UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS MATEMÁTICAS

TESIS DOCTORAL

Contrastes de Hipótesis Múltiples Bajo Dependencia con Aplicación a los Microarrays: Una Aproximación Bayesiana

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Elisa da Conceição José María

Directores

Luis Sanz San Miguel
María Isabel Salazar Mendoza

Madrid

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

TESIS DOCTORAL

Contrastes de Hipótesis Múltiples Bajo Dependencia con Aplicación a los Microarrays: Una Aproximación Bayesiana

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Elisa da Conceição José Maria

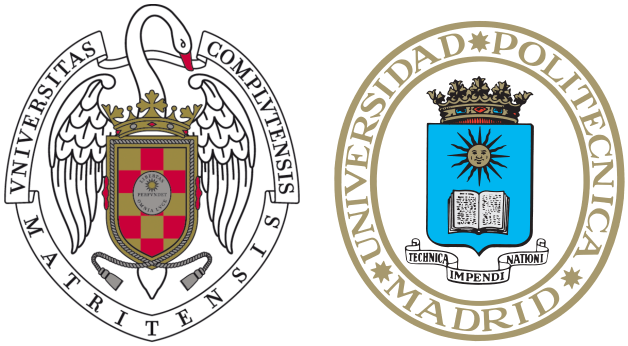
DIRECTORES

Luis Sanz San Miguel (UCM)

María Isabel Salazar Mendoza (UCM)

Año 2020

Programa de Doctorado en Ingeniería Matemática,
Estadística e Investigación Operativa por la
Universidad Complutense de Madrid y la
Universidad Politécnica de Madrid



Contrastes de Hipótesis Múltiples Bajo Dependencia con Aplicación a los Microarrays: Una Aproximación Bayesiana

Tesis Doctoral

Elisa da Conceição José Maria

Directores

Luis Sanz San Miguel (UCM)

María Isabel Salazar Mendoza (UCM)

Año 2020

Dedicatoria

A mi hijo Rayen,
a mis padres Carlos (paz a su alma) y Julieta
y a mis hermanos Breslau, Carlos, Luís y Ana.

Agradecimientos

Son tantas las personas que merecen mi agradecimiento que algunas de ellas no recibirán el debido reconocimiento, independientemente de lo que escriba. Aquí solo se reflejan algunas de las personas más destacadas que han jugado un papel importante en la configuración de mi carrera académica y profesional.

A mis directores de tesis, María Isabel Salazar Mendoza y Luiz Sanz San Miguel, personas excepcionales, de una paciencia inigualable y siempre de buen humor, quiero expresarles mi gratitud por compartir conmigo, de forma sabia e incansable, su experiencia para que este trabajo llegara a término, prestándome un gran apoyo tanto en el plano profesional como en el personal, transmitiéndome, durante el periodo predoctoral, la ilusión y el ánimo necesarios para lograr terminar este trabajo de investigación. Gracias de corazón, por su dedicación, por la confianza que han depositado en mí y, principalmente, por su paciencia frente a mis limitaciones. Toda la entrega y preocupación conmigo hacen de ellos, más que directores, amigos. Hago extensible esta gratitud al profesor Miguel Ángel Gómez Villegas, por brindarme durante el transcurso de esta investigación su gran experiencia, sus aportaciones y consejos, su pronta disponibilidad y el tiempo que dedicó para poder sacar adelante esta tesis.

Una mención especial se merece, sin duda alguna, la profesora Begoña Vitoriano por su apoyo incondicional, por sus ideas, su tiempo y por facilitarme todos los recursos a su alcance para convertirme en una investigadora. Nada de lo que escriba aquí reflejaría fielmente lo agradecida que me siento por todo lo que ha echo por mí.

Al Departamento de Estadística e Investigación Operativa y a todos los profesores que, directa o indirectamente, han facilitado mi estancia en Madrid. Entre ellos, quiero agradecer de forma especial a Tinguaro, Pilar, José, Rosa, Javier Martín y Juan Tejada por la motivación.

A todos mis compañeros del Departamento de Estadística e Investigación Operativa, Diana, por su disponibilidad y paciencia para enseñarme los misterios del R,

Javier, por su apoyo con Latex, Adán, Ahinara, Elena, Inma, Pablo y Víctor, por dejarme marearles con mis dudas, Bibiana, Camilo, Carlos y Paula, por los buenos ratos pasados en el despacho 205 y en las comidas, momentos que facilitaron mi estancia en Madrid. Compartir despacho con vosotros ha sido muy enriquecedor, tanto en lo científico como en lo personal.

Escribir sobre mi hijo Rayen, me rompe el corazón, genera en mí una mezcla de sentimientos, culpa y alegría, pero no puedo dejar de agradecerle la paciencia y el esfuerzo para comprender sin aceptar la ausencia de su mamá pese a las ganas de tenerme cerca, y a mi familia por cuidar de él durante mi ausencia.

Por supuesto, no puedo olvidarme de mis amigos Arantxa e Ivan y sus familias, que siempre han estado a mi lado dispuestos para ayudarme a escapar del estrés y demostrándome que en esta vida no solo es importante el trabajo, a Ana y Mina, a quienes he tenido la suerte de conocer recientemente, a mi cuñada Ángela y a mi tía Virginia, por la disponibilidad para escuchar mis desahogos.

A Sófi, por su cariño, a Iñaki y Guille, por haberme recibido con los brazos abiertos en sus casas.

Finalmente, quiero agradecer la financiación del proyecto European Mathematical Society-Simons for Africa for a Collaborative Researsh visit (type A2) and B Top-up grant for women, del proyecto de Cooperación al Desarrollo de la Universidad Complutense de Madrid (El máster de Estadística en Mozambique: Consolidación, Transferencia y Fortalecimiento Institucional), la ayuda recibida de la Universidad Complutense de Madrid a través de la financiación - GR29/20 y del grupo HUMLOG a través de la financiación - GR97064 y la beca de la Universidad Rovuma de Mozambique.

Índice general

Resumen ix

Abstract xiii

Prólogo xvii

1. Introducción 1

1.1. Tecnología Microarrays de ADN 2

1.2. El problema de los Contrastes de Hipótesis Múltiples Frecuentistas . 4

1.3. Paradigma Bayesiano 8

1.4. Modelos Ocultos de Markov (HMM) 12

1.5. Procedimientos de Agrupamiento 14

2. Contraste de Hipótesis Múltiples. Modelo de Dependencia Mediante Cópulas N-Variantes 17

2.1. Funciones Cópulas 18

2.1.1. Método de Inversión de Cópulas 20

2.1.2. Densidad de Probabilidad Asociada a las Cópulas 21

2.1.3. Cópulas y Dependencia 22

2.1.4. Funciones Cópulas Comunes 24

2.1.5. Estructura de Correlación Uniforme para la Cópula Gaussiana 27

2.2. Planteamiento del Problema y Modelo General con Cópulas N-variantes 29

2.3. Inferencia Basada en los Métodos de Cadenas de Markov Monte Carlo (MCMC) 33

3. Aplicación Mediante Cópulas Gaussianas y de Clayton **37**

3.1. Modelo con Cópulas Gaussianas N-variantes y Funciones de Distribución Marginales Normales 38

3.1.1. Distribuciones a Priori y a Posteriori para los Parámetros del Modelo 41

3.1.2. Determinación de Distribuciones Condicionadas a Posteriori . 44

3.1.3. Algoritmo MCMC 47

3.1.4. Simulación. Resultados 51

3.1.5. Aplicación a Datos Reales 56

3.2. Modelo con Cópulas de Clayton N-variantes y Funciones de Distribución Marginales Normales 59

3.2.1. Simulación. Resultados 62

3.2.2. Aplicación a Datos Reales 65

3.2.3. Selección de Modelos 66

4. Conclusiones y Extensiones **69**

4.1. Conclusiones 69

4.2. Futuras Investigaciones 71

Apéndice A. Algoritmo MCMC: Metropolis-Hastings-within-Gibbs **73**

Apéndice B. Distribuciones condicionadas a posteriori **81**

Bibliografía **89**

Resumen

Muchos experimentos requieren contrastar simultáneamente un elevado número de hipótesis. Un ejemplo son los experimentos con microarrays de ADN en el campo de la genómica, donde es habitual analizar simultáneamente un gran número de genes con la finalidad de identificar cuáles de ellos se expresan de manera diferencial bajo dos condiciones experimentales.

Uno de los problemas que se presentan en este contexto, además de la necesidad de contrastar simultáneamente un elevado número de hipótesis, uno para cada gen, es cómo modelar la dependencia que suele existir en el nivel de expresión entre los genes.

La literatura al respecto es extensa, sin embargo, los procedimientos propuestos, en gran parte desde un punto de vista frecuentista y bajo el supuesto de independencia, no resuelven, en general, todos los problemas planteados anteriormente.

El objetivo principal de esta tesis es la identificación de genes con expresión diferencial bajo dos condiciones de tratamiento distintas e independientes y bajo dependencia en el nivel de expresión de los genes. Para ello, se propone un procedimiento bayesiano en el que la dependencia se modela mediante funciones cópulas.

El procedimiento se aplica a contrastes de igualdad de medias y los datos se supone que provienen de distribuciones marginales normales considerando, para cada variable, varianzas iguales en las dos condiciones de tratamiento y distintas entre las variables. Igualmente, se considera la misma estructura de dependencia para las dos condiciones de tratamientos y se modela, en primer lugar, mediante la cópula Gaussiana de la familia Elíptica, considerando la matriz de correlación uniforme con la finalidad de reducir la alta dimensión del espacio paramétrico. También se modela

la dependencia mediante la cópula de Clayton, de la familia Arquimediana, que permite modelar distribuciones multivariantes con dependencia asimétrica, mediante una única función univariante, reduciendo también de esta manera el elevado número de parámetros.

La metodología que se propone se ilustra con datos simulados de una normal multivariante y con datos reales procedentes de experimentos con microarrays de ADN. En ambos casos, se comparan los modelos con cópulas Gaussianas y los modelos con cópulas de Clayton mediante el criterio de selección DIC, asimismo, se estima el valor del FDR con el objetivo de evaluar también la proporción de falsos positivos.

A partir de los resultados obtenidos podemos concluir que, en cuanto a la estimación del parámetro de dependencia de ambas copulas, el procedimiento es robusto frente a la elección de los parámetros de la distribución a priori de la probabilidad inicial de cada hipótesis nula. En efecto, se obtienen valores muy similares y próximos al valor del parámetro con el que se generaron los datos, en todos los ejemplos simulados y con ambos tipos de cópulas.

En cuanto al número de hipótesis nulas rechazadas y aceptadas, el procedimiento no es robusto respecto a la elección de dichas distribuciones, ya que este número resulta ser muy distinto según la distribución a priori utilizada, obteniendo un número mayor de aciertos cuando se utilizan distribuciones a priori betas sesgadas a la derecha.

De acuerdo con el criterio DIC, el modelo con cópulas Gaussianas resultó ser el más adecuado para los datos simulados, como se esperaba ya que los datos simulados se habían generado de una distribución normal. Asimismo, el número de hipótesis nulas rechazadas con este modelo es más próximo al verdadero número de hipótesis nulas falsas que el que se obtiene con el modelo utilizando cópulas de Clayton. Además, con el modelo de cópulas Gaussianas se rechaza un número de hipótesis nulas ligeramente superior al número de hipótesis nulas falsas, mientras que el valor estimado del FDR se mantiene en niveles aceptables y es significativamente menor que el obtenido con el modelo de cópulas de Clayton. Por tanto, el procedimiento propuesto funciona bien, cuando la dependencia se modela mediante la función cópula

más adecuada.

Por otro lado, el procedimiento que se propone es flexible en la medida en que puede utilizarse con otras matrices de correlación, o con otras funciones cópulas para modelar la dependencia, así como con otras funciones de distribuciones marginales.

Finalmente, la literatura en la que se utilizan las funciones cópulas, en el contexto de los contrastes de hipótesis múltiples, para modelar la dependencia es escasa, en la que solo hemos encontrado trabajos que utilizan estas funciones para modelar la dependencia entre un bajo número de variables y con un enfoque frecuentista, por lo que el procedimiento que se propone en esta tesis resulta fundamental, especialmente en el campo de la genómica.

Abstract

Many experiments require the simultaneous testing of a large number of hypotheses. An example of this, are the experiments with DNA microarrays in the field of genomics, where it is usual to simultaneously analyze a large number of genes in order to identify which of them are differentially expressed under two experimental conditions.

One of the problems that arise in this context, in addition to the need to simultaneously test a large number of hypotheses, one for each gene, is how to model the dependence that usually exists in the level of expression between genes.

The literature on the matter is extensive. However, the proposed procedures, most from a frequentist point of view and under the assumption of independence, do not usually solve the problems raised above.

The main objective of this thesis is the identification of genes with differential expression in two different and independent treatment conditions by considering the dependence on the expression level of the genes. To achieve this goal, a Bayesian procedure is proposed in which the dependency is modelled using copula functions.

The procedure is applied to tests of equality of means assuming that the data come from normal marginal distributions. Different variances for all variables are considered, but the same variance is assumed for each variable in the two treatment conditions. Likewise, the same dependency structure is considered for the two treatment conditions. Such dependency is modeled, firstly, by means of the Gaussian copula of the Elliptic family, considering the uniform correlation matrix in order to reduce the high dimension of the parametric space. Dependency is also modeled using the Clayton copula, of the Archimedean family, which allows modeling multivariate

distributions with asymmetric dependence through a single univariate function, thus reducing the high number of model parameters.

The proposed methodology is illustrated with simulated data from a multivariate normal distribution and also with real data from experiments with DNA microarrays. In both cases, Gaussian copula models and Clayton copula models are compared using the Deviance Information Criterion (DIC). Likewise, the value of the False Discovery Rate (FDR) is estimated in order to further evaluate the proportion of false positives.

From the results obtained it can be concluded that, regarding the estimation of the dependency parameter of both copulas, the procedure is robust to the choice of the parameters of the beta prior distribution for the initial probability of each null hypothesis. Indeed, in the simulated examples and with both types of copula, very similar values are obtained and close to the value of the parameter with which the data was generated.

However, regarding the number of rejected (accepted) null hypotheses , the procedure is not robust with respect to the choice of such beta priors, since the proportion of null hypotheses rejected (accepted) varies widely depending on the parameters chosen for the beta prior distribution. The main conclusion here is that a higher hit ratio is obtained when using beta priors skewed to the right.

As expected, according to the DIC criterion the model with Gaussian copulas turned out to be the most suitable for the simulated data, as they were generated from a normal distribution. Likewise, the number of rejected null hypotheses with this model is closer to the true number of false null hypotheses than that obtained with the model using Clayton copulas. In addition, with the Gaussian copula model, a number of null hypotheses slightly higher than the number of false null hypotheses is rejected, while the estimated value of the FDR remains at acceptable levels and is significantly lower than that obtained with the Clayton copula model. Therefore, the proposed procedure works well, when the dependency is modeled by the most suitable copula function.

On the other hand, the proposed procedure is flexible to the extent that it can

be used with other correlation matrices, or with other copula functions to model the dependency as well as with other marginal distributions. It is enough to adapt the algorithms appropriately.

Finally, in the context of multiple hypotheses tests, the literature on the use of copula functions to model dependency is scarce. In fact, we have only found works that use these functions to model the dependence between a small number of variables and under a frequentist approach. Therefore, the procedure proposed in this thesis is fundamental, especially in the field of genomics.

Prólogo

En las últimas décadas, debido a los avances tecnológicos, los procedimientos de contraste de hipótesis múltiples han pasado a ser una de las más importantes y modernas herramientas estadísticas empleadas para el análisis de datos procedentes de experimentos con microarrays de ADN. Sin embargo, la estructura compleja de estos datos, con miles de genes, ocasiona algunas complicaciones en los análisis estadísticos, debido a la necesidad de contrastar simultáneamente un gran número de hipótesis, una por cada gen. Desde una perspectiva frecuentista, si cada hipótesis se contrasta de forma individual a un nivel α, la probabilidad de cometer al menos un error de tipo I aumenta considerablemente con el número de hipótesis.

Además de la dificultad apuntada anteriormente, los análisis se complican por la existencia de dependencia entre el nivel de expresión de los genes, lo que contribuye a incrementar aún más el error en los resultados.

Se han propuesto muchos procedimientos, la mayoría desde un punto de vista frecuentista y bajo el supuesto de independencia, pero no se resuelven, en general, todos los problemas planteados anteriormente.

El objetivo principal de este trabajo es la identificación de genes con expresión diferencial bajo dos condiciones distintas de tratamiento. En ese sentido, la metodología base aplicada para lograr responder a esta cuestión son los procedimientos de contraste de hipótesis múltiples bajo dependencia. Para ello, se propone un procedimiento mediante un enfoque bayesiano, puesto que éste tiene la ventaja, sobre los métodos frecuentistas, de poder modelar la información que se tiene a priori sobre los parámetros que se pretende estimar. Por otra parte, la dependencia se modela mediante funciones cópulas. El concepto de cópula es muy atractivo porque

toda función de distribución conjunta se puede representar en forma de cópula y permite modelar una amplia gama de estructuras de dependencia. Asimismo, son poco frecuentes los procedimientos de inferencia mediante contrastes de hipótesis múltiples que recurren al uso de funciones cópulas para modelar la dependencia en los datos. Por lo que este trabajo resulta de gran interés, especialmente en el campo de la Genómica.

La tesis está organizada en cuatro capítulos. A continuación se resume el contenido de cada uno de ellos.

En el capítulo 1 se describe la tecnología de los microarrays de ADN, puesto que el análisis de los datos procedentes de estos experimentos ha sido la motivación de este trabajo, debido a la estructura compleja que presentan estos datos con miles de genes, existiendo además dependencia entre el nivel de expresión de los mismos. Asimismo, se introduce el problema de los contrastes de hipótesis múltiples desde las perspectivas frecuentista y bayesiana, analizando la literatura más relevante existente hasta la actualidad. Se analizan también algunos trabajos en los que se utilizan los Modelos Ocultos de Markov para modelar la estructura de dependencia en el contexto de los contrastes de hipótesis múltiples, así como los que utilizan procedimientos combinando técnicas de clustering y contrastes múltiples para identificar genes con expresión diferencial.

En el capítulo 2 se introducen las funciones cópulas, así como una revisión de la literatura sobre estas funciones. También se presentan algunas cópulas de las familias Elíptica y Arquimediana usadas para modelar la dependencia en esta tesis. Asimismo, se presenta el enfoque bayesiano que se propone para el problema de los contrastes de hipótesis múltiples, cuando se modela la dependencia mediante cópulas N-variantes. Se describe también, de forma general, el algoritmo MCMC necesario para realizar la inferencia bayesiana.

En el capítulo 3 se aplica la metodología bayesiana propuesta de contrastes de hipótesis múltiples, considerando distribuciones marginales normales y siendo las medias los parámetros de interés. La dependencia se modela, en primer lugar, mediante la cópula Gaussiana, utilizando la estructura de correlación uniforme

propuesta por Žežula (2009) y, en segundo lugar, mediante la cópula de Clayton (Clayton, 1978a; Nelsen, 2007). El procedimiento se ilustra en ambos casos con datos simulados y con datos reales procedentes de experimentos con microarrays de ADN, comparando ambos modelos mediante el criterio de selección DIC (Deviance Information Criterion).

