2}(x_2^i|\zeta_2))\ . \tag{7.2.5}$$

El priori de los parámetros marginales se define del mismo modo que para la cópula paramétrica, es decir, se utiliza un priori plano en una región adecuada para la magnitud que se modeliza según juicio experto (ver Sec. 6.1.1).

7.3. Tasa de ocurrencia de Poisson. Verosimilitud, priori y posteriori

La ocurrencia de los sucesos de interés se modeliza según un proceso de Poisson evaluado, de parámetro λ. El uso de procesos de Poisson

evaluados permite asegurar que la ocurrencia de los sucesos es independiente de su magnitud (Teorema 2.3.4). Por ello, se puede considerar que la tasa de ocurrencia de los sucesos (λ de Poisson) es independiente de los otros parámetros del modelo. La verosimilitud de λ se expresa como

$$\mathrm{L}(\lambda|D, \zeta_1, \alpha_1, \ldots, \alpha_k) = \mathrm{L}(\lambda|D) = \exp(-\lambda \cdot t)\frac{(\lambda \cdot t)^n}{n!} \ , \qquad (7.3.1)$$

donde n es el número de sucesos observados y t es el tiempo de observación.

El priori de la tasa de ocurrencia λ se define del mismo modo que para la cópula paramétrica (ver Sec. 6.1.3), es decir, se ha establecido un priori uniforme en $\log(\tau)$.

Dado el priori y la verosimilitud del parámetro de ocurrencia λ, su densidad a posteriori tiene una expresión sencilla:

$$f_\lambda(\lambda|D) \propto f_\lambda(\lambda) \cdot \mathrm{L}(\lambda|D) \ . \qquad (7.3.2)$$

7.4. Expresión del posteriori conjunto de los parámetros

La densidad conjunta a posteriori de los parámetros se obtiene a partir de los prioris y las verosimilitudes previamente determinados. Dada la independencia entre el parámetro λ de ocurrencia y los demás parámetros (marginales y de dependencia), la densidad conjunta a posteriori puede expresarse como el producto de los posterioris de ambos conjuntos de parámetros:

$$\begin{aligned} &f_{\lambda,\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J}(\lambda, \zeta_1, \zeta_2, \alpha_1, \ldots, \alpha_J|D) = \\ &f_{\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J}(\zeta_1, \zeta_2, \alpha_1, \ldots, \alpha_J|D) \cdot f_\lambda(\lambda|D) \ , \end{aligned}$$

donde $\alpha_1, \ldots, \alpha_J$ denotan los parámetros de la cópula CrEnC (Ec. 5.0.2); ζ_i, $i = 1, 2$, el par de parámetros (ξ_i, β_i) de las distribuciones marginales $GPD_X(\xi_{Xi}, \beta_{Xi})$, $i = 1, 2$ (Ec. 2.4.1) y D los datos de la muestra. La notación $f(D|\delta, \zeta_1, \zeta_2)$ en la Ec. 7.4.1 indica la densidad conjunta de los datos, dados los parámetros del modelo.

La expresión de la densidad conjunta a posteriori de λ es fácil de determinar (ver Ec. 7.3.2). Para los parámetros marginales y de dependencia, la expresión de la densidad conjunta a posteriori es un tanto más complicada:

$$\begin{aligned}
&f_{\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J}(\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J|D) \propto \\
&\mathrm{L}(\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J|D) f_{\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J}(\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J) = \\
&f(D|\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J) f_{\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J}(\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J)\ . \qquad (7.4.1)
\end{aligned}$$

El posteriori conjunto $f_{\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J}(\zeta_1,\zeta_2,\alpha_1,\ldots,\alpha_J|D)$ (Ec. 7.4.1) puede resultar complicado de calcular. Se aplica el método de Gibbs (Sec. 2.5.1) para simular una muestra del posteriori, que a efectos prácticos lo sustituirá. La muestra del posteriori se obtiene a partir de las densidades condicionales a posteriori de cada uno de los parámetros respecto de los demás. Para cada parámetro del modelo, estas densidades condicionales a posteriori se pueden expresar en función de los prioris y las verosimilitudes condicionales ya obtenidas:

- Para $\alpha_1,\ldots,\alpha_J$,

$$\begin{aligned}
&f_{\alpha_1,\ldots,\alpha_J|\zeta_1,\zeta_2}(\alpha_1,\ldots,\alpha_J|\zeta_1,\zeta_2,D) = \\
&\frac{\mathrm{L}(\alpha_1,\ldots,\alpha_J,\zeta_1,\zeta_2|D) f_{\alpha_1,\ldots,\alpha_J,\zeta_1,\zeta_2}(\alpha_1,\ldots,\alpha_J,\zeta_1,\zeta_2)}{f_{\zeta_1,\zeta_2}(\zeta_1,\zeta_2|D)} \propto \\
&\mathrm{L}(\alpha_1,\ldots,\alpha_J|\zeta_1,\zeta_2,D) f_{\alpha_1,\ldots,\alpha_J}(\alpha_1,\ldots,\alpha_J|\zeta_1,\zeta_2)\ ;
\end{aligned}$$

- Para $\zeta_1 = (\xi_1,\beta_1)$,

$$\begin{aligned}
&f_{\zeta_1|\alpha_1,\ldots,\alpha_J,\zeta_2}(\zeta_1|\alpha_1,\ldots,\alpha_J,\zeta_2,D) \propto \\
&\mathrm{L}(\zeta_1|\alpha_1,\ldots,\alpha_J,\zeta_2,D) f_{\zeta_1}(\zeta_1|\alpha_1,\ldots,\alpha_J,\zeta_2)\ ;
\end{aligned}$$

- Para $\zeta_2 = (\xi_2,\beta_2)$,

$$\begin{aligned}
&f_{\zeta_2|\alpha_1,\ldots,\alpha_J,\zeta_1}(\zeta_2|\alpha_1,\ldots,\alpha_J,\zeta_1,D) \propto \\
&\mathrm{L}(\zeta_2|\alpha_1,\ldots,\alpha_J,\zeta_1,D) f_{\zeta_2}(\zeta_1|\alpha_1,\ldots,\alpha_J,\zeta_1)\ ;
\end{aligned}$$

El proceso iterativo del muestreo de Gibbs se repite n veces hasta obtener la muestra deseada de los parámetros, $\xi^{(k)} = (\alpha_1^{(k)},\ldots,\alpha_J^{(k)},\zeta_1^{(k)},$

$\zeta_2^{(k)}$,$\lambda^{(k)}$), $k = 1, \ldots, n$. La convergencia del proceso y la independencia de las distintas cadenas se puede comprobar a partir de varios criterios de convergencia, como el criterio de Gelman (Robert and Casella, 2000). Debe observarse que el muestreo de Gibbs puede resultar computacionalmente intensivo, dado que el tamaño muestral debe ser grande para que la muestra pueda sustituir al posteriori de los parámetros. Una vez obtenida, a partir de esta muestra del posteriori pueden deducirse parámetros de *hazard* de interés como probabilidades de excedencia de valores de referencia marginales y conjuntos, periodos de retorno, entre otros (Ver Sec. 8.2).

7.5. Otros trabajos

Las cópulas CrEnC propuestas en esta tesis se basan en el trabajo de Rumsey y Posner (Rumsey Jr and Posner, 1965). Sin embargo, en los últimos años, otros autores han realizado desarrollos paralelos que en algunos casos tiene puntos en común con la propuesta presentada:

- Calsaverini and Vicente (2009) relaciona cópulas y teoría de la información (entropía y entropía cruzada). Introduce la importancia del uso de coeficientes de dependencia invariantes por transformaciones monótonas. Utiliza la información de cópula como elemento de selección entre diferentes modelos.

- Chen et al. (2013) analiza la dependencia entre caudales fluviales. Utiliza la entropía de cópula para seleccionar la cópula que mejor ajusta un conjunto de datos.

- Smith (2007) relaciona cópulas e información de Fisher, diferenciando la información proporcionada por los parámetros marginales y los de dependencia. Se remarca que si se relajan las hipótesis de continuidad, las propiedades de invariancia no tienen por qué mantenerse, dado que la cópula ya no será única.

- Dempster et al. (2007) utiliza la entropía cruzada para seleccionar la cópula más parecida a la cópula gaussiana, dadas ciertas restricciones. Se menciona la necesidad de aproximación discreta, debido a los problemas del soporte limitado de las marginales.

- Zhao and Lin (2011) utiliza una versión discretizada de la entropía de cópula, a partir de la cópula empírica, sin considerar restricciones. Aplica a series económicas, donde estima las marginales mediante kernel.

- Meeuwissen and Bedford (1997) hallan una expresión cerrada para la distribución bidimensional de mínima entropía cruzada con correlación de rangos dada, es decir, con una sola restricción en forma de momentos, y demuestra su existencia. Especifica que la estimación de este tipo de funciones por multiplicadores de Lagrange no es adecuada, y aproxima la función por una serie de funciones discretas. No discute el problema del soporte limitado, pero utiliza distribuciones uniformes en el intervalo $[-1/2, 1/2]$, en lugar del usual $[0, 1]$.

- Miller and Liu (2002) caracteriza la forma de la densidad conjunta de mínima entropía cruzada dadas restricciones en forma de momentos, y obtiene su expresión mediante cálculo de variaciones. Menciona las cópulas (básicamente la cópula normal) pero resuelve en el caso general.

- Ebrahimi et al. (2008) caracteriza la forma de la densidad de máxima entropía dadas restricciones en forma de momentos, obteniendo su representación a partir de una optimización mediante multiplicadores de Lagrange.

- Hao and Singh (2013) halla la expresión de la distribución de máxima entropía dadas unas restricciones en forma de momentos, que determina mediante multiplicadores de Lagrange y aplica a análisis bivariado de sequías.

- Cabe destacar el trabajo de Chu (2011), hallado recientemente. En este trabajo se halla la cópula de máxima entropía, dados los momentos (seleccionados invariantes por transformaciones monótonas) y las marginales (uniformes, por ser cópula). La cópula obtenida se aproxima mediante un método de cuadratura de Gauss-Legendre. Estas cópulas se aplican a *asset allocation*. Es innegable que existen numerosos puntos de contacto con el enfoque presentado en esta Tesis. No obstante, no se consideran los

problemas debidos al soporte limitado de las marginales, y no se considera la incertidumbre de las estimaciones, dado que no se realiza una estimación bayesiana de los parámetros.

Capítulo 8

Model Checking y cantidades predictivas

8.1. Model checking

En la Sección 2.5.2 se presentaron algunos de los p-valores propuestos en la literatura para evaluar la compatibilidad entre el modelo propuesto y los datos disponibles. La interpretación (frecuentista) usual de la significación de la muestra o p-valor asocia valores pequeños con incompatibilidad entre los datos y el modelo. Sin embargo, interpretar si un valor es *pequeño* no siempre es fácil. Uno de los problemas básicos de algunos de los p-valores definidos en la Sección 2.5.2 es la falta de uniformidad, sin la cual no se pueden interpretar adecuadamente los valores obtenidos. En este apartado se introducen los p-valores utilizados y se propone un nuevo p-valor uniforme con el que se solventan los problemas de interpretación.

8.1.1. p-valores implementados

El modelo propuesto (Cap. 3) está constituido por una elección de distribuciones marginales (GPD), consideraciones sobre el soporte limitado del fenómeno (GPD en el dominio de Weibull), y el modelo de dependencia entre estas marginales (cópula paramétrica de Gumbel o CrEnC). Se desea evaluar la compatibilidad de los distintos aspectos

del modelo propuesto con los datos y para ello se han escogido los enfoques de p-valor de discrepancia y de p-valor a posteriori.

Para cada n-pla de parámetros de la muestra del posteriori se ha simulado una remuestra del mismo tamaño que la muestra original de excesos. Los p-valores escogidos se evalúan en la muestra original y en las remuestras. El modelo se puede considerar compatible con los datos si el p-valor obtenido con éstos se encuentra aproximadamente en el centro de los valores obtenidos para las remuestras (interpretación visual mediante histograma) o bien si la proporción de p-valores de las remuestras es aproximadamente del 50 % . Si el valor correspondiente a la muestra original se encuentra en las colas de los valores de las remuestras, se debe dudar de la compatibilidad entre datos y modelo. El p-valor predictivo a posteriori corresponde al centro o a la media (dependiendo de la escala escogida) de los p-valores predictivos para cada n-pla de parámetros del posteriori.

La compatibilidad de los distintos aspectos del modelo propuesto con los datos se evalúa mediante diversos p-valores:

Contrastes marginales. Priori GPD Se realiza el contraste de Kolmogorov-Smirnov para evaluar la compatibilidad de los datos marginales con una distribución $GPD(\xi, \beta)$:

$$\begin{array}{ll} H_0: & X \sim \text{GPD}(\cdot,\cdot) \\ H_1: & X \text{ tiene otra distribución} \end{array} .$$

a) p-valor predictivo Kolmogorov-Smirnov:

Se ha implementado el cálculo del p-valor predictivo a posteriori. Se obtiene el p-valor correspondiente al contraste KS para cada n-pla de la muestra del posteriori. Se comprueba la compatibilidad de la $GPD(\xi_i, \beta_i),\ i = 1, \ldots, m$, con la correspondiente remuestra y con los excesos originales. Estos p-valores, $p_i, i = 1, \ldots, m$, se combinan para obtener el p-valor predictivo a posteriori deseado.

b) p-valor predictivo multinomial:

Se realiza el contraste de razón de verosimilitudes generalizado para evaluar la compatibilidad de una muestra con un conjunto de probabilidades multinomiales. En este caso,

la distribución multinomial corresponde a una distribución $GPD(\xi, \beta)$ discretizada en intervalos de igual probabilidad. Se ha implementado el cálculo del p-valor predictivo a posteriori.

Se ha obtenido el p-valor correspondiente al contraste multinomial para cada n-pla de la muestra del posteriori. Se ha comprobado la compatibilidad de la $GPD(\xi_i, \beta_i),\ i = 1, \ldots, m$, con la correspondiente remuestra ($prem_i$) y con los excesos originales ($porig_i$). Estos p-valores, $p_i, i = 1, \ldots, m$, se combinan para obtener el p-valor predictivo a posteriori deseado.

c) p-valor predictivo Aitchison:

Se evalúa la compatibilidad de una muestra con un conjunto de probabilidades multinomiales mediante la distancia de Aitchison al cuadrado. La distribución multinomial corresponde a una distribución $GPD(\xi, \beta)$ discretizada en intervalos de igual probabilidad. Se puede considerar que las proporciones tienen carácter composicional (Pawlowsky-Glahn and Egozcue, 2001), y por tanto las distancias entre ellas deben medirse mediante la distancia de Aitchison (Aitchison, 1986). Si $(p_1, \ldots, p_m)$ denotan las probabilidades de referencia ($GPD(\xi, \beta)$ de los m intervalos y $(n_1/n, \ldots, n_m/n)$ denotan las frecuencias muestrales relativas para esos mismos intervalos, la distancia de Aitchison al cuadrado entre ambos vectores D_A^2 es

$$D_A^2 = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{i-1} \left(\log \frac{p_i}{p_j} - \log \frac{n_i/n}{n_j/n} \right)^2 \tag{8.1.1}$$

Para evitar los problemas causados por las frecuencias muestrales nulas, la distancia implementada es una ligera modificación de la anterior Ec. 8.1.1:

$$D_A^2 = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{i-1} \left(\log \frac{p_i}{p_j} - \log \frac{(n_i + 1)/(n + m)}{(n_j + 1)/(n + m)} \right)^2 \tag{8.1.2}$$

Se ha implementado el cálculo del p-valor predictivo a posteriori: para cada n-pla de la muestra del posteriori se obtiene el p-valor correspondiente al contraste Aitchison y estos p-valores, $p_i, i = 1, \ldots, m$, se combinan para obtener el p-valor predictivo a posteriori deseado. La compatibilidad de la $GPD(\xi_i, \beta_i),\ i = 1, \ldots, m$, con la correspondiente remuestra y con los excesos originales, de modo que se obtienen dos p-valores predictivos diferentes.

Contrastes marginales. GPD**- Weibull** Para cada una de las marginales, se desea contrastar si el priori de dominio de atracción Weibull ($\xi < 0$) para la GPD establecido es coherente con los datos.

a) p-valor predictivo slope:

   Se ha implementado el cálculo del p-valor a posteriori:

   Dada una muestra (original o remuestras) se calculan los coeficientes de la recta de excesos esperados (Castillo, 1988). Se obtiene la pendiente predictiva a posteriori tanto para los excesos originales como para las remuestras.

b) p-valor discrepancia slope:

   Dada una muestra (se utilitzan la muestra de excesos original y las remuestras) se calculan los coeficientes de la recta de regresión de los excesos esperados. Se obtiene el p-valor de discrepancia, correspondiente a la proporción de pendientes predictivas que superan el valor observado del estadístico.

Bondad de ajuste global. Modelo marginal y dependencia

a) p- valor bayesiano basado en discrepancia de la probabilidad de excedencia

   Dada una muestra (original o remuestra) se calcula la probabilidad conjunta de excedencia de un par de valores de referencia. Se obtiene el p-valor de discrepancia, correspondiente a la proporción de muestras predictivas que superan el valor observado del estadístico para la muestra original.

Esta discrepancia permite evaluar el ajuste global del modelo a los datos.

b) p- valor bayesiano basado en discrepancia τ de Kendall
Dada una muestra (original o remuestras) se calcula el coeficiente τ de Kendall (Ec. 2.2.6) correspondiente. Se obtiene el p-valor de discrepancia, correspondiente a la proporción τ predictivas que superan el valor observado del estadístico para la muestra original. Esta discrepancia permite evaluar el ajuste global del modelo a los datos.

8.1.2. Intervalo donde el p-valor es uniforme

Se desea evaluar la compatibilidad del modelo de referencia $F_\vartheta(\cdot)$, ϑ desconocido, con los datos disponibles. Se selecciona un conjunto de posibles valores del parámetro $\vartheta_i,\ i = 1, \ldots, n$, posiblemente una muestra del posteriori, y para cada uno de ellos se realiza el contraste de bondad de ajuste elegido (Kolmogorov-Sminov, multinomial u otra selección), de manera que se obtiene el correspondiente p-valor, $p_i,\ i = 1, \ldots, n$. Por tanto, se dispone de un conjunto de p-valores

$$p_i = p_i(\vartheta_i) = \text{KSgof}(y|\vartheta_i),$$

o bien

$$p_i = p_i(\vartheta_i) = \text{ multgof}(y|\vartheta_i),\ etc.\ ,$$

cuyos soportes están contenidos en $[0, 1]$, aunque su distribución no es en general uniforme (Robins et al., 2000). Además, los p-valores a posteriori son combinaciones de estos p-valores. Incluso en el caso en que pudiéramos considerar que sus distribuciones son uniformes, los valores a posteriori no tendrían por qué presentar esa distribución.

Se desea obtener un intervalo, quizá más reducido que el intervalo unidad, donde poder interpretar los valores de p-valor obtenidos de la manera usual, es decir, considerando que su distribución es uniforme en ese intervalo.

Debe observarse que, dependiendo del método de selección de los parámetros del modelo, los correspondientes p-valores pueden presentar un cierto grado de dependencia. Por ejemplo, si se seleccionan secuencialmente los valores de un mapa del posteriori de los parámetros, es

probable que valores contiguos representen bondades de ajuste similares, y por tanto, haya dependencias entre ellos.

Se propone el siguiente algoritmo para *uniformizar* los p-valores, es decir, hallar un intervalo de valores en que este p-valor es uniforme:

1. Transformar la muestra de p-valores, p_i, mediante la transformación probit. De esta manera se obtiene una variable con soporte real. Se define la variable:

$$Q_i = \Phi^{-1}(p_i),$$

donde Φ^{-1} corresponde a la función de distribución inversa de la normal estándard. Cada Q_i está, por tanto, distribuida normalmente.

2. Definir una nueva variable como combinación lineal con pesos de las Q_i:

$$\Delta = \sum_{i=1}^{n} \psi_i Q_i = \sum_{i=1}^{n} \psi_i \ \Phi^{-1}(p_i) \quad , \tag{8.1.3}$$

donde ψ_i indica el posteriori del valor θ_i, que ha sido previamente calculada. La variable Δ es una combinación lineal de variables normales, por tanto tendrá distribución normal con esperanza y varianza conocidas:

$$\mathrm{E}[\Delta] = \mathrm{E}\left[\sum_{i=1}^{n} \psi_i Q_i\right] = \sum_{i=1}^{n} \psi_i \mathrm{E}[Q_i] = 0 \quad ,$$

$$\mathrm{Var}[\Delta] = \mathrm{Var}\left[\sum_{i=1}^{n} \psi_i Q_i\right] = \mathrm{E}\left[\sum_i \sum_j \psi_i Q_i \psi_j Q_j\right] =$$

$$\sum_i \sum_j \psi_i \psi_j \mathrm{E}[Q_i Q_j] = \sum_i \sum_j \psi_i \psi_j \rho_{ij} = \sum_i \psi_i^{\delta}, \quad 1 \leq \delta \leq 2 \quad .$$

Cabe remarcar que $1 \leq \delta \leq 2$ dado que

- si $\rho_{ij} = 0$, independencia entre las Q_i,

$$\sum_i \sum_j \psi_i \psi_j \rho_{ij} = \sum_i \psi_i^2,$$

- si $\rho_{ij} = 1$, dependencia total entre las Q_i

$$\sum_i \sum_j \psi_i \psi_j \rho_{ij} = \left(\sum_i \psi_i \right) \left(\sum_j \psi_j \right) = 1,$$

  dado que el posteriori suma 1,

- si $0 < \rho_{ij} < 1$,

$$\sum_i \sum_j \psi_i \psi_j \rho_{ij} = \sum_i \psi_i^{\delta}, \quad 1 < \delta < 2,$$

  dado que $\rho_{ij} = \frac{\mathrm{Cov}[Q_i, Q_j]}{\sqrt{\mathrm{Var}[Q_i]}\sqrt{\mathrm{Var}[Q_j]}}$.

3. Estandarizar la variable Δ, conocidas su esperanza y varianza:

$$T = \frac{\Delta}{\sqrt{\sum_i \psi_i^{\delta}}} = \frac{\sum_{i=1}^n \psi_i Q_i}{\sqrt{\sum_i \psi_i^{\delta}}} = \frac{\sum_{i=1}^n \psi_i \ \Phi^{-1}(p_i)}{\sqrt{\sum_i \psi_i^{\delta}}}, \ 1 \leq \delta \leq 2 \ .$$

   T es una variable normal estándar.

4. Aplicar la transformación probit inversa a T:

$$pp = \Phi \left(\frac{\sum_{i=1}^n \psi_i \ \Phi^{-1}(p_i)}{\sqrt{\sum_i \psi_i^{\delta}}} \right), \ 1 \leq \delta \leq 2, \tag{8.1.4}$$

   donde Φ corresponde a la función de distribución normal estándar. El nuevo p-valor pp es uniforme entre las dos cotas definidas por $\delta = 1$ y $\delta = 2$.

Con frecuencia este algoritmo se aplicará al caso particular en que los valores $\vartheta_i, \ i = 1, \ldots n$ son una muestra del posteriori del parámetro ϑ dados los datos. En este caso los valores ϑ_i se suponen equiprobables, y por tanto los pesos utilizados en la combinación lineal (Ec. 8.1.3) son todos iguales $(1/n)$. En este caso, las expresiones de las cotas de uniformidad obtenidas son más sencillas. Sustituyendo $\psi_i = 1/n$ en la Ec. 8.1.3 se obtiene:

2b. $\Delta = \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(p_i)$; $\mathrm{E}[\Delta] = 0$; $\mathrm{Var}[\Delta] = \sum_i \frac{1}{n^{\delta}} = \frac{1}{n^{\delta - 1}}$.

3b. Estandarizar la variable Δ, conocidas su esperanza y varianza:

$$T = \frac{\Delta}{\sqrt{\frac{1}{n^{\delta-1}}}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Q_i}{\frac{1}{n^{(\delta-1)/2}}} = n^{(\delta-3)/2} \sum_{i=1}^{n} \Phi^{-1}(p_i),\ 1 \leq \delta \leq 2\ .$$

T es una variable normal estándar.

4b. Aplicar la transformación probit inversa a T:

$$pp = \Phi\left(n^{(\delta-3)/2} \sum_{i=1}^{n} \Phi^{-1}(p_i) \right),\ 1 \leq \delta \leq 2, \tag{8.1.5}$$

donde Φ corresponde a la función de distribución normal estándar. El nuevo p-valor pp es uniforme entre las dos cotas definidas por $\delta = 1$ y $\delta = 2$:

$$\left[\Phi\left(\frac{1}{n} \sum_{i=1}^{n} \Phi^{-1}(p_i) \right), \Phi\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Phi^{-1}(p_i) \right) \right] \tag{8.1.6}$$

8.2. Algunas cantidades de interés a posteriori

El muestreo de Gibbs (Sec. 2.5.1) permite simular una muestra extensa de la densidad conjunta a posteriori de los parámetros del modelo (ocurrencia, tamaños y depedencia) a partir de las densidades condicionales unidimensionales a posteriori de cada uno de esos parámetros respecto de los demás y los datos.

La muestra simulada permite calcular de manera aproximada las cantidades de interés, coeficientes de *hazard*, que se obtendrían a partir de la densidad conjunta a posteriori. Estas cantidades de interés pueden ser de tipología muy diversa, pero podrían englobarse de manera muy general en las categorías siguientes:

a) Distribuciones de probabilidad de los parámetros o cálculo de probabilidades de sucesos definidos sobre los parámetros. Por ejemplo, cálculo de

- marginal del parámetro ξ_j de la distribución marginal $GPD(\xi_j, \beta_j)$, $j = 1\ i/o 2$
- probabilidad de que el parámetro ξ_{Xi} sea negativo, dados los datos, $\mathrm{P}[\xi_{Xi} < 0|D]$ (comprobación de la probabilidad de que la marginal GPD esté en el dominio de Weibull)
- probabilidad de que (ξ_{X1}, ξ_{X2}) sean negativos, dados los datos, $\mathrm{P}[\xi_{X1} < 0, \xi_{X2} < 0|D]$ (comprobación de la probabilidad conjunta de que las marginales GPD estén en el dominio de Weibull)

b) Distribuciones de probabilidad o probabilidades de sucesos definidos a partir de funciones de los parámetros. Por ejemplo,

- la distribución del periodo de retorno $\tau = 1/\lambda$
- periodo de retorno τ correspondiente a un valor de referencia de los excesos, x_1^0 o x_2^0
- periodo de retorno τ correspondiente a excesos entre dos valores seleccionados, (x_1^1, x_1^2) o (x_2^1, x_2^2)
- exceso x de X_1 o X_2 correspondiente a un periodo de retorno τ_0 seleccionado

c) Distribuciones predictivas de las variables o sucesos sobre ellas. Por ejemplo,

- probabilidades de no excedencia marginales para valores de referencia de los excesos de X_1 o X_2, (x_1^0 o x_2^0), que incluso pueden ser valores no observados en los datos
- probabilidades de no excedencia conjuntas para valores seleccionados de los excesos conjuntos (x_1^0, x_2^0), que incluso pueden ser valores no observados en los datos

d) Distribuciones de funciones de los parámetros que incluyen parámetros y distribuciones predictivas de las variables. Por ejemplo,

- distribución de $\lambda(x)$ por encima de un cierto umbral x, $\lambda(x) = \lambda(1 - F_1(x|\xi_X, \beta_X))$ (Prop. 2.3.5)

- distribución del periodo de retorno de sucesos por encima de un determinado umbral, $\tau(x) = \tau/(1 - F_1(x|\xi_X, \beta_X))$

e) etc.

Capítulo 9

Cópula CrEnC. Restricciones en forma de momentos: Momentos incluidos

El modelo definido en el Capítulo 3 describe la ocurrencia de los sucesos de interés, sus distribuciones marginales y su dependencia. Si se representa la dependencia entre variables mediante la cópula CrEnC, definida en el Capítulo 5, ésta depende de los parámetros $(\alpha_1, \ldots, \alpha_k)$ correspondientes a restricciones en forma de momentos (que escogeremos invariantes por transformaciones monótonas).

En los Capítulos 10, 11 y 12 se aplica este modelo a diversos conjuntos de datos. Se ha implementado un conjunto de medidas de asociación que representan diferentes tipos de dependencia (ver Sec. 2.2), y a los que corresponden los parámetros $\alpha_i, i = 1, \ldots, 7$ del modelo. Estos momentos son:

- μ_1: ρ de Spearman
- μ_2: Blest (I)
- μ_3: Blest simetrizado
- μ_4: γ de Gini
- μ_5: β de Blomquist

- μ_6: Spearman's footrule φ
- μ_7: τ de Kendall

Las Figuras 9.1, 9.2, 9.3, 9.4, 9.5 y 9.6 presentan los contornos de isodensidad en $\mathbb{R}^2$ de la cópula CrEnC correspondiente a cada uno de estos momentos, para diferentes valores del coeficiente (manteniendo nulos los demás). Dado un valor fijado del momento, se desea interpretar el siginificado del coeficiente que lo acompaña. En las Figuras se observan diferentes tipos de dependencia, incluyendo la independencia (Fig. 9.6). Para la mayoría de los momentos, los valores positivos del coeficiente corresponden a dependencia positiva, y los negativos a dependencia negativa, con excepción del coeficiente de Blest simetrizado, Fig. 9.3, donde la relación coeficiente y dependencia es la inversa.

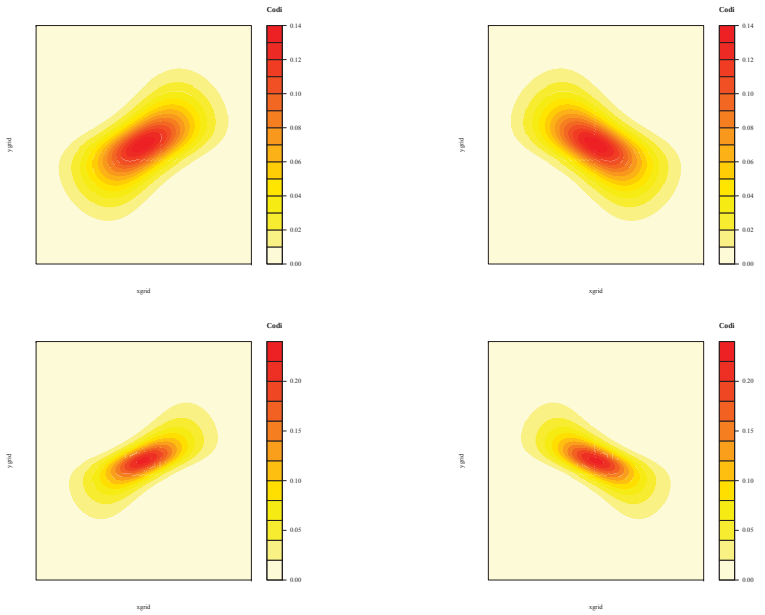

Figura 9.1: Contornos de isodensidad de la cópula CrEnC con momento μ_1 (ρ de Spearman) para diferentes valores del coeficiente: $\alpha_1 = 1, \alpha_1 = 2, \alpha_1 = -1, \alpha_1 = -2$ (de izq. a dcha. y de arriba a abajo).

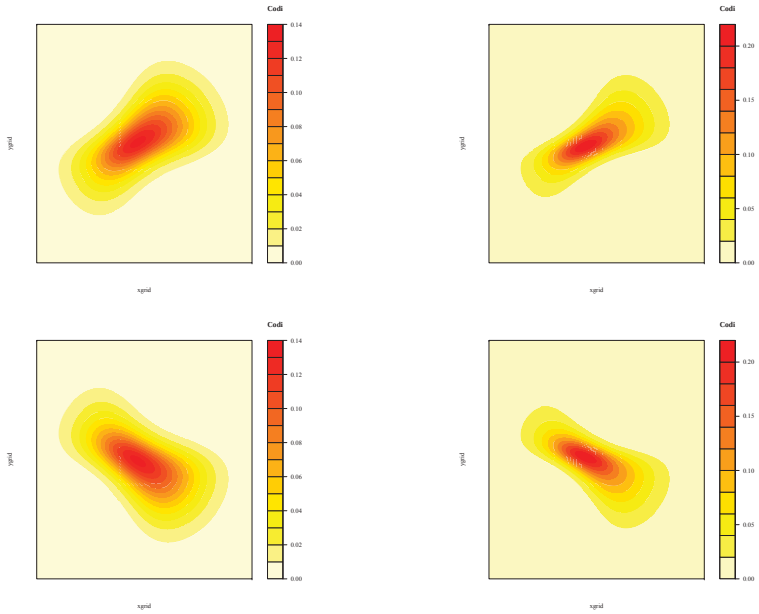

Figura 9.2: Contornos de isodensidad de la cópula CrEnC con momento μ_2 (Blest) para diferentes valores del coeficiente: $\alpha_2 = 1, \alpha_2 = 2, \alpha_2 = -1, \alpha_2 = -2$ (de izq. a dcha. y de arriba a abajo).

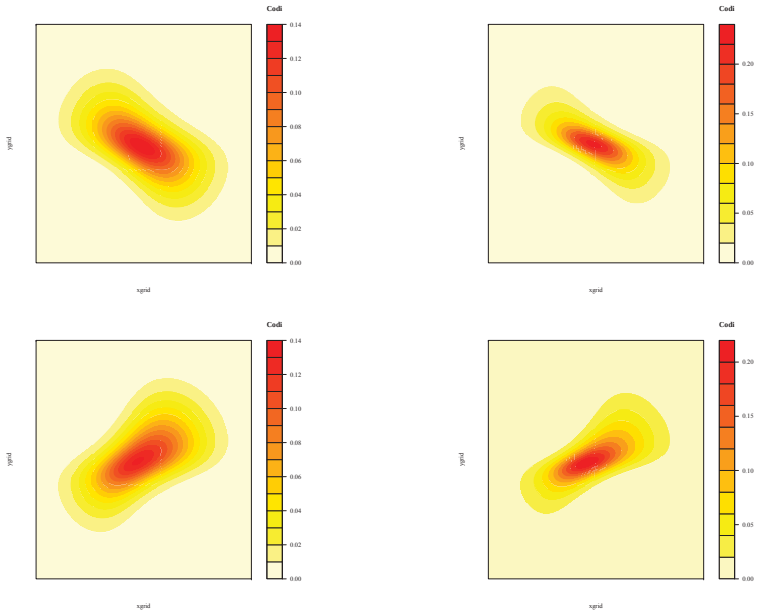

Figura 9.3: Contornos de isodensidad de la cópula CrEnC con momento μ_3 (Blest simetrizado) para diferentes valores del coeficiente: $\alpha_3 = 1, \alpha_3 = 2, \alpha_3 = -1, \alpha_3 = -2$ (de izq. a dcha. y de arriba a abajo).

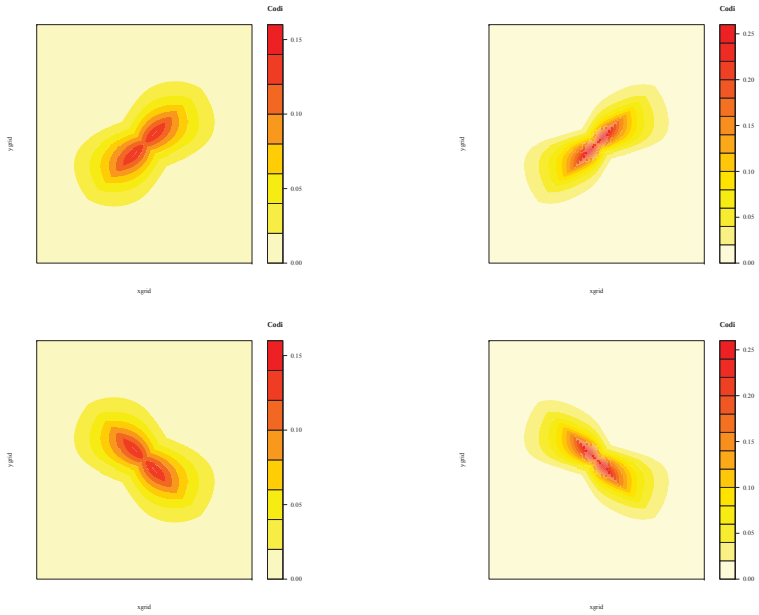

Figura 9.4: Contornos de isodensidad de la cópula CrEnC con momento μ_4 (γ de Gini) para diferentes valores del coeficiente: $\alpha_4 = 1, \alpha_4 = 2, \alpha_4 = -1, \alpha_4 = -2$ (de izq. a dcha. y de arriba a abajo).

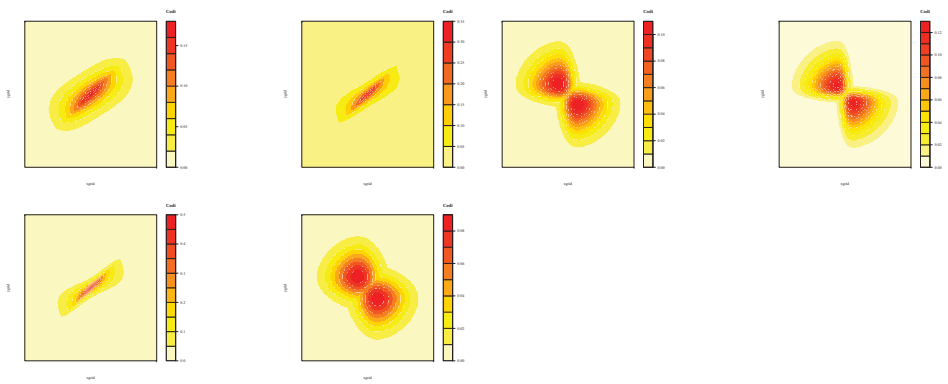

Figura 9.5: Contornos de isodensidad de la cópula CrEnC con momento μ_6 (Spearman's footrule φ) para diferentes valores del coeficiente: $\alpha_6 = 1, \alpha_6 = 2, \alpha_6 = 3, \alpha_6 = -1, \alpha_6 = -2, \alpha_6 = -3$ (de izq. a dcha. y de arriba a abajo).

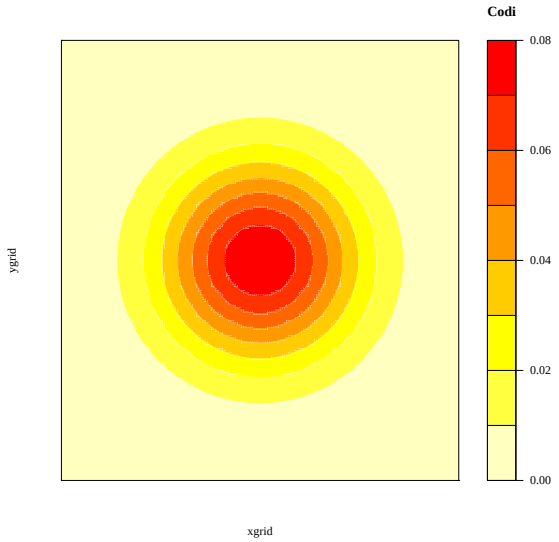

Figura 9.6: Contornos de isodensidad de la cópula CrEnC con momento μ_i para valor nulo del coeficiente: $\alpha_i = 0$.

Capítulo 10

Estudio de un registro de precipitación simulado

En los capítulos anteriores se ha presentado un modelo para la representación conjunta de cantidades extremales en dos ubicaciones (ocurrencia, marginales y dependencia), donde la dependencia puede representarse mediante una cópula paramétrica Gumbel o una cópula CrEnC. El modelo propuesto se aplica a un conjunto de datos simulados, para comprobar el buen desempeño del mismo.

10.1. Datos

Se dispone de dos series de precipitación diaria (mm) dependientes entre sí simuladas (cuyas distribuciones marginales son generalizadas de Pareto (GPD)), a las cuales se les ha aplicado el modelo propuesto en el Capítulo 3 y sucesivos. La log-precipitación sobre un umbral suficientemente alto (log(20)) se ha modelizado mediante una distribución GPD y la dependencia entre las dos series se ha modelizado utilizando una cópula CrEnC. Se ha simulado una muestra de 178 sucesos conjuntos de lluvia en dos ubicaciones. El diagrama de dispersión conjunto de estos excesos (Fig. 10.1) indica que existe una dependencia moderada de tipo lineal entre ambas series. El coeficiente de correlación de Pearson ($\rho_P = 0.7580$) y el de Spearman ($\rho_S = 0.7306$) confirman esta apreciación visual de dependencia moderada.

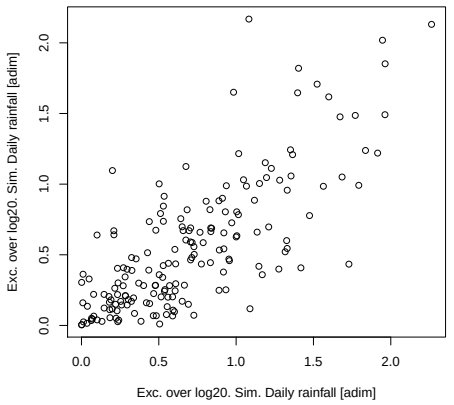

Figura 10.1: Excesos de log-precipitación sobre el umbral seleccionado (Sim.) . Escala $\mathbb{R}^{+2}$.

10.2. Priori de los parámetros marginales del modelo

Las distribuciones *GPD* de los excesos de log-precipitación sobre el umbral escogido en las dos ubicaciones se consideran similares y por tanto se ha establecido un mismo priori conjunto para los parámetros marginales en ambas ubicaciones (Fig. 10.2). El proceso de determinación de este priori es el descrito en la Sección 6.1.

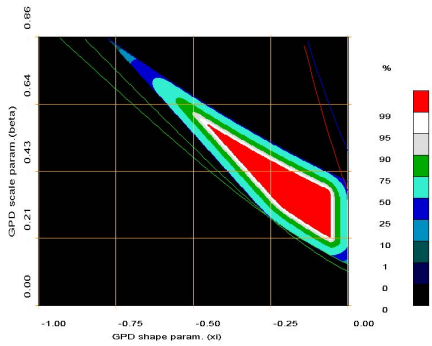

Figura 10.2: Priori conjunto de los parámetros de la distribución $GPD(\xi, \beta)$.

10.3. Ocurrencia de los sucesos y parámetros marginales

Se considera que la ocurrencia de los sucesos viene dada por un proceso de Poisson evaluado, con tamaños GPD. Las Figuras 10.3 y 10.4 muestran los histogramas de la muestra del posteriori obtenida para los parámetros marginales y de ocurrencia de este modelo. La Tabla 10.1 muestra percentiles seleccionados de las muestras para cada uno de los parámetros.

Una de las hipótesis a priori del modelo es que las distribuciones GPD marginales pertenecen al dominio de Weibull, es decir, se ha impuesto que el parámetro ξ sea negativo. La distribución de la muestra del posteriori de este parámetro para ambas ubicaciones se centra en valores lejanos al cero, por lo que la hipótesis Weibull es coherente con el conjunto de datos disponible.

En el caso de las distribuciones a posteriori de los parámetros β, puede apreciarse un corte en el histograma de la primera de las ubicaciones. Este corte es debido al recinto conjunto (ξ, β) establecido a priori, el cual restringe los valores admisibles del parámetro.

Las distribuciones a posteriori en la segunda ubicación presentan una dispersión mucho más elevada que las de la primera ubicación, por lo que las estimaciones a posteriori de parámetros compuestos,

como la cota superior de la distribución $(-\beta/\xi)$ del dominio GPD-Weibull, también presentarán una gran dispersión (Ver Tabla 10.2). Es más, si se observan los correspondientes valores en la escala usual (Tabla 10.3) se observa que los percentiles superiores de la cota superior de la distribución obtenido son mayores que las cotas superiores de precipitación que pudiéramos considerar razonables físicamente. Esto es debido a los valores de ξ cercanos a cero en la muestra del posteriori, valores que corresponden a valores con una cola finita pero mucho más pesada, más cercana al dominio GPD-Gumbel.

Tabla 10.1: Percentiles de la muestra del posteriori para los parámetros del modelo.

	2.5 %	25 %	50 %	75 %	97.5 %
ξ_{X1}	-0.35	-0.32	-0.30	-0.27	-0.20
β_{X1}	0.74	0.81	0.83	0.85	0.86
ξ_{X2}	-0.31	-0.24	-0.20	-0.15	-0.06
β_{X2}	0.56	0.62	0.67	0.71	0.80
λ	14.36	15.94	16.96	18.03	19.96

.

Tabla 10.2: Percentiles de las cotas superiores de la distribución GPD (dominio de Weibull) de los excesos de log-precipitación sobre el umbral seleccionado, correspondientes los parámetros estimados del modelo (Tabla 10.1).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior Ubicación1	2.40	2.59	2.77	3.05	4.11
cota superior Ubicación2	2.45	2.94	3.39	4.24	9.84

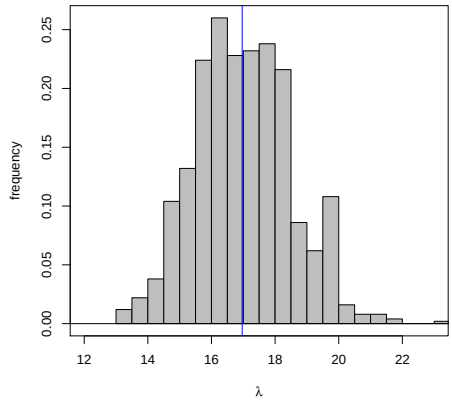

Figura 10.3: Histograma de la muestra del posteriori para el parámetro λ de Poisson (tasa de ocurrencia). Línea azul: mediana de la muestra.

Tabla 10.3: Percentiles de las cotas superiores estimados de precipitación correspondientes a la cota superior de los excesos sobre umbral (Tabla 10.2).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior Ubicación1 (mm)	221.44	266.60	318.22	423.78	1215.30
cota superior Ubicación1 (mm)	232.83	377.59	591.13	1394.85	375033.32

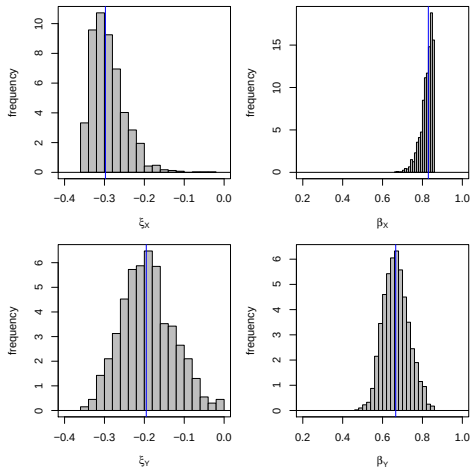

Figura 10.4: Histograma de la muestra del posteriori para los parámetros marginales ξ, β. La línea azul marca la mediana de la muestra del posteriori para cada parámetro.

10.4. Dependencia mediante cópula CrEnC

La dependencia entre las series de excesos de log-precipitación simuladas se ha modelizado mediante una cópula CrEnC (ver Capítulo 5). Para dar forma a la cópula CrEnC se ha implementado un conjunto de medidas de asociación que corresponden a diferentes tipos de dependencia (ver Cap. 9), y a los que corresponden los parámetros $\alpha_i, i = 1, \ldots, 7$ de la cópula.

La Figura 10.5 muestra los histogramas de la muestra del posteriori obtenida para los parámetros de la cópula. Para este conjunto de datos, se ha seleccionado un solo momento mediante el método de razón de verosimilitudes (Ver Sec. 7.1.2), que corresponde al coeficiente α_6, Spearman's footrule φ. La Tabla 10.4 muestra percentiles seleccionados de la muestra del posteriori para este parámetro.

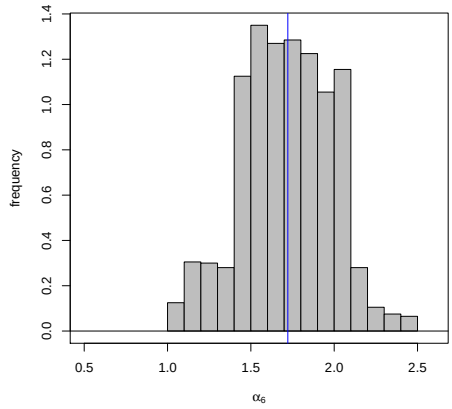

Figura 10.5: Histograma de la muestra del posteriori para el parámetro de la cópula CrEnC seleccionado, α_6. La línea azul marca la mediana de la muestra del posteriori para el parámetro.

En la Figura 10.6 se muestra la densidad estimada de la cópula CrEnC mediante contornos de isodensidad en $\mathbb{R}^2$.

Tabla 10.4: Percentiles de la muestra del posteriori para el parámetro seleccionado de la cópula CrEnC

	2.5 %	25 %	50 %	75 %	97.5 %
α_6	1.15	1.53	1.72	1.93	2.17

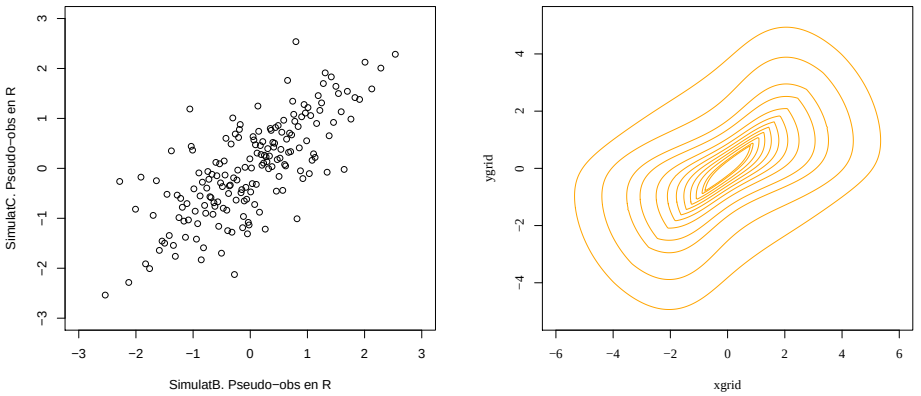

Figura 10.6: Cópula CrEnC en $\mathbb{R}^2$. Pseudo-observaciones en $\mathbb{R}^2$ (Izq.). Contornos de isodensidad a partir de la muestra del posteriori de los parámetros (Fig. 10.5) (Dcha.).

Bondad de ajuste. Dependencia mediante cópula CrEnC

Tal y como se presenta en la Sección 8.1, se desea contrastar la bondad de ajuste del modelo en diversos aspectos. Según los resultados mostrados en la Tabla 10.5, no puede rechazarse que la distribución marginal de las precipitación simulada sea GPD, pese a que el contraste Kolmogorov-Smirnov aparezca un tanto alterado para la primera de las ubicaciones. La comprobación de la hipótesis suplementaria de GPD en el dominio de Weibull ($\xi < 0$ y soporte limitado) mediante los diversos p-valores *Slope*, pendiente de la recta de regresión de los excesos esperados, Tabla 10.6, indica que la hipótesis es compatible

con los datos (como ya se podía apuntar al observar los histogramas de la muestra a posteriori de los parámetros marginales, Fig. 10.4). Finalmente, los estadísticos basados en τ, Tabla 10.7, indican que la dependencia global de la muestra inicial y la estimada es similar. Por tanto, el modelo es coherente globalmente con los datos.

Tabla 10.5: Bondad de ajuste marginal para la Ubicación1 y la Ubicación2. Percentiles seleccionados de los p-valores a posteriori.

	p-val.	2.5 %	50 %	97.5 %
Ubic1	K-S	0.00089	0.01666	0.05313
Ubic1	Multinomial	0.00745	0.05656	0.46954
Ubic2	K-S	0.09503	0.81804	1.00000
Ubic2	Multinomial	0.23208	0.77912	0.98892

Tabla 10.6: Pendientes predictivas de la recta de regresión de los excesos esperados para cada una de las marginales. p-valor de discrepancia correspondiente al contraste sobre la validez del priori GPD en DA-Weibull para cada marginal.

Value	X_1	X_2
Slope orig.	-0.286071	-0.143184
Slope predictive	-0.208869	-0.137417
Slope Discrepancy	0.947423	0.504124

Tabla 10.7: Contraste sobre el modelo global. Coeficiente τ original, coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. Los coeficientes muestran dependencias similares. El p-valor basado en la discrepancia τ no permite rechazar la hipótesis de validez del modelo global especificado.

τ original	τ PredictivoG	Discrepancia τ
0.542055	0.556579	0.681443

10.5. Comprobación de los resultados

Dado que se ha aplicado el modelo a una muestra simulada de precipitación de la cual se conocen los parámetros originales, parece adecuado hacer una comparación entre estos parámetros poblacionales y los parámetros estimados por el modelo, tanto marginales como de dependencia.

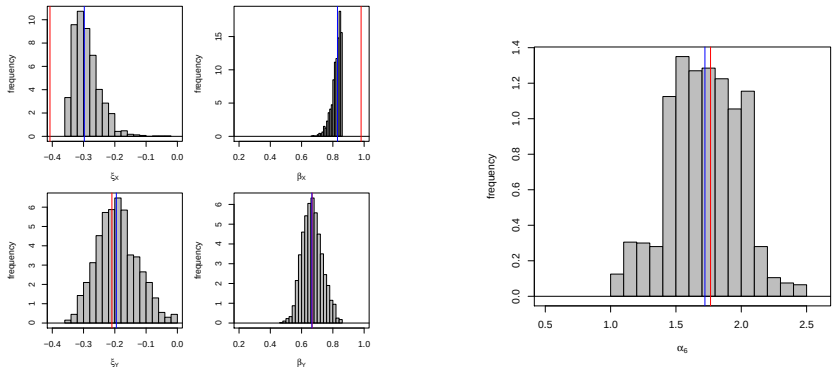

Figura 10.7: Histograma de la muestra del posteriori para los parámetros de la cópula seleccionados (derecha) y para los parámetros marginales ξ, β (izquierda). La línea azul marca la mediana de la muestra del posteriori para cada parámetro. En rojo, línea marcando el valor del parámetro original.

En la Figura 10.7, izquierda, se representan los valores de la muestra del posteriori obtenida con los valores de parámetros originales. Los resultados son muy diferentes para ambas ubicaciones. Para la segunda ubicación los valores centrales de la muestra del posteriori coinciden en gran medida con los valores originales (en el caso del parámetro β, son tan similares que una de las dos líneas no se aprecia correctamente). En cambio, para la primera de las ubicaciones los valores originales quedan lejos de los valores de la muestra del posteriori. Esto es debido a la restricción de la región conjunta del priori de los parámetros (Fig. 10.2). Los valores del posteriori simulados han de corresponder a distribuciones GPD-Weibull con cota superior mayor que el máximo de los valores de la muestra. Así, al restringir los valores de β, los correspondientes valores de ξ han de ser menores en valor absoluto, y eso se refleja en la correspondiente diferencia entre la muestra del posteriori

y el parámetro original. En el caso de los parámetros de dependencia, representada mediante cópula CrEnC, el parámetro seleccionado no sólo es el el mismo (α_6) que el original, sino que además los valores estimados son similares a los valores centrales de la muestra del posteriori (Fig. 10.7, derecha). Podemos concluir que el modelo ajusta bien a los datos, y las estimaciones obtenidas son coherentes con los valores originales utilizados para la simulación.

10.6. Sensibilidad al priori

En la Sección 10.3 se ha presentado la estimación de los parámetros marginales del modelo. La Figura 10.4 muestra los histogramas de la muestra del posteriori para estos parámetros. El posteriori del parámetro β aparece limitado en su extremo superior, a causa del priori establecido para el parámetro (Fig. 10.2). Es decir, este priori es informativo. Por tanto, parece conveniente analizar cómo afecta este priori a las cantidades a posteriori obtenidas a partir del modelo. No se pretende hacer un análisis exhaustivo de la sensibilidad al priori, pero sí hacer una primera aproximación a los efectos que pueden producir los cambios en él.

Se ha estimado el modelo y realizado cálculo de valores a posteriori en cinco situaciones diferentes, modificando en cada caso el priori de sólo uno de los dos parámetros ξ, β. Se ha ampliado el límite superior del priori de cada parámetro, dejando intacto el inferior. Esta decisión se basa en las apreciaciones visuales a partir de la Fig. 10.4: aparentes limitaciones del posteriori ($\beta < 0.857$) y en menor medida por $\xi < 0$.

En cada uno de los escenarios se ha calculado la probabilidad de no excedencia a posteriori de un valor de exceso de referencia marginal, 1.60940 (100 mm) y un valor de exceso de referencia conjunto (2.30260,2.01490), (200,150) mm, y se ha calculado la diferencia entre esta probabilidad y las probabilidades correspondientes calculadas para los parámetros de la Sec. 10.3:

$$d(p_1, p_2) = \frac{\log(p_1/p_2)}{par_1 - par_2} \ , \tag{10.6.1}$$

donde p_1, p_2 denotan las probabilidades de no excedencia a posteriori obtenidas para los valores par_1, par_2 del parámetro de interés.

- Modificación del límite superior del priori de β

  Se ha estimado el modelo modificando el límite superior del priori de β a valores $\beta_{sup} = 1.0$, $\beta_{sup} = 1.2$, $\beta_{sup} = 1.5$. Los histogramas de los parámetros marginales de las muestras del posteriori en cada caso se presentan en las Figuras 10.8, 10.9 y 10.10. En la Tabla 10.8 se presentan las diferencias de las probabilidades de referencia obtenidas al modificar este límite superior respecto a la probabilidad original, utilizando la diferencia en Ec. 10.6.1. En la Fig. 10.4 se observa que el priori limita el posteriori de β_1, pero no el de β_2, lo que se refleja en cambios en las probabilidades predictivas marginales de Y_1 y en las conjuntas, pero no en las de Y_2. El aumento del límite superior $\beta_{sup} = 1.0$ resulta insuficiente puesto que el priori sigue limitando el posteriori (Fig. 10.8). En el otro extremo, el aumento a $\beta_{sup} = 1.5$ afecta poco a la muestra del posteriori (Fig. 10.10). Las diferencias más acusadas entre probabilidades se dan en el primer cambio de límite superior, aunque son diferencias pequeñas. Los parámetros ξ y β están vinculados por el límite superior, y al ampliar el conjunto de valores de β, los valores de ξ obtenidos en el posteriori se modifican. Los valores obtenidos se encuentran más alejados del cero, lo que proporciona probabilidades de no excedencia marginales ligeramente inferiores. Para las probabilidades conjuntas, dado que solo varía una de las distribuciones marginales a posteriori el comportamiento es menos interpretable. Para $\beta_{sup} = 1.0$ las probabilidades conjuntas disminuyen ligeramente, mientras que para los valores no limitantes del priori $\beta_{sup} = 1.2$, $\beta_{sup} = 1.5$, las probabilidades de no excedencia conjunta aumentan ligeramente. En resumen, el priori establecido resulta informativo, pero los valores a posteriori obtenidos al modificar el límite superior de β no son demasiado diferentes de los originales, es decir, el posteriori no es demasiado sensible al priori establecido para este parámetro.

- Modificación del límite superior del priori de ξ

  Pese a que la hipótesis a priori de distribución marginal GPD en el dominio de Weibull ($\xi < 0$) parece ser consistente con los datos, el modelo se ha estimado también modificando el límite superior del priori de ξ a valores $\xi_{sup} = 0.2$, $\xi_{sup} = 0.5$. Los histogramas de

Tabla 10.8: Diferencia entre probabilidades de referencia respecto a la original para modificaciones del límite superior del priori de β.

Prob	$\beta_{sup} = 1.0$	$\beta_{sup} = 1.2$	$\beta_{sup} = 1.5$
$P[Y_1 \leq 1.60940]$	-0.07830	-0.03813	-0.02011
$P[Y_2 \leq 1.60940]$	0.00196	0.00115	-0.00162
$P[Y_1 \leq 2.30260, Y_2 \leq 2.01490]$	-0.04328	0.03293	0.01402

Tabla 10.9: Diferencia entre probabilidades de referencia respecto a la original para modificaciones del límite superior del priori de ξ.