Finalmente, en el capítulo 4, se presentan las conclusiones y las posibles extensiones de la tesis.

Capítulo 1

Introducción

El problema de los contrastes de hipótesis múltiples no es reciente. Fisher (1935) fue el primer autor en alertar sobre este problema y la literatura existente al respecto es extensa. Sin embargo, en las últimas décadas, este problema se ha evidenciado mucho más, debido al gran desarrollo de nuevas tecnologías en numerosas áreas del conocimiento y especialmente en el campo de la genómica, que han permitido la obtención de conjuntos de datos de gran volumen y complejidad, siendo en muchos casos necesario para su análisis estadístico el uso de contrastes de hipótesis múltiples, donde se requiere contrastar simultáneamente un elevado número de hipótesis entre las cuales existe, además, cierto grado de dependencia.

En este capítulo se describen el problema de los contrastes de hipótesis múltiples y la tecnología de los microarrays de ADN que ha motivado esta tesis, con la que se obtienen grandes cantidades de datos, que en muchos casos están fuertemente correlacionados, siendo necesario para su análisis contrastar simultáneamente un elevado número de hipótesis. También se describen diferentes enfoques para tratar este problema y se hace una revisión bibliográfica sobre los mismos.

El capítulo consta de 5 secciones. En la sección 1.1, se procede a una descripción de la tecnología de microarrays de ADN y su relevancia en diversos campos de aplicación. En la sección 1.2 se describe la problemática de los contrastes múltiples, desde una perspectiva frecuentista, cuando se contrastan simultáneamente un elevado número de hipótesis. Asimismo, se presenta una revisión bibliográfica sobre diversos

procedimientos propuestos en la literatura destinados a solucionar estos problemas. En la sección 1.3, se presenta el enfoque bayesiano para el problema de los contrastes múltiples y la revisión bibliográfica general relacionada con el tema. En la sección 1.4, se presentan algunos trabajos en los que se utilizan los Modelos Ocultos de Markov (HMM), bajo los enfoques frecuentista y bayesiano, como una alternativa para resolver el problema de dependencia en los casos de contrastes simultáneos con un elevado número de hipótesis. Finalmente, en la sección 1.5, se exponen también otros trabajos que combinan distintas técnicas destinadas al ajuste de la dependencia en los casos de contrastes múltiples con un elevado número de hipótesis.

1.1. Tecnología Microarrays de ADN

Los microarrays de ADN surgen en el campo de la genómica y sus aplicaciones en diversas áreas de conocimiento como la farmacología, la medicina, la zootecnia y la agronomía, entre otras, donde es necesario conocer el funcionamiento simultáneo de los genes y sus relaciones con determinadas características de un organismo.

Sanger and Coulson (1975) propusieron un nuevo método para la determinación de secuencias de nucleótidos en el ADN, lo que supuso una gran revolución en este campo, pero se hacía necesaria una tecnología que permitiera medir simultáneamente la expresión de todos los genes de un genoma, o al menos de una parte de éste, en lugar de estudiar un solo gen o unos pocos genes. A finales de los años 80 del siglo XX surge una tecnología innovadora para la determinación y cuantificación del ADN en muestras, que daría lugar a la primera plataforma de microarrays desarrollada por los científicos Lubert Stryer y Stephen Fodor, entre otros. Finalmente, Schena et al. (1995) publican el primer trabajo utilizando microarrays de ADN para medir los niveles de expresión genética en plantas. En la actualidad, esta tecnología es de gran importancia en la medida que se está aplicando en estudios relacionados con las más diversas enfermedades que afectan directa o indirectamente a la población humana. Un ejemplo son los estudios realizados con dos tejidos biológicos, cancerígeno y sano, donde uno de los objetivos es la identificación de genes con diferencias

significativas de expresión entre los dos tipos de tejido biológico estudiados. En todos los organismos, el contenido de ADN de todas sus células es idéntico y los genes se expresan dependiendo de algunos factores que regulan la expresión génica, tales como el estado de enfermedad o la especificidad temporal, entre otros, por tanto, la comparación de dos niveles de expresión de genes de diferentes tejidos puede llevar al entendimiento de diversos fenómenos presentes en un organismo. La monitorización de genes utilizando la tecnología de microarrys de ADN permite cuantificar el nivel de expresión génica de un determinado número de genes.

Un microarray de ADN es un componente electrónico miniaturizado que carga millones de transcriptores (figura 1.1). Se puede definir como una colección ordenada de sondas (fragmentos génicos conocidos, que se encuentran inmovilizados en una matriz sólida) donde cada una de ellas representa una única especie de ácido nucleico correspondiente al gen de interés.

Figura 1.1: Chip de ADN de Affymetrix, empleado para detectar expresión de genes de seres humanos (a la izquierda) y de ratón (a la derecha). Imagen tomada de https://commons.wikimedia.org/wiki/File:Affymetrix-microarray.jpg

El proceso de cuantificación de los valores de expresiones génicas que se obtienen de los experimentos con microarrays de ADN puede ser dividido en cuatro etapas: extracción de muestras, identificación (coloración), hibridación y, por último, lectura de los valores de las expresiones génicas. Básicamente, después de la extracción de los materiales destinados al análisis, cada una de las muestras se identifica debidamente por medio de procesos químicos que agregan una coloración particular a las muestras.

Seguidamente, las muestras identificadas se extienden sobre arrays (laminas de vidrio), donde son fijados los genes que van a ser analizados.

El principio técnico está basado principalmente en la propiedad de hibridación por complementariedad de ácidos nucleicos, es decir, en la propiedad que tienen dos cadenas homólogas de unir sus bases complementarias mediante la formación de puentes de hidrógeno. Un prerrequisito fundamental para cualquier tipo de microarray de ADN es la existencia de una posición individual para cada componente del microarray (Chaudhuri, 2005). La señal de hibridación producida en cada sonda corresponde al nivel de expresión de cada uno de los genes en una determinada muestra en el momento del estudio. De esa forma, las señales son detectadas, cuantificadas, integradas y normalizadas con softwares específicos (Niemeyer and Blohm, 1999; Van de Rijn and Gilks, 2004).

1.2. El problema de los Contrastes de Hipótesis Múltiples Frecuentistas

La tecnología de Microarays permite la cuantificación y la evaluación simultánea del nivel de expresión de millares de genes de un determinado organismo en diferentes condiciones, haciendo posible la comparación de muestras de tejido por niveles de expresión génica.

Por tanto, se hace necesario un procedimiento que permita contrastar simultáneamente un elevado número de hipótesis, una para cada gen. Desde una perspectiva frecuentista, si se contrasta cada hipótesis de manera individual a un nivel de significación α, asumiendo independencia, la probabilidad de rechazar erróneamente al menos una hipótesis, $[1-(1-\alpha)^m]$, aumenta rápidamente con m, siendo m el número de hipótesis. En este sentido, los procedimientos de contrastes de hipótesis múltiples frecuentistas, cuando el número de hipótesis es elevado, resultan inapropiados.

A lo largo de los años, la complejidad del problema de contrastes de hipótesis múltiples ha desencadenado una amplia y rica historia. Fisher (1935) fue quien,

por primera vez, advierte sobre la problemática del incremento del error cuando se efectúan contrastes múltiples. Su filosofía ha inspirado otros métodos debidos a Duncan (1955); Dunnett (1955) y Dunn (1961) o el conocido procedimiento de Bonferroni para la construcción simultánea de intervalos de confianza que se debe principalmente a Dunn (1961), entre otros.

Un procedimiento muy utilizado en el contexto de los contrates múltiples, para controlar el nivel de significación global, es el procedimiento de Bonferroni. Este procedimiento es sencillo de manejar, aunque resulta muy conservador, en el sentido de que se obtienen menos hipótesis nulas falsas rechazadas de las que realmente existen. El procedimiento consiste en especificar el nivel de significación global α deseado y contrastar cada hipótesis de forma individual al nivel $\alpha^* = \alpha/m$, donde m es el número total de contrastes realizados. Si m es muy grande estos niveles de significación individuales serán muy pequeños, lo que reduce drásticamente la potencia de los tests. Este procedimiento, así como los derivados del mismo, tienen la propriedad de controlar el family-wise error rate ($FWER$), que se define como la probabilidad de rechazar erróneamente al menos una hipótesis nula.

Procedimientos con la finalidad de mejorar la corrección de Bonferroni fueron propuestos por Holm (1979), Simes (1986) y Hochberg (1988), entre otros. Sin embargo, dichos procedimientos también se caracterizan por ser conservadores, lo que abre así el espacio para el desarrollo de nuevos métodos más robustos de contrastes múltiples. Un criterio menos restrictivo que los que controlan el $FWER$ es el criterio que controla el false discovery rate (FDR), este procedimiento es el más utilizado para contrastes múltiples con un gran número de hipótesis, propuesto por Benjamini and Hochberg (1995). El FDR se define como la proporción esperada de errores de tipo I entre las hipótesis rechazadas.

Consideremos el problema de contrastar simultáneamente m hipótesis nulas $H_i, i = 1, 2, \ldots, m$. El problema de los contrastes de hipótesis múltiples presenta una estructura simple, como se puede ver en la tabla 1.1 (Benjamini and Hochberg, 1995), donde V, S, U y T son variables aleatorias no observables, siendo V el número de hipótesis nulas verdaderas rechazadas, S el número de hipótesis nulas

Nº de hipótesis nulas	No Rechazadas	Rechazadas	Total
Verdaderas	U	V	m_0
Falsas	T	S	$m - m_0$
Total	$m - R$	R	m

Tabla 1.1: Resultados cuando se contrastan simultáneamente m hipótesis.

falsas rechazadas, U el número de hipótesis nulas verdaderas no rechazadas y T el número de hipótesis nulas falsas no rechazadas; R representa el número de hipótesis rechazadas por un determinado procedimiento y es una variable aleatoria observable y m_0 es el número desconocido de hipótesis nulas verdaderas. Entonces el FDR se define como sigue: $FDR = E(Q)$, donde

$$Q = \frac{V}{max(R,1)} = \begin{cases} \frac{V}{R} & \text{si } R > 0 \\ 0 & \text{en otro caso} \end{cases}$$

De esta manera, en vez de controlar la probabilidad de $V > 0$, como en el caso de la corrección de Bonferroni, Benjamini and Hochberg (1995) proponen controlar el FDR, es decir, la proporción de hipótesis nulas verdaderas rechazadas entre todas las hipótesis rechazadas para un experimento dado. El criterio propuesto por estos autores controla el FDR en un cierto nivel nominal, α, bajo el supuesto de independencia de los estadísticos para las diferentes hipótesis, y consiste en rechazar las hipótesis nulas correspondientes a los p-valores ordenados, $p_{(1)} \leq p_{(2)} \leq ..... \leq p_{(k)}$, donde $k = max\Big\{i : p_{(i)} \leq \frac{i}{N}\alpha\Big\}$.

Un tipo de error alternativo al FDR, introducido por Storey (2002), es el positive false discovery rate ($pFDR$) que es una modificación del FDR y se define como $pFDR = E\left[\frac{V}{R} \mid R > 0\right]$.

Otro procedimiento que también es muy común en la literatura consiste en utilizar el q-valor, que se define como el menor valor del $pFDR$ para el cual una determinada hipótesis sería rechazada (Storey, 2003).

En estudios destinados a la comparación de expresiones génicas entre grupos experimentales (sanos, enfermos), el número de hipótesis nulas rechazadas, R, ilustrado

en la tabla 1.1, representaría los genes detectados con expresión diferencial entre sano y enfermo de un total de m genes considerados.

La filosofía y el procedimiento de Benjamini and Hochberg (1995) ha inspirado muchas investigaciones sobre la problemática de los contrastes múltiples, tales como los procedimientos de puntos de corte basados en los p-valores (Genovese and Wasserman, 2004) o el p-valor adaptativo (Benjamini and Hochberg, 2000), entre otros. Estos procedimientos tienen la particularidad de garantizar el control del FDR bajo el supuesto de independencia de los estadísticos. Sin embargo, en muchas ocasiones el supuesto de independencia es poco realista y la mayoría de la literatura propuesta se realiza considerando independencia entre los estadísticos, básicamente porque asumir independencia hace el análisis más tratable, conduciendo así a posibles errores en los resultados.

Los contrastes de hipótesis múltiples se pueden ver muy afectados por la estructura de correlación entre las observaciones, reduciendo la potencia de los procedimientos que controlan el FDR. Clarke and Hall (2009) han observado que, en la práctica, el procedimiento de Benjamini-Hochberg es robusto frente a la dependencia cuando se aplica a contrastes simúltaneos con un elevado número de hipótesis. La violación de la suposición de independencia puede tener como consecuencia la pérdida de potencia de los tests, por ser excesivamente conservador o la pérdida de control del FDR, por ser excesivamente liberal (Sun and Cai, 2009).

Otros trabajos sobre el problema de contrastes de hipótesis múltiples, con la finalidad sobre todo de controlar el FDR, considerando el efecto de dependencia en los análisis, fueron desarrollados por Benjamini and Yekutieli (2001), Efron (2007) y Qiu et al. (2005), ente otros. Un procedimiento similar, con la particularidad de considerar dependencia positiva fue propuesto por Sarkar (2006) en un estudio dirigido a controlar el FDR, demostrando que dependencias fuertes afectan tanto al control del FDR como a la potencia del método. Asimismo, Benjamini and Yekutieli (2001) demuestran que el procedimiento de Benjamini and Hochberg (1995) también controla el FDR bajo ciertas estructuras de dependencia de los estadísticos de los contrastes y, además, proponen una modificación del procedimiento para estructuras

de dependencia arbitrarias.

1.3. Paradigma Bayesiano

La estadística bayesiana ha tenido en las últimas décadas una gran aceptación en muchas áreas de investigación científica. Bajo una perspectiva bayesiana toda fuente de incertidumbre debe describirse por medio de modelos probabilísticos. Este enfoque no está condicionado al tamaño de la muestra y utiliza la distribución a priori, esto es, la información previa a la proporcionada por los datos, aunque, en algunos casos, puede representar una dificultad plasmar mediante una distribución de probabilidad el conocimiento que se tiene sobre el suceso objeto de estudio.

Para tratar el problema de contrastes de hipótesis múltiples, este enfoque se basa en el cálculo de las probabilidades a posteriori de las hipótesis nulas contrastadas, evaluando la credibilidad de cada una de ellas en función de dicha probabilidad.

Este trabajo ha sido desarrollado mediante el enfoque bayesiano, puesto que ofrece la ventaja, frente al frecuentista, de no estar tan condicionado al tamaño de la muestra y por utilizar, además de la información muestral, la información previa que se tiene sobre los parámetros del modelo.

En el enfoque bayesiano toda la inferencia se hace a partir de la distribución a posteriori. Básicamente, el objetivo de un análisis bayesiano es obtener la distribución de probabilidad a posteriori de un parámetro o un vector de parámetros. La idea es la siguiente, el investigador tiene información previa sobre los parámetros que se pretenden estimar y que puede ser cuantificada por medio de una distribución de probabilidad antes de obtener las observaciones de la muestra. Esta información inicial se verá modificada en función de los datos observados, obteniendo así una distribución a posteriori que resumirá todo el conocimiento, el del investigador y el aportado por los datos.

La distribución a posteriori es simplemente la distribución a priori de los parámetros, $\pi(\theta)$, después de ser actualizada utilizando la verosimilitud de los datos

observados, $f(y\,|\theta)$. Esa actualización se obtiene mediante el Teorema de Bayes;

$$\pi(\theta\,|y) = \frac{f(y\,|\theta)\pi(\theta)}{g(y)} \tag{1.1}$$

Donde $g(y) = \sum_\theta f(y|\theta)\pi(\theta)$, si θ es discreto y $g(y) = \int_\Theta f(y\,|\theta)\,\pi(\theta)d\theta$, si θ es continuo.

La distribución a posteriori se utiliza para realizar la inferencia sobre θ. La obtención de la distribución a posteriori implica, en muchas ocasiones, la utilización de procedimientos complejos de cálculo numérico, métodos MCMC, etc.

Más general, dada una muestra $y_1, \ldots, y_k$ y la verosimilitud, denotada por

$$L(\theta \mid y) = f(y_1, \ldots, y_k\,|\theta)$$

la expresión base para la distribución de probabilidad a posteriori viene dada por:

$$\pi(\theta\,|y_1, \ldots, y_k) = \frac{f(y_1, \ldots, y_k\,|\theta)\pi(\theta)}{\int_\Theta f(y_1, \ldots, y_k\,|\theta)\pi(\theta)\,d\theta} \tag{1.2}$$

Una forma equivalente de (1.2) omite el denominador de la expresión, puesto que el denominador es constante con respecto a θ, por lo que, en algunos casos, no necesita ser calculado, y la distribución de probabilidad a posteriori se puede escribir como

$$\pi(\theta|y_1, \ldots, y_k) \propto f(y_1, \ldots, y_k|\theta)\pi(\theta) \tag{1.3}$$

Una parte fundamental del análisis bayesiano consiste en especificar las distribuciones a priori para todos los parámetros desconocidos del modelo, θ, sin embargo, la especificación de estas distribuciones constituye, en muchos casos, un gran desafío y ha sido durante mucho tiempo un tema de discusión en la comunidad estadística. En algunas aplicaciones existe una evidencia objetiva de la distribución a priori de los parámetros, es decir, se puede tratar de un caso en el que se dispone de una información sólida sobre los posibles valores del parámetro y sobre cuál es su distribución de probabilidad. En esos casos puede resultar cómodo y sencillo seleccionar un elemento de familias natural conjugadas (Box and Tiao, 2011).

La especificación subjetiva de las distribuciones a priori dependerá del conocimiento científico que se tenga del fenómeno, sin embargo, es de suma importancia

que haya interacción investigador-estadístico para llegar a un acuerdo sobre las distribuciones a priori adoptadas y las implicaciones en las distribuciones a posteriori.

En la literatura hay diversas formulaciones para las distribuciones a priori no informativas, tales como las propuestas por Jeffreys (1967) o la distribución uniforme, entre otras. En los casos donde hay total desconocimiento sobre la distribución a priori de los parámetros, se suele usar la regla de Jeffreys para elegir la distribución a priori, que consiste en utilizar la información de Fisher $I(\phi)$ como:

$$\pi(\phi) = \sqrt{detI(\phi)}$$

donde ϕ puede ser un parámetro o un vector de parámetros. Por otro lado, aunque existe notable controversia al respecto, una distribución que expresa poca o ninguna información puede ser representada mediante una distribución a priori Uniforme (Laplace, 1774),

$$\pi(\theta) = cte$$

Estas distribuciones, en algunos casos, pueden ser impropias, sin embargo, esto no tiene por qué representar un problema, siempre que la distribución a posteriori que se obtenga sea propia.

Una vez obtenida la distribución a posteriori, se puede realizar la inferencia sobre cada uno de los parámetros del modelo mediante las correspondientes marginales. De ahí que, en muchos casos, surjan integrales cuya expresión analítica no es factible obtener por métodos analíticos, lo que hace necesario el uso de métodos numéricos aproximados como los métodos basados en cadenas de Markov Monte Carlo (MCMC).

En el campo bayesiano, bajo el supuesto de independencia, cabe señalar los trabajos sobre contrastes de hipótesis múltiples de Waller and Duncan (1969), Berry and Hochberg (1999), Efron et al. (2001), Storey (2003), Scott and Berger (2006, 2010), Ausín et al. (2011) y Gómez-Villegas et al. (2014), entre otros.

Son pocos los trabajos en los cuales se emplea la modelización bajo la perspectiva bayesiana que tengan en cuenta la estructura de correlación de los datos. Este enfoque tiene la gran ventaja de permitir de forma natural la incorporación de fuentes de

variabilidad e incertidumbres no observadas. Trabajos considerando estructuras de dependencia en los datos, entre otros, Chi (2011), Rayaprolu and Chi (2014) y Qiu et al. (2005), estos últimos autores estudiaron el impacto de la correlación en el método bayesiano empírico.

Las metodologías basadas en modelos jerárquicos también se han mostrado bastante eficientes cuando se aplican al análisis con un elevado número de datos, precisamente debido a su capacidad para reducir la variabilidad presente en los mismos. Un ejemplo son, los experimentos con microarays de ADN que presentan muchas fuentes de variación sistemática, las cuales pueden afectar a la medida de los niveles de expresión génica. En consecuencia, es normal encontrar residuos asimétricos al analizar el ajuste de los modelos de datos derivados de estos experimentos (Durbin et al., 2002).

Para relajar las variaciones en los análisis generados por este tipo de experimentos destacan los modelos jerárquicos bayesianos, los cuales están siendo muy utilizados en diferentes áreas científicas. La modelización jerárquica bajo el enfoque bayesiano está basada en el simple hecho de que la distribución conjunta se puede descomponer en una serie de modelos condicionales. La distribución conjunta es difícil de especificar en procesos complejos, pero, en estos casos, se puede obtener mediante el producto de la serie de modelos condicionales relativamente más simples.

Zhao et al. (2008) aborda el problema utilizando un modelo jerárquico bayesiano multivariante para experimentos con datos replicados y compatible con la dependencia entre la media y la varianza en el modelo bayesiano univariante. Este modelo relaja el supuesto de coeficiente de variación constante entre medidas, al considerar una estructura de covarianzas. Para inferir los patrones de expresión de los genes desarrollan un test de la razón de verosimilitud generalizada (GLRT).

Ghosal and Roy (2012) utilizan un modelo jerárquico con un enfoque bayesiano. Los autores proponen abordar el problema mediante transformaciones probit de los p-valores correlacionados, para el desarrollo de un modelo multivariante bajo dependencia de las hipótesis. Describen un procedimiento bayesiano no paramétrico para contrastes múltiples, que modela directamente la distribución conjunta de los

p-valores mediante modelos mixtos flexibles. El procedimiento lo aplican al análisis de un elevado número de datos, específicamente a datos que proceden de experimentos con microarrays de ADN.

1.4. Modelos Ocultos de Markov (HMM)

Un modelo oculto de Markov (Hidden Markov Model) es una herramienta efectiva para modelar la estructura de dependencia, siendo estos modelos muy utilizados en áreas como el reconocimiento de la voz, el procesamiento de señales y el análisis de secuencias de ADN. En el contexto de los contrastes de hipótesis múltiples, los modelos HMM asumen que la secuencia de los estados no observables forman una cadena de Markov $(\theta_i)_1^m = (\theta_1, \theta_2, \ldots, \theta_m)$, donde $\theta_i = 1$ si la hipótesis i es no nula y $\theta_i = 0$ en otro caso y los datos observados se generan condicionalmente independientes a los estados ocultos $(\theta_i)_i^m$ (Sun and Cai, 2009).

Una versión relacionada con los modelos HMM son los modelos Hidden Markov random fields (HMRFs), presentada por Liu et al. (2014a) y François et al. (2006), entre otros, y una extensión de estos últimos son los modelos Markov-random-field-coupled mixture, estudiados por Liu et al. (2012), Liu et al. (2014b) y Liu et al. (2016) entre otros, muy utilizados en estudios de segmentación de imágenes.

El desarrollo de modelos ocultos de Markov (HMM) para soportar la estructura de dependencia de los datos, ha supuesto un gran impacto en la investigación, especialmente en el campo de la genómica. Esta herramienta se ha mostrado bastante efectiva para el control del FDR. A continuación, mencionamos algunos trabajos relativos a los contrastes de hipótesis múltiples que han usado esta herramienta para modelar la estructura de dependencia.

Sun and Cai (2009), Chi (2011) y Rayaprolu and Chi (2014), entre otros, exploran la potencialidad ofrecida por el modelo HMM para tratar la estructura de dependencia en los contrastes múltiples de hipótesis. Con la misma finalidad Liu et al. (2012), Liu et al. (2014b) y Liu et al. (2016) proponen un modelo parecido, el Markov-random-field-coupled mixture model.

Chi (2011) estudió el efecto de la estructura de dependencia de los estados finitos del modelo HMM en la razón de verosimilitudes, para contrastes múltiples óptimos en estados ocultos.

Sun and Cai (2009) explorando la estructura de dependencia a través de modelos HMM, demostraron la optimalidad del FDR bajo ciertas condiciones a un nivel α apropiado y la realización empírica de estos modelos. Fundamentalmente, su procedimiento está basado en la construcción de un nuevo estadístico para los contrastes en lugar del p-valor, el LIS (índice local de significancia) que tiene en cuenta las observaciones adyacentes a través de la exploración de la dependencia local en el modelo HMM. Los modelos HMM utilizados en su investigación, tenían la particularidad de considerar parámetros desconocidos asociados a las observaciones como variables de Bernoulli y distribuidos según una cadena de Markov.

Además, comparan su procedimiento basado en el estadístico LIS con los procedimientos basados en p-valores y con el procedimiento del FDR local de Efron et al. (2001). Los autores concluyen que, tanto los procedimientos basados en p-valores como el procedimiento del FDR local, son ineficientes cuando los estadísticos están correlacionados. Asimismo, concluyen que los procedimientos basados en p-valores son menos eficientes que el procedimiento del FDR local, en la medida que el p-valor considera las hipótesis separadamente a la hora de determinar el nivel de significación y el procedimiento del FDR local considera las hipótesis simultáneamente.

Liu et al. (2012), Liu et al. (2014b) y Liu et al. (2016) desarrollaron un procedimiento de contrastes múltiples basado en un modelo (Markov-random-field-coupled mixture model) que permite una estructura de dependencia arbitraria y parámetros dependientes heterogéneos. Este procedimiento aparece como una extensión del procedimiento de Sun and Cai (2009), que solamente permite representar una estructura de dependencia secuencial y en el que los parámetros de dependencia son homogéneos. Asimismo, Liu et al. (2014a) proponen un algoritmo eficaz para el aprendizaje de parámetros dependientes heterogéneos basado en los modelos Hidden Markov-random-field.

En general, los desarrollos basados en modelos gráficos, tanto bajo el enfoque

frecuentista como bajo el enfoque bayesiano, utilizan modelos mixtos. Estos modelos requieren que la función de densidad del estadístico bajo la hipótesis alternativa (f_1) sea estimada paramétricamente, sin embargo, es frecuente que no se pueda estimar como una simple distribución paramétrica.

Liu et al. (2014b) desarrollaron un procedimiento semiparamétrico para contrastes múltiples bajo dependencia, que generaliza el procedimiento del FDR local de Efron et al. (2001), centrado en la estimación de la función de densidad del estadístico bajo la hipótesis alternativa (f_1) de forma semiparamétrica, es decir, una parte de f_1 es estimada de forma paramétrica y la restante de forma no paramétrica. Desde el punto de vista del modelaje gráfico, estos autores han verificado que su método de estimación semiparamétrico captura mejor la dependencia entre las hipótesis que el procedimiento de modelaje gráfico totalmente paramétrico empleado por Sun and Cai (2009) y Liu et al. (2012).

1.5. Procedimientos de Agrupamiento

En la actualidad, uno de los principales desafíos relacionados con el problema de contrastes de hipótesis múltiples está relacionado con la necesidad de procedimientos capaces de modelizar la dependencia. En las referencias bibliográficas presentadas anteriormente se ha podido ver que los procedimientos para el problema de contrastes de hipótesis múltiples resultan bastante conservadores, debido a la cantidad y complejidad de datos y, sobre todo, a que la mayoría de los procedimientos se realizan bajo independencia de los tests. Sin embargo, se observa una mayor robustez y un aumento de la potencia cuando se considera la estructura de correlación de los datos.

Uno de los objetivos en el campo de la genómica, más concretamente en los experimentos procedentes de microarrays de ADN, es la identificación de genes con diferencias significativas de expresión entre los tejidos biológicos estudiados, por medio de la comparación de patrones de expresión génica entre grupos. Esta comparación, además de llevarse a cabo mediante contrastes múltiples, puede requerir de otros procedimientos complementarios de análisis estadísticos y bioinformáticos.

Avances recientes han demostrado la factibilidad de los procedimientos combinando técnicas de clustering y contrastes múltiples para identificar genes con expresión diferencial, principalmente con la finalidad de estudiar el comportamiento de distintos grupos de genes, estableciendo relaciones que ayuden a tomar decisiones de tipo biológico (Dahl and Newton, 2007; Marín and Rodríguez-Bernal, 2012; Yuan and Kendziorski, 2006).

Marín and Rodríguez-Bernal (2012) han considerado los contrastes múltiples con el objetivo de identificar diferencias, en el nivel medio de expresión, entre genes en diferentes condiciones de tratamiento. Utilizan también un método de análisis clúster basado en un modelo mixto. Esta metodología permite la comparación de más de dos tratamientos al mismo tiempo para un mismo gen. En su procedimiento, consideran una mixtura de la distribución t de Student no centrada con un número desconocido de componentes. Yuan and Kendziorski (2006) desarrollaron un procedimiento bajo un enfoque bayesiano empírico, destinado a agrupar simultáneamente los datos en base a técnicas clúster y detectar genes con expresión diferenciada. Dahl and Newton (2007) propusieron un modelo BEMMA (Bayesian Effect Model for Microarrays), que representa una metodología bayesiana empírica para acomodar la dependencia entre genes a través de grupos latentes, mostrando una mejora en la potencia de los métodos de contrates múltiples al agrupar observaciones correlacionadas.

En el capítulo siguiente se describe el procedimiento bayesiano de contrastes de hipótesis múltiples propuesto en esta tesis, en el que la dependencia existente entre los datos se modela mediante funciones cópulas.

Capítulo 2

Contraste de Hipótesis Múltiples. Modelo de Dependencia Mediante Cópulas N-Variantes

Nuestro principal objetivo es proponer un procedimiento de contraste de hipótesis múltiples, mediante un enfoque bayesiano, con la finalidad de contrastar un elevado número de hipótesis bajo dependencia. Debido a la naturaleza múltiple del problema, hemos considerado, para la construcción del modelo, introducir funciones cópulas N-variantes que reflejen la dependencia entre las variables del modelo. En el contexto de los contrastes de hipótesis múltiples y con un enfoque frecuentista, Dickhaus and Gierl (2012) utilizan la cópula de Clayton N-variante para modelar la dependencia en el análisis de expresión génica y en Bodnar and Dickhaus (2014) se puede encontrar una revisión detallada del problema de los contrastes de hipótesis múltiples, dirigido al control del FDR, en la que se utilizan cópulas Arquimedianas.

De acuerdo con el enfoque bayesiano, todos los parámetros desconocidos serán estimados a partir de la distribución a posteriori. Es decir, nuestro principal interés es la obtención de la distribución a posteriori de los parámetros, así como la probabilidad a posteriori de que cada hipótesis nula sea cierta, condicionada a los datos.

En este capítulo se presenta el procedimiento bayesiano adoptado para modelar grandes cantidades de datos teniendo en cuenta la estructura de dependencia. También

se exponen los conceptos básicos para la construcción del modelo mediante funciones cópulas.

El capítulo consta de 3 secciones. En la sección 2.1, se definen las funciones cópulas de un modo general junto con sus propiedades y una breve revisión de la literatura sobre estas funciones. También se describe la cópula Gaussiana de la familia Elíptica, así como la estructura de correlación adoptada para este modelo, y la cópula de Clayton de la familia Arquimediana, que serán las cópulas utilizadas en esta tesis para modelar la dependencia. En la sección 2.2, se presenta el enfoque bayesiano que se propone para el problema de los contrastes de hipótesis múltiples cuando la dependencia se modela a través de funciones cópulas multivariantes. Finalmente se describe el algoritmo MCMC necesario para realizar la inferencia bayesiana.

2.1. Funciones Cópulas

El estudio de dependencia entre dos o más variables representa uno de los aspectos más interesantes y relevantes en los análisis estadísticos, ya que permite analizar dichas relaciones y realizar inferencias más precisas. El concepto de cópula resulta bastante atractivo debido a que las cópulas abarcan un gran conjunto de estructuras de dependencia y permiten incorporar al modelo de forma eficiente la estructura de dependencia de los datos. El gran potencial y utilidad de las cópulas está relacionado con el hecho de que la combinación de diversas cópulas con diferentes distribuciones marginales permite una gran flexibilidad para modelar distribuciones conjuntas.

El concepto de función cópula fue introducido por Sklar (1959) y, posteriormente, han surgido muchos otros trabajos como los de Genest and MacKay (1986), Joe (1997), Cherubini et al. (2004) y Nelsen (2007), entre otros. En la actualidad, las cópulas se han convertido en una importante herramienta de modelado multivariante en estudios donde la dependencia entre las variables aleatorias es de gran interés. En la literatura hay evidencias empíricas, reportadas recientemente, de que el uso de cópulas permite una descripción más realista de la componente de dependencia entre series (Embrechts et al., 2001, 2002). Ausín and Lopes (2010) proponen una

metodología bayesiana para hacer inferencia y predicción en modelos de cópulas GARCH. Žežula (2009) presenta una discusión sobre la cópula Gaussiana. Embrechts et al. (2003) y Costa Dias (2004) propusieron algunas aplicaciones practicas de las cópulas en áreas de estudio como la banca y los seguros, Clayton (1978b) en análisis de supervivencia y Nikoloulopoulos and Karlis (2008) y Lambert and Vandenhende (2002) en estadística aplicada a la medicina. A pesar de existir una gran cantidad de literatura sobre las funciones cópulas, la mayoria de las aplicaciones prácticas se restringen a cópulas bivariantes. Joe (1997), Embrechts et al. (2003) y Díaz (2014) proponen algunas extensiones multivariantes de cópulas Arquimedianas.

Las cópulas son funciones de distribuciones multivariantes que permiten agregar un conjunto de funciones univariantes, con una determinada estructura de dependencia. De acuerdo con Genest and Favre (2007) las cópulas son funciones que miden la dependencia entre variables. La esencia del enfoque de cópulas es que una distribución conjunta de variables aleatorias, se puede representar mediante una distribución de las distribuciones marginales (Clemen and Reilly, 1999).

De manera más formal (Joe, 1997; Nelsen, 2007), una n-cópula es una función C del cubo unidad $[0,1]^n \to [0,1]$ que satisface las siguientes propiedades:

- $C(v_1, v_2, \ldots, v_n) = 0$; si $v_i = 0$ para algún $i = 1, 2, \ldots, n$
- $C(1, \ldots, 1, v_i, \ldots, 1) = v_i$ para todo $i = 1, 2, \ldots, n$ y $v_i \in [0,1]$
- La función C es n creciente.

A pesar de que la literatura sobre el tema es extensa, el siguiente teorema, de mediados del siglo XX es la base de la teoría de cópulas.

Teorema 1. *Sklar (1959). Sea $X = (X_1, X_2, \ldots, X_k)$ un vector de variables aleatorias con función de distribución conjunta $F(x_1, x_2, ..., x_k)$ y distribuciones marginales $F_1(x_1), \ldots, F_k(x_k)$. Entonces existe una cópula C tal que:*

$$\begin{aligned} F(x_1, x_2, ..., x_k) &= C(F_1(x_1), \ldots, F_k(x_k)) \\ &= C(u_1, \ldots, u_k) \end{aligned} \tag{2.1}$$

Donde $C(u_1, \ldots, u_k)$ es una función de distribución conjunta con marginales uniformes, $F_1(x_1) = u_1, \ldots, F_k(x_k) = u_k$. Si $F_1, F_2, \ldots, F_k$ son continuas, entonces C es única; en caso contrario C está unívocamente determinada en

$$RanF_1 \times \ldots \times RanF_k$$

Inversamente, si C es una cópula k-dimensional y $F_1, F_2, \ldots, F_k$ son funciones de distribución, entonces la función F definida anteriormente es una función de distribución k-dimensional con marginales $F_1, F_2, \ldots, F_k$.

Una demostración del teorma de Sklar puede verse en Schweizer and Sklar (1983). Una característica de los modelos con cópulas, es el hecho de que la densidad conjunta se puede separar en dos componentes, por un lado, el producto de las densidades marginales y, por otro, la estructura de dependencia, véase Ausín and Lopes (2010) y Smith (2011).

A continuación se detallan algunos conceptos relacionados con las funciones cópulas.

2.1.1. Método de Inversión de Cópulas

Como consecuencia del Teorema de Sklar, es posible obtener distribuciones multivariantes a partir de una función cópula y las marginales que se desea fijar. Este hecho supone una importante ventaja para la modelización y simulación de variables aleatorias. A lo largo de los años, muchos autores han desarrollado métodos que sirven para construir cópulas con ciertas características deseables, orientadas a identificar algún tipo particular de dependencia. El método de inversión es uno de los más comúnmente utilizados para la construcción de cópulas, véase Nelsen (2007). Este método permite obtener funciones cópulas mediante las inversas de las funciones de distribución.

Corolario 2.1. *Sea H una c.d.f de una distribución multivariada, y $F_1, F_2, \ldots, F_k$ las c.d.f marginales continuas, con ${F_1}^{-1}, {F_2}^{-2}, \ldots, {F_k}^{-1}$ sus respectivas funciones*

inversas y C la función cópula. Entonces, para todo $u = (u_1, u_2, \ldots, u_k) \in [0,1]^n$

$$C(u_1, u_2, \ldots, u_k) = H({F_1}^{-1}(x_1), {F_2}^{-1}(x_2), \ldots, {F_k}^{-1}(x_k))$$

Nótese que si $X_1, \ldots, X_k$ son variables aleatorias continuas con funciones de distribución $F_1, F_2, \ldots, F_k$, respectivamente, entonces C es la distribución conjunta de $u_1 = F_1(X_1), \ldots, u_k = F_k(X_k)$, ya que $F_1(x_1), \ldots, F_k(X_k)$ se distribuyen uniformemente en el intervalo $[0,1]$.

Dos ejemplos donde se emplea este método son la cópula Gaussiana y la cópula de la t de Student, que pueden ser escritas, respectivamente, como

$$\begin{aligned} C_G(u_1, u_2, \ldots, u_k) =& \Phi_k({\Phi_1}^{-1}(u_1), {\Phi_1}^{-1}(u_2), \ldots, {\Phi_1}^{-1}(u_k)) \\ C_T(u_1, u_2, \ldots, u_k) =& t_{w,k}(t_{w,1}^{-1}(u_1), t_{w,1}^{-1}(u_2), \ldots, t_{w,1}^{-1}(u_k)) \end{aligned}$$

donde Φ_k es la distribución normal k-variante y $t_{w,k}$ la distribución t-Student k-variante con w grados de libertad.