Prob	$\xi_{sup} = 0.2$	$\xi_{sup} = 0.5$
$P[Y_1 \leq 1.60940]$	-0.00300	0.00122
$P[Y_2 \leq 1.60940]$	0.00196	0.00193
$P[Y_1 \leq 2.30260, Y_2 \leq 2.01490]$	-0.03250	-0.01215

los parámetros marginales de las muestras del posteriori en cada caso se presentan en las Figuras 10.11 y 10.12. En la Tabla 10.9 se presentan las diferencias de las probabilidades de referencia obtenidas al modificar este límite superior respecto a la probabilidad original, utilizando la distancia en Ec. 10.6.1. Al aumentar el límite superior del parámetro (Fig. 10.4), los histogramas de la muestra del posteriori apenas presentan cambios respecto a los originales , y las probabilidades predictivas calculadas con estos cambios respecto a las originales apenas tienen variación. Por tanto, la hipótesis GPD-Weibull es coherente con los datos, lo que se traduce en que el posteriori no es sensible al aumento del límite superior del priori de este parámetro ξ.

En resumen, aunque se ha establecido un priori subjetivo basado en la opinión de especialista, que resulta ser informativo, el posteriori es poco sensible a los cambios en este priori.

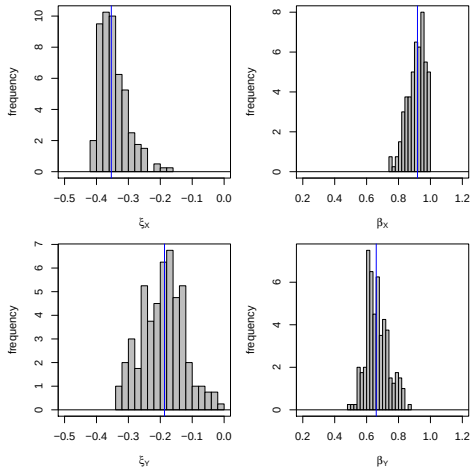

Figura 10.8: Histograma de la muestra del posteriori para los parámetros marginales ξ, β. La línea azul marca la mediana de la muestra del posteriori para el parámetro. Límite superior del priori de β modificado: $\beta_{sup} = 1.0$.

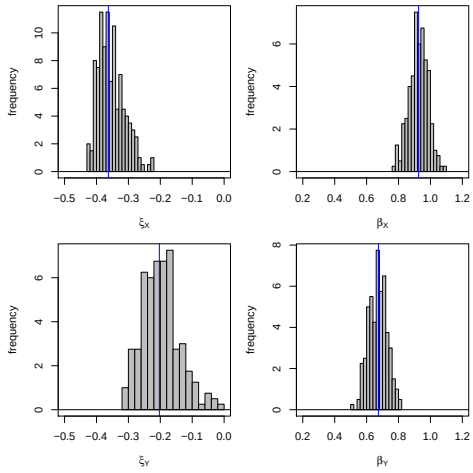

Figura 10.9: Histograma de la muestra del posteriori para los parámetros marginales ξ, β. La línea azul marca la mediana de la muestra del posteriori para el parámetro. Límite superior del priori de β modificado: $\beta_{sup} = 1.2$.

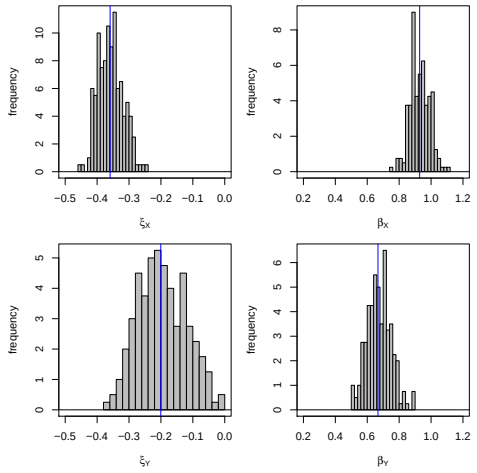

Figura 10.10: Histograma de la muestra del posteriori para los parámetros marginales ξ, β. La línea azul marca la mediana de la muestra del posteriori para el parámetro. Límite superior del priori de β modificado: $\beta_{sup} = 1.5$.

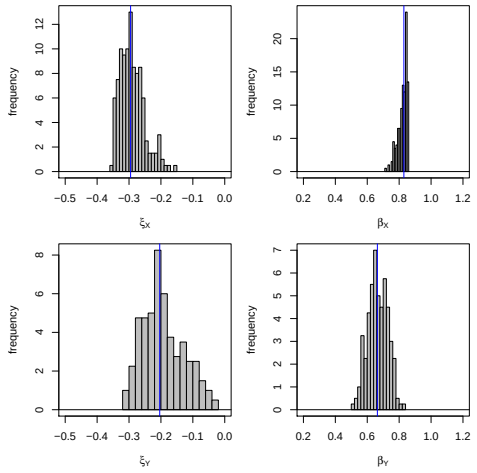

Figura 10.11: Histograma de la muestra del posteriori para los parámetros marginales ξ, β. La línea azul marca la mediana de la muestra del posteriori para el parámetro. Límite superior del priori de ξ modificado: $\xi_{sup} = 0.2$.

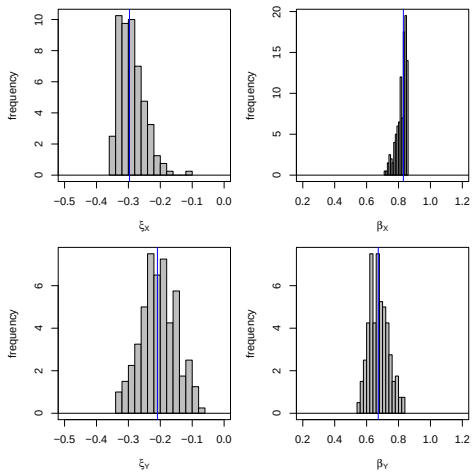

Figura 10.12: Histograma de la muestra del posteriori para los parámetros marginales ξ, β. La línea azul marca la mediana de la muestra del posteriori para el parámetro. Límite superior del priori de ξ modificado: $\xi_{sup} = 0.5$.

10.7. Discusión

Se ha modelizado un conjunto de datos simulado mediante el modelo propuesto, utilizando dependencia mediante cópula CrEnC. El modelo ajusta bien a los datos, global y marginalmente, y los resultados son comparables a los valores originales utilizados para simular los datos. Se ha comprobado que las estimaciones de los parámetros marginales se ven afectadas por el priori establecido mediante opinión de especialista. Este priori, pese a ser informativo, no afecta a los valores a posteriori obtenidos a partir del modelo.

Capítulo 11

Estudio de un registro de precipitación

En el capítulo anterior se ha aplicado el modelo propuesto en el Capítulo 3 a un conjunto de datos simulados. Una vez comprobado el buen ajuste de este modelo a los datos sintéticos, éste se aplica a diversos conjuntos de datos observados, correspondientes a registros de precipitación. A continuación se describen los datos (origen, tipología y análisis exploratorio) y se describen los resultados obtenidos al estimar los parámetros del modelo. Por último, se valora el ajuste del modelo y se proporcionan unos valores a posteriori de interés.

11.1. Datos

Se desea analizar la dependencia entre dos series de precipitación diaria registradas en localizaciones cercanas. Para ilustrar el procedimiento se han escogido tres parejas de localizaciones: Bolulla y Callosa de Ensarrià (Alicante), Vergel de Recons y Simat de Valldigna (Alicante) y Vall de Laguard Fontilles y Almudaina (Alicante), Fig. 11.1.

Se dispone de un registro de 30 años de precipitación diaria para cada una de las ubicaciones, descrito en Romero et al. (1998), y Egozcue and Ramis (2001). Se puede considerar que la variable precipitación diaria tiene una escala relativa (Egozcue et al., 2006; Tolosana-Delgado et al., 2010) y por tanto conviene transformala a escala real mediante

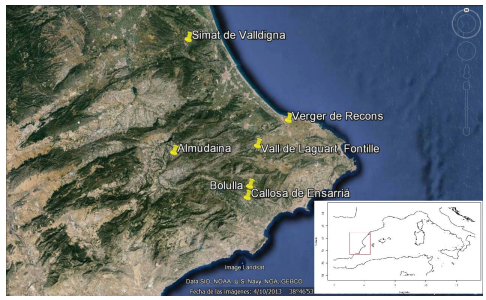

Figura 11.1: Localización de las tres parejas de observatorios (Alicante).

una transformación logarítmica. La ocurrencia de los sucesos de precipitación se modela mediante un proceso de Poisson evaluado homogéneo. A cada suceso se le asigna la log-precipitación máxima del suceso. Así, para cada una de las ubicaciones, la log-precipitación por encima de un umbral suficientemente alto (log(20)) se ha modelizado mediante una distribución Generalizada de Pareto, GPD(ξ_i, β_i), $i = 1, 2$ (Egozcue et al., 2006). La dependencia entre los excesos de log-precipitación en ambas localizaciones se ha modelizado utilizando dos herramientas diferentes: una cópula paramétrica, la cópula de Gumbel (Gumbel, 1960) (Sec. 2.1) y mediante la cópula CrEnc definida en el Capítulo 5.

Los parámetros del modelo (marginales, de ocurrencia y de cópula) han sido estimados utilizando un muestreo de Gibbs, con 2000 iteraciones y un *Burn In* del 50 % de las iteraciones (Gelman et al., 1995). Se ha verificado la convergencia de la cadena para cada uno de los parámetros utilizando el criterio de Gelman (Gelman et al., 1995). La muestra del posteriori obtenida mediante el muestreo de Gibbs permite representar la incertidumbre presente en las estimaciones de los parámetros.

Para dar forma a la cópula CrEnC se ha implementado un conjunto de medidas de asociación que corresponden a diferentes tipos de dependencia (ver Cap. 9), y a los que corresponden los parámetros $\alpha_i, i = 1, \ldots, 7$ de la cópula. La cópula CrEnC se ha estimado utilizando sólo aquellos momentos seleccionados mediante el método de razón de verosimilitudes (ver Sec. (7.1.2)). Adicionalmente, en una de las parejas de ubicaciones se han utilizado todos los momentos significativos individualmente para los datos. Los resultados muestran (Sección 11.2.3)

que la selección mediante razón de verosimilitudes resulta suficiente.

11.1.1. Selección del umbral GPD

Se dispone de un registro de 30 años de log-precipitación en cada una de las ubicaciones. En el tratamiento de la base de datos original (Romero et al., 1998; Egozcue and Ramis, 2001) se utiliza el umbral $\log(20)$, para definir el concepto de día lluvioso, y éste es el valor escogido como umbral absoluto. Se han obtenido los excesos de log-precipitación sobre este umbral, $h = \log(20)$ y se ha estudiado su adecuación como umbral 'suficientemente alto' para que la distribución marginal sea GPD (Teorema de Pickands, Sec. 2.4).

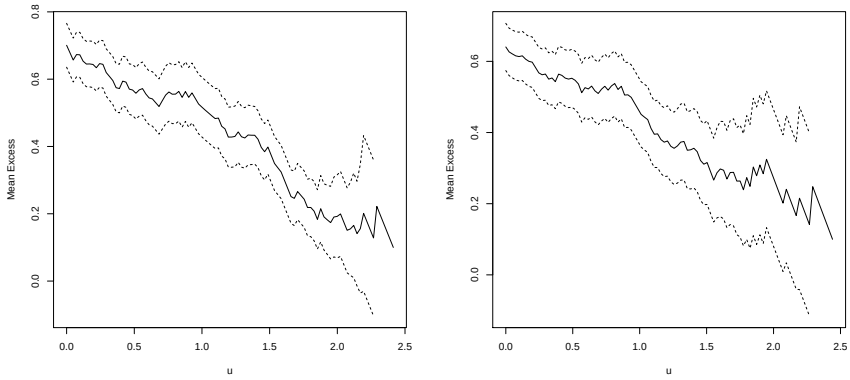

Figura 11.2: Mean excess plot de los excesos de log-precipitación por encima de log(20). Izquierda: Bolulla; derecha: Callosa. Diagnóstico mediante paquete Ismev de R .

La Figura 11.2 muestra las gráficas del exceso medio (*Mean excess plot*) para las ubicaciones de Bolulla y Callosa. Esta gráfica de diagnóstico permite detectar visualmente el umbral h_0 a partir del cual se puede considerar que la distribución de los excesos sobre h_0 corresponde a una $GPD(\xi, \beta)$ (Castillo, 1988). En ambas gráficas se observa que el umbral h_0 es 'suficientemente alto', y por tanto, se puede considerar que los excesos de log-precipitación sobre $\log(20)$ pueden considerarse

distribuidos GPD. La bondad de ajuste del modelo marginal a los datos se comprueba visualmente (ver Fig. 11.3 y Fig. 11.4). Se presentan los QQplot de los datos respecto la distribución de referencia y un ajuste de la densidad a los datos. Estas dos gráficas han sido elaboradas mediante el paquete Ismev de R (Heffernan and Stephenson, 2012).

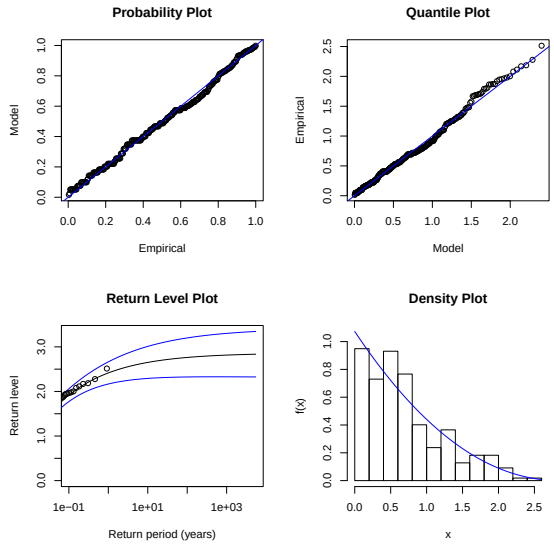

Figura 11.3: Bondad de ajuste a la distribución $GPD(\xi, \beta)$ de los excesos de log-precipitación por encima de log(20) en Bolulla. Diagnóstico mediante paquete Ismev de R.

Los resultados de las comprobaciones marginales para las demás ubicaciones son similares y se omiten por simplicidad. Podemos considerar que las distribuciones marginales de los datos, excesos de log-precipitación sobre log(20) en cada una de las ubicaciones, corresponden a una distribución GPD.

11.1.2. Priori de los parámetros marginales del modelo

Las distribuciones GPD de los excesos de log-precipitación sobre el umbral escogido se consideran similares en las seis ubicaciones y por

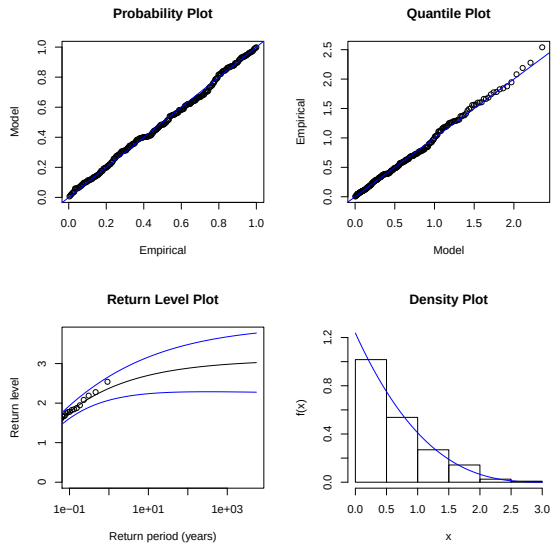

Figura 11.4: Bondad de ajuste a la distribución $GPD(\xi, \beta)$ de los excesos de log-precipitación por encima de log(20) en Callosa. Diagnóstico mediante paquete Ismev de R.

tanto se ha establecido un mismo priori conjunto para los parámetros marginales en todas las ubicaciones, mostrado en la Fig. 11.5.

El proceso de determinación de este priori es el descrito en la sección 6.1. No se ha hecho un análisis sistemático de la sensibilidad al priori, pero sí una primera aproximación a la influencia de los cambios de los límites superiores del priori para cada uno de los parámetros marginales (Ver Sec. 10.6).

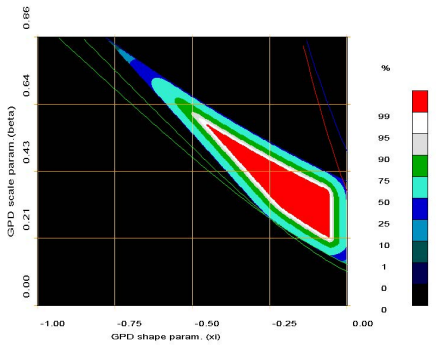

Figura 11.5: Priori conjunto de los parámetros de la distribución $GPD(\xi, \beta)$.

11.2. Bolulla y Callosa de Ensarrià (Alicante)

En esta Sección se analiza la dependencia entre la precipitación diaria en Bolulla y Callosa de Ensarrià (Alicante), (-0.113O, 38.678N y -0.121O, 38.650N respectivamente, Fig. 11.6). La log-precipitación sobre un umbral suficientemente alto (log(20)), Fig. 11.7, se ha modelizado mediante una distribución generalizada de Pareto (GPD). Se dispone de una muestra de 334 sucesos de lluvia en alguna de las dos ubicaciones, de los cuales, 178 son sucesos conjuntos (Fig. 11.8). El diagrama de dispersión conjunto de los excesos indica que existe una dependencia moderada de tipo lineal entre ambas series. El coeficiente de correlación de Pearson ($\rho_P = 0.7503$) y el de Spearman ($\rho_S = 0.4689$) confirman esta apreciación visual de dependencia moderada.

11.2.1. Ocurrencia de los sucesos y parámetros marginales

Las Figuras 11.9 y 11.10 muestran los histogramas de la muestra del posteriori obtenida para los parámetros marginales y de ocurrencia del modelo y la Tabla 11.1 muestra algunos percentiles seleccionados de las muestras para cada uno de los parámetros. La mediana predictiva

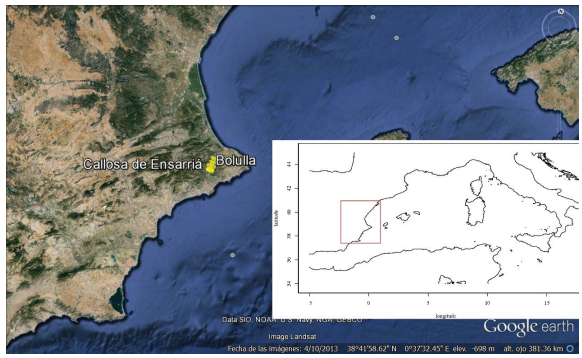

Figura 11.6: Localización de Bolulla y Callosa de Ensarrià (Alicante).

del parámetro λ se halla entorno a los 17 sucesos anuales (ver Tabla 11.1), en concordancia con el análisis realizado en Egozcue and Ramis (2001).

Los histogramas de la muestra del posterior de los parámetros β para ambas ubicaciones muestran cortes debidos al priori establecido. En cambio, la hipótesis de que la distribución marginal sea GPD- Weibull, es decir, $\xi < 0$ no afecta a las muestras del posteriori, dado que se observan valores del parámetro ξ claramente separados del cero.

A partir de estos parámetros marginales pueden obtenerse cantidades a posteriori, como las cotas superiores correspondientes a las distribuciones marginales GPD-Weibull ($-\beta/\xi$, Tabla 11.2). Los valores de estas cotas superiores expresados en la escala usual (mm), Tabla 11.3, muestran un rango de valores coherente con las observaciones y con la idea de la cota superior física de la variable analizada (precipitación diaria).

11.2.2. Dependencia mediante cópula paramétrica Gumbel

La dependencia entre los excesos de log-precipitación se ha modelizado mediante una cópula paramétrica de la familia Gumbel (Gumbel, 1960). La Figura 11.11 corresponde al histograma de la muestra del posteriori del parámetro δ de la cópula.

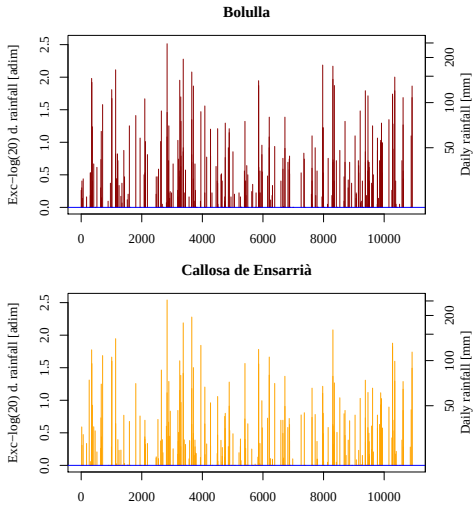

Figura 11.7: Excesos de log-precipitación diaria por encima de log(20) en Bolulla y Callosa de Ensarrià.

La Tabla 11.4 presenta descriptivos seleccionados de la muestra de este parámetro δ, con valores correspondientes a una dependencia moderada, coherente con los valores de ρ_P de Pearson i ρ_S de Spearman. La densidad de la cópula Gumbel estimada para la mediana de estos valores, en $[0,1]^2$, se muestra mediante contornos de isodensidad en la Fig. 11.12. La densidad en $\mathbb{R}^2$ se ha representado en la Fig. 11.13.

Bondad de ajuste. Dependencia mediante cópula Gumbel

Se valora la coherencia de la representación de la dependencia entre las series mediante la cópula Gumbel. Para ello se utilizan los estadísticos de contraste presentados en la Sección 8.1. En la Tabla 11.5 se muestra el contraste respecto al ajuste global del modelo. La dependencia presente en los excesos originales es similar a la presente en las remuestras generadas a partir del posteriori, aspecto que se confirma observando los valores predictivos a posteriori del parámetro δ (Tabla 11.6). El modelo es globalmente coherente con los datos y representa

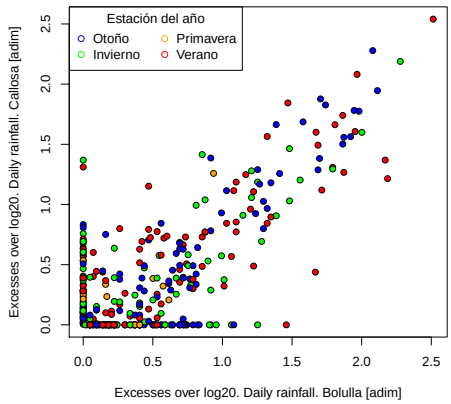

Figura 11.8: Excesos de log-precipitación sobre el umbral seleccionado, en Bolulla y Callosa. Escala $\mathbb{R}^{+2}$.

correctamente tanto el comportamiento marginal como de dependencia.

11.2.3. Dependencia mediante cópula CrEnC

La dependencia entre las series de excesos de log-precipitación en Bolulla y Callosa se ha modelizado también mediante una cópula CrEnC (Capítulo 5), utilizando los momentos establecidos en el Capítulo 9. A continuación se presentan tanto los resultados correspondientes a la cópula con los momentos seleccionados mediante el método de razón

Tabla 11.1: Percentiles de la muestra del posteriori para los parámetros del modelo. Bolulla y Callosa.

	2.5 %	25 %	50 %	75 %	97.5 %
ξ_{X1}	-0.30	-0.26	-0.24	-0.21	-0.15
β_{X1}	0.75	0.80	0.83	0.84	0.86
ξ_{X2}	-0.30	-0.26	-0.23	-0.19	-0.11
β_{X2}	0.67	0.74	0.78	0.81	0.85
λ	14.09	15.91	16.98	18.06	19.98

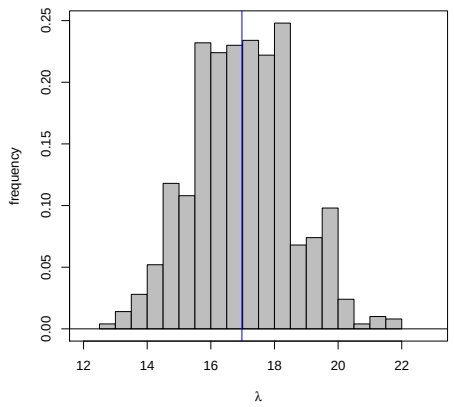

Figura 11.9: Histograma de la muestra del posteriori para el parámetro λ de Poisson (tasa de ocurrencia). Línea azul: mediana de la muestra. Bolulla y Callosa.

de verosimilitudes como los resultados obtenidos utilizando todos los momentos seleccionados individualmente en ese proceso.

Dependencia mediante cópula CrEnC: momento seleccionado por razón de verosimilitudes

Se desea modelizar la dependencia entre las series de excesos de log-precipitación mediante una cópula CrEnC. Se ha aplicado la selección de momentos mediante el método de razón de verosimilitudes.

Tabla 11.2: Percentiles de las cotas superiores de la distribución *GPD* (dominio de Weibull) de los excesos de log-precipitación sobre el umbral seleccionado, correspondientes los parámetros estimados del modelo (Tabla 11.1).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior Bolulla	2.77	3.14	3.40	3.80	5.22
cota superior Callosa	2.75	3.07	3.38	3.88	6.17

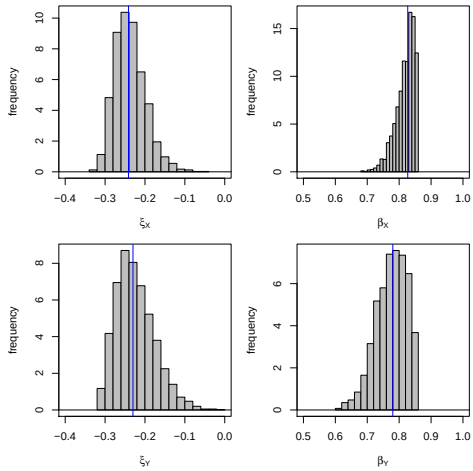

Figura 11.10: Histograma de la muestra del posteriori para los parámetros de la cópula seleccionados (derecha) y para los parámetros marginales ξ, β (izquierda). La línea azul marca la mediana de la muestra del posteriori para cada parámetro. Bolulla y Callosa.

Ha resultado seleccionado un único momento, α_6. En la Figura 11.14 se representan los histogramas de la muestra del posteriori obtenida para este parámetro de la cópula. La Tabla 11.7 muestra percentiles seleccionados de la muestra de este parámetro, que corresponden a dependencias moderadas.

En la Fig. 11.15 se muestra la densidad de la cópula CrEnC mediante contornos de isodensidad en $\mathbb{R}^2$, correspondientes a la mediana de la muestra del posteriori del parámetro. Se observa que la densidad ajusta visualmente a las pseudo-observaciones (en escala $\mathbb{R}^2$). Debe advertirse

Tabla 11.3: Percentiles de las cotas superiores estimados precipitación correspondientes a la cota superior de los excesos sobre umbral (Tabla 11.2).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior Bolulla (mm)	318.48	460.83	602.23	895.41	3689.93
cota superior Callosa (mm)	312.12	431.53	586.76	966.04	9536.57

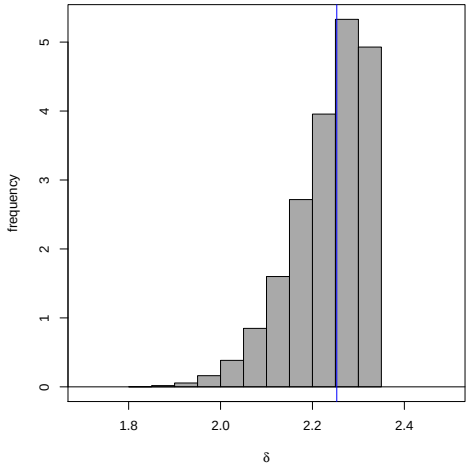

Figura 11.11: Histograma de la muestra del posteriori para el parámetro δ de la cópula Gumbel. La línea azul marca la mediana de la muestra del posteriori para el parámetro.

Tabla 11.4: Percentiles de la muestra del posteriori para el parámetro δ de la cópula Gumbel

	2.5 %	25 %	50 %	75 %	97.5 %
δ	2.04	2.19	2.25	2.30	2.33

que estos ajustes visuales deben ser interpretados con cautela: suelen ser complicados de apreciar, y por tanto es conveniente complementarlos con un contraste de bondad de ajuste; además, en este caso se presenta el contorno correspondiente a uno solo de los valores de la muestra del posteriori (Fig. 11.15). Dado que se dispone de una muestra completa del posteriori, las conclusiones a partir de la interpretación visual basada en un único valor de la muestra pueden ser parciales.

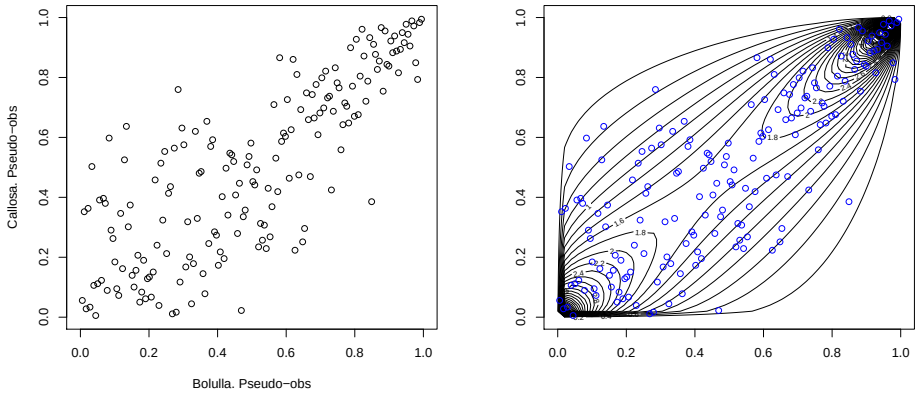

Figura 11.12: Pseudo-observaciones de la muestra (Bolulla y Callosa). Cópula Gumbel en $[0,1]^2$. Contornos de isodensidad a partir de la muestra del posteriori de los parámetros (Fig. 11.11).

Tabla 11.5: Contraste sobre el modelo global. Coeficiente τ original y coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. Los coeficientes muestran dependencias similares. El p-valor basado en la discrepancia τ no permite rechazar la hipótesis de validez del modelo global especificado.

τ original	τ PredictivoG	Discrepancia τ
0.59868	0.55149	0.11469

Dependencia mediante cópula CrEnC: todos los momentos seleccionado por razón de verosimilitudes

La selección de momentos que intervienen en la cópula CrEnC con frecuencia deja fuera del modelo algunos parámetros que han resultado significativos individualmente en el contraste de razón de verosimilitudes, pero que no son seleccionados conjuntamente en los sucesivos pasos del proceso. Para estos datos, los parámetros $\alpha_1, \alpha_2, \alpha_3, \alpha_4, y\ \alpha_6$ (Ver Sec. 11.2.3) son significativos individualmente, y de ellos se escoge el parámetro α_6. Sin embargo, en el proceso *forward* de selección se descarta incluir ningún parámetro más al modelo.

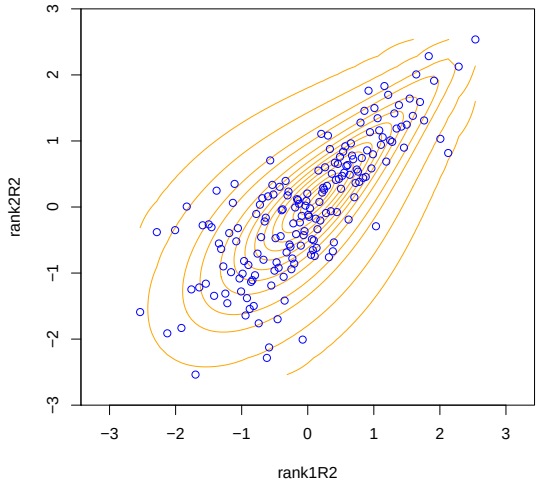

Figura 11.13: Ajuste de la Cópula Gumbel en $\mathbb{R}^2$ para Bolulla y Callosa. Contornos de isodensidad a partir de la muestra del posteriori de los parámetros.

En esta Sección se han considerado todos los momentos significativos individualmente. La Figura 11.16 muestra los histogramas correspondientes a la muestra del posteriori de los parámetros de la cópula con todos los parámetros, y la Tabla 11.8 percentiles seleccionados de la muestra. Los resultados son consistentes con los obtenidos en la Sección 11.2.3. Los parámetros no seleccionados en el proceso de razón de verosimilitudes presentan distribuciones muy simétricas y centradas en el cero. Son nulos en media, y por tanto los valores a posteriori que

Tabla 11.6: Contraste sobre el modelo global. Coeficiente δ original y coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. Los coeficientes muestran dependencias similares.

δ original	δ PredictivoG
2.4918	2.2477

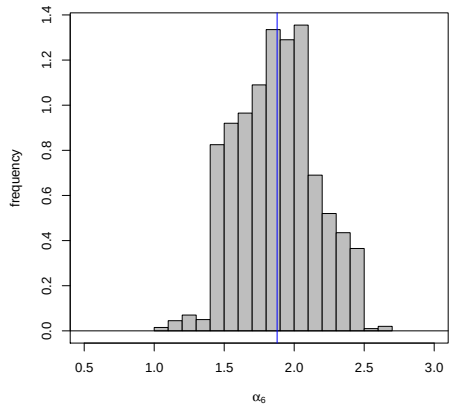

Figura 11.14: Histograma de la muestra del posteriori para el parámetro de cópula CrEnC seleccionado, α_6. La línea azul marca la mediana de la muestra del posteriori para el parámetro.

Tabla 11.7: Percentiles de la muestra del posteriori para el parámetro de la cópula CrEnC seleccionado.

	2.5 %	25 %	50 %	75 %	97.5 %
α_6	1.42	1.66	1.88	2.06	2.43

se obtendrían usando todos estos parámetros no diferirían de los obtenidos en la Sección 11.2.3. Aplicando el principio de parsimonia, por razones de simplicidad, en el resto de ejemplo optaremos por el modelo correspondiente al momento seleccionado por el método de razón de verosimilitudes.

Bondad de ajuste marginal y global. Dependencia mediante cópula CrEnC

Se desea valorar la coherencia de varios aspectos del modelo con los datos mediante los estadísticos de contraste presentados en la Sección

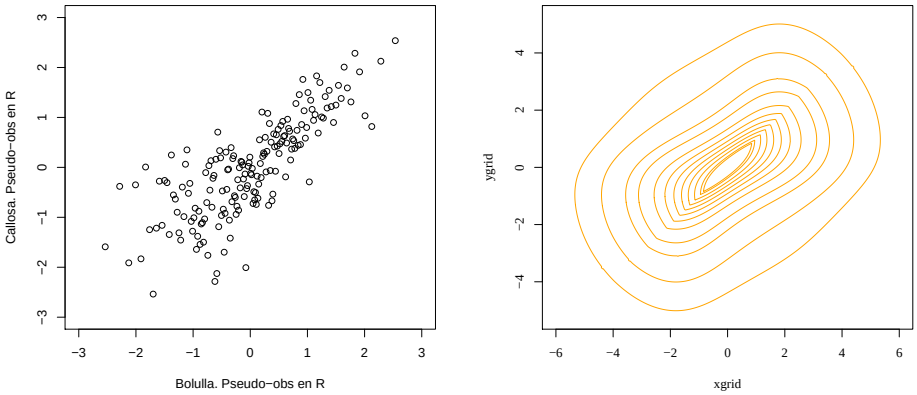

Figura 11.15: Cópula CrEnC en $\mathbb{R}^2$. Pseudo-observaciones en $\mathbb{R}^2$ (Izq.). Contornos de isodensidad a partir de la muestra del posteriori de los parámetros (Fig. 11.14) (Dcha.).

8.1. En la Tabla 11.9 se presentan percentiles seleccionados de los p-valores a posteriori correspondientes a la bondad de ajuste del modelo GPD marginal (Kolmogorov-Smirnov y multinomial). Los p-valores a posteriori para Callosa indican que no se puede rechazar que la distribución marginal GPD sea adecuada para esta ubicación. En el caso de Bolulla, la decisión sobre la bondad de ajuste depende del p-valor utilizado, lo que sugiere que quizá el umbral escogido no es suficientemente alto para que el ajuste GPD sea adecuado (ver Fig.

Tabla 11.8: Percentiles de la muestra del posteriori para los parámetros de la cópula CrEnC (parámetros significativos individualmente).

	2.5 %	25 %	50 %	75 %	97.5 %
α_1	-2.08	-0.65	0.03	0.66	2.01
α_2	-2.02	-0.66	0.03	0.67	2.02
α_3	-2.01	-0.71	-0.04	0.67	2.02
α_4	-2.02	-0.65	0.00	0.70	2.08
α_6	1.42	1.66	1.87	2.05	2.42

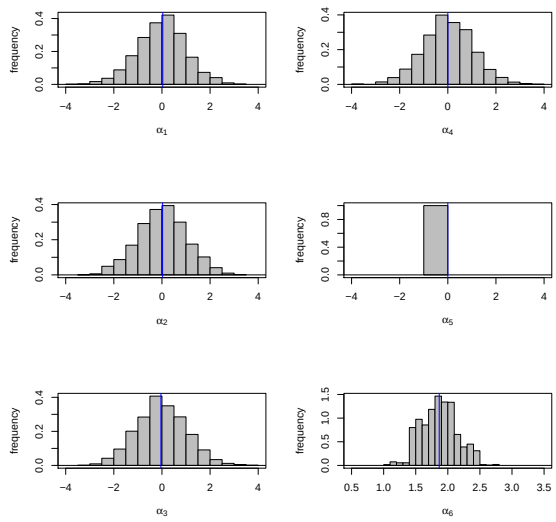

Figura 11.16: Histograma de la muestra del posteriori para los parámetros de la cópula significativos individualmente. La línea azul marca la mediana de la muestra del posteriori para cada parámetro.

11.17), pese a que el diagnóstico visual mediante gráfico de exceso esperado indicaba lo contrario (Fig. 11.3). En situaciones como esta, con p-valores a posteriori en la frontera, es importante poder realizar una interpretación 'usual' de estos p-valores, asumiendo que son uniformes en un intervalo, posiblemente más reducido que $[0, 1]$. Se han calculado las cotas de uniformidad presentadas en la Sección (8.1.2) (Tabla 11.10), las cuales proporcionan un intervalo de uniformidad sensiblemente más reducido que el intervalo $[0, 1]$. Se proporciona también el valor del centro de la distribución de p-valores, que sirve de referencia para tomar una decisión sobre la bondad de ajuste. No obstante, se debe profundizar en la interpretación de estos intervalos, dado que dependen del valor *verdadero* del grado de dependencia δ existente entre los elementos de la muestra de p-valores. Para estas dos ubicaciones el centro de la distribución de p-valores se encuentra más cerca de la cota correspondiente a $\delta = 1$ (dependencia) que de la cota correspondiente a la independencia $\delta = 2$ (Fig. 11.18).

Por lo que respecta a los p-valores predictivos a posteriori, Tabla 11.9, éstos indican que no se puede rechazar que el el modelo GPD marginal sea adecuado para las remuestras predictivas, tanto para Bolulla como para Callosa, lo que indica que el modelo global ajusta bien a los datos.

La hipótesis adicional de distribución en el dominio de Weibull se verifica mediante los estadísticos de *Slope*, pendiente de la recta de excesos esperados, Tabla 11.11, que no descartan la coherencia de esta hipótesis con los excesos originales y con las correspondientes remuestras. Respecto al ajuste global del modelo, la Tabla 11.12 muestra que la dependencia presente en los excesos originales es similar a la presente en las remuestras generadas a partir del posteriori. Por tanto, en términos generales, el modelo es globalmente coherente con los datos y representa correctamente tanto el comportamiento marginal como de dependencia.

Tabla 11.9: Bondad de ajuste marginal para Bolulla y Callosa. Percentiles seleccionados de los p- valores a posteriori (Sup.). Percentiles seleccionados de los p-valores predictivos a posteriori (Inf.)

	p-val.	2.5 %	50 %	97.5 %
Bolulla	K-S	0.00802	0.13510	0.30084
Bolulla	Multinomial	0.00291	0.02389	0.25189
Resample Bolulla	K-S	0.00830	0.42736	1.00000
Resample Bolulla	Multinomial	0.01072	0.38315	0.95910
Callosa	K-S	0.13718	0.76085	1.00000
Callosa	Multinomial	0.05600	0.52614	0.94570
Resample Callosa	K-S	0.00001	0.03194	0.88452
Resample Callosa	Multinomial	0.00026	0.13913	0.88102

Tabla 11.10: Bondad de ajuste marginal para Bolulla y Callosa. Cota inferior y superior de uniformidad y centro de los p- valores a posteriori (ec. 8.1.5).

	p-val.	pvalcotainf	centro	pvalcotasup
Bolulla	K-S	0	0.10612	0.11307
Bolulla	Multinomial	0	0.03402	0.03824
Callosa	K-S	1	1	1
Callosa	Multinomial	0.55251	0.55756	0.99816

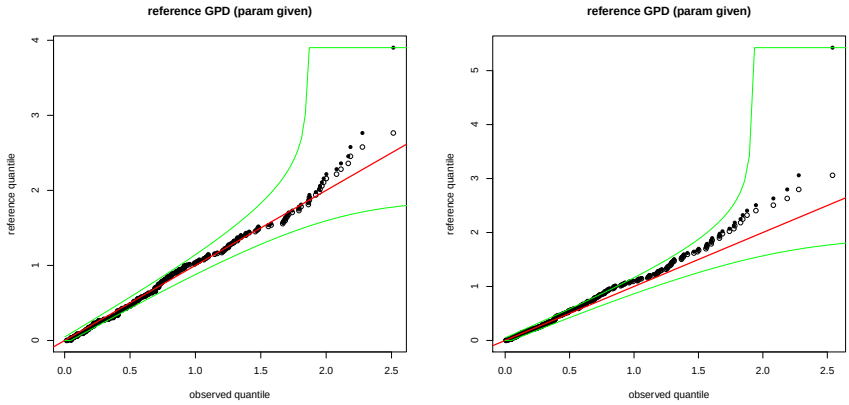

Figura 11.17: QQplot de los excesos de log-precipitación por encima de log(20) respecto a la distribución GPD con parámetros estimados por máxima verosimilitud (estima preliminar). Izquierda: Bolulla; derecha: Callosa.

11.2.4. Valores de interés a posteriori para Bolulla y Callosa.

Una vez descritos los resultados de la estimación de los parámetros del modelo e interpretada su bondad de ajuste a los datos, podemos obtener cantidades predictivas a posteriori de interés, tanto marginales como conjuntas.

Dado un conjunto de valores de referencia de precipitación diaria, se han obtenido sus probabilidades a posteriori y los cuantiles de es-

Tabla 11.11: Pendientes predictivas de la recta de regresión de los excesos esperados para cada una de las marginales. p-valor de discrepancia correspondiente al contraste sobre la validez del priori GPD en DA-Weibull para cada marginal.

Value	X_1	X_2
Slope orig.	-0.219610	-0.192074
Slope predictive	-0.209718	-0.138509
Slope Discrepancy	0.578351	0.798969

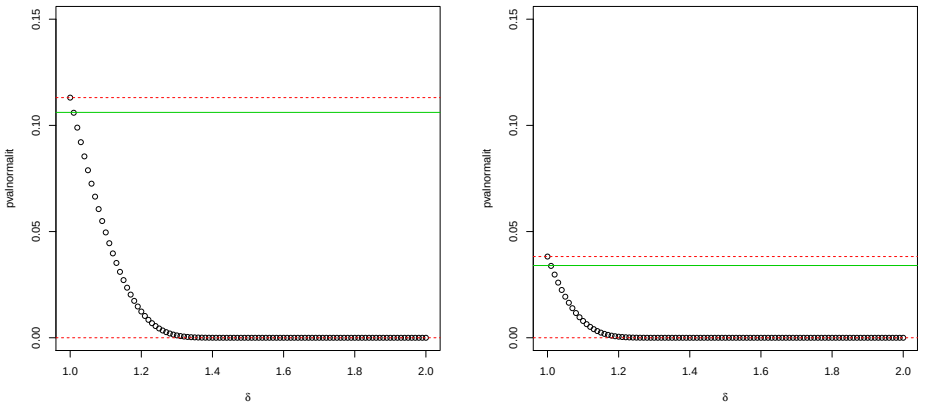

Figura 11.18: Bondad de ajuste marginal para Bolulla. Cálculo del intervalo donde el p-valor a posteriori es uniforme para el p-valor Kolmogorov-Smirnov (izq.) y el Multinomial (dcha.) según el grado de dependencia δ. Los extremos $\delta = 1$ y $\delta = 2$ corresponden respectivamente a una total dependencia e independencia entre las realizaciones de p-valor. Las líneas discontínuas (rojo) muestran estas cotas, y la línea continua (verde), el valor del centro de la muestra de p-valores.

tas probabilidades (Tablas 11.13 y 11.14); sus periodos de retorno a posteriori (Tablas 11.15 y 11.16)y los valores a posteriori asociados a periodos de retorno de referencia (Tabla 11.17 y 11.18).

Además se han calculado probabilidades de no excedencia conjuntas a posteriori de algunos valores de referencia y cuantiles de estas probabilidades (Tabla 11.19). Se observa que el modelo no solo estima correctamente probabilidades de pares de valores observados, sino que también proporciona estimaciones de probabilidades de combinaciones de valores no observados en la serie.

11.2.5. Discusión

Se ha modelizado la serie de log-precipitación registrada en Bolulla y Callosa mediante el modelo propuesto, donde la dependencia se modela mediante una cópula Gumbel y mediante una cópula CrEnC. Se ha contrastado la bondad de ajuste del modelo, tanto marginal como global. Se observa que la dependencia queda bien descrita tanto por la cópula

Tabla 11.12: Contraste sobre el modelo global. Coeficiente τ original y coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. Los coeficientes muestran dependencias similares. El p-valor basado en la discrepancia τ no permite rechazar la hipótesis de validez del modelo global especificado.

Tau original	Tau PredictivoG	Discrepancia Tau
0.598680	0.556579	0.145361

Tabla 11.13: Probabilidades de no excedencia a posteriori para valores seleccionados de los excesos en Bolulla. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1ref	Mean	GeomMean	q0.05	q0.5	q0.95
0.405 (40mm)	0.4091	0.4091	0.3957	0.4063	0.4310
0.916 (50mm)	0.7279	0.7282	0.7076	0.7257	0.7534
1.500 (90mm)	0.9103	0.9111	0.8923	0.9099	0.9286
1.610 (100mm)	0.9299	0.9308	0.9124	0.9296	0.9466
2.020 (150mm)	0.9758	0.9772	0.9621	0.9760	0.9866
2.305 (200mm)	0.9903	0.9919	0.9808	0.9908	0.9969

Gumbel como por la cópula CrEnC, aunque el ajuste de esta última es ligeramente mejor. Se han obtenido valores a posteriori de interés para ambas ubicaciones utilizando el modelo estimado. Adicionalmente, se ha verificado que el conjunto de estadísticos seleccionado mediante el método de razón de verosimilitudes basta para describir la dependencia de la precipitación en estas dos ubicaciones.

Tabla 11.14: Probabilidades de no excedencia a posteriori para valores seleccionados de los excesos en Callosa. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x2ref	Mean	GeomMean	q0.05	q0.5	q0.95
0.405 (40mm)	0.4284	0.4283	0.4029	0.4259	0.4614
0.916 (50mm)	0.7478	0.7483	0.7217	0.7456	0.7800
1.610 (100mm)	0.9388	0.9396	0.9208	0.9390	0.9544
1.500 (90mm)	0.9209	0.9217	0.9017	0.9209	0.9387
2.020 (150mm)	0.9795	0.9806	0.9666	0.9806	0.9887
2.305 (200mm)	0.9919	0.9931	0.9826	0.9929	0.9974

Tabla 11.15: Periodo de retorno a posteriori correspondiente a valores seleccionados de los excesos en Bolulla. Periodo de retorno a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1tau	Mean	GeomMean	q0.05	q0.5	q0.95
0.500 (33mm)	1.9355	1.9350	1.8802	1.9242	2.0269
0.916 (50mm)	3.6857	3.6805	3.4198	3.6456	4.0552
1.500 (90mm)	11.3778	11.2557	9.2814	11.0940	14.0116
1.610 (100mm)	14.6959	14.4599	11.4152	14.2007	18.7225
2.020 (150mm)	47.9334	43.8360	26.3766	41.6172	74.5914
2.302 (200mm)	208.0822	122.0501	51.7904	107.6377	320.6003

Tabla 11.16: Periodo de retorno a posteriori correspondiente a valores seleccionados de los excesos en Callosa. Periodo de retorno a posteriori: media, centro y cuantiles a posteriori seleccionados.

ytau	Mean	GeomMean	q0.05	q0.5	q0.95
0.500 (33mm)	2.0184	2.0168	1.9097	2.0031	2.1623
0.916 (50mm)	3.9866	3.9758	3.5934	3.9307	4.5448
1.500 (90mm)	12.9312	12.7868	10.1734	12.6414	16.3155
1.610 (100mm)	16.8307	16.5761	12.6215	16.4044	21.9295
2.020 (150mm)	54.6840	51.6697	29.9180	51.4842	88.8448
2.302 (200mm)	170.7327	143.3776	57.0790	139.6587	375.7993

Tabla 11.17: Excesos en Bolulla correspondientes a periodos de retorno seleccionados. Excesos a posteriori: media, centro y cuantiles seleccionados.

tiempo	Mean	GeomMean	q0.05	q0.5	q0.95
10.	1.4505	1.4494	1.3579	1.4517	1.5392
20.	1.7524	1.7504	1.6320	1.7505	1.8902
50.	2.0824	2.0790	1.9148	2.0797	2.2843
100.	2.2886	2.2839	2.0865	2.2815	2.5404
400.	2.6130	2.6050	2.3292	2.5991	2.9711
500.	2.6561	2.6476	2.3561	2.6420	3.0271

Tabla 11.18: Excesos en Callosa correspondientes a periodos de retorno seleccionados. Excesos a posteriori: media, centro y cuantiles seleccionados.

tiempo	Mean	GeomMean	q0.05	q0.5	q0.95
10.	1.3935	1.3922	1.2978	1.3947	1.4899
20.	1.6936	1.6916	1.5729	1.6882	1.8347
50.	2.0277	2.0242	1.8685	2.0092	2.2412
100.	2.2406	2.2356	2.0473	2.2134	2.5399
400.	2.5841	2.5746	2.3117	2.5385	2.9998
500.	2.6308	2.6204	2.3482	2.5877	3.0618

Tabla 11.19: Probabilidades de excedencia conjuntas a posteriori para valores seleccionados de los excesos en Bolulla y Callosa. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1refconj	x2refconj	Mean	GeomMean	q0.05	q0.5	q0.95
0.405 (40mm)	0.405 (40mm)	0.4253	0.4235	0.3659	0.4274	0.4888
0.916 (50mm)	0.916 (50mm)	0.1494	0.1468	0.1042	0.1480	0.1983
1.500 (90mm)	1.610 (100mm)	0.0339	0.0309	0.0140	0.0307	0.0587
1.610 (100mm)	1.500 (90mm)	0.0369	0.0339	0.0140	0.0363	0.0587
2.020 (150mm)	2.305 (200mm)	0.0065	0.0050	0.0028	0.0028	0.0140
2.305 (200mm)	2.020 (150mm)	0.0071	0.0055	0.0028	0.0028	0.0195

11.3. Vall de Laguard Fontilles y Almudaina (Alicante)

Se desea analizar la dependencia de la precipitación diaria registrada en dos localizaciones cercanas (Vall de Laguard-Fontilles (38°46'40.38"N, 0° 5'18.03.°) y Almudaina (38°45'38.97"N, 0°21'14.22.°), Fig. 11.19). Se ha modelado la log-precipitación sobre un umbral suficientemente alto (log(20)) en cada ubicación, Fig. 11.20, mediante una distribución generalizada de Pareto (GPD). Se dispone de una muestra de 480 sucesos de lluvia en alguna de las dos ubicaciones, de los cuales, 180 son sucesos conjuntos (Fig. 11.21). Visualmente, los datos presentan una dependencia baja. Los coeficientes de correlación de Pearson y Spearman son respectivamente $\rho_P = 0.3798$ y $\rho_S = 0.1041$, valores que confirman la apreciación visual.

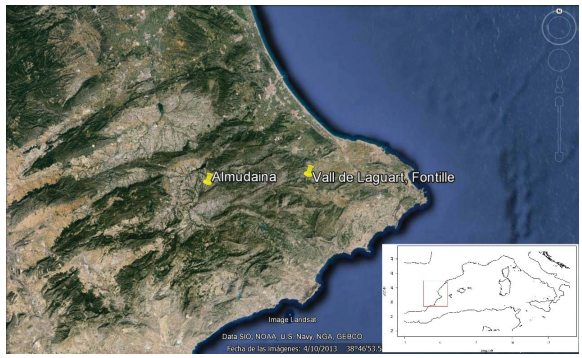

Figura 11.19: Localización de Vall de Laguard Fontilles y Almudaina.

11.3.1. Ocurrencia de los sucesos y parámetros marginales

La información contenida en la muestra del posteriori de los parámetros marginales y de ocurrencia de los sucesos de precipitación se resume en las Figuras 11.22 y 11.23, histogramas de estos parámetros. La Tabla 11.20 muestra percentiles seleccionados de las muestras del posteriori

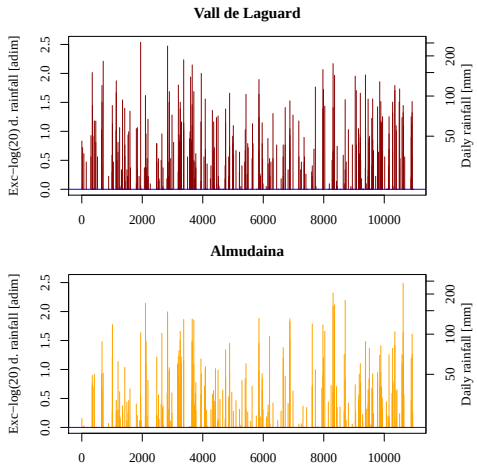

Figura 11.20: Excesos de log-precipitación diaria sobre log(20) en Vall de Laguard Fontilles y Almudaina.

de cada uno de estos parámetros. La mediana predictiva del parámetro λ se halla entorno a los 17 sucesos anuales (ver Tabla 11.20), en concordancia con el análisis realizado en Egozcue and Ramis (2001).

Se observa que el priori establecido para el parámetro β afecta claramente las estimaciones de estos parámetros marginales, dado que los histogramas aparecen cortados en su extremo derecho. En cambio, la hipótesis GPD-Weibull ($\xi < 0$) no resulta restrictiva, dado que los valores de ξ se encuentran claramente separados del cero, con valores centrales entorno al $\xi = -0.3$ (Fig. 11.23). A partir de estas muestras del posteriori de los parámetros se pueden obtener otros valores a posteriori, como las estimaciones de la cota superior de las distribuciones Weibull-GPD (Tabla 11.21). Si traducimos estos valores a la escala usual de precipitación diaria (mm), se observa que los valores de cotas superiores obtenidos (Tabla 11.22) son coherentes con los datos y con la idea de cota superior como límite físico de la magnitud estudiada en estas localizaciones.

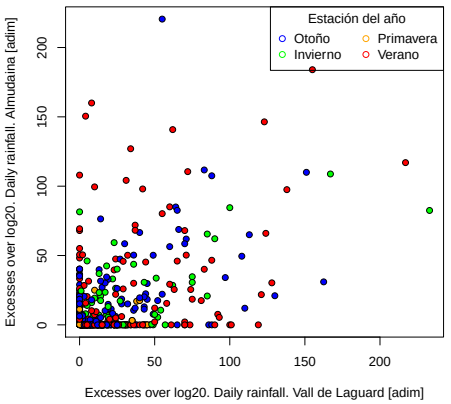

Figura 11.21: Excesos de log-precipitación sobre el umbral seleccionado. Escala $\mathbb{R}^{+2}$. Vall de Laguard y Almudaina.

11.3.2. Dependencia mediante cópula paramétrica Gumbel

La dependencia entre las series de excesos de log-precipitación en Vall de Laguard y Almudaina se modeliza en primer lugar mediante una cópula paramétrica de la familia Gumbel. En la Fig. 11.25 se muestra una representación gráfica de la densidad de la cópula Gumbel estimada, en $[0,1]^2$ mediante contornos de isodensidad a partir de los valores estimados (mediana) del parámetro δ de la cópula (Fig. 11.24 y

Tabla 11.20: Percentiles de la muestra del posteriori para los parámetros del modelo. Vall de Laguard y Almudaina.

	2.5 %	25 %	50 %	75 %	97.5 %
ξ_{X1}	-0.31	-0.28	-0.26	-0.24	-0.19
β_{X1}	0.77	0.83	0.85	0.85	0.86
ξ_{X2}	-0.31	-0.27	-0.25	-0.21	-0.14
β_{X2}	0.72	0.78	0.81	0.84	0.85
λ	14.49	16.12	17.10	18.11	20.02

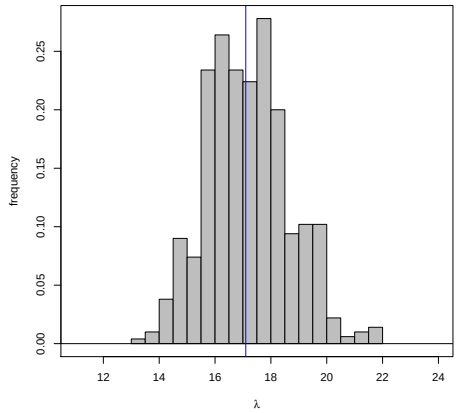

Figura 11.22: Histograma de la muestra del posteriori para el parámetro λ de Poisson (tasa de ocurrencia). Línea azul: mediana de la muestra. Vall de Laguard y Almudaina.

Tabla 11.23). La Fig. 11.26 muestra los correspondientes contornos en $\mathbb{R}^2$. Los valores estimados corresponden a una dependencia moderada, en correspondencia con los valores de ρ_P de Pearson i ρ_S de Spearman presentados al inicio del ejemplo.

Tabla 11.21: Percentiles de las cotas superiores de la distribución *GPD* (dominio de Weibull) de los excesos de log-precipitación sobre el umbral seleccionado, correspondientes los parámetros estimados del modelo (Tabla 11.20).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior Vall de Laguard	2.54	3.02	3.23	3.49	4.48
cota superior Almudaina	2.72	3.04	3.30	3.70	5.29

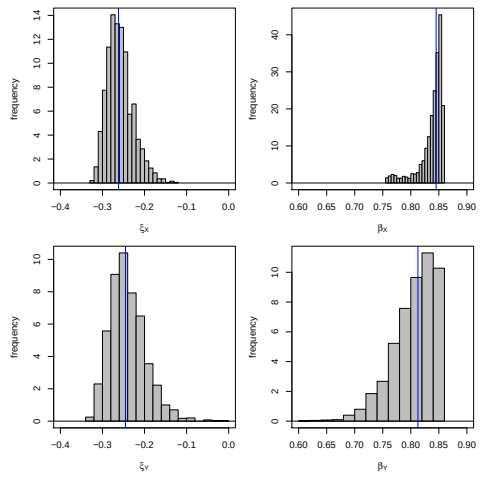

Figura 11.23: Histograma de la muestra del posteriori para los parámetros marginales ξ, β. La línea azul marca la mediana de la muestra del posteriori para cada parámetro. Vall de Laguard y Almudaina.