2.1.2. Densidad de Probabilidad Asociada a las Cópulas

La función cópula puede interpretarse como una función para ligar distribuciones multivariantes con sus respectivas distribuciones marginales univariantes.

Asumiendo que $F_1(x_1), .., F_k(x_k)$ y C son funciones diferenciables, entonces la función de densidad conjunta $f(x_1, x_2, ..., x_k)$ de $X = (X_1, X_2, ..., X_k)$, se puede escribir en función del producto de las funciones de densidad de probabilidad marginales (Smith, 2011):

$$\begin{aligned} f(x_1, x_2, \ldots, x_k) &= c(F_1(x_1) \ldots, F_k(x_k)) \prod_{i=1}^{k} f_i(x_i) \\ &= c(u_1, u_2, \ldots, u_k) \prod_{i=1}^{k} f_i(x_i) \end{aligned} \tag{2.2}$$

donde $f_i(x_i)$, es la función de densidad de $F_i(x_i)$, para $i = 1, \ldots, k$, y la función de densidad de la cópula es

$$c(F_1(x_1), .., F_k(x_k)) = \frac{\partial^k}{\partial F_1(x_1) \ldots \partial F_k(x_k)} C(F_1(x_1), F_2(x_2), \ldots, F_k(x_k))$$

Así, combinando el hecho de que cualquier variable aleatoria continua puede ser transformada, por su función de distribución acumulada, en una variable aleatoria con distribución U(0,1), las cópulas pueden ser utilizadas para proporcionar una estructura de dependencia multivariante separadamente de las distribuciones marginales. Por tanto, en (2.2) se descompone la función de densidad conjunta en dos partes, $c(F_1(x_1), .., F_k(x_k))$ que describe la estructura de dependencia y $\prod_{i=1}^{k} f_i(x_i)$ el comportamiento marginal de cada una de las componentes. Por esa razón, con frecuencia a $c(u_1, \ldots, u_k)$ se la llama función de dependencia.

En este trabajo nos hemos centrado en casos donde las distribuciones marginales son continuas, sin embargo, se puede combinar cualquier tipo de cópula con cualquier tipo de distribuciones marginales, es decir, cualquier distribución se puede escribir en forma de cópula, a su vez, podemos deducir una cópula a partir de cualquier función de distribución multivariante.

2.1.3. Cópulas y Dependencia

El uso de modelos con funciones cópulas para construir una distribución conjunta de variables aleatorias continuas, puede verse como una versión de función de distribución conjunta libre de las marginales que es capaz de capturar estructuras de dependencia. La relación de dependencia entre variables aleatorias puede ser cuantificada por medio de distintas medidas, que se presentan a continuación.

Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es la forma más habitual de cuantificar la relación entre dos variables aleatorias. Se define como:

$$\rho_{XY} = \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)Var(Y)}}$$

Existen otros tipos de medidas de dependencia, también conocidas como medidas de concordancia. Algunas de estas medidas, tales como la τ de Kendall y el ρ_s de Spearman, definidas a continuación, permiten evitar algunas limitaciones del coeficiente de correlación lineal, puesto que son medidas robustas e invariantes

por transformaciones estrictamente crecientes de las variables aleatorias, de hecho Schweizer and Wolff (1981) estudian la relación entre las medidas de concordancia y las funciones cópulas.

Tau de Kendall

Sean (X_1, Y_1) y (X_2, Y_2) dos vectores aleatorios independientes e idénticamente distribuidos, con funciones de distribución H_1 y H_2, respectivamente, y con marginales F (para X_1 y X_2) y G (para Y_1 e Y_2). Se define la medida τ de Kendall como la diferencia entre las probabilidades de concordancia y discordancia entre los vectores (X_1, Y_1) y (X_2, Y_2).

$$\begin{aligned}\tau &= P\left((X_1 - X_2)(Y_1 - Y_2) > 0\right) - P\left((X_1 - X_2)(Y_1 - Y_2) < 0\right) \\ &= 2P\left((X_1 - X_2)(Y_1 - Y_2) > 0\right) - 1\end{aligned}$$

La versión de la τ de Kendall en términos de una cópula C, se puede escribir como sigue (Schweizer and Wolff, 1981; Kruskal, 1958; Nelsen, 2007):

$$\tau_C = 4\int_0^1 \int_0^1 C(v_1, v_2)dC(v_1, v_2) - 1 \tag{2.3}$$

La ecuación (2.3) muestra que el coeficiente τ de Kendall está completamente determinado por la cópula.

Rho de Spearman

Sean (X_1, Y_1), (X_2, Y_2) y (X_3, Y_3) tres vectores aleatorios independientes con función de distribución conjunta común H y sea C una cópula asociada a H. Se define el coeficiente ρ_S de Spearman como

$$\rho_S = P\left((X_1 - X_2)(Y_1 - Y_3) > 0\right) - P\left((X_1 - X_2)(Y_1 - Y_3) < 0\right)$$

La versión de la ρ_S de Spearman en término de la cópula C, puede ser escrita como (Schweizer and Wolff, 1981):

$$\rho_S = 12\int_0^1 \int_0^1 \left[C(v_1, v_2) - v_1 v_2\right] dv_1 dv_2 = 12\int_0^1 \int_0^1 C(v_1, v_2)dv_1 dv_2 - 3$$

Existen distintos tipos de cópulas para modelar la dependencia entre variables, de las cuales, dependiendo de la estructura de dependencia captada, las más usadas son las cópulas Elípticas y las Arquimedianas.

2.1.4. Funciones Cópulas Comunes

Por lo general, cuando se habla de los diferentes tipos de cópulas existentes, en realidad se hace referencia a diferentes familias. Todas las cópulas que pertenecen a una misma familia tienen la misma estructura matemática que depende de un cierto número de parámetros. De manera que, para cada uno de los valores de esos parámetros se obtiene un miembro de esa familia.

En esta subsección describiremos algunas de las familias de cópulas más comunes.

Cópulas Elípticas

Las cópulas de la familia Elíptica se caracterizan por compartir propiedades de la distribución normal multivariante, tales como la simetría y porque la dependencia está totalmente determinada por la matriz de correlaciones (Embrechts et al., 2001), es decir, las cópulas Elípticas están asociadas a las distribuciones Elípticas y se caracterizan por representar la relación de dependencia simétrica. Ejemplos de cópulas Elípticas más comunes son la cópula Gaussiana, asociada a la distribución normal, y la cópula de la t-Student, asociada a la distribución t-Student. La cópula Gaussiana presenta la ventaja de ser invariante bajo transformaciones estrictamente crecientes de las variables aleatorias. Esta cópula es muy utilizada por permitir grados de dependencia positiva o negativa por igual, sin embargo, en muchos casos los datos reales no cumplen la propiedad de simetria de las cópulas Gaussianas. Como solución a estos problemas se utilizan otras cópulas que permiten asimetria en los datos, como son las cópulas de la familia Arquimediana.

Cópulas Arquimedianas

Estas cópulas abarcan un gran número de cópulas con características diferen-

tes, que resultan difíciles de clasificar mediante un tipo específico de dependencia, a diferencia de las Elípticas que reflejan la dependencia simétrica, sin embargo, estas cópulas se caracterizan por permitir modelar distribuciones multivariantes mediante una única función univariante, simplificando los cálculos de las medidas de dependencia. A continuación, se presenta la definición de las cópulas de la familia Arquimediana para el caso de las cópulas bivariantes dada por Genest and Rivest (1993).

Definición 2.1 (Cópula Arquimediana)**.** *Sea* Φ *el conjunto de funciones continuas, estrictamente decrecientes y convexas de la forma* $\varphi : [0,1] \to [0,\infty]$*, donde* $\varphi(1) = 0$*. Cada elemento de* Φ *genera una cópula* C *a partir de la siguiente relación*

$$C(v_1, v_2) = \varphi^{-1}\left(\varphi(v_1) + \varphi(v_2)\right), \quad (v_1, v_2)^T \in (0,1)^2$$

La función φ en la definición 2.1 es conocida como el generador de la cópula. Cuando $\varphi(0) = \infty$, φ se denomina generador estricto. Estas cópulas proporcionan una estructura general para modelar distribuciones bivariantes.

La versión multivariante (Embrechts et al., 2001, 2003; Joe, 1997) está representada por:

$$C(v_1, \ldots, v_k) = \varphi^{-1}\left[\sum_{i=1}^{k} \varphi(u_i)\right] \tag{2.4}$$

con función de densidad

$$c(v_1, \ldots, v_k) = \varphi_k^{-1}\left[\sum_{i=1}^{k} \varphi(u_i)\right] \prod_{i=1}^{k} \varphi\prime(u_i)$$

donde φ es el generador de la cópula C y φ_k^{-1} denota la derivada de orden k de la inversa de la función generador. Embrechts et al. (2003) presentan un teorema donde se muestran las condiciones necesarias y suficientes para que (2.4) sea una cópula.

En este trabajo, para modelar la estructura de dependencia entre las variables de estudio, hemos considerado las cópulas Gaussiana y de Clayton de las familias Elípticas y Arquimediana respectivamente. A continuación se describen estas cópulas. Una revisión más completa de los diferentes tipos de cópulas puede verse en Joe (1997) y Nelsen (2007).

Cópula Gaussiana

El nombre Gaussiana, se debe al hecho de que la expresión coincide con la función de distribución normal bivariante estándar. Esta cópula permite por igual, grados de dependencia positiva o negativa. Las cópulas gaussianas no admiten una fórmula explicita, pero se pueden expresar en forma de una integral.

Considerando el coeficiente de correlación de Pearson, $-1 \leq \rho \leq 1$, la cópula Gaussiana bivariante se puede escribir

$$\begin{aligned} C_\rho(u_1, u_2) =& \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \\ =& \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi(1-\rho)^{\frac{1}{2}}} exp\left\{ \frac{-(\xi_1^2 - 2\rho\xi_1\xi_2 + \xi_2^2)}{2(1-\rho^2)} \right\} d\xi_1 d\xi_2 \end{aligned}$$

donde Φ es la c.d.f de la distribución normal estándar.

Diferenciando la expresión anterior, se obtiene la expresión de la densidad de la cópula Gaussiana:

$$c_\rho(u_1, u_2) = \frac{1}{2\pi(1-\rho)^{\frac{1}{2}}} exp\left\{ \frac{-(\xi_1^2 - 2\rho\xi_1\xi_2 + \xi_2^2)}{2(1-\rho^2)} \right\}$$

Por definición, las funciones de distribución marginales coinciden con la normal estándar. En el caso multivariante, la cópula Gaussiana se escribe:

$$C_\Sigma(u_1, \ldots, u_k) = \Phi_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_k))$$

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

La expresión de la densidad de la cópula Gaussiana multivariante puede escribirse como:

$$c_X\left(u_X; \Sigma_X\right) = \frac{1}{\sqrt{|\Sigma_X|}} exp\left\{ -\frac{1}{2} \xi_X' (\Sigma_X^{-1} - I_N) \xi_X) \right\} \tag{2.5}$$

La ρ_S de Spearman y la τ de Kendall para la cópula Gaussiana vienen dadas por (Žežula, 2009; Silva and Lopes, 2008):

$$\rho_S = \frac{6}{\pi} arcsen(\frac{\rho}{2}) \qquad y \qquad \tau = \frac{2}{\pi} arcsen(\rho)$$

Cópula de Clayton

Es un cópula Arquimediana, donde el generador es $\varphi = (\frac{1}{\theta_c})(t^{-\theta_c} - 1)$, $\theta_c > 0$, por lo que, en términos bidimensionales la función cópula de Clayton viene dada por:

$$C(u_1, u_2) = \left(u_1^{-\theta_c} + u_2^{-\theta_c} - 1\right)^{\frac{-1}{\theta_c}}$$

y para $k \geq 2$, la función cópula de Clayton tiene la forma:

$$C(u_1, \ldots, u_k) = \left(1 + \sum_{i=1}^{k} (u_i^{-\theta_c} - 1)\right)^{-\frac{1}{\theta_c}}$$

de manera que la densidad de la cópula de Clayton es

$$c(u_1, \ldots, u_k) = \left(1 - k + \sum_{i=1}^{k} u_i^{-\theta_c}\right)^{-k-(\frac{1}{\theta_c})} \prod_{l=1}^{k} \left[u_l^{-\theta_c - 1}(\theta_c(l-1) + 1)\right]$$

La cópula de Clayton coincide con la cópula de independencia cuando $\theta_c \to 0$ y coincide con la cópula de dependencia positiva perfecta cuando $\theta_c \to \infty$. El valor de la τ de Kendall, para esta cópula, se puede obtener en función del parámetro, θ_c, de la cópula como sigue (Silva and Lopes, 2008):

$$\tau(\theta_c) = \frac{\theta_c}{\theta_c + 2}$$

2.1.5. Estructura de Correlación Uniforme para la Cópula Gaussiana

El propósito de utilizar cópulas en la construcción del modelo, se debe al hecho de usarlas como herramientas para describir la relación de dependencia entre las variables. Debido a la naturaleza múltiple del problema planteado en el capítulo 1 y la necesidad de estimar una gran cantidad de parámetros utilizaremos, cuando la dependencia sea modelada mediante la cópula Gaussiana, una estructura especial de la matriz de correlación propuesta por Žežula (2009), destinada a reducir el número de parámetros en la construcción de la cópula Gaussiana. La relación de dependencia entre variables aleatorias puede ser cuantificada mediante diversas medidas, siendo la más conocida el coeficiente de correlación de Pearson, sin embargo, esta medida es la

más limitada puesto que refleja únicamente un tipo de dependencia, la dependencia lineal, lo que hace que no sea una medida apropiada en muchas situaciones como, por ejemplo, en los casos donde se trabaja con distribuciones no Elípticas.

En cualquier caso, para diseñar nuestro modelo, utilizaremos en primer lugar la cópula Gaussiana N-variante y la relación de dependencia será cuantificada a través del coeficiente de correlación de Pearson. Además, debido a la dimensión y a la complejidad de la estructura del tipo de datos procedentes de microarrays de ADN, que implica la estimación de una gran cantidad de parámetros, usaremos la estructura de correlación propuesta por Žežula (2009):

$$\Sigma = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} = (1-\rho)I_N + \rho 11' \tag{2.6}$$

donde $\rho \in \left[\frac{-1}{N-1}; 1\right]$. Si $\rho \neq 1$ y $\rho \neq \frac{-1}{N-1}$, entonces el determinante $|\Sigma| = [1+(N-1)\rho](1-\rho)^{N-1}$ es distinto de cero y existe la matriz inversa de Σ, cuya expresión es $\Sigma^{-1} = \frac{1}{1-\rho}\left(I_N - \frac{\rho}{1+(N-1)\rho}11'\right)$, siendo 1 un vector N-dimensional con todas sus componentes igual a uno.

Por tanto, la densidad de la cópula Gaussiana, de acuerdo con la estructura de correlación uniforme, tiene la siguiente expresión (Žežula, 2009):

$$c(u_X;\rho) = \frac{1}{|\Sigma|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}\xi'\left(\Sigma^{-1}-I\right)\xi\right\} \tag{2.7}$$

$$c(u_X;\rho) = \frac{1}{\left\{[1+(N-1)\rho](1-\rho)^{N-1}\right\}^{\frac{1}{2}}} exp\left\{-\frac{\rho}{2(1-\rho)}\xi'\left(I_N - \frac{1}{1+(N-1)\rho}11'\right)\xi\right\}$$

$$\begin{aligned} c(u_X;\rho) &= \frac{1}{\left\{[1+(N-1)\rho](1-\rho)^{N-1}\right\}^{\frac{1}{2}}} \\ &\times exp\left\{-\frac{\rho}{2(1-\rho)}\frac{1}{[1+(N-1)\rho]}\left((N-1)\rho\sum_{i=1}^{N}\xi_i^2 - 2\sum_{i<j}\sum\xi_i\xi_j\right)\right\} \end{aligned}$$

La ventaja de usar la cópula Gaussiana es que permite una estructura flexible de dependencia paramétrica. Sin embargo, esta cópula considera solamente la dependencia de pares entre componentes individuales de variables aleatorias, es decir, no

incorpora estructuras de dependencia más complejas, pero a cambio depende de un solo parámetro.

Escoger una cópula para ajustar un conjunto de datos implica un extenso trabajo de ajuste, puesto que existen muchas familias de funciones para elegir y se tendría que estimar el parámetro o el conjunto de parámetros de cada familia, por lo que es importante saber cómo seleccionar un número reducido de familias, entre las cuales esperamos elegir la cópula que mejor describe nuestros datos.

En la literatura se han propuesto varios métodos de selección de cópulas. Genest and Rivest (1993) propusieron un procedimiento no paramétrico para la selección de cópulas de la familia Arquimediana que proporciona el mejor ajuste para datos pareados y Silva and Lopes (2008) utilizaron los métodos MCMC para estimar funciones cópulas con un número reducido de parámetros.

2.2. Planteamiento del Problema y Modelo General con Cópulas N-variantes

Consideramos N variables aleatorias dependientes, medidas en dos situaciones distintas de tratamiento independientes. Es decir, sea $X = (X_1, X_2, \ldots, X_N)$ un vector de N variables aleatorias dependientes, medidas bajo una situación, e $Y = (Y_1, Y_2, \ldots, Y_N)$ el vector de N variables aleatorias dependientes, medidas bajo la otra situación, y tales que X e Y son independientes con distribución $F_X\left(X|\Theta_X, \lambda_X\right)$ y $F_Y\left(Y|\Theta_Y, \lambda_Y\right)$, respectivamente, donde $\Theta_X = (\theta_{X_1}, \ldots, \theta_{X_N})$ y $\Theta_Y = (\theta_{Y_1}, \ldots, \theta_{Y_N})$ son los vectores de parámetros sobre los que se realizarán los contrastes y $\Lambda = (\lambda_X, \lambda_Y)$ otro grupo de vectores de parámetros del modelo.

Deseamos comparar, para cada variable, el parámetro θ_{X_i} y θ_{Y_i} de las dos situaciones de tratamiento, para lo que consideramos, debido a la naturaleza múltiple del problema, el siguiente contraste múltiple de hipótesis:

$$H_{0i} : \theta_{X_i} = \theta_{Y_i} \quad \text{frente a} \quad H_{1i} : \theta_{X_i} \neq \theta_{Y_i}, i = 1, 2, \ldots, N \tag{2.8}$$

Nuestro objetivo es decidir qué hipótesis nulas se rechazan y cuales se aceptan por

medio de las probabilidades a posteriori de cada una de las hipótesis nulas. Por tanto, para obtener estas probabilidades y para realizar la inferencia sobre los parámetros, cuando las variables X_i e Y_i han sido observadas, hace falta determinar el modelo de distribución de probabilidad de las variables X e Y para obtener la función de verosimilitud, así como la distribución a posteriori de los parámetros desconocidos del modelo.

La función de densidad conjunta de X e Y se puede escribir como el producto de las densidades de X e Y, por considerar independencia entre las dos situaciones de tratamiento, así,

$$f\left(X,Y|\Theta\right)=f\left(X|\Theta_X,\lambda_X\right)f\left(Y|\Theta_Y,\lambda_Y\right) \tag{2.9}$$

donde $\Theta=(\Theta_X,\lambda_X,\Theta_Y,\lambda_Y)$ es el vector de parámetros del modelo.

Sin embargo, las N variables aleatorias en cada situación de tratamiento son dependientes entre si, es decir, tanto las variables X_i del vector $X=(X_1,\ldots,X_N)$ como las variables Y_i del vector $Y=(Y_1,\ldots,Y_N)$ son dependientes entre si, por lo que resulta necesario definir la función de densidad conjunta de las N variables en cada una de las dos situaciones de tratamiento. Para ello, consideramos una función cópula N-variante para cada una de las funciones de densidad $f\left(X|\Theta_X,\lambda_X\right)$ y $f\left(Y|\Theta_Y,\lambda_Y\right)$.

Con la finalidad de reflejar la dependencia entre las variables de cada situación, de acuerdo con la ecuación (2.2), a continuación se describe la función de densidad conjunta de X como el producto de la densidad de la cópula y el producto de las funciones de densidad marginales y de forma análoga para la variable Y.

- $f\left(X|\Theta_X,\lambda_X\right)=c_X(u_{X_1},...,u_{X_N};\omega_X)\prod_{i=1}^{N}f_i(X_i|\theta_{X_i},\lambda_{X_i})$ representa la función de densidad conjunta de las variables dependientes $X_i, i=1,2,...,N$, del vector $X=(X_1,...,X_N)$, definida por medio de una cópula $c_X(u_{X_1},...,u_{X_N};\omega_X)$, donde $u_{X_i}=F_i\left(x_i\right)$, $i=1,\ldots,N$, y ω_X es el vector de parámetros de la función de densidad de la cópula para la condición X, siendo $f_i(X_i|\theta_{X_i},\lambda_{X_i})$ la densidad marginal de X_i.

- $f\left(Y|\Theta_Y,\lambda_Y\right)=c_Y(u_{Y_1},...,u_{Y_N};\omega_Y)\prod_{i=1}^{N}f_i(Y_i|\theta_{Y_i},\lambda_{Y_i})$ representa la función de densidad conjunta de las variables dependientes $Y_i, i=1,2,...,N$, del vector $Y=(Y_1,...,Y_N)$, definida por medio de una cópula $c_Y(u_{Y_1},...,u_{Y_N};\omega_Y)$, donde $u_{Y_i}=F_i\left(y_i\right)$, $i=1,\ldots,N$, y ω_Y es el vector de parámetros de la función de densidad de la cópula para la condición Y, siendo $f_i(Y_i|\theta_{Y_i},\lambda_{Y_i})$ la densidad marginal de Y_i.

Por tanto, la función de densidad conjunta (2.9) para X e Y se puede escribir como sigue:

$$\begin{aligned} f\left(x_1,\ldots,x_N;y_1,\ldots,y_N|\Theta\right)=c_X\left(u_{X_1},u_{X_2},\ldots,u_{X_N};\omega_X\right)\prod_{i=1}^{N}f_i\left(x_i|\theta_{X_i},\lambda_{X_i}\right) \\ \times c_Y\left(u_{Y_1},u_{Y_2},\ldots,u_{Y_N};\omega_Y\right)\prod_{i=1}^{N}f_i\left(y_i|\theta_{Y_i},\lambda_{Y_i}\right) \end{aligned} \tag{2.10}$$

De esta manera, $\Theta=(\Theta_X,\Theta_Y,\lambda_X,\lambda_Y,\omega_X,\omega_Y)$ es el vector de parámetros actualizado. Por simplicidad, a lo largo de esta tesis usaremos la notación $c_X\left(u_X;\omega_X\right)$ y $c_Y\left(u_Y;\omega_Y\right)$ en lugar de $c_X\left(u_{X_1},u_{X_2},\ldots,u_{X_N};\omega_X\right)$ y $c_Y\left(u_{Y_1},u_{Y_2},\ldots,u_{Y_N};\omega_Y\right)$, respectivamente.

Desde la perspectiva bayesiana, para proceder a la inferencia, todas las cantidades desconocidas de Θ deben ser estimadas a partir de las distribución a posteriori.

$$\pi\left(\Theta\middle|x_{\cdot 1},\ldots,x_{\cdot n_x};y_{\cdot 1}\ldots,y_{\cdot n_y}\right)\propto\pi\left(\Theta\right)L\left(\Theta\middle|x_{\cdot 1},\ldots,x_{\cdot n_x};y_{\cdot 1}\ldots,y_{\cdot n_y}\right)$$

Siendo $x_{\cdot j}=\left(x_{1j},x_{2j},\ldots,x_{Nj}\right), j=1,2,\ldots,n_x$ e $y_{\cdot k}=\left(y_{1k},y_{2k},\ldots y_{Nk}\right), k=1,2,\ldots,n_y$ muestras de $X=(X_1,\ldots,X_N)$ e $Y=(Y_1,..,Y_N)$, donde n_x y n_y son los tamaños muestrales para X e Y, respectivamente.

Teniendo en cuenta (2.10), la verosimilitud se puede escribir de la forma siguiente:

$$\begin{aligned} L\left(\Theta\middle|x_{\cdot 1},\ldots,x_{\cdot n_x};y_{\cdot 1}\ldots,y_{\cdot n_y}\right)=\prod_{j=1}^{n_x}c_X\left(u_{X_{\cdot j}};\omega_X\right)\prod_{i=1}^{N}f_i\left(x_{ij}|\theta_{X_i},\lambda_{X_i}\right) \\ \times\prod_{k=1}^{n_y}c_Y\left(u_{Y_{\cdot k}};\omega_Y\right)\prod_{i=1}^{N}f_i\left(y_{ik}|\theta_{Y_i},\lambda_{Y_i}\right) \end{aligned} \tag{2.11}$$

Como se puede ver, la verosimilitud tiene una forma compleja puesto que depende de H_{0_i} y H_{1i} definidas en (2.8). Con objeto de hacerla más tratable, introducimos N

variables latentes independientes τ_i (Diebolt and Robert, 1994) definidas según la distribución de $Bernoulli\,(1-p_i)$, para $i=1,2,\ldots,N$, donde p_i es la probabilidad inicial de cada hipótesis nula,

$$\tau_i = \begin{cases} 0 & si \quad \theta_{X_i} = \theta_{Y_i} \\ 1 & si \quad \theta_{X_i} \neq \theta_{Y_i} \end{cases} \tag{2.12}$$

De ahí que, $Pr(\tau_i = 0|p_i) = p_i$ y $Pr(\tau_i = 1|p_i) = 1 - p_i$. Por lo que consideramos, para cada i, que el vector de observaciones $(x_i., y_i.)$ procede de la distribución bajo H_{0i} cuando $\tau_i = 0$ y de la distribución bajo H_{1i} cuando $\tau_i = 1$, siendo $x_i. = (x_{i1}, ..., x_{in_x})$ e $y_i. = (y_{i1}, ..., y_{in_y})$. Entonces,

$$\begin{aligned} X_{ij}|\tau_i = 0, \Theta \quad = \quad & X_{ij}|\tau_i = 1, \Theta \sim f_i\left(x_{ij}|\theta_{X_i}, \lambda_{X_i}\right), i = 1, \ldots, N, j = 1, \ldots, n_x \\ & Y_{ik}|\tau_i = 0, \Theta \sim f_i(y_{ik}|\theta_{X_i}, \lambda_{Y_i}), i = 1, \ldots, N, k = 1, \ldots, n_y \\ & Y_{ik}|\tau_i = 1, \Theta \sim f_i(y_{ik}|\theta_{Y_i}, \lambda_{Y_i}), i = 1, \ldots, N, k = 1, \ldots, n_y \end{aligned} \tag{2.13}$$

Desde el punto de vista bayesiano, las variables latentes $\tau = (\tau_i, \ldots, \tau_N)$ se pueden considerar como un conjunto adicional de parámetros. De esta manera, la verosimilitud (2.11) se puede expresar como sigue:

$$\begin{aligned} L\left(\Theta, \tau|x_{\cdot 1}, \ldots, x_{\cdot n_x}; y_{\cdot 1} \ldots, y_{\cdot n_y}\right) = & \prod_{j=1}^{n_x} c_X\left(u_{X_{\cdot j}}; \omega_X\right) \prod_{i=1}^{N} f_i\left(x_{ij}|\theta_{X_i}, \lambda_{X_i}\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}}; \omega_Y\right) \\ & \times \prod_{i:\tau_i=0} f_i\left(y_{ik}|\theta_{X_i}, \lambda_{Y_i}\right) \prod_{i:\tau_i=1} f_i\left(y_{ik}|\theta_{Y_i}, \lambda_{Y_i}\right) \end{aligned} \tag{2.14}$$

donde $\Theta = (\Theta_X, \Theta_Y, \lambda_X, \lambda_Y, \omega_X, \omega_Y, p)$ es el vector de parámetros, siendo $p = (p_1, \ldots, p_N)$ el vector de las probabilidades iniciales de cada hipótesis nula y $\tau = (\tau_1, \ldots, \tau_N)$ el vector de variables latentes.

Entonces, dada una distribución a priori para (Θ, τ), $\pi\left(\Theta, \tau\right)$, la distribución a posteriori se puede obtener de la siguiente manera:

$$\pi\left(\Theta, \tau\middle|x_{\cdot 1}, \ldots, x_{\cdot n_x}; y_{\cdot 1} \ldots, y_{\cdot n_y}\right) \propto \pi\left(\Theta, \tau\right) L\left(\Theta, \tau\middle|x_{\cdot 1}, \ldots, x_{\cdot n_x}; y_{\cdot 1} \ldots, y_{\cdot n_y}\right)$$

Esta distribución no siempre se puede obtener de forma analítica, pero puede ser aproximada mediante los métodos de cadenas de Markov Monte Carlo (MCMC).

En la siguiente sección se describe, de forma general, el algoritmo MCMC que será utilizado posteriormente en el capítulo 3.

A partir de la distribución a posteriori, se pueden estimar cada uno de los parámetros del modelo mediante las correspondientes marginales y, puesto que $Pr\left(\tau_i = 0 \middle| x_{\cdot 1}, \ldots, x_{\cdot n_x}; y_{\cdot 1} \ldots, y_{\cdot n_y}\right) = Pr\left(H_{0i} \middle| x_{\cdot 1}, \ldots, x_{\cdot n_x}; y_{\cdot 1} \ldots, y_{\cdot n_y}\right)$, también se puede estimar la probabilidad a posteriori de cada hipótesis nula, a través de la correspondiente distribución marginal de τ, con la finalidad de decidir, en función de estas probabilidades, las hipótesis nulas que se aceptan y las que se rechazan, siendo éste el objetivo principal del procedimiento.

2.3. Inferencia Basada en los Métodos de Cadenas de Markov Monte Carlo (MCMC)

La necesidad de integrar funciones, en la mayoría de las veces complejas y multidimensionales, es extremamente importante en la inferencia bayesiana. Sin embargo, no siempre se puede obtener la expresión analítica de estas integrales, pero pueden ser aproximadas mediante los métodos basados en cadenas de Markov Monte Carlo (MCMC). De esta manera, la inferencia se basa en muestras obtenidas por medio de distribuciones condicionadas a posteriori.

En este trabajo utilizaremos un enfoque bayesiano, en el cual los parámetros serán estimados conjuntamente a través de los métodos MCMC. Utilizaremos especialmente el algoritmo de Gibbs para distribuciones condicionadas a posteriori con forma conocida y el algoritmo de Metropolis-Hastings para distribuciones condicionadas a posteriori desconocidas. La combinación de estos dos algoritmos es conocido como Algoritmo Metropolis-Hastings-within-Gibbs.

Algoritmo Metropolis-Hastings

Este algoritmo permite generar una muestra de la distribución conjunta a posteriori, $\pi(\theta|X)$, a partir de las distribuciones condicionadas a posteriori, que pueden

tener forma conocida o no (Chib and Greenberg, 1995).

La idea básica es simular una cadena de Markov, $\theta^1, \theta^2, \ldots, \theta^M$, cuya distribución estacionaria sea $\pi(\theta|X)$, la distribución de interés en el problema.

Cuando las distribuciones condicionadas a posteriori no tienen una forma conocida, se puede utilizar el algoritmo Metropolis que genera el valor del parámetro a partir de una distribución propuesta $q(.|.)$ y la probabilidad de aceptación del valor generado tiene la forma siguiente (Chib and Greenberg, 1995; Robert and Casella, 2013):

$$\alpha(\theta, \theta^c) = min\left\{1, \frac{\pi(\theta^c|X)q(\theta|\theta^c)}{\pi(\theta|X)q(\theta^c|\theta)}\right\}$$

Una característica importante de $\alpha(\theta, \theta^c)$ es que la constante normalizadora de la distribución a posteriori no necesita ser conocida, es decir, α no depende de la constante de integración de la distribución a posteriori.

Sea $\{\theta^1, \ldots, \theta^M, \ldots\}$ una cadena de Markov. De acuerdo con el algoritmo Metropolis-Hastings, dado el valor actual θ^t, el valor de θ para el próximo estado, θ^{t+1}, se escoge muestreando un candidato θ^c de la distribución propuesta $q(\theta^c|\theta)$. De este modo, el punto candidato θ^c se acepta con probabilidad $\alpha(\theta, \theta^c)$.

Cuando $q(\theta^c|\theta) = q(\theta|\theta^c)$, la distribución propuesta es simétrica y el término $\frac{q(\theta|\theta^c)}{q(\theta^c|\theta)}$ se cancela, entonces la probabilidad de aceptación se reduce a:

$$\alpha(\theta, \theta^c) = min\left\{1, \frac{\pi(\theta^c|X)}{\pi(\theta|X)}\right\}$$

Este es el algoritmo original desarrollado por Metropolis et al. (1953), que posteriormente fue generalizado por Hastings (1970).

Cuando $q(\theta^c|\theta) = q(\theta^c)$, la distribución propuesta es independiente de θ y el algoritmo se conoce como Metropolis–Hastings de Independencia.

Un caso especial del algoritmo Metropolis-Hasting es el algoritmo Metropolis-Hastings-within-Gibbs, que permite muestrear de distribuciones conocidas y no conocidas en un mismo algoritmo. A continuación se describe este algoritmo.

Algoritmo Metropolis-Hastings-within-Gibbs (Patz and Junker, 1999):

Sea $\pi(\theta, \beta|X)$ la distribución estacionaria deseada y $q_\theta(\theta^0, \theta^1)$ y $q_\beta(\beta^0, \beta^1)$ dos distribuciones diferentes generadoras de candidatos (θ^c, β^c).

1. Deseamos generar $\theta^k \sim p(\theta|\beta^{k-1}, X)$:

   a) Genera $\theta^c \sim q_\theta(\theta^{k-1}, \theta)$

   b) Acepta $\theta^k = \theta^c$ con probabilidad

$$\alpha(\theta^{k-1}, \theta^c) = min\{1, \frac{p(X|\theta^c, \beta^{k-1})p(\theta^c, \beta^{k-1})q_\theta(\theta^c, \theta^{k-1})}{p(X|\theta^{k-1}, \beta^{k-1})p(\theta^{k-1}, \beta^{k-1})q_\theta(\theta^{k-1}, \theta^c)}\}$$

   caso contrario $\theta^k = \theta^{k-1}$

2. Deseamos generar $\beta^k \sim p(\beta|\theta^k, X)$:

   a) Genera $\beta^c \sim q_\beta(\beta^{k-1}, \beta)$

   b) Acepta $\beta^k = \beta^c$ con probabilidad

$$\alpha(\beta^{k-1}, \beta^c) = min\{1, \frac{p(X|\theta^k, \beta^c)p(\theta^k, \beta^c)q_\beta(\beta^c, \beta^{k-1})}{p(X|\theta^k, \beta^{k-1})p(\theta^k, \beta^{k-1})q_\beta(\beta^{k-1}, \beta^c)}\}$$

   caso contrario $\beta^k = \beta^{k-1}$

La cadena de Markov resultante tiene una distribución estacionaria $p(\theta, \beta|X)$.

Para una información completa sobre estos métodos se pueden consultar Casella and George (1992), Chib and Greenberg (1995), Patz and Junker (1999), Gelfand et al. (2003), Robert and Casella (2013) y sus referencias.

En el capítulo siguiente se aplica esta metodología cuando se considera que los parámetros de interés son las medias. La dependencia se modela, en primer lugar, mediante la cópula Gaussiana y, en segundo lugar, mediante la cópula de Clayton, considerando en ambos casos distribuciones marginales normales.

Capítulo 3

Aplicación Mediante Cópulas Gaussianas y de Clayton

El principal objetivo de este trabajo es la identificación de genes con expresión diferencial bajo dos condiciones experimentales en situaciones de dependencia, las más habituales en la práctica. Para ello, suponemos que se observan N variables aleatorias medidas en dos situaciones distintas e independientes de tratamiento. Puesto que el nivel de expresión suele estar correlacionado entre los genes, suponemos que estas variables, en cada condición, son dependientes.

En el campo de la genómica es habitual modelar los datos de expresiones génica mediante la distribución normal, como puede verse en muchos de los trabajos publicados al respecto (Scott and Berger, 2006; Zhao et al., 2008; Salazar, 2011; Gómez-Villegas et al., 2014), entre otros. Por lo que asumimos que estas variables tienen por distribución una normal y para modelar la dependencia, en primer lugar, consideramos la cópula Gaussiana de la familia Elíptica, por su flexibilidad y porque también lo aplicaremos a un modelo con funciones de distribución marginales normales y, en segundo lugar, utilizamos la cópula de Clayton de la familia Arquimediana, debido a que los datos reales que disponemos muestran correlación positiva alta, mientras que la correlación negativa es menos significativa, siendo ésta una característica de la cópula de Clayton (Silva and Lopes, 2008).

Este capítulo consta de 2 secciones. En la sección 3.1, se aplica la metodología

bayesiana propuesta de contrastes de hipótesis múltiples, modelando la dependencia mediante la cópula Gaussiana y considerando distribuciones marginales normales. Se definen las distribuciones a priori y se obtiene la distribución a posteriori, así como las distribuciones condicionadas a posteriori que serán utilizadas en el algoritmo MCMC, Metropolis-Hastings-within-Gibbs, necesario para obtener una muestra de la distribución a posteriori conjunta, a partir de la cual se estiman los parámetros del modelo y la probabilidad de que cada hipótesis nula sea cierta. El procedimiento se ilustra con datos simulados, estimando el valor del FDR con el objetivo de evaluar la proporción de falsos positivos. El procedimiento se aplica también a datos reales procedentes de un experimento con microarrays de ADN. En la sección 3.2, se aplica la metodología bayesiana propuesta de contrastes de hipótesis múltiples, considerando también distribuciones marginales normales pero la dependencia se modela mediante cópulas de Clayton. Igualmente el procedimiento se ilustra con los mismos datos, simulados y reales, que en la sección anterior. Por último, se utiliza el criterio de selección DIC (Deviance Information Criterion) para comparar ambos modelos.