Bondad de ajuste. Dependencia mediante cópula paramétrica Gumbel

Se valora la coherencia del ajuste de la dependencia entre la log-precipitación en Vall de Laguard y Almudaina mediante una cópula Gumbel mediante los estadísticos de contraste presentados en la Sección 8.1. En la Tabla 11.24 se muestra el contraste respecto al ajuste global del modelo. La dependencia presente en los excesos originales es inferior

Tabla 11.22: Cota superior de precipitación (mm) en Vall de Laguard y Almudaina. Percentiles correspondientes a la cota superior de los excesos sobre umbral (Tabla 11.21).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior Vall de Laguard (mm)	252.70	409.38	506.58	654.58	1759.15
cota superior Almudaina (mm)	304.01	419.03	544.87	809.85	3948.40

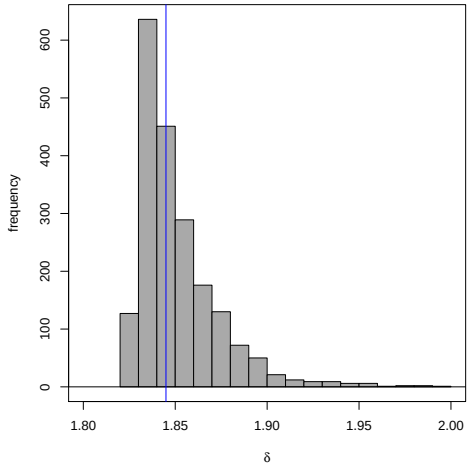

Figura 11.24: Histograma de la muestra del posteriori para el parámetro δ de la cópula Gumbel. La línea azul marca la mediana de la muestra del posteriori para el parámetro.

Tabla 11.23: Percentiles de la muestra del posteriori para el parámetro de la cópula Gumbel

	2.5 %	25 %	50 %	75 %	97.5 %
δ	1.83	1.84	1.84	1.86	1.91

a la presente en las remuestras generadas a partir del posteriori, aspecto que se confirma observando los valores a posteriori del parámetro δ (Tabla 11.25). Por tanto, el modelo con dependencia Gumbel exagera la dependencia global presente en los datos.

11.3.3. Dependencia mediante cópula CrEnC

La dependencia entre las series de excesos de log-precipitación en Vall de Laguard y Almudaina ha sido modelizada también mediante una cópula CrEnC. Se ha aplicado el método de razón de verosimilitudes de selección de momentos. Ha resultado seleccionado un único momento, el correspondiente al coeficiente α_6 (Spearman's footrule φ). La Fig.

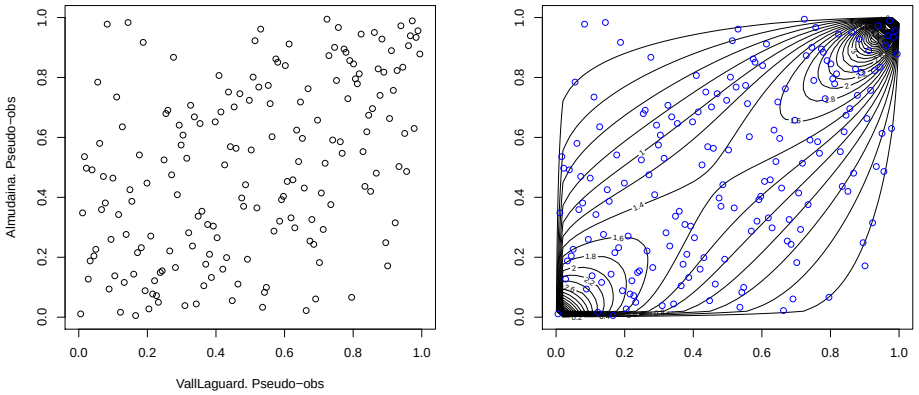

Figura 11.25: Pseudo-observaciones de la muestra (Vall de Laguard y Almudaina). Cópula Gumbel en $[0,1]^2$. Contornos de isodensidad a partir de la muestra del posteriori del parámetro (Fig. 11.24).

11.27 presenta el histograma de la muestra del posteriori del parámetro y la Tabla 11.26 muestra percentiles seleccionados de esta muestra. Se aprecia una pequeña asimetría en valores pequeños, pero la mayoría de los valores se encuentran entorno a 0.8, indicando una dependencia moderada de las series.

En la Fig. 11.28 se muestra una representación gráfica de la densidad de la cópula CrEnC en $\mathbb{R}^2$ mediante los contornos de isodensidad correspondientes a la mediana de la muestra del posteriori del parámetro seleccionado.

Tabla 11.24: Contraste sobre el modelo global. Coeficiente τ original, coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. El coeficiente predictivo muestra dependencia superior a la original. El p-valor basado en la discrepancia τ sugiere rechazar la hipótesis de validez del modelo de dependencia Gumbel especificado.

τ original	τ PredictivoG	Discrepancia τ
0.29758	0.45835	0.999002

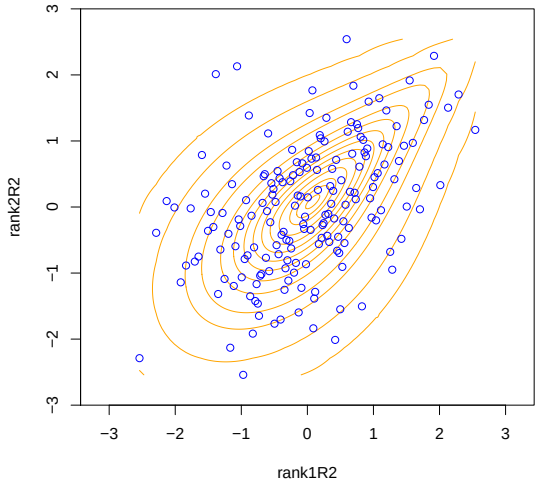

Figura 11.26: Ajuste de la Cópula Gumbel para Vall de Laguard y Almudaina en $\mathbb{R}^2$. Contornos de isodensidad a partir de la muestra del posteriori del parámetro.

Los resultados a posteriori obtenidos a partir de estas estimaciones se presentan en el Annexo B: Bondad de ajuste del modelo (Sección B.1) y valores a posteriori (Sección B.2).

11.3.4. Discusión

Se ha ajustado el modelo propuesto a los datos de log-precipitación en Vall de Laguard y Almudaina. Se ha modelizado la dependencia

Tabla 11.25: Contraste sobre el modelo global. Coeficiente δ original, coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. El coeficiente predictivo muestra una dependencia superior a la original.

δ original	δ PredictivoG
1.42365	1.85759

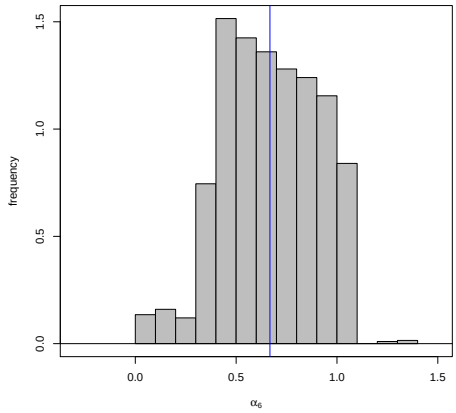

Figura 11.27: Histograma de la muestra del posteriori para el parámetro de la cópula CrEnc seleccionado (α_6). La línea azul marca la mediana de la muestra del posteriori para cada parámetro.

Tabla 11.26: Percentiles de la muestra del posteriori para el parámetro seleccionado de la cópula CrEnC.

	2.5 %	25 %	50 %	75 %	97.5 %
α_6	0.18	0.49	0.67	0.86	1.04

entre las series utilizando una cópula Gumbel y una cópula CrEnC. Se observa que al ajustar la dependencia mediante una cópula Gumbel se obtienen niveles de dependencia superiores a los originales, mientras que la cópula CrEnC proporciona un dependencia similar a la de los datos originales. A partir del modelo estimado se han obtenido valores a posteriori de interés, tanto marginales como conjuntos.

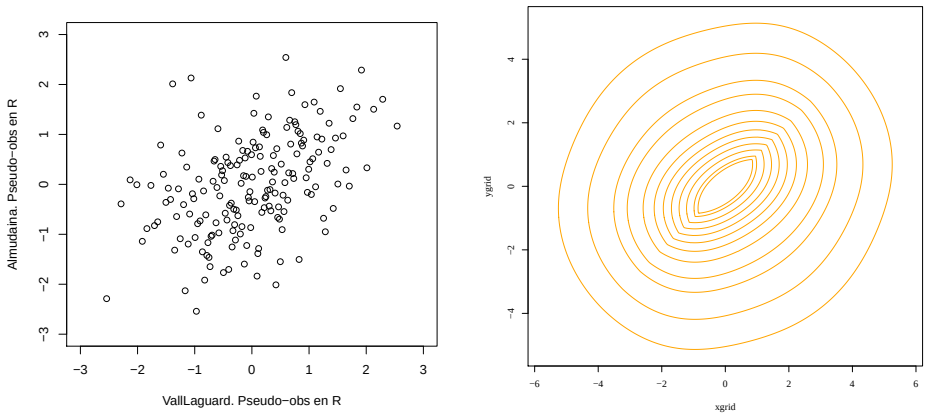

Figura 11.28: Pseudo-observaciones en $\mathbb{R}^2$ (Izq.). Contornos de isodensidad de la Cópula CrEnC en $\mathbb{R}^2$ a partir de la muestra del posteriori de los parámetros (Fig. 11.27)(Dcha.).

11.4. Vergel de Recons y Simat de Valldigna (Alicante)

Se desea analizar la dependencia de la precipitación diaria registrada en dos localizaciones cercanas (Vergel de Recons (38°50'26.97"N, 0° 0'36.15."E) y Simat de Valldigna (39° 2'26.77"N, 0°18'35.27."O), Fig. 11.29. Para cada localización se ha modelado la log-precipitación sobre un umbral de log(20), Fig. 11.30, mediante una distribución generalizada de Pareto (GPD). Se dispone de una muestra de 532 sucesos de lluvia (log-precipitación superior a log(20)) en alguna de las dos ubicaciones, de los cuales, 146 son sucesos conjuntos (Fig. 11.31). Visualmente se aprecia poca dependencia entre ambas series. Los coeficientes de correlación de Pearson y Spearman indican también poca dependencia entre ambas series ($\rho_P = 0.1397$ y $\rho_S = -0.1444$ respectivamente).

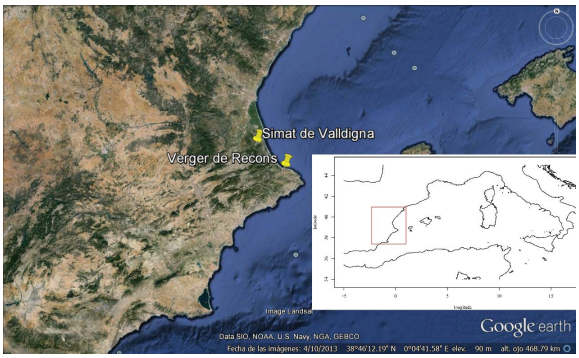

Figura 11.29: Localización de Vergel de Recons y Simat de Valldigna.

11.4.1. Ocurrencia de los sucesos y parámetros marginales

Se considera que los sucesos de precipitación se pueden modelizar según un proceso de Poisson evaluado, con tamaños GPD. Las Figuras 11.32 y 11.33 muestran los histogramas de la muestra del posteriori de los parámetros de ocurrencia y marginales del modelo y la Tabla 11.27 muestra percentiles seleccionados de las muestras del posteriori para cada uno de los parámetros. El percentil del 50 % a posteriori para el parámetro λ se halla entorno a los 14 sucesos anuales (ver Tabla 11.27), en concordancia con el análisis realizado en Egozcue and Ramis (2001).

La hipótesis a priori de modelo marginal GPD-Weibull, ($\xi < 0$) aparentemente no es restrictiva. Los valores de ξ para ambas marginales están claramente separados de cero, Fig. 11.33, por lo que la hipótesis es coherente con los datos. En cambio, el recinto del priori establecido para β sí limita los valores mayores del parámetro, lo que visualmente corresponde a un corte en el histograma. A partir de estas estimaciones se pueden obtener otros valores a posteriori, com la cota superior de las distribuciones Weibull-GPD (Tabla 11.28). Si se traducen estos valores a la escala usual de precipitación (mm) (Tabla 11.29), se observa que los valores de cotas superiores estimadas son coherentes la idea de límite físico de la precipitación diaria estudiada en estas ubicaciones.

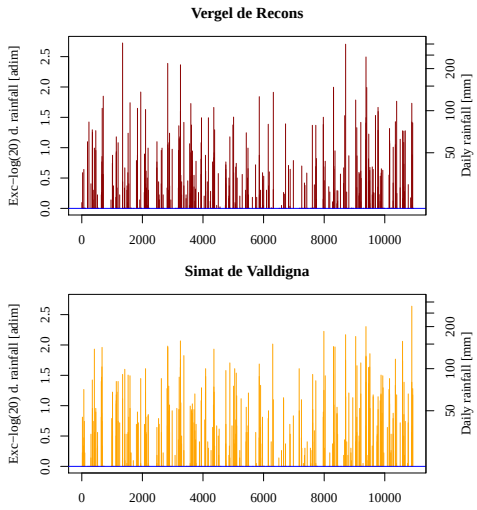

Figura 11.30: Excesos de log-precipitación diaria sobre log(20) en Vergel de Recons y Simat de Valldigna.

11.4.2. Dependencia mediante cópula paramétrica Gumbel

La dependencia entre las dos series de excesos de log-precipitación en Vergel y Simat se ha modelizado en primer lugar utilizando una familia paramétrica de cópula, la familia de Gumbel. La dependencia entre ambas series es baja, por lo que se ha considerado adecuado realizar un contraste de independencia antes de ajustar la cópula. Para optimizar recursos se ha utilizado el contraste implementado en el paquete *copula* del software *R* (Hofert et al., 2014; Yan, J., 2007). Pese a que la dependencia es baja, la hipótesis primaria de independencia es rechazada (Global Cramer-von Mises statistic: 0.2253 con p-valor 0.000499).

Una vez rechazada la independencia entre ambas series de excesos, se ha estimado la distribución a posteriori del modelo conjunto de marginales, ocurrencia y dependencia. En la Tabla 11.30 y la Fig. 11.34 se muestran el histograma y algunos descriptivos de la muestra del posteriori del parámetro δ de la cópula de Gumbel.

En la Fig. 11.35 se muestra una representación gráfica de la densidad

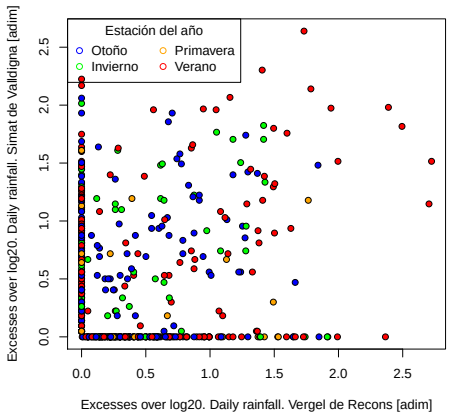

Figura 11.31: Excesos de log-precipitación sobre log(20). Escala $\mathbb{R}^{+2}$. Vergel de Recons y Simat de Valldigna.

de la cópula Gumbel mediante contornos de isodensidad en $[0,1]^2$ correspondientes a la mediana de la muestra del posteriori del parámetro. El ajuste en $\mathbb{R}^2$ se muestra en la Fig. 11.36.

Tabla 11.27: Percentiles de la muestra del posteriori para los parámetros del modelo. Vergel de Recons y Simat de Valldigna.

	2.5 %	25 %	50 %	75 %	97.5 %
ξ_{X1}	-0.28	-0.25	-0.23	-0.21	-0.15
β_{X1}	0.75	0.80	0.82	0.84	0.86
ξ_{X2}	-0.30	-0.28	-0.27	-0.25	-0.20
β_{X2}	0.82	0.84	0.85	0.85	0.86
λ	11.39	12.93	13.78	14.87	16.73

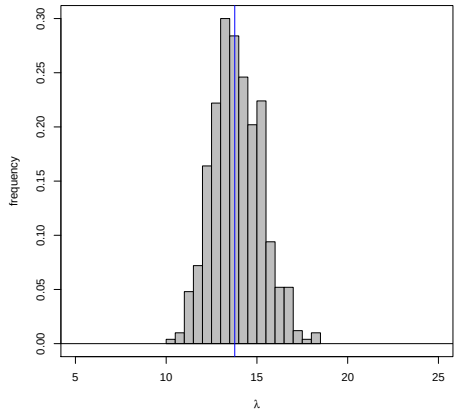

Figura 11.32: Histograma de la muestra del posteriori para el parámetro λ de Poisson (tasa de ocurrencia). Línea azul: mediana de la muestra. Vergel de Recons y Simat de Valldigna.

Bondad de ajuste. Dependencia mediante cópula paramétrica Gumbel

Los estadísticos de contraste presentados en la Sección 8.1 se utilizan para valorar la coherencia de la representación de dependencia mediante cópula de Gumbel con los datos de log-precipitación en Vergel de Recons y Simat de Valldigna. En la Tabla 11.31 se muestra el contraste respecto al ajuste global del modelo. La dependencia presente en los excesos originales es superior a la presente en las remuestras generadas a partir

Tabla 11.28: Percentiles de las cotas superiores de la distribución *GPD* (dominio de Weibull) de los excesos de log-precipitación sobre el umbral seleccionado, correspondientes los parámetros estimados del modelo (Tabla 11.27).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior Vergel	2.72	3.33	3.59	3.94	5.15
cota superior Simat	2.79	3.00	3.18	3.41	4.11

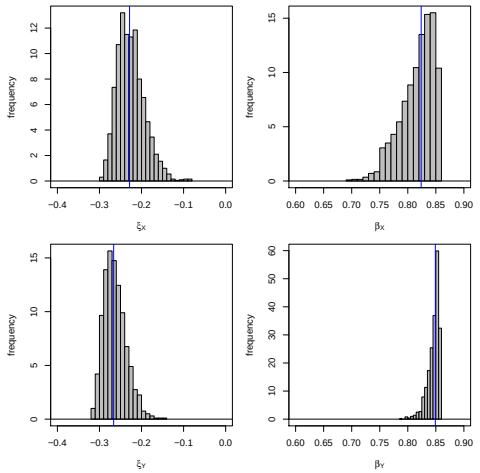

Figura 11.33: Histograma de la muestra del posteriori para los parámetros marginales ξ, β. La línea azul marca la mediana de la muestra del posteriori para cada parámetro. Vergel de Recons y Simat de Valldigna.

Tabla 11.29: Percentiles de las cotas superiores estimados precipitación correspondientes a la cota superior de los excesos sobre umbral (Tabla 11.28).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior Vergel (mm)	304.83	560.79	726.73	1025.51	3440.44
cota superior Simat (mm)	325.22	403.07	480.62	607.19	1223.25

del posteriori, aspecto que se confirma observando los valores predictivos a posteriori del parámetro δ (Tabla 11.32). Por tanto, el modelo con dependencia Gumbel subestima la dependencia global presente en los datos.

11.4.3. Dependencia mediante cópula CrEnC

Dada la baja dependencia observada entre las series, en la Sección 11.4.2 se ha realizado un contraste de independencia previo al ajuste de cópula, en el cual se ha rechazado la hipótesis primaria de independencia. Una vez descartada, se ha procedido a estimar la dependencia entre

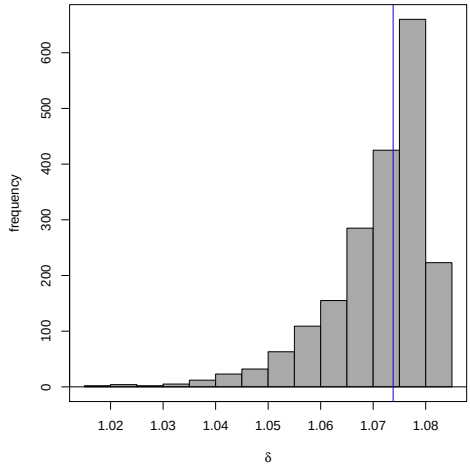

Figura 11.34: Histograma de la muestra del posteriori para el parámetro δ de la cópula de Gumbel. La línea azul marca la mediana de la muestra del posteriori para cada parámetro.

Tabla 11.30: Percentiles de la muestra del posteriori para el parámetro de la cópula de Gumbel.

	2.5 %	25 %	50 %	75 %	97.5 %
δ	1.00	1.00	1.00	1.00	1.00

las series de log-precipitación en Vergel y Simat mediante una cópula CrEnC. En el proceso de razón de verosimilitudes se ha seleccionado un solo momento, el correspondiente al coeficiente de Spearman φ, α_6. La Tabla 11.33 y la Fig. 11.37 muestran un resumen de la muestra del posteriori del parámetro, con valores que indican una baja dependencia entre las series.

En la Fig. 11.38 se muestra una representación gráfica de la densidad de la cópula CrEnC en $\mathbb{R}^2$ mediante contornos de isodensidad, correspondiente a la mediana de la muestra a posteriori del parámetro. Visualmente los contornos parecen ajustar correctamente las pseudo-muestras (en escala $\mathbb{R}^2$), pero deben interpretarse con cautela: por un

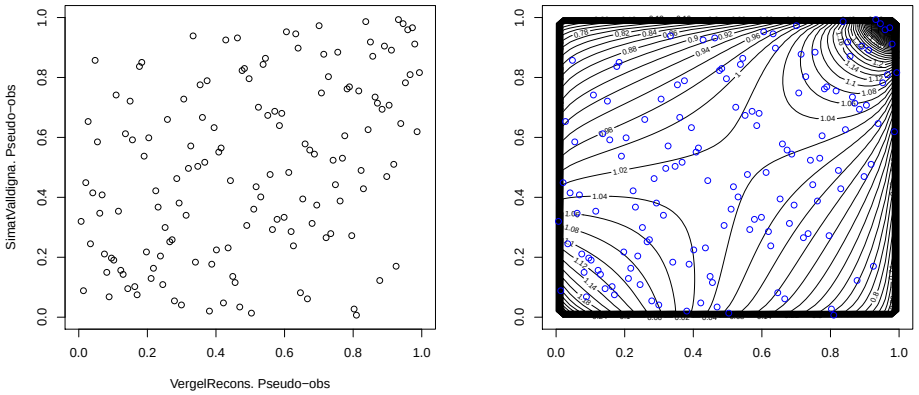

Figura 11.35: Ajuste de la Cópula Gumbel en Vergel de Recons y Simat en $[0,1]^2$. Pseudo-observaciones (Izq.). Contornos de isodensidad a partir de la muestra del posteriori de los parámetros (Fig. 11.34) (Dcha.).

lado, la apreciación de este ajuste visual es difícil; por otro, se dispone de una amplia muestra del posteriori y limitar la interpretación a un solo ajuste visual parece inapropiado. Por tanto, la interpretación debe complementarse con contrastes de bondad de ajuste a posteriori como los propuestos.

Los resultados a posteriori obtenidos a partir de estas estimaciones se presentan en el Annexo B: Bondad de ajuste del modelo (Sección B.3) y valores a posteriori (Sección B.4).

Tabla 11.31: Contraste sobre el modelo global. Coeficiente τ original y coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. El coeficiente predictivo muestra dependencia inferior a la original. El p-valor basado en la discrepancia τ sugiere rechazar la hipótesis de validez del modelo de dependencia Gumbel especificado.

τ original	τ PredictivoG	Discrepancia τ
0.24308	0.07035	0.00299

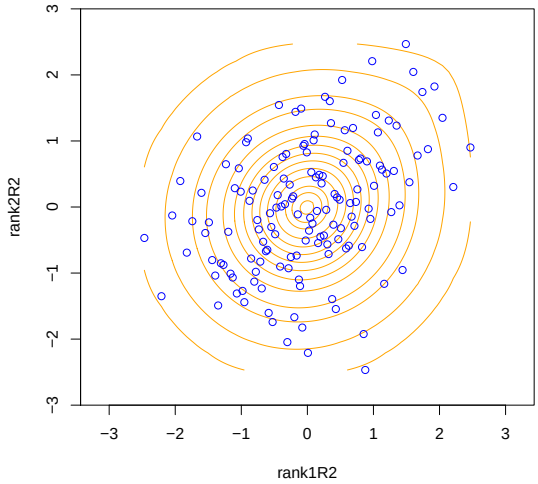

Figura 11.36: Ajuste de la Cópula Gumbel en Vergel de Recons y Simat de Valldigna en $\mathbb{R}^2$. Contornos de isodensidad a partir de la muestra del posteriori de los parámetros (Tabla 11.30).

11.4.4. Discusión

Se ha ajustado el modelo propuesto, cuya dependencia se representa mediante una cópula Gumbel y una cópula CrEnC, a una serie de log-precipitación registrada en Vergel de Recons y Simat de Valldigna. Ambas series presentan poca dependencia, por lo que previamente se ha realizado un contraste de independencia. La hipótesis primaria de independencia se rechaza, por lo que se procede al ajuste usual de las cópulas y la estimación de sus parámetros. El ajuste mediante cópula Gumbel subestima la dependencia, mientras que la cópula CrEnC proporciona un mejor ajuste. Se han obtenido valores a posteriori de interés a partir del modelo estimado.

Tabla 11.32: Contraste sobre el modelo global. Coeficiente δ original y coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. El coeficiente predictivo muestra una dependencia inferior a la original.

δ original	δ PredictivoG
1.32114	1.08026

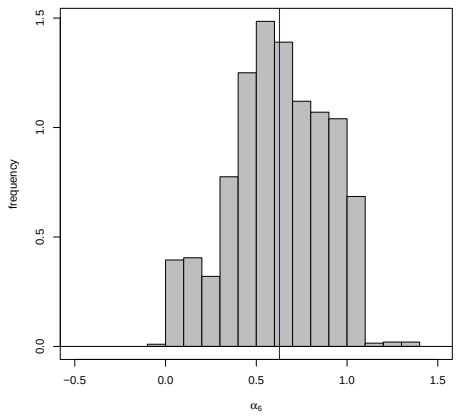

Figura 11.37: Histograma de la muestra del posteriori para el parámetro de la cópula seleccionado, α_6. La línea azul marca la mediana de la muestra del posteriori para el parámetro.

Tabla 11.33: Percentiles de la muestra del posteriori para el parámetro seleccionado de la cópula CrEnC

	2.5 %	25 %	50 %	75 %	97.5 %
α_6	0.07	0.45	0.63	0.84	1.04

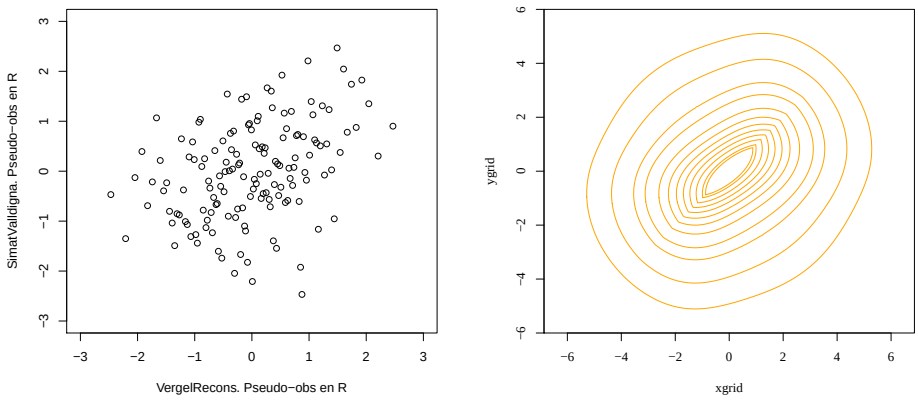

Figura 11.38: Cópula CrEnC de Vergel de Recons y Simat de Valldigna en $\mathbb{R}^2$. Pseudo-observaciones (Izq.). Contornos de isodensidad a partir de la muestra del posteriori de los parámetros (Dcha.).

Capítulo 12

Estudio de un registro de altura de ola (HIPOCAS-Boya)

El modelo propuesto en el Capítulo 3 y sucesivos se ha aplicado a diversos conjuntos de datos de precipitación diaria, tanto simulados (Capítulo 10) como registrados (Capítulo 11). En este Capítulo el modelo se aplica a un nuevo conjunto de datos observados, correspondientes a un registro de altura de ola significante. A continuación se describen los datos, se les aplica el modelo, valorando su bondad de ajuste y se obtienen algunos valores a posteriori de interés.

12.1. Datos

Se desea comparar las propiedades estadísticas de una serie de observaciones y un modelo de hindcast. Se dispone de una serie de altura de ola significante (H_s, promedio del tercio superior de las alturas de ola, proporcional al total de energía de la ola), combinando una serie de altura de ola de hindcast (proporcionado por el proyecto HIPOCAS) y datos de altura de ola medidos en una boya.

La variable altura de ola significante tiene una escala relativa (Egozcue et al., 2006; Tolosana-Delgado et al., 2010). Teniendo en cuenta esta escala relativa, se modelizará el logaritmo de la variable (log-altura

de ola significante). El proyecto HIPOCAS (Sotillo et al., 2005; Guedes Soares et al., 2002) utilizó campos de viento diario promedio de un modelo REMO (Jacob and Podzun, 1997) combinado con un sistema de generación de ola WAM (WAMDIGroup, 1988) para generar campos de altura de ola durante el periodo 1958-2001. Se considera la serie de altura significante de ola en el nodo HIPOCAS 2056046 (longitud 40.75 N, latitud 1.00 W), que denominaremos serie h, en el periodo 03/01/1958/01/03 a 31/12/2001. Se considera también la serie (serie b) de las medidas de altura significante de ola en la boya de Tortosa (red XIOM, longitud 40.72 N, latitud 0.98 W) entre el 16/06/1990 y el 31/12/2008 (Fig. 12.1 y Fig. 12.2).

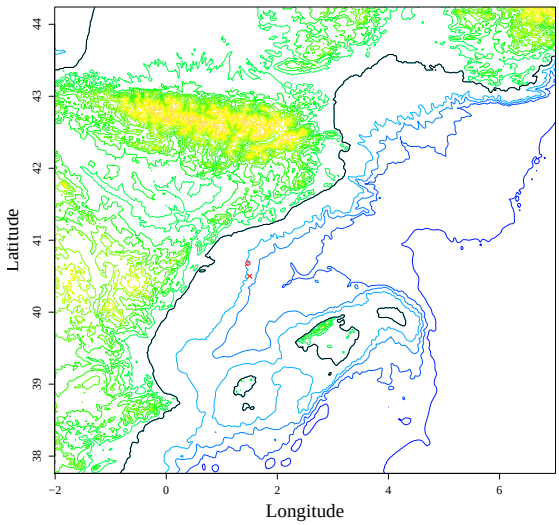

Figura 12.1: Ubicación del nodo HIPOCAS (cruz) y de la boya (círculo).