3.1. Modelo con Cópulas Gaussianas N-variantes y Funciones de Distribución Marginales Normales

Sean X_i, $i = 1, 2, \ldots, N$, las variables aleatorias dependientes medidas bajo una condición experimental e Y_i, $i = 1, 2, \ldots, N$, las variables aleatorias dependientes medidas bajo la otra condición experimental, con X_i e Y_i independientes, $i = 1, 2, \ldots, N$.

Para decidir si, para cada variable, existen diferencias entre los dos tratamientos, consideramos el siguiente contraste de hipótesis múltiples:

$$H_{0i} : \mu_{X_i} = \mu_{Y_i} \quad \text{frente a} \quad H_{1i} : \mu_{X_i} \neq \mu_{Y_i}, \quad i = 1, 2, \ldots, N$$

donde μ_{X_i} y μ_{Y_i} son las medias de las variables X_i e Y_i, respectivamete, $i = 1, \ldots, N$.

Asumimos que las variables X_i e Y_i, $i = 1, \ldots, N$, tienen distribución normal. Concretamente, suponemos que $X_i \sim N(\mu_{X_i}, \sigma_i^2)$ e $Y_i \sim N(\mu_{Y_i}, \sigma_i^2)$, es decir, consideramos varianzas iguales para cada variable en las dos condiciones de tratamiento, $\sigma_i^2 = \sigma_{X_i}^2 = \sigma_{Y_i}^2$, para $i = 1, \ldots, N$, por simplicidad, pero el procedimiento también se puede aplicar considerando, para cada variable, distintas varianzas en cada condición de tratamiento, mientras que suponemos varianzas distintas para todas las variables.

Para modelar la dependencia entre las variables consideramos para la distribución de densidad conjunta, en cada condición de tratamiento, una cópula Gaussiana N-dimensional, ya que utiliza la correlación entre pares de variables precisamente de la misma forma que la distribución normal multivariante y también porque esta cópula permite cualquier distribución marginal. De manera que, se utilizan las cópulas Gaussianas N-dimensionales $c_X\left(u_X; \Sigma_X\right)$ y $c_Y\left(u_Y; \Sigma_Y\right)$ para X e Y respectivamente, definidas en (2.5):

$$\begin{aligned} c_X\left(u_X; \Sigma_X\right) &= \frac{1}{\sqrt{|\Sigma_X|}} exp\left\{-\frac{1}{2}\xi_X'(\Sigma_X^{-1} - I_N)\xi_X)\right\} \\ c_Y\left(u_Y; \Sigma_Y\right) &= \frac{1}{\sqrt{|\Sigma_Y|}} exp\left\{-\frac{1}{2}\psi_Y'\left(\Sigma_Y^{-1} - I_N\right)\psi_Y\right\} \end{aligned} \tag{3.1}$$

donde $\xi_X = (\xi_{X_1} = \Phi^{-1}(u_{X_1}), ..., \xi_{X_N} = \Phi^{-1}(u_{X_N}))$ y $\psi_Y = (\psi_{Y_1} = \Phi^{-1}(u_{Y_1}), .., \psi_{Y_N} = \Phi^{-1}(u_{Y_N}))$ son los cuantiles de orden u_{X_i} y u_{Y_i}, respectivamente, de la distribución $N(0,1)$, $i = 1, 2, \ldots, N$, y Σ_X y Σ_Y son las matrices de correlación de las cópulas para las condiciones X e Y, respectivamente, donde $X = (X_1, \ldots, X_N)$ e $Y = (Y_1, \ldots, Y_N)$. Sin pérdida de generalidad, consideramos la misma estructura de dependencia para las dos condiciones de tratamientos, por tanto, $\Sigma_X = \Sigma_Y = \Sigma$.

Entonces, la densidad conjunta de X e Y, definida en (2.10), viene dada por:

$$\begin{aligned} f\left(x_1, \ldots, x_N; y_1, \ldots, y_N|\Theta\right) = c_X\left(u_X; \Sigma\right) \prod_{i=1}^{N} f_i\left(x_i\middle|\mu_{X_i}, \sigma_i^2\right) \\ \times\, c_Y\left(u_Y; \Sigma\right) \prod_{i=1}^{N} f_i\left(y_i\middle|\mu_{Y_i}, \sigma_i^2\right) \end{aligned}$$

donde $f_i\left(x_i|\mu_{X_i}, \sigma_i^2\right) \sim N(\mu_{X_i}, \sigma_i^2)$, $f_i\left(y_i|\mu_{Y_i}, \sigma_i^2\right) \sim N(\mu_{Y_i}, \sigma_i^2)$ y $\Theta = (\mu_X, \mu_Y, \sigma^2, \Sigma)$, siendo $\mu_X = (\mu_{X_1}, \ldots, \mu_{X_N})$, $\mu_Y = (\mu_{Y_1}, \ldots, \mu_{Y_N})$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_N^2)$ y Σ la matriz de correlación (2.6) de las cópulas que se especifica más adelante.

Para obtener la verosimilitud, al igual que en la sección anterior, consideramos las variables latentes definidas en (2.12), que en este caso, se tiene:

$$\begin{aligned} X_{ij}|\tau_i = 0, \Theta = X_{ij}|\tau_i = 1, \Theta &\sim N(\mu_{X_i}, \sigma_i^2) \\ Y_{ik}|\tau_i = 0 &\sim N(\mu_{X_i}, \sigma_i^2) \\ Y_{ik}|\tau_i = 1 &\sim N(\mu_{Y_i}, \sigma_i^2) \end{aligned}$$

donde $\Theta = (\mu_X, \mu_Y, \sigma, p, \sum)$, con $p = (p_1, \ldots, p_N)$, siendo $p_i = Pr(\tau_i = 0|p_i)$, $i = i, \ldots, N$. De esta manera, consideramos que, para cada $i = 1, \ldots, N$, el vector de observaciones $(x_i., y_i.)$ procede de una distribución bajo H_{0i} cuando $\tau_i = 0$ y, por tanto, consideramos las densidades marginales de $x_i.$ e $y_i.$ definidas según la misma ley $N(\mu_{X_i}, \sigma_i^2)$ para las dos situaciones de tratamiento. Por otro lado, cuando se observa el valor de la variable latente $\tau_i = 1$, consideramos que el vector de observaciones $(x_i., y_i.)$, $i = 1, \ldots, N$, procede de una distribución bajo H_{1i} y, por tanto, las densidades marginales del vector $x_i.$, para la primera situación de tratamiento, estarán definidas por la ley $N(\mu_{X_i}, \sigma_i^2)$, mientras que para el vector $y_i.$, para la segunda situación de tratamiento, estarán definidas por la ley $N(\mu_{Y_i}, \sigma_i^2)$.

Entonces, dadas $x._j = (x_{1j}, x_{2j}, ..., x_{Nj})$, $j = 1, 2, ..., n_x$ e $y._k = (y_{1k}, y_{2k}, ...y_{Nk})$, $k = 1, 2, ..., n_y$, muestras aleatorias de $X = (X_1, ..., X_N)$ e de $Y = (Y_1, .., Y_N)$, respectivamente, la verosimilitud se puede escribir como sigue:

$$\begin{aligned} L\left(\Theta, \tau | x._1, \ldots, x._{n_x}; y._1 \ldots, y._{n_y}\right) = &\prod_{j=1}^{n_x} c_X\left(u_{X._j}; \Sigma\right) \prod_{i=1}^{N} f_i\left(x_{ij} \middle| \mu_{X_i}, \sigma_i^2\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y._k}; \Sigma\right) \\ &\times \prod_{i:\tau_i=0} f_i\left(y_{ik} \middle| \mu_{X_i}, \sigma_i^2\right) \prod_{i:\tau_i=1} f_i\left(y_{ik} \middle| \mu_{Y_i}, \sigma_i^2\right) \end{aligned} \tag{3.2}$$

Debido a la naturaleza múltiple del problema y la necesidad de estimación de una gran cantidad de parámetros, ya que, además de μ_X, μ_Y, σ^2, p, y τ, todos ellos vectores de dimensión N, se tiene que Σ es una matriz de parámetros de dimensión $N \times N$, utilizaremos para esta matriz la estructura de correlación uniforme (2.6), propuesta por Žežula (2009), destinada a la construcción de la cópula Gaussiana con pocos parámetros, puesto que Σ depende únicamente del parámetro de correlación ρ.

Así, partiendo de (2.7), las cópulas se pueden escribir de la forma siguiente:

$$c_X(u_{X_{\cdot j}};\rho) = \frac{1}{\left\{\left[1+(N-1)\rho\right](1-\rho)^{N-1}\right\}^{\frac{1}{2}}} \tag{3.3}$$
$$\times\, exp\left\{-\frac{\rho}{2(1-\rho)}\frac{1}{\left[1+(N-1)\rho\right]}\left((N-1)\rho\sum_{i=1}^{N}\xi_{ij}^2 - 2\sum_{i=1}^{N}\sum_{m>i}\xi_{ij}\xi_{mj}\right)\right\}$$
$$c_Y\left(u_{Y_{\cdot k}};\rho\right) = \frac{1}{\left\{\left[1+(N-1)\,\rho\right](1-\rho)^{N-1}\right\}^{\frac{1}{2}}} \tag{3.4}$$
$$\times\, exp\left\{-\frac{\rho}{2\,(1-\rho)}\frac{1}{\left[1+(N-1)\,\rho\right]}\left((N-1)\,\rho\sum_{i=1}^{N}\psi_{ik}^2 - 2\sum_{i=1}^{N}\sum_{m>i}\psi_{ik}\psi_{mk}\right)\right\}$$

donde $\rho \in \left[\frac{-1}{N-1};1\right]$, es el coeficiente de correlación de Pearson, $\xi_X = (\xi_{X_1} = \Phi^{-1}(u_{X_1}),\ldots,\xi_{X_N} = \Phi^{-1}(u_{X_N}))$ y $\psi_Y = (\ \psi_{Y1} = \Phi^{-1}\left(u_{Y_1}\right),\ldots,\ \psi_{YN} = \Phi^{-1}\left(u_{Y_N}\right))$ son los cuantiles de orden u_{X_i} y u_{Y_i}, respectivamente, de la distribución $N(0,1)$, $i = 1,2,\ldots,N$.

Con esta estructura, se ha utilizado la cópula como una herramienta para describir la dependencia entre las variables, mientras que, para cuantificar la relación de dependencia entre las variables, se ha utilizado el coeficiente de correlación de Pearson, puesto que las densidades Gaussianas se pueden parametrizar utilizando este coeficiente (Žežula, 2009).

Una vez especificada la verosimilitud, se puede obtener la distribución a posteriori como se indica a continuación:

$$\pi\left(\Theta,\tau\middle|x_{\cdot 1},\ldots,x_{\cdot n_x};y_{\cdot 1}\ldots,y_{\cdot n_y}\right) \propto \pi\left(\Theta,\tau\right) L\left(\Theta,\tau\middle|x_{\cdot 1},\ldots,x_{\cdot n_x};y_{\cdot 1}\ldots,y_{\cdot n_y}\right)$$

Con la finalidad de estimar los parámetros, $\Theta = (\mu_X,\mu_Y,\sigma^2,p,\rho)$, y la probabilidad a posteriori de cada hipótesis nula, mediante la correspondiente distribución marginal de τ, es necesario especificar la distribución a priori $\pi(\Theta,\tau)$, lo que se aborda en la siguiente subsección.

3.1.1. Distribuciones a Priori y a Posteriori para los Parámetros del Modelo

En el contexto bayesiano, los parámetros son considerados variables aleatorias, por lo que en esta subsección especificaremos funciones de distribución a priori para los parámetros del modelo.

Consideramos distribuciones a priori independientes para todos los parámetros del modelo como en Silva and Lopes (2008), excepto para τ, definido en (2.12), cuya distribución depende de p, puesto que la función de densidad conjunta, en los modelos con cópulas, se puede separar en dos partes, una parte referente a la estructura de dependencia, representada por la cópula, y la otra parte referente a la estructura de independencia, representada por el producto de las funciones de densidad marginales, por lo que hemos considerado especificar las distribuciones a priori teniendo en cuenta esta característica. Es decir, vamos a considerar la distribución a priori de ρ, el parámetro de la cópula, independiente de la distribución a priori de los parámetros de las marginales $(\mu_X, \mu_Y, \sigma^2, p)$ y para estos últimos, consideramos también distribuciones a priori independientes, puesto que son parámetros de la parte referente a la estructura de independencia. Por otro lado, considerando distribuciones a priori independientes se puede reducir la complejidad del modelo, que en si mismo es complejo debido al elevado número de parámetros. Sin embargo, es posible considerar distribuciones a priori que reflejan algún tipo de dependencia entre los parámetros, como por ejemplo en Scott and Berger (2006) y Ausín et al. (2011).

Así, la densidad a priori conjunta para los parámetros del modelo $(\Theta, \tau) = (\mu_X, \mu_Y, \sigma^2, p, \rho, \tau)$, donde $\mu_X = (\mu_{X_1}, .., \mu_{X_N}), \mu_Y = (\mu_{Y_1}, ..., \mu_{Y_N}), \sigma^2 = (\sigma_1^2, ..., \sigma_N^2)$, $p = (p_1, .., p_N)$ y $\tau = (\tau_1, .., \tau_N)$, se define de la siguiente manera:

$$\pi(\Theta, \tau) = \pi(\rho) \prod_{i=1}^{N} \pi(\mu_{X_i}) \pi(\mu_{Y_i}) \pi(\sigma_i^2) \pi(p_i) \pi(\tau_i | p_i) \tag{3.5}$$

Con el objetivo de proporcionar el mayor peso posible a la información aportada por los datos a la hora de tomar una decisión, hemos considerado, para la mayoría de los parámetros, distribuciones a priori no informativas.

Las funciones de distribución a priori para las medias bajo las dos condiciones de tratamiento, μ_{X_i} y μ_{Y_i}, se suponen procedentes de densidades uniformes con rangos $[a_{X_i}, b_{X_i}]$ y $[a_{Y_i}, b_{Y_i}]$, para $i = 1, 2, \ldots, N$, como en Broët et al. (2002).

Para la distribución a priori de las varianzas, σ_i^2, $i = 1, 2, \ldots, N$, se considera la

función de densidad a priori de Jeffreys como en Scott and Berger (2006):

$$\pi(\sigma_i^2) = \frac{1}{\sigma_i^2} \quad i = 1, 2, \ldots, N$$

Los parámetros τ_i, $i = 1, \ldots, N$, han sido incorporados en el modelo como provenientes de una distribución de Bernoulli, $\tau_i | p_i \sim Benoulli(1 - p_i)$, donde p_i es la probabilidad inicial de cada hipótesis nula, $i = 1, 2, \ldots, N$. Para este parámetro se asume como distribución a priori una distribución beta, $p_i \sim Beta(\alpha_i, \beta_i)$, $i = 1, 2, \ldots, N$, como en Ausín et al. (2011) y Salazar (2011).

Finalmente, para parametrizar la estructura de dependencia entre las variables del modelo mediante la cópula Gaussiana N-variante, hemos considerado el coeficiente de correlación de Pearson como la medida de dependencia entre las variables, siendo éste el parámetro de la cópula Gaussiana a estimar. Para este parámetro hemos considerado la distribución uniforme definida en el rango $[a, b]$, como en Silva and Lopes (2008) y Daniels and Pourahmadi (2009).

Por tanto, la distribución a priori conjunta de los parámetros (3.5) se puede escribir como sigue:

$$\pi(\Theta, \tau) = \frac{1}{b-a} \prod_{i=1}^{N} \frac{1}{b_{X_i} - a_{X_i}} \frac{1}{b_{Y_i} - a_{Y_i}} \frac{1}{\sigma_i^2} p_i^{1-\tau_i} (1 - p_i)^{\tau_i} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} p_i^{\alpha_i - 1} (1 - p_i)^{\beta_i - 1} \tag{3.6}$$

Dadas las observaciones de las variables en estudio, la distribución a priori conjunta (3.6) y la verosimilitud (3.2), la distribución a posteriori del conjunto de parámetros $(\Theta, \tau) = (\mu_X, \mu_Y, \sigma, p, \rho, \tau)$ resulta:

$$\begin{aligned} \pi\left(\Theta, \tau \middle| x_{\cdot 1}, \ldots, x_{\cdot n_x}; y_{\cdot 1} \ldots, y_{\cdot n_y}\right) \propto & \frac{1}{b-a} \prod_{i=1}^{N} \frac{1}{b_{X_i} - a_{X_i}} \frac{1}{b_{Y_i} - a_{Y_i}} \frac{1}{\sigma_i^2} p_i^{1-\tau_i} (1 - p_i)^{\tau_i} \\ & \times \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} p_i^{\alpha_i - 1} (1 - p_i)^{\beta_i - 1} \\ & \times \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; \rho) \prod_{i=1}^{N} f_i(x_{ij} | \mu_{X_i}, {\sigma_i}^2) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; \rho) \\ & \times \prod_{i:\tau_i=0} f_i(y_{ik} | \mu_{X_i}, {\sigma_i}^2) \prod_{i:\tau_i=1} f_i(y_{ik} | \mu_{Y_i}, {\sigma_i}^2). \end{aligned} \tag{3.7}$$

donde $c_X(u_{X_{\cdot j}}; \rho)$ y $c_Y(u_{Y_{\cdot k}}; \rho)$ son las cópulas Gaussianas definidas en (3.3) y

(3.4), respectivamente, $f_i\left(x_{ij}|\mu_{X_i},\sigma_i^2\right) \sim N(\mu_{X_i},\sigma_i^2)$, $f_i\left(y_{ik}|\mu_{X_i},\sigma_i^2\right) \sim N(\mu_{X_i},\sigma_i^2)$ y $f_i\left(y_{ik}|\mu_{Y_i},\sigma_i^2\right) \sim N(\mu_{Y_i},\sigma_i^2)$.

Como puede verse, la distribución a posteriori conjunta de los parámetros (3.7) es compleja, no tiene una forma conocida y, consecuentemente, tampoco se pueden obtener las distribuciones marginales de forma analítica. Sin embargo, la inferencia bayesiana se puede realizar utilizando los métodos de Cadenas de Markov Monte Carlo (MCMC). Mediante estos métodos, se puede simular una Cadena de Markov $\left\{(\Theta^{(l)},\tau^{(l)}) : l = 1,\ldots,M\right\}$ que converja a la distribución a posteriori (3.7). Así, podemos estimar los parámetros a partir de la muestra generada, por ejemplo, mediante las medias marginales de dicha muestra. En particular usaremos el algoritmo Metropolis-Hastings-within-Gibbs.

3.1.2. Determinación de Distribuciones Condicionadas a Posteriori

Para aplicar el algoritmo MCMC que se expone en la subseccion 3.1.3, es necesario determinar la distribución a posteriori de cada parámetro condicionada al resto de parámetros. A continuación se describen estas distribuciones, la obtención de las mismas de forma más detallada se presenta en el apéndice B, donde $c_X(u_{X.j};\rho)$ y $c_Y(u_{Y.k};\rho)$ son las cópulas definidas en (3.3) y (3.4), respectivamente, $\tau_{-i} = (\tau_1,\ldots,\tau_{i-1},\tau_{i+1},\ldots,\tau_N)$ y $\Theta_{-\theta_i}$ es el vector de parámetros $\Theta = (\mu_X,\mu_Y,\sigma^2,p,\rho)$ sin el parámetro indicado en el subíndice, siendo θ_i cualquier parámetro de ese vector.

Distribución condicionada a posteriori para τ_i, $i = 1,\ldots,N$, dadas las observaciones y el resto de parámetros.

La probabilidad obtenida de que τ_i tome el valor 0, condicionada por el resto de

parámetros y las observaciones es:

$$P_r\left(\tau_i = 0|\Theta, \tau_{-i}, X, Y\right) = \frac{p_i \prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}}; \rho\right) exp\left\{-\frac{1}{2{\sigma_i}^2} S_{\mu_{X_i}}\right\}}{p_i \prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}}; \rho\right) exp\left\{-\frac{1}{2{\sigma_i}^2} S_{\mu_{X_i}}\right\} + (1-p_i) \prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}}; \rho\right) exp\left\{-\frac{1}{2{\sigma_i}^2} S_{\mu_{Y_i}}\right\}} \tag{3.8}$$

donde $S_{\mu_{X_i}} = \left[n_y\left(\overline{y}_i. - \mu_{X_i}\right)^2\right]$ y $S_{\mu_{Y_i}} = \left[n_y\left(\overline{y}_i. - \mu_{Y_i}\right)^2\right]$

La probabilidad a posteriori de $\tau_i = 1$, $i = 1, 2, ..., N$, es:

$$Pr\left(\tau_i = 1|\Theta, \tau_{-i}, X, Y\right) = 1 - Pr\left(\tau_i = 0|\Theta, \tau_{-i}, X, Y\right)$$

Distribución condicionada a posteriori para p_i, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros.

La distribución condicionada a posteriori para la probabilidad inicial de cada hipótesis nula, p_i, dadas las observaciones y el resto de parámetros, que se obtiene es:

$$\pi(p_i|X, Y, \Theta_{-pi}, \tau) \sim Beta(\alpha_i + 1 - \tau_i, \beta_i + \tau_i) \tag{3.9}$$

Distribución condicionada a posteriori de μ_{X_i}, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros, cuando $\tau_i = 0$ y $\tau_i = 1$.

Las distribuciones condicionadas a posteriori obtenidas para las medias μ_{X_i}, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros, cuando $\tau_i = 0$ y $\tau_i = 1$, respectivamente son:

$$\pi(\mu_{X_i}|X, Y, \tau_i = 0, \tau_{-i}, \Theta_{-\mu_{X_i}}) = \frac{\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; \rho) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; \rho) f_i(\mu_{X_i})}{E_{f_i(\mu_{X_i})}\left[\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; \rho) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; \rho)\right]} \tag{3.10}$$

donde

$$f_i(\mu_{X_i}) \sim N\left(\frac{n_x \overline{x}_i. + n_y \overline{y}_i.}{n_x + n_y}, \frac{\sigma_i}{\sqrt{n_x + n_y}}\right) \tag{3.11}$$

$$\pi(\mu_{X_i}|X,Y,\tau_i=1,\tau_{-i},\Theta_{-\mu_{X_i}}) = \frac{\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}};\rho) f_i(\mu_{X_i})}{E_{f_i(\mu_{X_i})}\left[\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}};\rho)\right]} \tag{3.12}$$

donde

$$f_i(\mu_{X_i}) \sim N\left(\overline{x}_{i\cdot}, \frac{\sigma_i}{\sqrt{n_x}}\right) \tag{3.13}$$

Distribución condicionada a posteriori para μ_{Y_i}, cuando $\tau_i = 1$, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros.

Las distribuciones condicionadas a posteriori obtenidas para las medias μ_{Y_i}, cuando $\tau_i = 1$, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros, es la siguiente:

$$\pi(\mu_{Y_i}|X,Y,\tau_i=1,\tau_{-i},\Theta_{-\mu_{Y_i}}) = \frac{\prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}};\rho) f_i(\mu_{Y_i})}{E_{f_i(\mu_{Y_i})}\left[\prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}};\rho)\right]} \tag{3.14}$$

donde

$$f_i(\mu_{Y_i}) \sim N\left(\overline{y}_{i\cdot}, \frac{\sigma_i}{\sqrt{n_y}}\right) \tag{3.15}$$

Distribución condicionada a posteriori para ${\sigma_i}^2$, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros, cuando $\tau_i = 0$ y $\tau_i = 1$

A continuación se describen las distribuciones condicionadas a posteriori obtenidas para las varianzas ${\sigma_i}^2$, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros, cuando $\tau_i = 0$ y $\tau_i = 1$, respectivamente:

$$\pi\left({\sigma_i}^2 \middle| X, Y, \tau_i = v, \tau_{-i}, \Theta_{-{\sigma_i}^2}\right) = \frac{\prod_{j=1}^{n_x} c_X\left(u_{X.j}; \rho\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; \rho\right) f_i\left(\sigma_i^2\right)}{E_{f_i\left(\sigma_i^2\right)}\left[\prod_{j=1}^{n_x} c_X\left(u_{X.j}; \rho\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; \rho\right)\right]} \tag{3.16}$$

donde $v \in \{0, 1\}$ y

$$f_i\left(\sigma_i^2\right) \sim \text{Gamma Inversa}\left(\frac{n_x + n_y}{2}, \frac{A}{2}\right) \text{con A} = \sum_j \left(x_{ij} - \mu_{X_i}\right)^2 + \sum_k \left(y_{ik} - \mu_{X_i}\right)^2, \text{ si } v = 0 \tag{3.17}$$

$$f_i\left(\sigma_i^2\right) \sim \text{Gamma Inversa}\left(\frac{n_x + n_y}{2}, \frac{B}{2}\right) \text{con B} = \sum_k \left(x_{ij} - \mu_{Xi}\right)^2 + \sum_k \left(y_{ik} - \mu_{Y_i}\right)^2, \text{ si } v = 1 \tag{3.18}$$

Distribución condicionada a posteriori para ρ, dadas las observaciones y el resto de parámetros

Finalmente, la distribución condicionada a posteriori obtenida para el parámetro de dependencia, ρ, dadas las observaciones y el resto de parámetros es:

$$\pi\left(\rho | X, Y, \Theta_{-\rho}, \tau_i = v, \tau_{-i}\right) = \frac{\prod_{j=1}^{n_x} c_X\left(u_{X.j}; \rho\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; \rho\right)}{\int \prod_{j=1}^{n_x} c_X\left(u_{X.j}; \rho\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; \rho\right) d_\rho}, \quad v = \{0, 1\} \tag{3.19}$$

3.1.3. Algoritmo MCMC

Como ya se ha comentado anteriormente, la distribución a posteriori no se puede obtener de forma analítica, por lo que la inferencia se basará en muestras obtenidas por medio de las distribuciones condicionadas a posteriori, descritas en la subsección anterior, utilizando el algoritmo de Metropolis-Hastings-within-Gibbs, para obtener una muestra de la distribución a posteriori conjunta (3.7).

El algoritmo se implementa como sigue:

En primer lugar se eligen valores iniciales para cada una de las variables latentes y para todos los parámetros del modelo. Posteriormente se actualizan, en primer lugar, cada una de las variables latentes, mediante el algoritmo 1 del apéndice A, generando un valor de la distribución dada en (3.8). Seguidamente se actualizan cada uno de los parámetros p_i, para $i = 1, 2, \ldots, N$, generando un valor de la distribución Beta dada en (3.9), utilizando el algoritmo 2 del apéndice A. A continuación se actualizan $\mu_{Y_i}, \mu_{Y_i}, \sigma_i^2$, con $i = 1, 2, \ldots, N$ y el parámetro ρ. Los valores se generan, para μ_{X_i} de las distribuciones dadas en (3.10) y (3.12) cuando $\tau_i = 0$ y $\tau_i = 1$, respectivamente, para μ_{Y_i} de la distribución dada en (3.14), para σ_i^2 de la distribución dada en (3.16) y finalmente, para el parámetro ρ de la distribución dada en (3.19).

Todas estas distribuciones no son conocidas, por lo que los valores se generan utilizando el algoritmo Metropolis-Hastings (algoritmos 3, 4, 5, y 6 del apéndice A, respectivamente). Para aplicar estos últimos algoritmos, se consideran, como distribuciones generadoras de candidatos, para μ_{X_i} las distribuciones dadas en (3.11) y (3.13) cuando $\tau_i = 0$ y $\tau_i = 1$, respectivamente, para μ_{Y_i} la distribución dada en (3.15), para σ_i^2 las distribuciones dadas en (3.17) y (3.18), cuando $\tau_i = 0$ y $\tau_i = 1$, respectivamente y, por último, una distribución uniforme para el parámetro ρ.

En cada actualización se utilizan los valores de las variables latentes y de todos los parámetros que ya habían sido actualizados y los valores obtenidos en el paso anterior para el resto de las variables latentes y de parámetros que aún no han sido actualizados.

A continuación se presenta, en el cuadro 1, la estructura del algoritmo MCMC propuesto. El algoritmo detallado se describe en el apéndice A.

Cuadro 1: Estructura general del algoritmo MCMC.

Algorithm Algoritmo MCMC

Require: initial values $(\Theta^{(0)}, \tau^{(0)}) = (\mu_X^{(0)}, \mu_Y^{(0)}, \sigma^{2(0)}, p^{(0)}, \rho^{(0)}, \tau^{(0)})$. Where $\tau^{(0)} = (\tau_1^{(0)}, ..., \tau_N^{(0)})$, $p^{(0)} = (p_1^{(0)}, .., p_N^{(0)})$, $\mu_X^{(0)} = (\mu_{X_1}^{(0)}, ..., \mu_{X_N}^{(0)})$, $\mu_Y^{(0)} = (\mu_{Y_1}^{(0)}, ..., \mu_{Y_N}^{(0)})$, $\sigma^{2(0)} = (\sigma_1^{2(0)}, .., \sigma_N^{2(0)})$

Procedure

1: Let the current state of the Markov chain be $(\Theta^{(l)}, \tau^{(l)}) = (\mu_X^{(l)}, \mu_Y^{(l)}, \sigma^{(l)}, p^{(l)}, \rho^{(l)}, \tau^{(l)})$

2: **for** $l \in 1 : M$ **do**

3: Update $\tau_i^{(l)}$, for $i = 1, \ldots, N$ ▷ by sampling from (3.8)

4: Update $p_i^{(l)}$, for $i = 1, \ldots, N$ ▷ by sampling from (3.9)

5: Update $\mu_{X_i}^{(l)}$, for $i = 1, \ldots, N$ ▷ by sampling from (3.10) and (3.12) when $\tau_i^{(l+1)} = 0$ and $\tau_i^{(l+1)} = 1$, respectively

6: Update $\mu_{Y_i}^{(l)}$, for $i = 1, \ldots, N$ ▷ by sampling from (3.14)

7: Update $\sigma_i^{2(l)}$, for $i = 1, \ldots, N$ ▷ by sampling from (3.16) with (3.17) when $\tau_i^{(l+1)} = 0$ and with (3.18) when $\tau_i^{(l+1)} = 1$

8: Update $\rho^{(l)}$ ▷ by sampling from (3.19).

9: **end for**

End Procedure : Return $\left\{\left(\Theta^{(l)}, \tau^{(l)}\right) : l = 1, .., M\right\}$

Dada una muestra de la cadena de Markov Monte Carlo en equilibrio, obtenida mediante la aplicación del algoritmo Metropolis-Hasting-within-Gibbs (cuadro 1),

$$\left\{\mu_X^{(l)}, \mu_Y^{(l)}, \sigma^{2(l)}, p^{(l)}, \rho^{(l)}, \tau^{(l)}\right\}_{l=1}^{M}$$

con $\mu_X = (\mu_{X_1}, \ldots, \mu_{X_N})$, $\mu_Y = (\mu_{Y_1}, \ldots, \mu_{Y_N})$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_N^2)$, $p = (p_1, \ldots, p_N)$, y $\tau = (\tau_1, \ldots, \tau_N)$, es posible obtener estimadores de los parámetros mediante las medias marginales como sigue:

$$\hat{\mu}_{X_i} = E\left[\mu_{X_i} \middle| X, Y\right] \approx \frac{1}{M}\sum_{l=1}^{M} \mu_{X_i}^{(l)} \tag{3.20}$$

$$\hat{\mu}_{Y_i} = E\left[\mu_{Y_i} \middle| X, Y\right] \approx \frac{1}{M}\sum_{l=1}^{M} \mu_{Y_i}^{(l)} \tag{3.21}$$

$$\hat{\sigma}_i^2 = E\left[\sigma_i^2 \middle| X, Y\right] \approx \frac{1}{M}\sum_{l=1}^{M} \sigma_i^{2(l)} \tag{3.22}$$

$$\hat{p}_i = E\left[p_i \middle| X, Y\right] \approx \frac{1}{M}\sum_{l=1}^{M} p_i^{(l)} \tag{3.23}$$

$$\hat{\rho} = E\left[\rho \middle| X, Y\right] \approx \frac{1}{M}\sum_{l=1}^{M} \rho^{(l)} \tag{3.24}$$

para cada $i = 1, 2, \ldots, N$.

También es posible estimar la probabilidad a posteriori de cada hipótesis nula por

$$\hat{Pr}\left(H_{0i}|X, Y\right) = 1 - \hat{Pr}\left(H_{1i}|X, Y\right) = 1 - \frac{1}{M}\sum_{l=1}^{M} I(\tau_i^{(l)} = 1) \tag{3.25}$$

para $i = 1, 2, \ldots, N$.

Estas probabilidades permitirán resolver el problema del contraste de hipótesis múltiples, decidiendo, en función de las mismas, las hipótesis que se aceptan y las que se rechazan, para lo que, en esta tesis, se ha utilizado un criterio de decisión ya clásico, vease Duncan (1965) y Lewis and Thayer (2004), utilizando la regla bayes cuando se considera, para cada acción conjunta, una función de pérdida aditiva y, para cada acción individual, la función de pérdida 0-1 generalizada con costes iguales. De esta manera se rechazarán todas las hipótesis nulas tales que $\hat{Pr}\left(H_{0i}|X, Y\right) \leq 0{,}5$, aceptando el resto.

3.1.4. Simulación. Resultados

En las secciones anteriores se ha presentado el procedimiento bayesiano de contraste de hipótesis múltiples, propuesto en este trabajo, para contrastar simultáneamente un elevado número de hipótesis bajo dependencia, donde se propone utilizar un modelo con funciones de distribución marginales normales y en el que la dependencia se ha representado mediante cópulas Gaussianas. Con la finalidad de evaluar el comportamiento del procedimiento propuesto, a continuación se realiza un estudio con datos simulados.

Se simularon tres conjuntos de datos con distribución normal multivariante y con la estructura de correlación uniforme definida por Žežula (2009), de la forma siguiente:

Para cada conjunto de datos, se realizó una simulación con $N = 50$ hipótesis y con $n = 17$ observaciones por hipótesis/gen, de las cuales 7 corresponden a las observaciones para la primera condición de tratamiento X y 10 para la segunda condición Y. De manera que $n_x = 7$ y $n_y = 10$, en consonancia con lo que es habitual en los datos procedentes de experimentos con microarrays de ADN, donde el número de muestras suele ser pequeño en relación al número de hipótesis/genes, debido a que estos experimentos son muy costosos (Müller et al., 2004).

Los tres conjuntos de datos, correspondientes a las dos condiciones experimentales, X e Y, se generaron a partir de las distribuciones Gaussianas multivariantes $N_{50}(\mu_X, \Sigma)$ y $N_{50}(\mu_Y, \Sigma)$, repectivamente, con la particularidad de tener la misma matriz de varianzas-covarianzas Σ, donde el coeficiente de correlación $\rho = 0{,}8$. La simulación se realizó forzando un 80 % de hipótesis nulas verdaderas para el primer conjunto de datos, un 50 % para el segundo y un 20 % para el tercero.

Las componentes del vector de medias μ_X para la primera condición se eligieron en el rango (1150,1160) y para las componentes del vector de medias μ_Y de la segunda condición en el rango (1165,1180). Por último, las componentes del vector de desviaciones típicas σ se eligieron en el rango (8,16).

Mediante el algoritmo propuesto, se consideraron actualizaciones simultáneas

del vector de parámetros del modelo, $(\Theta, \tau) = (\mu_X, \mu_Y, \sigma^2, p, \tau, \rho)$, realizando 15000 iteraciones y descartando las primeras 7500 iteraciones de la salida del MCMC.

Para implementar el algoritmo Metropolis-Hastings-within-Gibbs (cuadro 1), se consideraron para μ_X, μ_Y, σ^2 y ρ, distribuciones generadoras de candidatos independientes del estado actual de la cadena y con forma próxima a las distribuciones a posteriori definidas (3.10), (3.12), (3.14), (3.16) y (3.19), con la finalidad de garantizar una convergencia más rápida del algoritmo. Concretamente, se consideraron para μ_{X_i}, cuando $\tau_i = 0$ y $\tau_i = 1$, las distribuciones descritas en (3.11) y (3.13), respectivamente, para μ_{Y_i} la distribución descrita en (3.15) y para σ_i^2, cuando $\tau_i = 0$ y $\tau_i = 1$, las distribuciones (3.17) y (3.18), respectivamente. Para el parámetro de dependencia ρ, se estableció una distribución generadora de candidatos uniforme en el rango $(0{,}6; 0{,}9)$, es decir, suponemos que hay una correlación fuerte entre las variables en estudio, ya que los datos se generaron con un coeficiente de correlación $\rho = 0{,}8$.

Para la distribución a priori de p_i, $Beta\,(\alpha_i, \beta_i)$, consideramos, para los tres conjuntos de datos, los mismos parámetros para todas las hipótesis por simplicidad, es decir, $p_i \sim Beta\,(\alpha, \beta)$, $i = 1, \ldots, N$, y con la finalidad de realizar un análisis de sensibilidad, se consideraron los siguientes valores para dichos parámetros $(\alpha; \beta)$: $(0{,}5; 1)\,, (1; 1)$, $(1; 0{,}5)$ y $(2; 0{,}5)$, puesto que con éstos parámetros se obtienen distribuciones a priori muy diferentes, en el sentido de que unas distribuciones son sesgadas a la izquierda, otras sesgadas a la derecha y también se obtiene la distribución uniforme.

Para decidir las hipótesis nulas que se rechazan, utilizamos la estimación de la probabilidad a posteriori de cada hipótesis nula, rechazando todas aquellas hipótesis con esta probabilidad menor o igual que 0,5:

$$\hat{Pr}\,(H_{0i}|X, Y) = 1 - \hat{Pr}\,(H_{1i}|X, Y) = 1 - \frac{1}{M}\sum_{l=1}^{M} I(\tau_i^{(l)} = 1) \leq 0{,}5$$

para $i = 1, \ldots, N$.

Finalmente, con el objetivo de evaluar también la proporción de falsos positivos, se obtuvo el FDR, como en Müller et al. (2004), Do et al. (2005) y Gómez-Villegas

et al. (2014), a partir del expected false discovery rate introducido por Genovese and Wasserman (2002, 2003):

$$FDR = \frac{\sum_{i=1}^{N} Pr(H_{0i}|X,Y)\delta_i}{\sum_{i=1}^{N} \delta_i}$$

donde $\delta_i = 1$ si la hipótesis nula H_{0i} es rechazada y $\delta_i = 0$ si es aceptada. El FDR fue estimado también por medio de una muestra obtenida mediante el mismo algoritmo Metropolis-Hastings-within-Gibbs.