Cada serie ha sido modelada con un enfoque Peak-Over-Threshold, donde los sucesos de tormenta se definen como un periodo de más de 6 horas con H_s sobre un umbral de 2m, con un tiempo de 3 días como mínimo entre sucesos consecutivos. La ocurrencia de estos sucesos se modela mediante un proceso de Poisson evaluado. A cada suceso se le asigna la log-altura significante de ola máxima del suceso, que se modela mediante una distribución Generalizada de Pareto (GPD).

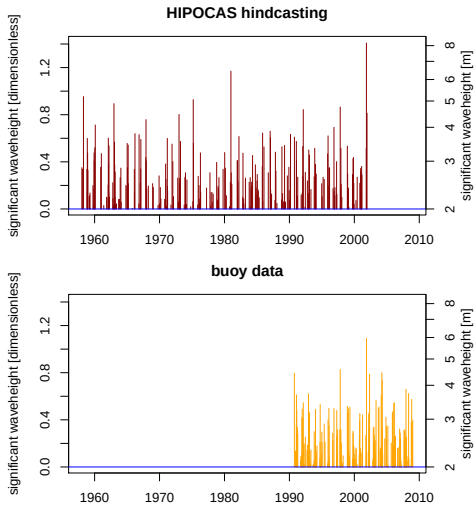

Figura 12.2: Altura significante de ola: HIPOCAS y boya en Tortosa

Se desea estudiar el comportamiento conjunto de la log-altura de ola significante registrada mediante boya y la procedente de hindcasting. En el intervalo de tiempo donde ambas series se superponen, sólo 36 eventos coinciden. Las magnitudes de esos 36 sucesos muestran visualmente una dependencia débil entre ellos (Fig. 12.3), con correlaciones de Pearson $\rho_P = 0.5041$ y de Spearman $\rho_S = 0.3896$ respectivamente.

La dependencia entre las magnitudes marginales se ha modelizado usando dos herramientas diferentes: una cópula de Gumbel y una cópula CrEnC. Se ha obtenido una muestra del posteriori de los parámetros del modelo (de ocurrencia, marginales y de cópula) mediante 10000 iteraciones de Gibbs. El 50 % de las iteraciones se ha descartado (*Burn In*) (Gelman et al., 1995) y, para cada uno de los parámetros, se ha verificado la convergencia de la cadena utilizando el criterio de Gelman (Gelman et al., 1995). La muestra del posteriori de los parámetros obtenida permite representar la incertidumbre en la estimación de los parámetros del modelo, que en este caso es elevada debido al reducido número de datos. Los resultados de la estimación se muestran a continuación.

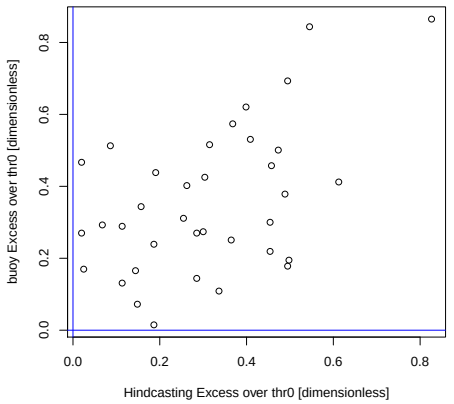

Figura 12.3: Excesos conjuntos de log-altura significante de ola sobre log(2): HIPOCAS y boya en Tortosa.

12.1.1. Selección del umbral GPD

Se dispone de un registro de log-altura de ola en series de boya e HIPOCAS. Se ha establecido el umbral log(2) y se han obtenido los excesos de log-altura de ola sobre este umbral absoluto. Se desea determinar si este umbral es suficientemente alto para que la distribución marginal sea GPD.

La Figura 12.4 muestra las gráficas del exceso medio (*Mean excess plot*) para ambas series. Esta gráfica de diagnóstico permite detectar visualmente el umbral h_0 a partir del cual se puede considerar que la distribución de los excesos sobre h_0 corresponde a una distribución $GPD(\xi, \beta)$ (Castillo, 1988). En ambas gráficas se observa que el umbral h_0 es suficientemente alto, y por tanto, se puede considerar que los excesos de log-altura de ola sobre log(2) pueden considerarse distribuidos GPD. Las Figuras 12.5 y 12.6 muestran diagnósticos de bondad de ajuste marginal del modelo a los datos (QQplot de los datos respecto la distribución de referencia y ajuste de la densidad a los datos). Los gráficos de diagnóstico 12.4, 12.5 y 12.6 han sido elaborados utilizando el paquete Ismev de R (Heffernan and Stephenson, 2012).

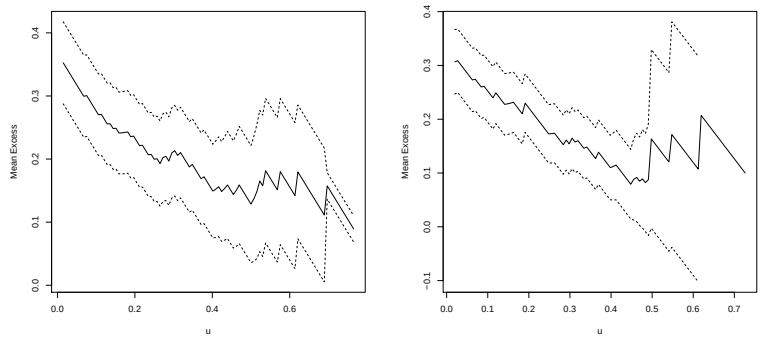

Figura 12.4: *Mean excess plot* de los excesos de log-altura de ola por encima de log(2). Izquierda: HIPOCAS; derecha: Boya. Diagnóstico mediante paquete Ismev de R.

12.1.2. Priori de los parámetros marginales del modelo

Se ha considerado que las distribuciones GPD de los excesos de log-altura de ola sobre el umbral $h_0 = \log(2)$ para las dos series (HIPOCAS y boya) son similares y por tanto se ha establecido un mismo priori conjunto para los parámetros marginales de las dos series, que ha sido determinado siguiendo el proceso descrito en la Sección 6.1. El priori resultante se muestra en la Fig. 12.7.

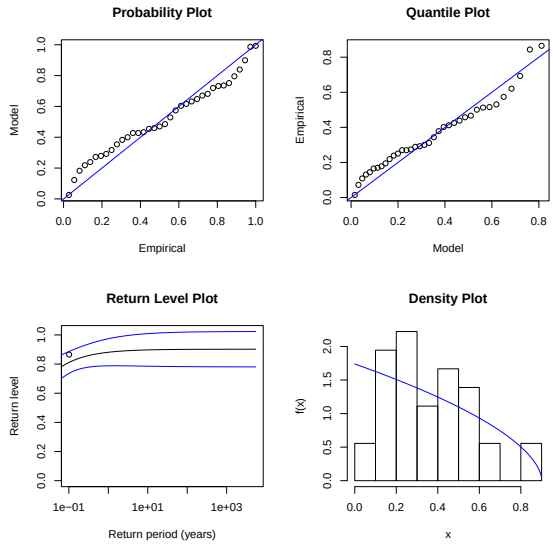

Figura 12.5: Bondad de ajuste a la distribución $GPD(\xi, \beta)$ de los excesos de log-altura de ola por encima de log(2) de HIPOCAS. Diagnóstico mediante paquete Ismev de R.

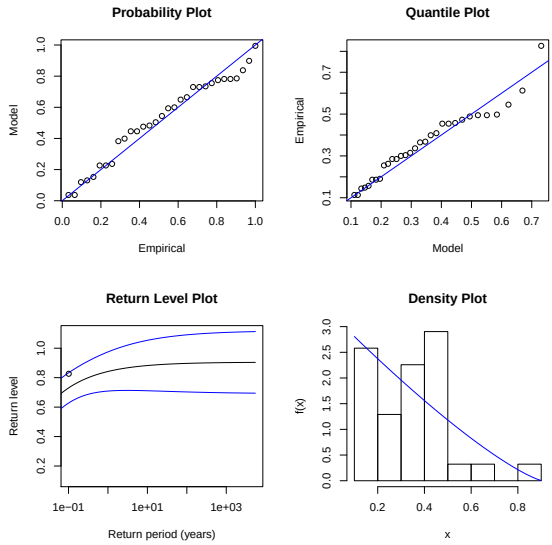

Figura 12.6: Bondad de ajuste a la distribución $GPD(\xi, \beta)$ de los excesos de log-altura de ola por encima de log(2) de Boya. Diagnóstico mediante paquete Ismev de R.

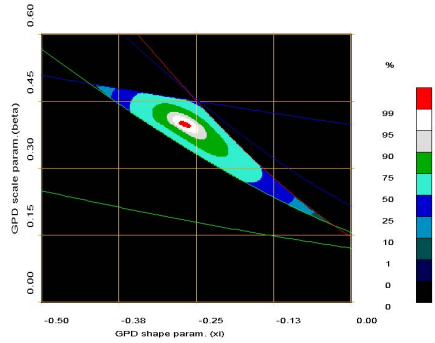

Figura 12.7: Priori conjunto de los parámetros de la distribución $GPD(\xi, \beta)$ para HIPOCAS y Boya.

12.2. Ocurrencia de los sucesos y parámetros marginales

Las magnitudes marginales (excesos sobre log(2) de log-altura de ola significante) se han modelizado mediante una distribución Generalizada de Pareto, $GPD(\xi_i, \beta_i), i = 1, 2$, y la ocurrencia de los sucesos mediante un proceso de Poisson homogéneo de parámetro λ. La información proporcionada por la muestra del posteriori de estos parámetros se ha resumido en las Figuras 12.8 y 12.9, histogramas de estos parámetros de ocurrencia y marginales del modelo. La mediana a posteriori del parámetro λ para los sucesos conjuntos se halla entorno a los tres sucesos anuales (ver Tabla 12.1), en concordancia con lo observado en otras ubicaciones cercanas de características similares.

Se considera que la log-altura de ola es una magnitud limitada, y por ello se ha establecido un priori que impone el dominio de Weibull ($\xi < 0$) para ambas marginales GPD (ver Fig. 12.7). La distribución GPD en ese dominio presenta una cota superior. La Tabla 12.2 muestra los percentiles a posteriori de la cota superior de la distribución de los excesos para la muestra de parámetros y la Tabla 12.3 muestra los percentiles a posteriori para la cota superior de la altura de ola en la escala usual (m). La cota superior de la distribución corresponde a un límite físico de la magnitud, elevado. Los percentiles superiores de la cota superior mostrados en la Tabla 12.3 presentan valores elevados. Estos valores son debidos tanto a la incertidumbre en las estimaciones (muy elevada, debido al tamaño reducido de la muestra conjunta) como a la proporción de muestras de ξ con valores cercanos a cero, Fig. 12.9, que corresponden a distribuciones Weibull con cotas superiores más elevadas, por su cercanía con el dominio Gumbel, de soporte ilimitado.

Tabla 12.1: Percentiles de la muestra del posteriori para los parámetros del modelo.

	2.5 %	25 %	50 %	75 %	97.5 %
ξ_{X1}	-0.26	-0.20	-0.16	-0.11	-0.02
β_{X1}	0.29	0.35	0.39	0.43	0.53
ξ_{X2}	-0.24	-0.18	-0.14	-0.10	-0.02
β_{X2}	0.25	0.31	0.34	0.38	0.47
λ	2.21	2.82	3.18	3.57	4.36

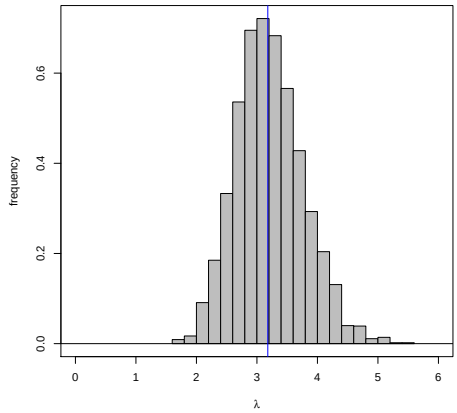

Figura 12.8: Histograma de la muestra del posteriori para el parámetro λ de Poisson (tasa de ocurrencia). Línea azul: mediana de la muestra.

Tabla 12.2: Percentiles de las cotas superiores de la distribución GPD (dominio de Weibull) de los excesos de log-altura significante de ola sobre el umbral seleccionado, correspondientes los parámetros estimados del modelo (Tabla 12.1).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior HIPOCAS	1.46	1.98	2.48	3.49	18.83
cota superior Boya	1.38	1.91	2.43	3.59	19.12

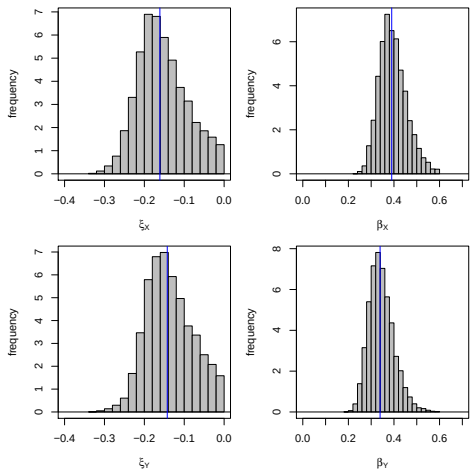

Figura 12.9: Histograma de la muestra del posteriori para los parámetros marginales ξ, β. La línea azul marca la mediana de la muestra del posteriori para cada parámetro.

Tabla 12.3: Percentiles de las cotas superiores estimados de altura significante de ola correspondientes a la cota superior de los excesos sobre umbral (Tabla 12.2).

	2.5 %	25 %	50 %	75 %	97.5 %
cota superior HIPOCAS (m)	8.58	14.45	23.83	65.55	302362808.00
cota superior Boya (m)	7.92	13.55	22.75	72.44	402675697.49

12.3. Dependencia mediante cópula Gumbel

La dependencia conjunta de los excesos de log-altura sobre log(2) registrados mediante boya e HIPOCAS se ha modelizado mediante una cópula paramétrica de la familia Gumbel (Sec. 2.1). Se trata de una cópula arquimediana, para las que existe una conexión entre el parámetro δ de la cópula y el coeficiente de correlación de Spearman ρ_S. La Figura 12.10 y la Tabla 12.4 muestran el resumen de la muestra del posteriori para el parámetro δ. Existe mucha incertidumbre en las estimas, con valores centrales entorno a $\delta = 1.6$, valor que corresponde a una dependencia baja.

En la Fig. 12.11 se muestra una representación gráfica de la densi-

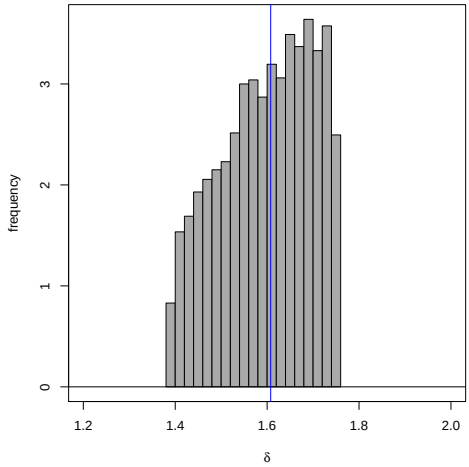

Figura 12.10: Histograma de la muestra del posteriori para el parámetro δ de la cópula de Gumbel. La línea azul marca la mediana de la muestra del posteriori para el parámetro.

Tabla 12.4: Percentiles de la muestra del posteriori para el parámetro de la cópula de Gumbel.

	2.5 %	25 %	50 %	75 %	97.5 %
δ	1.41	1.52	1.61	1.68	1.75

dad de la cópula Gumbel mediante contornos de isodensidad en $[0,1]^2$, para la mediana del valor δ estimado (Tabla 12.4). La representación de la densidad en $\mathbb{R}^2$, Fig. 12.12, facilita la interpretación del ajuste en la escala adecuada. El ajuste visual a las pseudo-observaciones correspondientes a la muestra es bueno, aunque debe tomarse con cautela: dado que se dispone de toda una muestra del posteriori, basar la interpretación en una sola representación gráfica (mediana) resulta simplista.

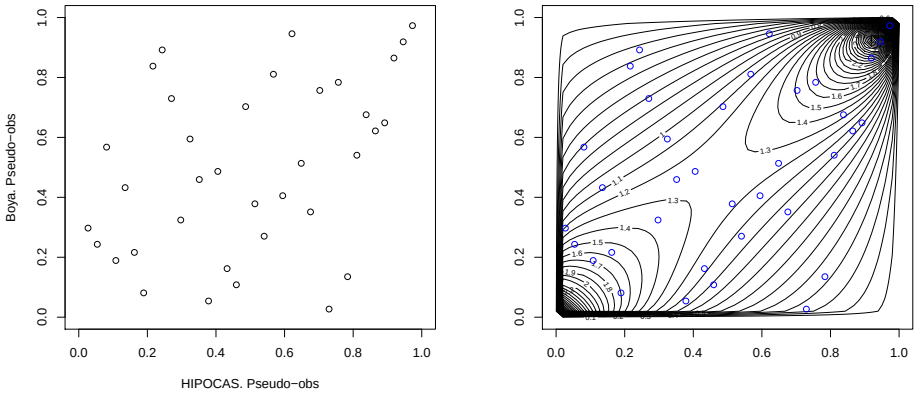

Figura 12.11: Ajuste de la Cópula Gumbel para HIPOCAS y Boya en $[0,1]^2$. Contornos de isodensidad a partir de la muestra del posteriori del parámetro (Fig. 12.10).

Bondad de ajuste. Dependencia mediante cópula paramétrica Gumbel

Se ha valorado la coherencia de la representación de la dependencia entre las dos series de altura de ola mediante una cópula paramétrica de Gumbel con los datos mediante los estadísticos de contraste presentados en la Sección 8.1. En la Tabla 12.5 se muestra el contraste respecto al ajuste global del modelo. La dependencia presente en los excesos originales es algo inferior a la presente en las remuestras generadas a partir del posteriori, aunque del mismo orden, aspecto que se confirma observando los valores a posteriori del parámetro δ (Tabla 12.6). El modelo con dependencia Gumbel representa una dependencia global similar a la presente en los datos, aunque la incertidumbre en los resultados es elevada.

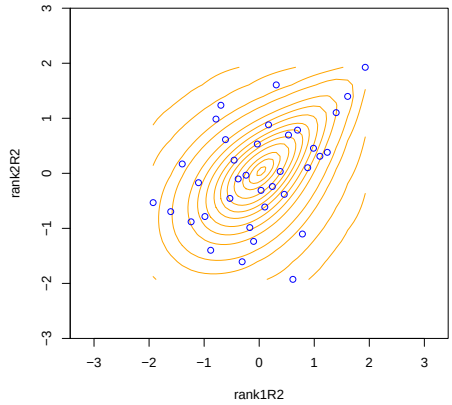

Figura 12.12: Ajuste de la Cópula Gumbel para HIPOCAS y Boya en $\mathbb{R}^2$. Contornos de isodensidad a partir de la muestra del posteriori de los parámetros.

Tabla 12.5: Contraste sobre el modelo global. Coeficiente τ original y coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. El coeficiente predictivo muestra dependencia similar a la original. El p-valor basado en la discrepancia τ sugiere que no se debe rechazar la hipótesis de validez del modelo de dependencia Gumbel especificado.

τ original	τ PredictivoG	Discrepancia τ
0.26032	0.36378	0.8163

Tabla 12.6: Contraste sobre el modelo global. Coeficiente δ original y coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. El coeficiente predictivo muestra una dependencia similar a la original.

δ original	δ PredictivoG
1.35193	1.6233

12.4. Dependencia mediante cópula CrEnC

La dependencia conjunta de los excesos de log-altura significante registrados mediante boya e HIPOCAS se ha modelizado también mediante una cópula CrEnC (ver Capítulo 5). Se ha implementado un conjunto de medidas de asociación (ver Capítulo 9) que corresponden a diferentes tipos de dependencia (ver Sec. 2.2), y a los que corresponden los parámetros $\alpha_i, i = 1, \ldots, 7$ del modelo.

La cópula CrEnC se ha estimado utilizando el momento seleccionado mediante el método de razón de verosimilitudes (ver Sec. 7.1.2). Para este conjunto de datos de log-altura de ola, el único momento seleccionado es el correspondiente a α_6 (Spearman's footrule φ). La muestra del posteriori obtenida permite representar la incertidumbre en la estimación de estos parámetros de dependencia. La Figura 12.13 muestra el histograma del parámetro α_6 seleccionado para la cópula. Las estimaciones presentan bastante dispersión, en gran parte debida al tamaño reducido de la muestra disponible, presentando sobretodo valores positivos del coeficiente. La Tabla 12.7 muestra percentiles seleccionados de la muestra del parámetro.

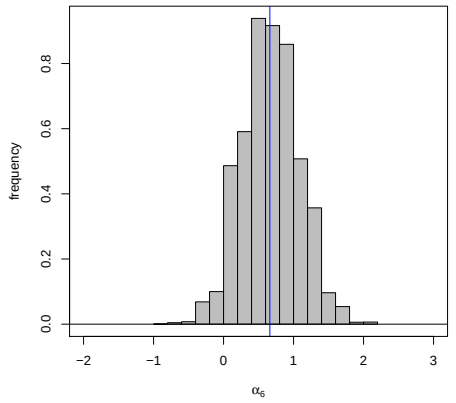

Figura 12.13: Histograma de la muestra del posteriori para el parámetro de la cópula seleccionado, α_6. La línea azul marca la mediana de la muestra del posteriori para el parámetro.

Tabla 12.7: Percentiles de la muestra del posteriori para el parámetro de cópula CrEnC selecionado.

	2.5 %	25 %	50 %	75 %	97.5 %
α_6	-0.11	0.40	0.66	0.94	1.44

En la Fig. 12.14 se muestra una representación gráfica de la densidad de la cópula CrEnC mediante contornos de isodensidad en $\mathbb{R}^2$, para la mediana del coeficiente α_6. Debido a la dispersión en las estimas, debe observarse que al realizar la representación de los contornos para otros valores de la muestra (p.ej. para el primer cuartil de la muestra), se obtendrían contornos muy diferentes. El ajuste visual por tanto resulta insuficiente, y debe complementarse con un contraste de bondad de ajuste a posteriori (Sec. 12.4).

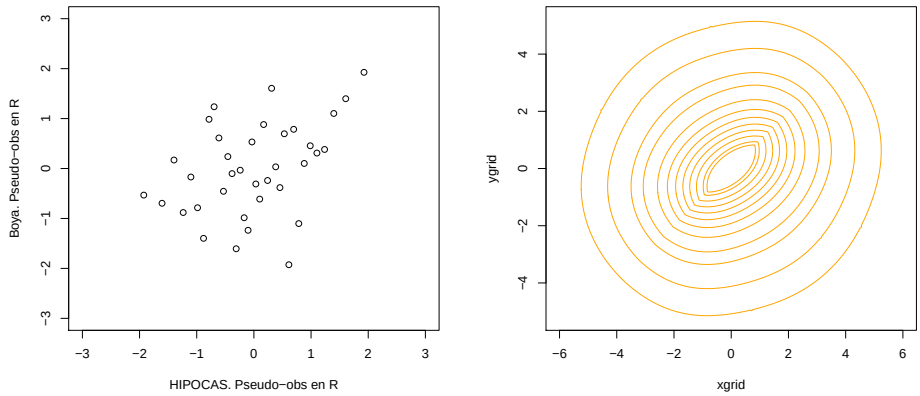

Figura 12.14: Cópula CrEnC para HIPOCAS y Boya en $\mathbb{R}^2$. Contornos de isodensidad a partir de la muestra del posteriori de los parámetros (Fig. 12.13).

Bondad de ajuste. Dependencia con cópula CrEnC

Se desea comprobar la coherencia de diferentes aspectos del modelo (hipótesis marginales y ajuste global) con los datos de altura de ola

registrados mediante Boya y los hindcast. En la Tabla 12.8 se presentan percentiles a posteriori seleccionados de los estadísticos de bondad de ajuste (Kolmogorov-Smirnov y Multinomial), los cuales permiten valorar la hipótesis de distribución marginal GPD de los excesos de log-altura de ola para las series originales y para las remuestas generadas a partir del modelo. Los p-valores a posteriori no permiten rechazar la hipótesis de marginal GPD para ninguna de las series, aunque el p-valor multinomial (afectado por el reducido tamaño de la muestra), sugiere que sí debería rechazarse. Los p-valores predictivos de las remuestras indican también un buen ajuste marginal.

La hipótesis suplementaria de distribución GPD en el dominio de Weibull, es decir, GPD con parámetro $\xi < 0$ y soporte acotado, puede comprobarse mediante los estadísticos *Slope*, pendiente de la recta de regresión del exceso esperado (Tabla 12.9). Los resultados coinciden con la apreciación visual de los histogramas de la muestra del posteriori del parámetro ξ (Fig. 12.9): pese a que los valores centrales de la muestra se encuentran lejos del cero, existe una gran dispersión y la hipótesis $\xi < 0$ corta el histograma. Esto se traduce en valores de pendiente (slope) ligeramente diferentes para la muestra original y las remuestras, y unos estadísticos de discrepancia que rozan el rechazo. Observando las Figuras 12.5 y 12.6, quizá debería tomarse un umbral más alto para extraer excesos y realizar el análisis, pero con los pocos datos de que se dispone en la serie actual, este tratamiento no es posible.

Tabla 12.8: Bondad de ajuste marginal para HIPOCAS y Boya. Percentiles seleccionados de los p- valores a posteriori (Sup.). Percentiles seleccionados de los p-valores predictivos a posteriori (Inf.)

	p-val.	2.5 %	50 %	97.5 %
HIPOCAS	K-S	0.00211	0.05934	0.26964
HIPOCAS	Multinomial	0.00199	0.02277	0.08444
Resample HIPOCAS	K-S	0.00089	0.28044	1.00000
Resample HIPOCAS	Multinomial	0.00104	0.30500	0.95755
Boya	K-S	0.00358	0.10776	0.34874
Boya	Multinomial	0.01045	0.04412	0.33306
Resample Boya	K-S	0.00165	0.31469	1.00000
Resample Boya	Multinomial	0.00102	0.34673	0.98101

Los estadísticos basados en τ permiten valorar el ajuste global del

Tabla 12.9: Contraste sobre la validez del priori GPD en DA-Weibull para cada marginal. Las pendientes de la recta de regresión del exceso esperado son similares en la muestra original y para las remuestras correspondientes a la muestra del posteriori.

Value	X_1	X_2
Slope orig.	-0.395747	-0.432202
Slope predictive	-0.150586	-0.129857
Slope Discrepancy	0.978351	0.996907

modelo a los datos. En la Tabla 12.11 se aprecia que la dependencia global es similar en los excesos originales y en el modelo ajustado, por lo que el modelo propuesto es coherente globalmente con los datos. Las probabilidades de exceso de valores de referencia (Tabla 12.10) en la muestra original y en el modelo ajustado sirven también como *proxy* de este ajuste global. Estas probabilidades son similares, confirmando el buen ajuste del modelo a los datos.

Tabla 12.10: Proporción original de excesos sobre log(2) que superan los valores de referencia de altura de ola en las dos series. Probabilidades de excedencia conjuntas predictivas (Mediana) para esos valores de referencia.

Ref. x_1^0	Ref. x_1^0	Exc. $\log(x_1^0)$ sobre log(2)	Exc. $\log(x_2^0)$ sobre log(2)	Proporción exced. orig.	Mediana prob. exced.
2.5m	2.5m	0.2231	0.2231	0.5000	0.3497
3.0m	3.0m	0.4055	0.4055	0.2027	0.1590
3.5m	3.5m	0.5596	0.5596	0.0405	0.0819
4.0m	4.0m	0.6931	0.6931	0.0405	0.0469
4.5m	4.5m	0.8109	0.8109	0.0405	0.0315

Tabla 12.11: Contraste sobre el modelo global. Coeficiente τ original, coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. Los coeficientes muestran dependencias similares. El p-valor basado en la discrepancia τ no permite rechazar la hipótesis de validez del modelo global especificado.

τ original	τ PredictivoG	Discrepancia τ
0.260317	0.374439	0.840206

12.5. Valores a posteriori de interés

Una vez descritos los resultados de la estimación de los parámetros del modelo e interpretada su bondad de ajuste a los datos, podemos obtener cantidades a posteriori y predictivas de interés, tanto marginales como conjuntas.

Dado un conjunto de valores de log-altura de ola de referencia se han obtenido sus probabilidades a posteriori y los correspondientes cuantiles a posteriori (Tablas 12.12 y 12.13); sus periodos de retorno a posteriori (Tablas 12.14 y 12.15) y los valores a posteriori asociados a periodos de retorno de referencia (Tabla 12.16 y 12.17).

Tabla 12.12: Probabilidades de no excedencia a posteriori para valores seleccionados de los excesos HIPOCAS. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1ref	Mean	GeomMean	q0.05	q0.5	q0.95
0.2231 (2.5m)	0.4497	0.4492	0.3636	0.4492	0.5255
0.4055 (3.0m)	0.6745	0.6773	0.5662	0.6754	0.7593
0.5596 (3.5m)	0.7970	0.8027	0.6953	0.8017	0.8780
0.6931 (4.0m)	0.8682	0.8757	0.7742	0.8739	0.9366
0.8109 (4.5m)	0.9116	0.9201	0.8304	0.9181	0.9671

Además se han calculado probabilidades de no excedencia conjuntas a posteriori de valores de referencia y sus percentiles (Tabla 12.18). Se observa que el modelo no sólo proporciona estimaciones de probabilidades para parejas de valores observados, sino que también proporciona

Tabla 12.13: Probabilidades de no excedencia para valores seleccionados de los excesos Boya. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x2ref	Mean	GeomMean	q0.05	q0.5	q0.95
0.2231 (2.5m)	0.4997	0.4997	0.4128	0.4996	0.5817
0.4055 (3.0m)	0.7287	0.7323	0.6301	0.7318	0.8180
0.5596 (3.5m)	0.8438	0.8500	0.7575	0.8487	0.9203
0.6931 (4.0m)	0.9059	0.9134	0.8316	0.9118	0.9639
0.8109 (4.5m)	0.9410	0.9491	0.8826	0.9470	0.9831

Tabla 12.14: Periodo de retorno a posteriori correspondiente a valores seleccionados de los excesos HIPOCAS. Periodo de retorno a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1tau	Mean	GeomMean	q0.05	q0.5	q0.95
0.2231 (2.5m)	1.8333	1.8252	1.5714	1.8156	2.1073
0.4055 (3.0m)	3.1797	3.1242	2.3049	3.0811	4.1552
0.5596 (3.5m)	5.3296	5.1170	3.2819	5.0436	8.1986
0.6931 (4.0m)	8.7839	8.1306	4.4291	7.9333	15.7695
0.8109 (4.5m)	14.4822	12.6530	5.8978	12.2087	30.3621

estimaciones de probabilidades de combinaciones de valores no observados en la serie registrada.

Tabla 12.15: Periodo de retorno correspondiente a valores seleccionados de los excesos Boya. Periodo de retorno a posteriori: media, centro y cuantiles a posteriori seleccionados.

x2tau	Mean	GeomMean	q0.05	q0.5	q0.95
0.2231 (2.5m)	2.0204	2.0096	1.7031	1.9983	2.3906
0.4055 (3.0m)	3.8484	3.7646	2.7031	3.7290	5.4953
0.5596 (3.5m)	7.0976	6.7279	4.1230	6.6083	12.5439
0.6931 (4.0m)	12.9971	11.6578	5.9393	11.3353	27.7387
0.8109 (4.5m)	24.4687	19.8184	8.5197	18.8618	59.1145

Tabla 12.16: Excesos HIPOCAS a posteriori correspondientes a periodos de retorno seleccionados. Excesos a posteriori: media, centro y cuantiles seleccionados.

tiempo	Mean	GeomMean	q0.05	q0.5	q0.95
10.	0.7748	0.7650	0.6036	0.7576	1.0277
50.	1.1832	1.1648	0.8925	1.1435	1.6019
100.	1.3331	1.3102	0.9897	1.2820	1.8268
500.	1.6325	1.5968	1.1761	1.5593	2.3182

12.6. Discusión

Se ha ajustado el modelo propuesto a una serie bivariada de datos de altura de ola significante provenientes de un modelo de hincasting y de una boya situada en el litoral catalán. La serie conjunta tiene pocos datos, por lo que las estimaciones de los parámetros presentan una incertidumbre considerable. La cópula CrEnC propuesta ajusta mejor que la cópula Gumbel. A partir del modelo estimado se han obtenido cantidades a posteriori de interés.

Tabla 12.17: Excesos Boya correspondientes a periodos de retorno seleccionados. Excesos a posterioris: media, centro y cuantiles seleccionados.

tiempo	Mean	GeomMean	q0.05	q0.5	q0.95
10.	0.6772	0.6695	0.5178	0.6605	0.8644
50.	1.0424	1.0277	0.7753	1.0105	1.3687
100.	1.1781	1.1597	0.8741	1.1423	1.5713
500.	1.4523	1.4233	1.0441	1.4007	2.0182

Tabla 12.18: Probabilidades de excedencia conjuntas a posteriori para valores seleccionados de los excesos de HIPOCAS y Boya. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1refconj	x2refconj	Mean	GeomMean	q0.05	q0.5	q0.95
0.2231 (2.5m)	0.2231 (2.5m)	0.3549	0.3435	0.2027	0.3378	0.5000
0.4055 (3.0m)	0.4055 (3.0m)	0.1722	0.1559	0.0676	0.1757	0.3108
0.5596 (3.5m)	0.5596 (3.5m)	0.0964	0.0805	0.0135	0.0946	0.2027
0.6931 (4.0m)	0.6931 (4.0m)	0.0597	0.0462	0.0135	0.0405	0.1216
0.8109 (4.5m)	0.8109 (4.5m)	0.0407	0.0311	0.0135	0.0405	0.0946

Capítulo 13

Conclusiones

El estudio conjunto de dos o más variables extremales es un problema común en la mayoría de disciplinas científicas. En un contexto de estudio de peligrosidad de fenómenos naturales, *hazard*, se pretende determinar la probabilidad de ocurrencia de sucesos extremales de interés, o cantidades derivadas de ellos: probabilidad de excedencia de un caudal de río que pudiera resultar peligroso para un puente situado en el mismo; probabilidad de excedencia de un valor de precipitación diaria que pudiera provocar inundaciones; periodos de retorno de esas cantidades; probabilidades de excedencia conjuntas de ambas cantidades, etc. Los sucesos extremales que nos interesa estudiar son escasos, y por tanto las series registradas suelen tener pocos datos. En la actualidad se puede complementar las series registradas con otras sintéticas, como las de hindcast (Ortego et al., 2012, 2014), pero en ausencia de este tipo de datos, conviene complementar la información disponible con otra similar, por ejemplo la registrada en otras estaciones cercanas. Además, la mayoría de variables de interés no deben estudiarse de manera aislada, sino conjuntamente con otras variables que intervengan en el proceso. Por ejemplo, para estudiar situaciones de *hazard* de temporales de lluvia, es conveniente no estudiar solamente la precipitación total de un suceso, sino que se debe relacionar esta variable con otras como la intensidad en intervalos de tiempo breves (5, 10, 20minutos), puesto que es la combinación de estos factores la que puede producir situaciones que pudieran resultar peligrosas. O bien, para estudiar la precipitación en una cuenca hidrográfica será necesario estudiar los

sucesos de lluvia registrados en dos o más de las estaciones situadas en esa cuenca.

En este contexto, se ha definido un modelo que representa la ocurrencia de sucesos de *hazard* mediante un procesos de Poisson evaluado. Se ha presentado un modelo marginal que puede incluir valores por defecto. Aunque la representación conjunta de este tipo de variables presenta particularidades, si el objeto de estudio son los excesos de estas marginales sobre un umbral suficientemente alto, los valores por defecto desaparecen. Dado que las situaciones de interés son de tipo extremal, los tamaños (excesos sobre un umbral suficientemente alto) se modelizan mediante distribuciones Generalizadas de Pareto (GPD), y la dependencia entre éstas se modeliza mediante funciones cópula. Existen multitud de cópulas paramétricas que permiten modelizar la dependencia entre dos variables. En la literatura se han usado en ámbitos muy diversos, pero no siempre es sencillo hallar la más adecuada ni analizar la bondad de su ajuste a los datos. En el desarrollo de la Tesis se ha explicitado el caso bidimensional, pero el modelo puede ampliarse a más de dos tamaños.

Dado un par de variables de interés, con frecuencia se dispone de información parcial sobre ellas, como los modelos marginales e información vaga sobre la dependencia entre ellas, en forma de momentos. Se ha introducido una nueva tipología de cópulas, las cópulas CrEnC, que son las de mínima información mutua dadas las marginales y un conjunto de momentos. Por su relación con las funciones cópula, se escogen momentos invariantes por transformaciones monótonas. De esta manera, la dependencia se conservará sean cuales sean las marginales escogidas. Para evitar los problemas de representación en un recinto acotado como $[0,1]^2$ y mejorar tanto la estimación como la visualización del ajuste de la cópula, se ha escogido la representación de las cópulas CrEnC en $\mathbb{R}^2$. Estas cópulas son flexibles, pero para su estimación se requiere el ajuste numérico de las funciones de normalización, las cuales se han aproximado mediante el método de Montecarlo. Se proporciona el algoritmo de estimación de estas cópulas CrEnC.

Se ha implementado un proceso de estimación bayesiana de los parámetros del modelo, que considera la incertidumbre en las estimaciones. Se ha obtenido una muestra extensa del posteriori de los parámetros mediante un muestreo de Gibbs, la cual permite calcular

cantidades predictivas a posteriori. Además se ha valorado la bondad de ajuste del modelo mediante p-valores bayesianos (a posteriori y basados en discrepancia), los cuales permiten valorar diferentes aspectos del modelo, como el ajuste marginal de la distribución GPD, la adecuación de la hipótesis de GPD en el dominio de Weibull o el ajuste global del modelo. La combinación de la representación de las cópulas en $\mathbb{R}^2$ y el uso de p-valores bayesianos mejora la interpretación de la bondad de ajuste de las cópulas, y por tanto, mejoran la selección del modelo más adecuado a los datos disponibles. Estos p-valores en general no son uniformes, y por ello se ha introducido el cálculo de un intervalo de valores entre los cuales el p-valor sí puede considerarse uniforme. Para facilitar los cálculos de cantidades de interés extremales, se han introducido también las transformaciones de cópulas por cambio de umbral o extracción de extremos.

El proceso de estimación se ha implementado en un programa Fortran generado íntegramente para esta Tesis. El programa de estimación consta de tres partes: el establecimiento de los prioris, la estimación de parámetros propiamente dicha y el postproceso, donde se obtienen estadísticos de bondad de ajuste del modelo y valores a posteriori de interés. Cabe destacar que los prioris sobre los parámetros (ξ, β) de la distribución GPD de los tamaños marginales se establecen por *expert assessment*: a partir de las especificaciones del experto se define una región conjunta para (ξ, β) donde se define una función suave, que es la utilizada como priori. Para el resto de parámetros se utiliza un priori uniforme en la escala adecuada del parámetro.

El modelo se ha aplicado a tres conjuntos de datos: un conjunto simulado, que permite valorar el buen funcionamiento del modelo y el programa; un conjunto de datos de precipitación diaria en dos ubicaciones y un conjunto de altura significante de ola en dos series diferentes (boya e HIPOCAS).

Para el conjunto de datos simulados se ha aplicado el modelo con representación de la dependencia entre series mediante una cópula CrEnC. Los momentos incluidos en la cópula CrEnC se seleccionan entre los diferentes momentos disponibles, mediante un procedimiento hacia delante. Para este conjunto de datos solamente resulta seleccionado uno de los momentos, pero el ajuste es visualmente correcto, y los estadísticos de bondad de ajuste obtenidos refuerzan esta interpreta-

ción visual. La estimación de los parámetros marginales parece afectada por el priori establecido, en particular por la restricción a priori de modelo GPD en el dominio de Weibull, por lo que se ha realizado una primera aproximación a la sensibilidad a este aspecto. Los resultados indican que las cantidades predictivas apenas se ven afectadas por esta limitación en el priori.

En el segundo ejemplo se dispone de tres conjuntos de datos de precipitación en dos ubicaciones, que presentan dependencias ligeramente diferentes. La dependencia para estos conjuntos se ha representado mediante una cópula de la familia Gumbel y también mediante una cópula CrEnC. Para los tres conjuntos de datos, en la cópula CrEnC ha sido seleccionado un único momento, el mismo para los tres conjuntos, aunque los valores de los coeficientes son diferentes y las cópulas obtenidas presentan contornos de isodensidad muy diferentes. Los contrastes de bondad de ajuste indican que el ajuste marginal no es bueno para algunas de las estaciones. Esto se debe a que al inicio del proceso de estimación se ha fijado un único umbral de referencia para los excesos GPD, y aunque las gráficas de excesos esperados de todas las estaciones indicaban que el umbral escogido era suficientemente alto, los resultados de los contrastes de bondad de ajuste indican que para algunas de las estaciones el umbral quizá debería ser más elevado. La dependencia entre las series queda mejor representada por la cópula CrEnC propuesta que por la cópula de la familia Gumbel, ya que esta en ocasiones sobreestima, y en otras subestima la dependencia existente. Se ha comprobado adicionalmente si sería necesario añadir al modelo los momentos que resultan significativos individualmente en el proceso de selección hacia delante, pero que son descartados en los pasos sucesivos. Aplicando el principio de parsimonia, se ha decidido utilizar sólo aquellos seleccionados en el proceso de selección hacia adelante, que en este caso es uno solo.

En el último ejemplo se trata un conjunto de datos de altura significante de ola proveniente de un registro de boya en la costa catalana y de una serie de hindcast HIPOCAS en un nodo cercano. Este es un registro típico extremal, donde se dispone de muy pocos datos. En este caso los datos provienen de dos series de tipología diferente para poder ampliar la información disponible. La dependencia entre estas series es escasa y por tanto ha sido necesario contrastar la indepen-

dencia de las series antes de ajustar un modelo de cópula, que ha sido rechazada. Existe dependencia, pero es escasa, y por tanto al ajustar la dependencia mediante una cópula Gumbel, los valores del parámetro δ obtenido se encuentran muy cerca del borde de su dominio de definición, produciendo un ajuste no demasiado bueno. Por lo que respecta a la representación de dependencia mediante cópula CrEnC, sólo un momento resulta escogido, pero se obtiene un mejor ajuste a los datos que con la cópula Gumbel. El ajuste GPD marginal no es bueno para una de las series, probablemente porque el umbral GPD resulta insuficiente: pese a que en las gráficas de exceso esperado se observa que el umbral escogido es suficiente, los estadísticos de bondad de ajuste indican que el ajuste no es bueno. No obstante, en este caso el conjunto de datos tiene un tamaño reducido, y aumentar el umbral no es factible, dado que significaría reducir el conjunto de datos más aún.

Finalmente, se han obtenido valores predictivos de interés para todos los conjuntos de datos: probabilidades de no excedencia de valores marginales y conjuntos; periodos de retorno de sucesos; periodos de retorno de sucesos en un intervalo, etc.

Globalmente, las cópulas CrEnC son una buena alternativa a las cópulas más usuales, ya que incluyen información en forma de restricciones y el modelo propuesto proporciona buenos resultados. En futuras líneas de trabajo se desean introducir más tipos de restricciones en forma de momentos. Además, se podrían describir en profundidad dos generalizaciones de los momentos invariantes por transformaciones monótonas, que en los desarrollos actuales se han denominado coeficiente Tipo1 y Tipo2. Estos nuevos momentos pueden mejorar la estimación de dependencia en situaciones no simétricas, como la que se obtiene al modelizar la dependencia entre intensidades de lluvia en dos intervalos diferentes de tiempo, una dependencia con una forma muy asimétrica debido a la relación funcional entre ambas magnitudes, que generalmente queda mal representada por las cópulas más usuales.

Bibliografía

Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21(4):1–21. 11.4.2

Abdous, B., Genest, C., and Rémillard, B. (2005). Dependence properties of meta-elliptical distributions. In Duchesne, P. and Rémillard, B., editors, *Statistical Modeling and Analysis for Complex Data Problems*, pages 1–15, New York. Springer. 1.2.2, 2.1.1

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p. c

Ali, M. M., Mikhail, N. N., and Haq, M. S. (1978). A class of bivariate distributions including the bivariate logistic. *J. Multivariate Anal.*, 8:405–412. 1.2.2, b

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999). *Conditional Specification of Statistical Models*. Springer, NY, USA. 1.2.7

Bayarri, M. J. and Berger, J. O. (2000). P-values for composite null models. *J. of the Am. Stat. Ass.*, 95:1127–1142. 2.5.2, A.0.2, A.0.3

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of extremes. Theory and Applications*. Wiley, Chichester, GB. 490 p. 1.2.5

Berg, D. and Bakken, H. (2006). Copula goodness-of-fit tests: A comparative study. Unpublished. www.danielberg.no/publications/CopulaGOF.pdf(Feb.2011). a

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester, GB. 608 p. 2.5

Blest, D. C. (2000). Rank correlation. an alternative measure. *Austral. and New Zealand J. Statist.*, 42(1):101–111. 2.1

Blomqvist, N. (1950). Rank correlation. an alternative measure. *Ann. Math. Statist.*, 21(4):539–600. 2.2.2, 2.1

Box, R. J. (1980). Sampling and Bayes inference in scientific modeling and robustness. *J. of the Roy. Stat. Soc., Ser. A*, 143:383–430. 2.5.2, A.0.1

Breymann, W., Dias, A., and Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quant. Finance*, 3(1):1–14. a

Calsaverini, R. S. and Vicente, R. (2009). An information-theoretic approach to statistical dependence: Copula information. *Europhysics Letters*, 88(6):1–6. 1.2.7, 7.5

Castillo, E. (1988). *Extreme value theory in engineering*. Academic Press, San Diego, Cal. USA. 389 p. 1.2.5, 1.3, 2.4, 2.4, a, 11.1.1, 12.1.1

Castillo, E., Hadi, A. S., Balakrishnan, N., and Sarabia, J. M. (2004). *Extreme value and related models with Applications in Engineering and Science*. Wiley, London, GB. 384 p. 1.2.5, 1.3, 2.4

Chen, L., Singh, V., and Guo, S. (2013). Measure of correlation between river flows using the copula-entropy method. *J. Hydrol. Eng.*, 18(12):1591–1606. 7.5

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York, NY, USA. 386 p. 2.5.1, 5.2

Chu, B. (2011). Recovering copulas from limited information and an application to asset allocation. *J. Bank. Financ.*, 35(7):1824 – 1842. 7.5

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151. 1.2.2

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, GB. 208 p. 1.3

Coles, S. G., Heffernan, J., and Tawn, J. A. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365. 1.2.6, 4.3

Cook, R. D. and Johnson, M. E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data. *J. Roy. Statist. Soc. B*, 43(2):210–218. 1.2.2

Cox, D. R. and Isham, V. (1980). *Point Processes*. Chapman & Hall CRC, London, GB. 188 p. 1.2.4

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Volume 1.* Springer Verlag, New York, NY, USA. 464 p. 1.2.4, 2.3

Dall'Aglio, G. (1991). Fréchet classes: the beginnings. In Dall'Aglio et al. (1991), pages 1–12. 2.2.7

Dall'Aglio, G., Kotz, S., and Salinetti, G., editors (1991). Dordrecht, Germany. Kluwer Academic Publ. 13

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *J. Roy. Statist. Soc. B*, 52:393–442. 1.3, 2.4

De Finetti, B. (1974). *Theory of Probability*. Wiley, Chichester, GB. 675 p. 1.2.3

de Haan, L. (1976). Sample extremes: an elementary introduction. *Statist. Neerlandica*, 30:161–172. 1.2.5

Dempster, M. A., Medova, E. A., and Yang, S. W. (2007). Empirical copulas for CDO trance pricing using relative entropy. *Int. J. Theoretical Appl. Finance*, 10(4):679–701. Erratum: ibid. vol. 10, num. 7, p. 1255–1260 (2007). 7.5

Dickinson Gibbons, J. (1993). *Nonparametric measures of association*. Sage University Paper Series, Newbury Park, CA, USA. – p. 1.2.6

Dobrić, J. and Schmid, F. (2007). A goodness of fit test for copulas based on rosenblatt s transformation. *Comput. Stat. Data Anal.*, 51:4633–4642. a

Drouet-Mari, D. and Kotz, S. (2001). *Correlation and dependence*. Imperial College Press, London, UK. 219 p. 1.2.6

Dupuis, D. J. (2007). Using copulas in hydrology: Benefits, cautions and issues. *J. Hydrol. Eng.*, 12(4):381–393. 2.2.2

Ebrahimi, N., Soofi, E. H., and Soyer, R. (2008). Multivariate maximum entropy identification, transformation and dependence. *J. Multiv. Anal.*, 99:1217–1231. 1.2.7, 7.5

Edwards, H. H., Mikusiński, P., and Taylor, M. D. (2004). Measures of concordance determines by d4-invariant copulas. *IJMMS*, 70:3867 – 3875. 2.2.2, 2.2.10

Egozcue, J. J., Pawlowsky-Glahn, V., Ortego, M. I., and Tolosana - Delgado, R. (2006). The effect of scale in daily precipitation hazard assessment. *Nat. Hazards Earth Syst. Sci.*, 6:459–470. 11.1, 12.1

Egozcue, J. J. and Ramis, C. (2001). Bayesian hazard analysis of heavy precipitation in eastern spain. *Int. J. Climatol.*, 21:1263–1279. 11.1, 11.1.1, 11.2.1, 11.3.1, 11.4.1

Embrechts, P. (2009). Copulas: A personal view. *J. Risk Ins.*, 76(3):639–650. 2.1.4

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Springer-Verlag, Berlin, Germany. 663 p. 1.2.4, 1.2.5, 1.3, 2.3, 2.4, 2.4

Embrechts, P., McNeil, A. J., and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. In Dempster, M. A. H., editor, *Risk Management: Value at Risk and Beyond.*, pages 223–233. Cambridge University Press, New York, USA. xiv + 274 p. 1.2.6

Fang, H. B., Fang, K. T., and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *J. Multiv. Anal.*, 82:1–16. 1.2.2, 2.1.1

Feller, W. (1968a). *An introduction to probability theory and its applications, Vol. 1.* Wiley, New York, USA, 3rd ed. edition. 528 p. 1.2.6, 3

Feller, W. (1968b). *An introduction to probability theory and its applications, Vol. 2.* Wiley, New York, NY, USA, 3rd ed. edition. 704 p. 1.2.6

Fermanian, J. D. (2005). Goodness-of-fit tests for copulas. *J. Multiv. Anal.*, 95:119–152. b

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222:309–368. 1.2.5

Fisher, R. A. and Tippet, L. H. C. (1928). Limiting forms of the frequence distribution of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.*, 24:180–190. 1.2.5

Frank, M. J. (1979). On the simultaneous associativity of f(x, y) and x+y-f(x, y). *Aequationes Math.*, 19:194–226. 1.2.2, c

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann de la Soc. Polonaise de Math.*, 6:93–116. 1.2.5

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon, Sect. A*, 9:53–77. 1.2.1, 1.2.7

Galambos, J. (1987). *The asymptotic theory of extreme order statistics.* Krieger, Malabar, Fl, USA, 2 edition. 1.2.5, 1.3

Galeano, E. (1998). *Patas arriba: la escuela del mundo al revés*. Siglo XXI, Madrid, ES. (document)

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proc. Roy. Soc.*, 45:135–145. 1.2.6

Galton, F. (1890). Kinship and correlation. *North Am. Rev.*, 150:419–431. 1.2.6

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, 85:398–409. 2.5.1

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Wiley, New York, NY,USA. – p. 2.5.1, 2.5.2, 2.5.5, 11.1, 12.1

Gelman, A., Meng, X. L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6:733–807. 2.5.2, 2.5.5

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, 74(3):549–555. 1.2.2, c

Genest, C., Carabarín-Aguirre, A., and Harvey, F. (2013). Copula parameter estimation using blomqvist's beta. *J. SFdS*, 154(1):5–24. 2.2.2

Genest, C. and Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.*, 12(4):347–368. 2.2.2, 2.2.2, 2.2.11

Genest, C. and MacKay, J. (1986a). The joy of copulas: bivariate distributions with uniform marginals. *Am. Statist.*, 40(4):280–283. 1.2.2, 2.2.2

Genest, C. and MacKay, R. J. (1986b). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canad. J. Statist.*, 14:145–159. 1.2.1, 1.2.2, 2.2.2

Genest, C. and Plante, J. F. (2003). On blest's measure of rank correlation. *Canad. J. Statist.*, 31(1):35–52. 1.2.6, 2.2.4, 2.2.2, 2.1

Genest, C., Quessy, J. F., and Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scand. J. Statist.*, 33:337–366. c

Genest, C. and Rémillard, B. (2006). Discussion of copulas: Tales and facts, by thomas mikosch. *Extremes*, 9:27–36. 2.1.4

Genest, C. and Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Ann. Inst. Henri Poincare-Probab. Stat.*, 44(6):1096–1127. c

Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insur. Math. Econ.*, 44:199–213. 2.1.3

Genest, C. and Rivest, L. P. (1993). Statistical inference procedures for bivariate archimedean copulas. *J. Am. Stat. Assoc.*, 88(423):1034–1043. 1.2.2

Gnedenko, B. (1943). Sur la distribution limit du terme maximum d'une série aléatoire. *Ann. Math*, 44:423–453. 1.2.5

Gokhale, D. V. (1999). On joint and conditional entropies. *Entropy*, 1:21–24. 1.2.7

Gómez Villegas, M. A. (2005). *Inferencia Estadística*. Ed. Díaz de Santos., Madrid, Spain. 7.1.2

Grandell, J. (1997). *Mixed Poisson processes*. Chapman & Hall, London, GB. 268 p. 2.3, 2.3

Guedes Soares, C., Carretero Albiach, J. C., Weisse, R., and Alvarez-Fanjul, E. (2002). A 40 years hindcast of wind, sea level and waves in european waters. In *Proceedings of the 21st International Conference on Offshore Mechanics and Arctic Engineering*, pages 669–675, Oslo, Norway. 12.1

Gumbel, E. J. (1933). Das alter des methusalem. *Z. Schweizerische Statistik un Volkwirtschaft*, 69:516–530. 1.2.5

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia University Press, New York, NY, USA. 375 p. 1.2.5

Gumbel, E. J. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Pub. Inst. Statist. Univ. Paris*, 9:171–173. 1.2.2, a, 11.1, 11.2.2

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. of the Roy. Stat. Soc., Ser. B*, 29:83–100. 2.5.2, 2.5.4

Hao, Z. and Singh, V. (2013). Entropy-based method for bivariate drought analysis. *J. Hydrol. Eng.*, 18(7):780–786. 7.5

Heffernan, J. E. and Stephenson, A. G. (2012). *ismev: An Introduction to Statistical Modeling of Extreme Values*. R package version 1.39. 11.1.1, 12.1.1

Hoeffding, W. (1940). Masstabinvariante korrelationstheorie. *Schrif. Math. Seminars Inst. Angew. Math. Univ. Berlin*, 5(3):179–233. Reprinted as "Scale-invariant correlation theoryïn *The Collected Works of Wassily Hoeffding*, N.I Fisher and P.K. Sen, editors (Springer-Verlag, New York), 57-107. 1.2.1, 1.2.6, 2.2.2

Hoeffding, W. (1941). Masstabinvariante korrelationsmasse für diskontinuerliche verteilungen. *Ark. Math. Wirtshaften und Sozialforschung*, 7:49–70. Reprinted as "Scale-invariant correlation measures for discontinous distributionsïn *The Collected Works of Wassily Hoeffding*, N.I Fisher and P.K. Sen, editors (Springer-Verlag, New York), 109-133. 1.2.1

Hoeffding, W. (1947). On the distribution of the rank correlation coefficient ô when the variates are not independent. *Biometrika*, 34(3/4):183–196. 2.2.2

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2014). *copula: Multivariate Dependence with Copulas*. R package version 0.999-9. 11.4.2

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71(1):75–83. 1.2.2

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396. 1.2.2

Hult, H. and Lindskog, F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Adv. Appl. Probab.*, 34(3):587–608. 2.1.1

Jacob, D. and Podzun, R. (1997). Sensitivity studies with the regional climate model REMO. *Meteorol. Atmos. Phys.*, 63:119–129. 12.1

Joe, H. (1989). Relative entropy measures of multivariate dependence. *J. Amer. Statist. Assoc.*, 84(405):157–164. 2.2.9, 2.2.4, 2.2.5, 2.2.2

Joe, H. (1990). Multivariate concordance. *J. Mult. Anal.*, 35:12–30. 2.2.2

Joe, H. (1997). *Multivariate models and dependence concepts*. Chapman & Hall, London, GB. 399 p. 1.2.6, 2.1.2

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30:81–93. 1.2.6, 2.2.6, 2.1

Klüppelberg, C. and Resnick, S. I. (2008). The Pareto copula, aggregation of risks, and the emperor's socks. *J. Appl. Probab.*, 45(1):67–84. 2.1.4

Kotz, S. and Nadarajah, S. (2000). *Extreme value distributions. Theory and applications*. Imperial College Press, London, GB. 185 p. 1.2.5, 2.4, 2.4

Kruskal, W. H. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.*, 53(284):814–861. 1.2.6, 2.1.1, 2.2.5, 2.2.12

Kullback, S. (1968). *Information Theory and Statistics*. Dover, NY, USA. 1.2.7, 2.6, 2.6.2

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86. 2.1, 2.6.2

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes.* Springer, New York, NY, USA. 336 p. 1.2.4, 1.2.5

Lee, P. M. (1997). *Bayesian Statistics. An introduction.* Arnold, London, GB. 344 p. 1.2.3, 2.5, 2.5

Lehmann, E. L. (1966). Some concepts of dependence. *Ann. of Math. Statist.*, 37(5):1137–1153. 1.2.6, 2.2.1

Lindskog, F., McNeil, A., and Schmock, U. (2003). Kendall's tau for elliptical distributions. In Bol, G., Nakhaeizadeh, G., Rachev, S., Ridder, T., and Vollmer, K.-H., editors, *Credit Risk: Measurement, Evaluation and Management*, pages 149–156. Physica-Verlag, A Springer-Verlag Company, Heidelberg, Germany. 2.1.1

Marshall, A. W. and Olkin, I. (1988). Families of multivariate distributions. *J. Am. Stat. Assoc.*, 83(403):834–841. 1.2.2

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. J. (2013). The Normal distribution in some constrained sample spaces. *SORT - Statistics and Operations Research Transactions*, 37(1):29–56. 5.1

Meeuwissen, A. M. H. and Bedford, T. (1997). Minimally informative distributions with given rank correlation for use in uncertainty analysis. *J. Stat. Comput. Simul.*, 57:143–174. 1.2.7, 7.5

Meng, X. L. (1994). Posterior predictive p-values. *Ann. Stat.*, 22:1142–1160. 2.5.2, 2.5.5

Mikosch, T. (2006). Copulas: Tales and facts. *Extremes*, 9:3–20. 2.1.4

Miller, D. J. and Liu, W.-H. (2002). On the recovery of joint distributions from limited information. *J. Econom.*, 107:259–274. 1.2.7, 7.5

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the theory of statistics*. McGraw-Hill, New York, NY, USA. 564 p. 1.2.6

Nataf, A. (1962). Détermination des distributions de probabilités dont les marges sont données. *C. R. Acad. Sci. Paris*, 255:42–42. 2.6.3

Nelsen, R. B. (1986). Properties of a one-parameter family bivariate distributions with specified marginals. *Commun. Statist. - Theory Meth.*, 15(11):3277–3285. 1.2.2

Nelsen, R. B. (1991). Copulas and association. In Dall'Aglio et al. (1991), pages 51–74. 1.2.7

Nelsen, R. B. (1996). Nonparametric measures of multivariate association. In Rüschendorf et al. (1996), pages 223–232. 2.2.2

Nelsen, R. B. (1998). Concordance and gini's measure of association. *Journal of Nonparametric Statistics*, 9:227 – 238. 2.2.3, 2.1

Nelsen, R. B. (1999). *An introduction to copulas*. Springer-Verlag, New York, NY, USA. 216 p. 1.2.1, 1.2.6, 2.1, 2.1.1, 2.1.2, 2.2.2, 4.1

Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables. *J. Multiv. Anal.*, 98:544–567. 2.2.2

Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Am. Stat. Assoc.*, 84(406):487–493. 1.2.2

Ortego, M. I., Egozcue, J. J., and Tolosana Delgado, R. (2014). Bayesian trend analysis of extreme wind using observed and hindcast series off catalan coast, NW mediterranean sea. *Nat. Hazards Earth Syst. Sci. Disc.*, 2:799–824. 13

Ortego, M. I. and Mateu-Figueras, G. (2006). Densidades de cópulas considerando la estructura de su espacio soporte. In Sicilia-Rodríguez, J., González-Martín, C., González-Sierra, M., and Alcaide-López, D., editors, *Actas XXIX Congreso Nacional de Estadística e Investigación Operativa*, pages 705–706, Tenerife (Spain). 5.1

Ortego, M. I., Tolosana-Delgado, R., Gibergans-Báguena, J., Egozcue, J. J., and Sánchez-Arcilla, A. (2012). Assessing wavestorm hazard evolution in the NW Mediterranean with hindcast and buoy data. *Climatic Change*, 113:713–731. 13

Panchenko, V. (2005). Goodness-of-fit test for copulas. *Physica A*, 355:176–182. b

Pasha, E. and Mansoury, S. (2008). Determination of maximum entropy multivariate probability distribution under some constraints. *Appl. Math. Sci.*, 2(57):2843–2849. 1.2.7

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stoch. Env. Res. Risk A. (SERRA)*, 15(5):384–398. 5.1, c

Pickands III, J. (1967). Sample sequences of maxima. *Ann. Math. Statist.*, 38:1570–1574. 1.2.5

Pickands III, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.*, 3(1):119–131. 2.4

Pickands III, J. (1986). The continous and differentiable domains of attractions of the extreme value distributions. *Ann. Probab.*, 14:996–1004. 1.2.5

Plackett, R. L. (1965). A class of bivariate distributions. *J. Am. Stat. Assoc.*, 60(310):516–522. 1.2.2

Reiss, R. D. (1989). *Approximate distributions of order statistics. With applications to nonparametric statistics*. Springer Verlag, Berlin, Germany. 355 p. 1.2.5

Reiss, R. D. (1993). *A Course on Point Processes*. Springer Verlag, New York, NY, USA. 1.2.4

Reiss, R. D. and Thomas, M. (1997). *Statistical Analysis of Extreme Values*. Birkhäuser, Germany. 462 p. 1.2.5, 1.3

Robert, C. P. (1994). *The Bayesian Choice. A decision-theoretic motivation*. Springer-Verlag, New York, NY, USA. 436 p. 2.5, 2.5

Robert, C. P. and Casella, G. (2000). *Monte Carlo Statistical Methods*. Springer, New York, NY, USA. 507 p. 2.5.1, 2.5.1, 6.3, 7.4

Robins, J. M., van der Vaart, A., and Ventura, V. (2000). Asymptotic distribution of p-values in composite null models. *J. Amer. Statist. Assoc.*, 95(452):1143–1156. 2.5.2, 8.1.2

Rodriguez-Iturbe, I., Gupta, V. K., and Waymire, E. (1984). Scale considerations in the modelling of temporal rainfall. *Water Resour. Res.*, 20(11):1611–1619. 1.2.4

Romero, R., J. A., G., Ramis, C., and Alonso, S. (1998). A 30-years (1964-93) daily rainfall data base for the spanish mediterranean regions: first exploratory study. *Int. J. Climatol.*, 18:541–560. 11.1, 11.1.1

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.*, 23:470–472. a

Rubin, D. B. (1984). Bayesianly justicable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12:1151–1172. 2.5.2, 2.5.4

Rumsey Jr, H. and Posner, E. C. (1965). Joint distributions with prescribed moments. *Ann. Math. Statist.*, 1(36):286–298. 1.2.7, 2.6, 2.6.3, 5.2, 7.5

Rüschendorf, L., Schweizer, B., and Taylor, M. D., editors (1996). *Distributions with Fixed Marginals and Related Topics*, volume 28 of *Lecture notes - Monograph Series*. Institute of Mathematical Statistics. 2.1.2, 3, 13

Savage, L. J. (1972). *The Foundations of Statistics*. Dover, New York, NY, USA. 310 p. 1.2.3

Scaillet, O. (2007). Kernel-based goodness-of-fit tests for copulas with fixed smoothing parameters. *J. Multiv. Anal.*, 98:533–543. b

Scarsini, M. (1984). On measures of concordance. *Stochastica*, 8(3):201–218. 1.2.6, 2.2.2, 2.2.3, 2.2.2

Schmid, F. and Schmidt, R. (2007). Nonparametric inference on multivariate versions of blomqvist's beta and related measures of tail dependence. *Metrika*, 66(3):323–354. 2.2.2

Schmid, F., Schmidt, R., Blumentritt, T., Gaißer, S., and Ruppert, M. (2010). Copula-based measures of multivariate association. In Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., editors, *Copula Theory and Its Applications.*, pages 209–236, Berlin. Springer-Verlag. 2.2.2, 5.1

Schweizer, B. (1991). Thirty years of copulas. In Dall'Aglio et al. (1991), pages 13–50. 2.2.7

Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *Ann. Statist.*, 9(4):870–885. 1.2.6, 2.1, 2.2.2, 2.2.2, 2.2.2, 2.2.2, 2.2.8, 2.2.2, 2.2.2

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System. Tech. J.*, (27):379–423. 2.6.1, 2.6.4

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8(1):229–231. 1.2.1, 2.1, 2.1

Smith, M. D. (2007). Invariance theorems for fisher information. *Commun. Statist. - Theory Methods*, 36(12):2213 – 2222. 7.5

Sotillo, M. G., Ratsimandresy, A. W., Carretero, J. C., Bentamy, A., Valero, F., and Gonzalez-Rouco, F. (2005). A high-resolution 44-year atmospheric hindcast for the Mediterranean basin: contribution to the regional improvement of global reanalysis. *Clim. Dyn.*, 25:219–236. 12.1

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.*, 15(1):72–101. 1.2.6, 2.2.5, 2.1

Tanner, M. A. (1993). *Tools for Statistical Inference*. Springer-Verlag, New York, NY, USA. 152 p. 2.5.1

Tawn, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3):397–415. 1.2.2

Taylor, M. D. (2007). Multivariate measures of concordance. *AISM*, 59:789–806. 2.2.2

Tiago de Oliveira, J. (1984). *Statistical Extremes and Applications*. Dordrecht Reidel, Boston, MA, USA. 692 p. 1.2.5

Tolosana-Delgado, R., Ortego, M. I., Egozcue, J. J., and Sánchez-Arcilla, A. (2010). Climate change in a point-over-threshold model: an example on ocean-wave-storm hazard in NE spain. *Adv. in Geosciences*, 26:113–117. 11.1, 12.1

Von Mises, R. (1923). Über der Variationsbreite einer Beobachtungsreihe. *Sitzungsberichte Berliner Math. Ges.*, 22:3–8. 1.2.5

Von Mises, R. (1936). La distribution de la plus grande de n valeurs. In *Selected Papers of Richard Von Mises*, volume 2, pages 141–160, USA. Amer. Math. Soc. 1.2.5

WAMDIGroup (1988). The WAM model : a third generation ocean wave prediction model. *J. Phys. Oceanogr.*, 18:1775–1810. 12.1

Whitt, W. (1976). Bivariate distributions with given marginals. *Ann. Statist.*, 4:1280–1289. 2.2.1, 2.6

Zhao, N. and Lin, W. T. (2011). A copula entropy approach to correlation measurement at the country level. *Appl. Math. Comput.*, 218(2):628 – 642. 7.5

Apéndice A

Otros p-valores de interés

A continuación se presentan algunos p-valores desarrollados en la literatura. Pese a ser de interés, no se han implementado en los desarrollos posteriores de este trabajo.

Definición A.0.1. *p-valor predictivo a priori* (p_{prior}) (Box, 1980):

$$p_{prior} = \mathrm{P}^{m(\cdot)}[t(\mathbf{X} \geq t(\mathbf{x}_{obs})] \ ,$$

donde $m(\mathbf{x})$ es la *distribución predictiva a priori*,

$$m(\mathbf{x}) = \int f(\mathbf{x};\theta)\pi(\theta)d\theta \ ,$$

y $\pi(\theta)$ es la *distribución a priori* establecida para el parámetro θ.

Definición A.0.2. *partial posterior predictive p-value* (p_{ppost}) (Bayarri and Berger, 2000):

$$p_{ppost} = \mathrm{P}^{m_{post}(\cdot|\mathbf{x}_{obs}\backslash t_{obs})}[T \geq t(\mathbf{x}_{obs})] \ ,$$

donde $m(t|\mathbf{x}_{obs}\backslash t_{obs})$ is

$$m(t|\mathbf{x}_{obs}\backslash t_{obs}) = \int f(t|\theta)\pi(\theta|\mathbf{x_{obs}}\backslash t_{obs})d\theta \ ,$$

y

$$\pi(\theta|\mathbf{x}_{obs}\backslash t_{obs}) \propto f(\mathbf{x}_{obs}|t_{obs};\theta)\pi(\theta) \propto \frac{f(\mathbf{x}_{obs};\theta)\pi(\theta)}{f(t_{obs};\theta)}.$$

es el *posteriori parcial*.

Definición A.0.3. *U-conditional predictive p-value* ($p_{cpred}(u)$), (Bayarri and Berger, 2000):

para algún estadístico condicional $U = u(\mathbf{x})$

$$p_{cpred(u)} = \mathrm{P}^{m(\cdot|u_{obs})}[T \geq t(\mathbf{x}_{obs})] \quad ,$$

donde $u_{obs} = u(\mathbf{x}_{obs})$, $t_{obs} = t(\mathbf{x}_{obs})$ y $m(t|u)$ es

$$m(t|u) = \int f(t|u;\theta)\pi(\theta|u)d\theta \ ,$$

suponiendo que

$$\pi(\theta|u) = \frac{f(u;\theta)\pi(\theta)}{\int f(u;\theta)}\pi(\theta)d\theta$$

es propio. Cabe notar que $f(t|u;\theta)$ y $f(u;\theta)$ están definidas como las densidades condicional y marginal de T y U bajo H_0 respectivamente.

Un caso particular, es el siguiente: para datos continuos, se escoge como U al estimador máximo verosímil condicional de θ, dado $t(\mathbf{x}) = t$:

$$\widehat{\theta}_{cMLE}(\mathbf{x}) = \arg\max f(\mathbf{x}\ t, \theta) = \arg\max \frac{f(\mathbf{x};\theta)}{f(t;\theta)}$$

El p-valor resultante recibe el nombre de *simply conditional predictive* p-value.

Cabe remarcar que cuando T es condicionalmente independiente de $\widehat{\theta}_{cMLE}$ y $(T, \widehat{\theta}_{cMLE})$ son suficientes conjuntamente, entonces $p_{ppost} = p_{cpred}$.

Apéndice B

Lluvia en dos ubicaciones. Valores de interés a posteriori.

B.1. Vall de Laguard y Almudaina. Dependencia con cópula CrEnC. Bondad de ajuste marginal y global.

Se evalúa la coherencia de distintos aspectos del modelo con los datos de precipitación diaria en Vall de Laguard y Almudaina mediante los estadísticos de contraste implementados y sus correspondientes p-valores a posteriori, descritos en la Sec. 8.1. En la Tabla B.1 se presentan algunos percentiles seleccionados de los p-valores a posteriori correspondientes a la medida de bondad de ajuste del modelo GPD marginal. La hipótesis de marginal GPD no puede rechazarse para los excesos de precipitación en Almudaina, aunque existen dudas sobre su validez para los excesos originales de Vall de Laguard. Se ha escogido un único umbral de referencia para todas las estaciones, pero éste parece no ser suficientemente alto para que la distribución GPD ajuste correctamente los datos de la estación Vall de Laguard (ver Fig. B.1). En cambio, los p-valores predictivos a posteriori (ver Tabla B.1) indican que el modelo GPD sí es adecuado para las remuestras de

ambas ubicaciones, proporcionando una valoración indirecta del buen funcionamiento global del modelo.

Los estadísticos de *Slope*, pendiente de la recta de excesos esperados (Tabla B.2) permiten verificar la hipótesis adicional de distribución GPD en el dominio de Weibull, $\xi < 0$. No se descarta la coherencia de esta hipótesis con los excesos originales y con las correspondientes remuestras. La Tabla B.3 muestra que la dependencia presente en los excesos originales es ligeramente inferior a la presente en las remuestras generadas a partir del posteriori. Por tanto, el modelo es globalmente coherente con los datos y representa correctamente tanto el comportamiento marginal como el de dependencia.

Tabla B.1: Bondad de ajuste marginal para Vall de Laguard y Almudaina. Percentiles seleccionados de los p- valores a posteriori (Sup.). Percentiles seleccionados de los p-valores predictivos a posteriori (Inf.)

	p-val.	2.5 %	50 %	97.5 %
Vall de Laguard	K-S	0e+00	5e-05	2e-04
Vall de Laguard	Multinomial	0.00000	0.02104	0.05876
Resample Vall de Laguard	K-S	0.01201	0.44744	1.00000
Resample Vall de Laguard	Multinomial	0.00367	0.42780	0.94101
Almudaina	K-S	0.03844	0.62705	0.98082
Almudaina	Multinomial	0.12614	0.34342	0.82059
Resample Almudaina	K-S	0.00060	0.21270	1.00000
Resample Almudaina	Multinomial	0.00268	0.35277	0.94471

Tabla B.2: Contraste sobre la validez del priori GPD en DA-Weibull para cada marginal. Las pendientes de la recta de regresión del exceso esperado son similares en la muestra original y para las remuestras correspondientes a la muestra del posteriori.

Value	X_1	X_2
Slope orig.	-0.290276	-0.225290
Slope predictive	-0.182274	-0.113882
Slope Discrepancy	0.978351	0.951546

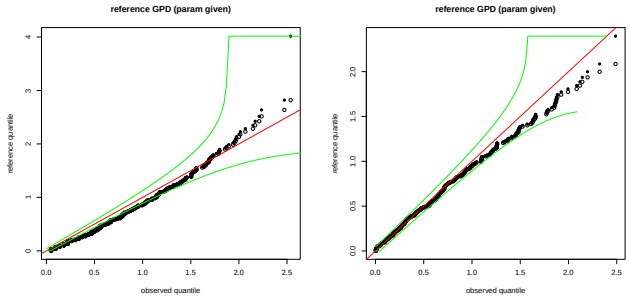

Figura B.1: QQ plot de los excesos de log-precipitación por encima de log(20) respecto a la distribución GPD con parámetros estimados por máxima verosimilitud (estima preliminar). Izquierda: Vall de Laguard; derecha: Almudaina.

Tabla B.3: Contraste sobre el modelo global. Coeficiente τ original, coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. Los coeficientes muestran una dependencia ligeramente inferior en los datos originales que en las remuestras. El p-valor basado en la discrepancia τ no permite rechazar la hipótesis de validez del modelo global especificado.

τ original	τ PredictivoG	Discrepancia τ
0.297579	0.461554	0.998969

B.2. Vall de Laguard y Almudaina. Valores de precipitación a posteriori

Una vez descritos los resultados de la estimación de los parámetros del modelo e interpretada su bondad de ajuste a los datos, podemos obtener cantidades a posteriori y predictivos a posteriori de interés, tanto marginales como conjuntas.

Dados valores de precipitación diaria de referencia se han obtenido las probabilidades predictivas a posteriori y sus correspondientes cuantiles (Tablas B.4 y B.5), sus periodos de retorno a posteriori (Tablas B.6 y B.7) y los valores a posteriori asociados a periodos de retorno de referencia (Tabla B.8 y B.9).

Además se han calculado probabilidades de no excedencia conjuntas predictivas a posteriori de valores de referencia y sus percentiles (Tabla B.10). Se observa que el modelo proporciona estimaciones de probabilidades de combinaciones de valores de probabilidad no observados en la serie registrada.

Tabla B.4: Probabilidades de no excedencia a posteriori para valores seleccionados de los excesos en Vall de Laguard. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1ref	Mean	GeomMean	q0.05	q0.5	q0.95
0.405 (40mm)	0.4038	0.4038	0.3944	0.4006	0.4302
0.916 (50mm)	0.7236	0.7239	0.7054	0.7208	0.7651
1.500 (90mm)	0.9093	0.9104	0.8881	0.9076	0.9446
1.610 (100mm)	0.9293	0.9307	0.9089	0.9284	0.9615
2.020 (150mm)	0.9762	0.9782	0.9623	0.9763	0.9942
2.305 (200mm)	0.9908	0.9930	0.9818	0.9915	0.9996

Tabla B.5: Probabilidades de no excedencia a posteriori para valores seleccionados de los excesos en Almudaina. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x2ref	Mean	GeomMean	q0.05	q0.5	q0.95
0.405 (40mm)	0.4159	0.4158	0.3980	0.4137	0.4430
0.916 (50mm)	0.7353	0.7356	0.7133	0.7344	0.7614
1.500 (90mm)	0.9146	0.9152	0.8951	0.9149	0.9322
1.610 (100mm)	0.9336	0.9343	0.9153	0.9345	0.9484
2.020 (150mm)	0.9776	0.9786	0.9646	0.9782	0.9869
2.305 (200mm)	0.9912	0.9924	0.9822	0.9919	0.9972

Tabla B.6: Periodo de retorno correspondiente a valores seleccionados de los excesos en Vall de Laguard. Periodo de retorno a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1tau	Mean	GeomMean	q0.05	q0.5	q0.95
0.500 (33mm)	1.9149	1.9144	1.8731	1.9012	2.0338
0.916 (50mm)	3.6308	3.6241	3.3950	3.5816	4.2574
1.500 (90mm)	11.3783	11.1825	8.9358	10.8231	18.0556
1.610 (100mm)	14.8356	14.4483	10.9757	13.9612	25.9494
2.020 (150mm)	53.4331	45.9732	26.5419	42.2158	171.5708
2.302 (200mm)	309.0918	141.3130	54.3611	116.1326	2200.4593

Tabla B.7: Periodo de retorno correspondiente a valores seleccionados de los excesos en Almudaina. Periodo de retorno a posteriori: media, centro y cuantiles predictivoa a posteriori seleccionados.

x2tau	Mean	GeomMean	q0.05	q0.5	q0.95
0.500 (33mm)	1.9641	1.9633	1.8888	1.9541	2.0788
0.916 (50mm)	3.7893	3.7835	3.4884	3.7653	4.1912
1.500 (90mm)	11.8961	11.8051	9.5334	11.7573	14.7601
1.610 (100mm)	15.3928	15.2251	11.8004	15.2623	19.3808
2.020 (150mm)	49.1015	46.7620	28.2150	45.8990	76.4336
2.302 (200mm)	155.0483	129.7296	55.9062	121.3462	346.2649

Tabla B.8: Excesos en Vall de Laguard correspondientes a periodos de retorno seleccionados. Excesos a posteriori: media, centro y cuantiles seleccionados.

tiempo	Mean	GeomMean	q0.05	q0.5	q0.95
10.	1.4568	1.4552	1.2912	1.4627	1.5566
20.	1.7526	1.7501	1.5313	1.7529	1.8941
50.	2.0721	2.0678	1.7792	2.0727	2.2700
100.	2.2692	2.2635	1.9253	2.2646	2.5160
400.	2.5741	2.5652	2.1384	2.5638	2.9174
500.	2.6141	2.6046	2.1649	2.6023	2.9708

Tabla B.9: Excesos en Almudaina correspondientes a periodos de retorno seleccionados. Excesos a posteriori: media, centro y cuantiles seleccionados.

tiempo	Mean	GeomMean	q0.05	q0.5	q0.95
10.	1.4279	1.4269	1.3353	1.4251	1.5237
20.	1.7279	1.7263	1.6199	1.7182	1.8662
50.	2.0575	2.0547	1.9127	2.0477	2.2573
100.	2.2648	2.2607	2.0767	2.2512	2.5199
400.	2.5934	2.5856	2.3207	2.5670	2.9502
500.	2.6374	2.6289	2.3496	2.6093	3.0087

Tabla B.10: Probabilidades de excedencia conjuntas a posteriori para valores seleccionados de los excesos en Vall de Laguard y Almudaina. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1refconj	x2refconj	Mean	GeomMean	q0.05	q0.5	q0.95
0.405 (40mm)	0.405 (40mm)	0.4358	0.4341	0.3729	0.4392	0.5047
0.916 (50mm)	0.916 (50mm)	0.1672	0.1647	0.1196	0.1630	0.2182
1.500 (90mm)	1.610 (100mm)	0.0476	0.0443	0.0193	0.0470	0.0746
1.610 (100mm)	1.500 (90mm)	0.0492	0.0461	0.0249	0.0470	0.0746
2.020 (150mm)	2.305 (200mm)	0.0135	0.0106	0.0028	0.0138	0.0304
2.305 (200mm)	2.020 (150mm)	0.0134	0.0106	0.0028	0.0138	0.0304

B.3. Vergel de Recons y Simat de Valldigna. Dependencia con cópula CrEnc. Bondad de ajuste marginal y global.

Se valora la coherencia de varios aspectos del modelo con los datos de precipitación diaria en Vergel de Recons y Simat de Valldigna mediante los estadísticos de contraste presentados en la Sección 8.1. En la Tabla B.11 se presentan percentiles a posteriori seleccionados de los p-valores correspondientes a la medida de bondad de ajuste del modelo GPD marginal. Para los excesos de Vergel de Recons no puede rechazarse la hipótesis de marginal GPD, pero este modelo debe descartarse para los datos de Simat de Valldigna. El umbral único escogido para todas las ubicaciones no es suficientemente alto para que la distribución sea GPD en esta ubicación. Los p-valores predictivos a posteriori indican que la distribución sí es adecuada para las remuestras de ambas ubicaciones, indicando indirectamente que el modelo establecido funciona globalmente bien. La hipótesis adicional de distribución GPD en el dominio de Weibull ($\xi < 0$ y soporte acotado) se verifica mediante los estadísticos de *Slope*, pendiente de la recta de excesos esperados (Tabla B.12), mediante los cuales no puede descartarse que esta hipótesis sea coherente para los excesos originales y sus correspondientes remuestras. Respecto al ajuste global del modelo, la Tabla B.13 muestra que la dependencia presente en los excesos originales es ligeramente superior a la presente en las remuestras generadas a partir del posteriori. El modelo es globalmente coherente con los datos, aunque mejorable en algunos aspectos.

Tabla B.11: Bondad de ajuste marginal para Vergel de Recons y Simat. Percentiles seleccionados de los p- valores a posteriori (Sup.). Percentiles seleccionados de los p-valores predictivos a posteriori (Inf.)

	p-val.	2.5 %	50 %	97.5 %
Vergel de Recons	K-S	0.01298	0.08856	0.16337
Vergel de Recons	Multinomial	0.02295	0.11601	0.29257
Resample Vergel de Recons	K-S	0.01181	0.46284	1.00000
Resample Vergel de Recons	Multinomial	0.01712	0.43978	0.97320
Simat	K-S	0e+00	3e-05	1e-04
Simat	Multinomial	0.00000	0.00002	0.00072
Resample Simat	K-S	0.04009	0.49518	1.00000
Resample Simat	Multinomial	0.00941	0.47337	0.95166

Tabla B.12: Pendientes originales y predictivas de la recta de regresión de los excesos esperados para cada una de las marginales. p-valor de discrepancia correspondiente al contraste sobre la validez del priori GPD en DA-Weibull para cada marginal.

Value	X_1	X_2
Slope orig.	-0.195068	-0.316375
Slope predictive	-0.192244	-0.214675
Slope Discrepancy	0.510309	0.978351

Tabla B.13: Contraste sobre el modelo global. Coeficiente τ original, coeficiente predictivo obtenido mediante la muestra del posteriori de Gibbs. Los coeficientes muestran una dependencia ligeramente superior en los datos originales que en las remuestras. El p-valor basado en la discrepancia τ no permite rechazar la hipótesis de validez del modelo global especificado.

τ original	τ PredictivoG	Discrepancia τ
0.243269	0.069737	0.001031

B.4. Vergel de Recons y Simat de Valldigna. Valores de precipitación a posteriori

Una vez descritos los resultados de la estimación de los parámetros del modelo e interpretada su bondad de ajuste a los datos, podemos obtener cantidades a posteriori y predictivas a posteriori de interés, tanto marginales como conjuntas.

Dado un conjunto de valores de precipitación diaria de referencia se han obtenido sus probabilidades a posteriori y los correspondientes cuantiles (Tablas B.14 y B.15), sus periodos de retorno a posteriori (Tablas B.16 y B.17) y los valores a posteriori asociados a periodos de retorno de referencia (Tabla B.18 y B.19).

Además se han calculado probabilidades de no excedencia conjuntas a posteriori de valores de referencia y sus percentiles (Tabla B.20). Se observa que el modelo proporciona estimaciones de probabilidades de combinaciones de valores de probabilidad no observados en la serie registrada.

Tabla B.14: Probabilidades de no excedencia a posteriori para valores seleccionados de los excesos en Vergel. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1ref	Mean	GeomMean	q0.05	q0.5	q0.95
0.405 (40mm)	0.4094	0.4093	0.3957	0.4065	0.4336
0.916 (50mm)	0.7259	0.7262	0.7076	0.7235	0.7519
1.500 (90mm)	0.9068	0.9075	0.8893	0.9062	0.9248
1.610 (100mm)	0.9264	0.9271	0.9097	0.9258	0.9433
2.020 (150mm)	0.9730	0.9740	0.9600	0.9732	0.9833
2.305 (200mm)	0.9884	0.9896	0.9784	0.9890	0.9947

Tabla B.15: Probabilidades de no excedencia a posteriori para valores seleccionados de los excesos en Simat. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x2ref	Mean	GeomMean	q0.05	q0.5	q0.95
0.405 (40mm)	0.4007	0.4007	0.3955	0.3999	0.4081
0.916 (50mm)	0.7205	0.7205	0.7085	0.7208	0.7320
1.610 (100mm)	0.9285	0.9289	0.9134	0.9292	0.9399
1.500 (90mm)	0.9081	0.9084	0.8927	0.9088	0.9201
2.020 (150mm)	0.9764	0.9771	0.9651	0.9773	0.9846
2.305 (200mm)	0.9913	0.9921	0.9835	0.9921	0.9963

Tabla B.16: Periodo de retorno a posteriori correspondiente a valores seleccionados de los excesos en Vergel. Periodo de retorno a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1tau	Mean	GeomMean	q0.05	q0.5	q0.95
0.500 (33mm)	1.9350	1.9350	1.8790	1.9230	2.0390
0.916 (50mm)	3.6590	3.6540	3.4200	3.6160	4.0310
1.500 (90mm)	10.9100	10.8100	9.0330	10.6600	13.3000
1.610 (100mm)	13.9100	13.7300	11.0800	13.4800	17.6400
2.020 (150mm)	40.7300	38.5600	25.0000	37.2800	59.9500
2.302 (200mm)	115.1000	94.8400	46.0600	89.6700	184.8000

Tabla B.17: Periodo de retorno a posteriori correspondiente a valores seleccionados de los excesos en Simat. Periodo de retorno a posteriori: media, centro y cuantiles a posteriori seleccionados.

x2tau	Mean	GeomMean	q0.05	q0.5	q0.95
0.500 (33mm)	1.902	1.902	1.877	1.899	1.932
0.916 (50mm)	3.580	3.579	3.430	3.582	3.731
1.500 (90mm)	10.960	10.920	9.321	10.960	12.520
1.610 (100mm)	14.150	14.070	11.550	14.120	16.640
2.020 (150mm)	45.020	43.700	28.650	44.120	64.800
2.302 (200mm)	138.700	125.500	60.010	124.800	264.100

Tabla B.18: Excesos a posteriori en Vergel correspondientes a periodos de retorno seleccionados. Excesos a posteriori: media, centro y cuantiles seleccionados.

tiempo	Mean	Geommea	q0.05	q0.5	q0.95
10.	1.4670	1.4660	1.3670	1.4680	1.5540
20.	1.7790	1.7780	1.6540	1.7810	1.9100
50.	2.1250	2.1220	1.9660	2.1230	2.3310
100.	2.3440	2.3390	2.1570	2.3330	2.6030
400.	2.6920	2.6850	2.4370	2.6690	3.0580
500.	2.7390	2.7320	2.4700	2.7140	3.1210

Tabla B.19: Excesos a posteriori en Simat correspondientes a periodos de retorno seleccionados. Excesos a posteriori: media, centro y cuantiles seleccionados.

tiempo	Mean	GeomMean	q0.05	q0.5	q0.95
10.	1.4620	1.4620	1.4070	1.4580	1.5360
20.	1.7560	1.7550	1.6750	1.7480	1.8670
50.	2.0720	2.0700	1.9530	2.0590	2.2340
100.	2.2650	2.2630	2.1190	2.2460	2.4680
400.	2.5630	2.5580	2.3640	2.5390	2.8430
500.	2.6010	2.5970	2.3950	2.5780	2.8930

Tabla B.20: Probabilidades de excedencia conjuntas a posteriori para valores seleccionados de los excesos en Vergel y Simat. Probabilidades a posteriori: media, centro y cuantiles a posteriori seleccionados.

x1refconj	x2refconj	Mean	GeomMean	q0.05	q0.5	q0.95
0.405 (40mm)	0.405 (40mm)	0.3667	0.3641	0.2959	0.3707	0.4320
0.916 (50mm)	0.916 (50mm)	0.0932	0.0897	0.0510	0.0918	0.1326
1.500 (90mm)	1.610 (100mm)	0.0158	0.0127	0.0034	0.0170	0.0306
1.610 (100mm)	1.500 (90mm)	0.0162	0.0132	0.0034	0.0170	0.0306
2.020 (150mm)	2.305 (200mm)	0.0047	0.0041	0.0034	0.0034	0.0102
2.305 (200mm)	2.020 (150mm)	0.0047	0.0042	0.0034	0.0034	0.0102