En las tablas (3.1), (3.2) y (3.3) se presentan los resultados obtenidos de los tres conjuntos de datos simulados con un 80 %, un 50 % y un 20 % de hipótesis nulas verdaderas, respectivamente.

Tabla 3.1: Resultados para el modelo con cópulas Gaussianas y para el conjunto de datos correspondientes al 80 % de hipótesis nulas verdaderas, para diferentes valores de los parámetros (α, β) de la distribución a priori de p_i

$(\alpha;\beta)$	**(0,5; 1)**		**(1; 1)**		**(1; 0,5)**		**(2; 0,5)**		
$\widehat{\rho}$	0,73		0,77		0,752		0,76		
$\widehat{FDR}$	0,219		0,298		0,045		0,023		
	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Total
Verdaderas	0	39	14	25	37	2	38	1	39
Falsas	0	11	0	11	0	11	0	11	11
Total	0	50	14	36	37	13	38	12	50

Tabla 3.2: Resultados para el modelo con cópulas Gaussianas y para el conjunto de datos correspondientes al 50 % de hipótesis nulas verdaderas, para diferentes valores de los parámetros (α, β) de la distribución a priori de p_i

$(\alpha;\beta)$	**(0,5; 1)**		**(1; 1)**		**(1; 0,5)**		**(2; 0,5)**		
$\widehat{\rho}$	0,803		0,813		0,804		0,773		
$\widehat{FDR}$	0,089		0,17		0,079		0,065		
	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Total
verdaderas	0	25	1	24	19	6	22	3	25
Falsas	0	25	0	25	0	25	0	25	25
Total	0	50	1	49	19	31	22	28	50

Como se muestra en las tablas (3.1) y (3.2), para ambas estructuras de datos,

el procedimiento es muy sensible a la elección de los parámetros α y β para la distribución a priori de p_i, en cuanto al número de hipótesis nulas rechazadas y aceptadas, ya que se pueden observar diferencias muy significativas en los resultados, no siendo así para la estimación del parámetro de dependencia ρ para el que, en todos los casos, se obtienen valores muy similares y muy próximos al valor con el que se generaron los datos ($\rho = 0{,}8$). A la vista de los resultados, las distribuciones a priori de p_i que muestran mejores resultados son las distribuciones sesgadas a la derecha, es decir, cuando asumimos las distribuciones a priori $Beta\,(1; 0{,}5)$ y $Beta\,(2; 0{,}5)$, donde se obtienen valores estimados del FDR en unos niveles aceptables y significativamente inferiores a los obtenidos con las otras distribuciones a priori, siendo el valor de FDR más alto 0,079 para los datos correspondientes al 50 % de hipótesis nulas verdaderas, utilizando la distribución a priori $Beta\,(1; 0{,}5)$.

Las tablas (3.1) y (3.2), resultado de la simulación con porcentajes del 80 % y 50 % respectivamente de hipótesis nulas verdaderas, muestran que las distribuciones a priori beta para p_i sesgadas a la derecha son las más apropiadas para alcanzar altas tasas de acierto. Con el fin de comprobar si dichas distribuciones a priori, sesgadas a la derecha, para p_i también son adecuadas en el caso de bajo porcentaje de hipótesis nulas verdaderas, se simuló un conjunto de datos con un 20 % de hipótesis nulas verdaderas, cuyos resultados se muestran en la tabla (3.3). Se puede apreciar que estos resultados son similares a los obtenidos en las tablas (3.1) y (3.2). Es decir, el número de hipótesis nulas aceptadas y rechazadas está más cerca del número de hipótesis nulas verdaderas y falsas cuando la distribución a priori del parámetro p_i está sesgada a la derecha, incluso con un bajo porcentaje de hipótesis nulas verdaderas. Por tanto, con nuestro procedimiento, se obtienen buenos resultados siempre que se utilice una distribución a priori de p_i sesgada a la derecha, independientemente del número de hipótesis nulas verdaderas.

Así, a partir de los resultados obtenidos en todos los ejemplos simulados, podemos concluir que, en cuanto a la estimación del parámetro de dependencia ρ, el procedimiento es robusto a la elección de los parámetros de la distribución a priori de la probabilidad inicial de cada hipótesis nula, p_i. Además, el procedimiento ajusta bien

Tabla 3.3: Resultados para el modelo con cópulas Gaussianas y para el conjunto de datos correspondientes al 20 % de hipótesis nulas verdaderas, para diferentes valores de los parámetros (α, β) de la distribución a priori de p_i

$(\boldsymbol{\alpha};\boldsymbol{\beta})$	**(0,5; 1)**		**(1; 1)**		**(1; 0,5)**		**(2; 0,5)**		
$\widehat{\rho}$	0,773		0,78		0,775		0,795		
$\widehat{FDR}$	0,04		0,085		0,054		0,07		
	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Total
verdaderas	0	12	0	12	8	4	11	1	12
Falsas	0	38	0	38	0	38	1	37	38
Total	0	50	0	50	8	42	12	38	50

la dependencia, puesto que el coeficiente de correlación estimado está muy próximo al valor real con el que se generaron los datos.

Sin embargo, en relación al número de hipótesis nulas rechazadas, el procedimiento no es robusto respecto a la elección de dichos parámetros. En cualquier caso, en los ejemplos simulados, cuando se utiliza una distribución a priori beta sesgada a la derecha para el parámetro p_i de la distribución de Bernoulli, el número de hipótesis nulas rechazadas y aceptadas está muy próximo al número de hipótesis nulas falsas y verdaderas, rechazando un número de hipótesis nulas ligeramente superior al número de hipótesis nulas falsas, a la vez que el FDR se mantiene en unos niveles aceptables, por lo que el procedimiento propuesto resulta ser un procedimiento de contraste de hipótesis múltiple con una elevada potencia, puesto que rechaza un alto número de hipótesis nulas falsas.

Por tanto, el procedimiento que se propone de contraste de hipótesis múltiples resulta muy adecuado en el contexto de los experimentos con microarrays de ADN, ya que, el objetivo principal, en muchos de estos estudios, es obtener el mayor número posible de genes potencialmente expresados para, posteriormente, llevar a cabo estudios más detallados. De modo que, en esta fase de análisis, se pueden tolerar más falsos positivos para obtener el mayor número posible de genes que potencialmente se expresen de manera diferencial (Dudoit et al., 2003).

3.1.5. Aplicación a Datos Reales

A continuación se aplica el procedimiento, descrito en las secciones anteriores, a un conjunto de datos procedentes de microarrays de ADN. Este conjunto de datos consta de 38 genes obtenidos a partir de tejidos de biopsias duodenales, realizadas en 13 niños con enfermedad celíaca y de edad media 5,6 ($\pm$0,6) años y 7 niños controles de edad media 8,1 ($\pm$2,2) años, pertenecientes a parte del estudio realizado en Pascual et al. (2016). Los datos están disponibles en NCBI-GEO datasets (The National Center for Biotechnology Information-Gene Expression Omnibus) a través del número de acesso GSE76168.

La matriz de datos consta de 38 filas (genes) y 20 columnas, donde las primeras 7 columnas corresponden a las muestras de los niños controles y las otras 13 columnas a las muestras de los niños con enfermedad celíaca.

Todos los genes estudiados están asociados a la enfermedad celíaca pero se desconoce si todos cambian su nivel de expresión, por lo que el objetivo principal es identificar los genes con expresión diferencial. Para ello, se pretende contrastar, para cada gen, si existen diferencias significativas en el nivel medio de expresión entre las muestras de tejidos procedentes de pacientes celíacos y las muestras procedentes de los controles.

Los datos se modelaron mediante cópulas Gaussianas, utilizando la matriz de correlación uniforme definida por Žežula (2009), y distribuciones marginales normales, por lo que la verosimilitud es la definida en (3.2) con las cópulas (3.3) y (3.4).

En cuanto a las distribuciones a priori, se consideraron las mismas que en la subsección (3.1.1). Es decir, distribuciones uniformes para la media de la variable de nivel de expresión de cada gen, tanto en pacientes con enfermedad celíaca como en controles y la densidad de Jeffreys para la varianza de cada gen, considerando la misma en pacientes con enfermedad celíaca y controles. Por otro lado, como se puede ver en la figura 3.1, donde se ha realizado un mapa de calor para la correlación entre cada par de genes, existen correlaciones negativas, por lo que, para el parámetro de dependencia ρ, sería lo más adecuado considerar la distribución a priori uniforme

en el intervalo (-1,1), pero se optó por el intervalo (-0,027,1) debido a la restricción $\rho \in \left[\frac{-1}{N-1}; 1\right]$ expuesta en la subsección 2.1.5. del capítulo 2, considerando también esta distribución en el algoritmo 6 de apéndice A como distribución generadora de candidatos para ρ. Para la probabilidad inicial de cada hipótesis nula, p_i, se consideró, igualmente, una distribución beta con los mismos parámetros para todos los genes, como en el análisis con datos simulados.

De acuerdo con los resultados del análisis con datos simulados, en el que se observó que existía sensibilidad a la elección de los parámetros α y β de la distribución a priori de p_i, el modelo mediante el cual se identificaban mejor las hipótesis nulas ciertas y falsas correspondía al de las distribuciones a priori de p_i sesgadas a la derecha, principalmente al modelo con la distribución a priori $Beta\,(2; 0{,}5)$, por lo que, para el análisis de datos reales, se asume esta distribución para p_i, $i = 1, \ldots, 38$.

Se aplicó el algoritmo propuesto ejecutando 40000 iteraciones y descartando las 20000 primeras. La estimación del coeficiente de correlación es $\hat{\rho} = 0.173$ y para decidir qué hipótesis se rechazan, se estimó la probabilidad a posteriori de cada hipótesis nula, rechazando aquellas hipótesis con esta probabilidad menor o igual que 0,5, obteniendo, al igual que en Pascual et al. (2016), 16 genes con expresión diferencial pero con algunas diferencias en cuanto a los genes identificados.

Mediante nuestro procedimiento se identifican, al igual que en Pascual et al. (2016), los siguientes genes: $C1orf106$, $C2orf74$, $CCR4$, $CCR6$, $FASLG$, $ICOSLG$, $IL18R1$, $IL23A$, $JAK2$, $PLEK$, $TAGAP$, $TNFSF18$, sin embargo, no identifica los genes: $IL18RAP$, $IL6$ y $UBE2L3$, identificados en Pascual et al. (2016), e identifica los siguientes genes $FBX048$, $PTPN2$, $TNFAIP3$ y $TYK2$, no identificados por dichos autores.

Como se puede ver en la figura 3.1, existe mucha correlación entre algunos genes, sin embargo, el procedimiento utilizado en Pascual et al. (2016) se realiza bajo el supuesto de independencia entre los genes, por lo que nuestro modelo resulta más adecuado. Sin embargo, el valor estimado para $\hat{\rho} = 0{,}173$ es muy bajo teniendo en cuenta que hay genes con correlación alta positiva y genes con correlación alta negativa, como puede verse en la figura 3.1, lo que puede ser debido al hecho de

considerar el mismo coeficiente de correlación entre todos los pares de genes, con el objetivo de reducir el número de parámetros en el modelo, lo que es poco realista.

En la siguiente sección se modela la dependencia, para este mismo conjunto de datos, mediante la cópula de Clayton, como alternativa a la cópula Gaussiana, que permite modelar distribuciones multivariantes mediante una única función univariante y con dependencia asimétrica, puesto que estos datos muestran más correlaciones altas positivas que negativas.

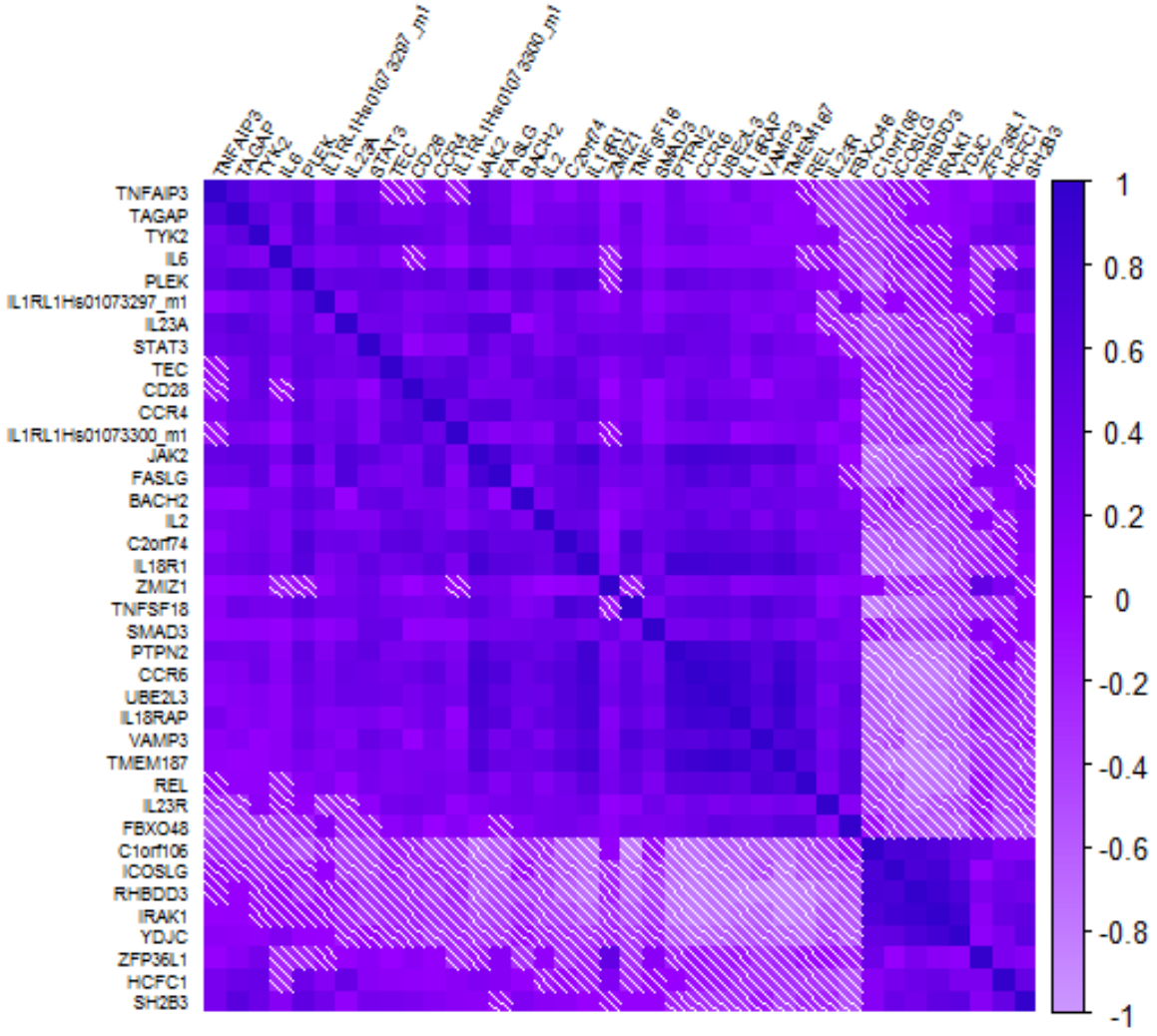


Figura 3.1: Gráfico de correlación entre cada par de genes para los datos de Pascual et al. (2016).

En el campo de la genómica, la distribución normal se usa ampliamente para modelar datos de expresiones génicas. Así, adoptamos distribuciones marginales

normales y modelamos la estructura de dependencia a través de la cópula Gaussiana, que comparte las propiedades de una distribución normal multivariante. Se ha optado por utilizar la matriz de correlación uniforme con la finalidad de reducir la dimensionalidad de los parámetros. Sin embargo, el enfoque propuesto es flexible en la medida en que puede usarse con otras matrices de correlación más realistas, aunque, el modelo podría resultar muy complejo, debido al elevado número de parámetros, si se consideran distintos coeficientes de correlación para cada par de variables. También podrían usarse otras funciones cópulas para modelar la dependencia, como se verá en la siguiente sección, así como con otras distribuciones marginales.

3.2. Modelo con Cópulas de Clayton N-variantes y Funciones de Distribución Marginales Normales

Para modelar la dependencia asimétrica en un conjunto de datos, se pueden utilizar las cópulas de la familia Arquimediana como alternativa a las cópulas de la familia Elíptica, como la cópula Gaussiana utilizada en la sección anterior.

Las cópulas de las familia Arquimediana abarcan un gran numero de cópulas con particularidades diferentes, caracterizándose por permitir modelar distribuciones multivariantes mediante una única función univariante, simplificando de esta manera los cálculos.

Suponer que una distribución multivariante es normal porque las marginales son normales, no siempre es lo más adecuado (Feller, 1966; Kowalski, 1973; Gelman and Meng, 1991), por lo que la cópula Gaussiana no siempre es la más adecuada para modelar la dependencia en un conjunto de datos donde las marginales son normales, aunque comparta las mismas propiedades de la normal multivariante para modelar la dependencia, es decir, aunque utilice la correlación entre pares de variables para modelar la dependencia, de la misma manera que la distribución normal multivariante.

En esta sección, los datos se modelan mediante distribuciones marginales normales, al igual que en la sección anterior, pero para representar la dependencia entre los datos se utiliza la cópula de Clayton de la familia Arquimediana.

Como se describe en la subsección 2.1.4 del capítulo 2, la cópula de Clayton, para $k \geq 2$, viene dada por:

$$C(u_1, \ldots, u_k) = \left(1 + \sum_{i=1}^{k} (u_i^{-\theta_c} - 1)\right)^{-\frac{1}{\theta_c}}$$

y la densidad de la cópula por:

$$c(u_1, \ldots, u_k) = \left(1 - k + \sum_{i=1}^{k} u_i^{-\theta_c}\right)^{-k-(\frac{1}{\theta_c})} \prod_{l=1}^{k} \left[u_l^{-\theta_c - 1}(\theta_c(l-1)+1)\right]$$

En esta sección, se aplica la metodología bayesiana de contraste de hipótesis múltiples propuesta en el capítulo 2, considerando, como en la sección anterior, distribuciones marginales normales para las observaciones, pero la dependencia se representa ahora mediante la cópula de Clayton, como alternativa a la cópula Gaussiana. Por tanto, la verosimilitud es la definida en (3.2), siendo $c_X\left(u_{X_{\cdot j}}; \theta_c\right)$ y $c_Y\left(u_{Y_{\cdot k}}; \theta_c\right)$ las cópulas de Clayton definidas a continuación:

$$\begin{aligned} c_X\left(u_{X_{\cdot j}}; \theta_c\right) &= \left(1 - N + \sum_{i=1}^{N} u_{X_{ij}}^{-\theta_c}\right)^{-N-(\frac{1}{\theta_c})} \prod_{l=1}^{N} \left[u_{X_{lj}}^{-\theta_c - 1}(\theta_c(l-1)+1)\right] \\ c_Y\left(u_{Y_{\cdot k}}; \theta_c\right) &= \left(1 - N + \sum_{i=1}^{N} u_{Y_{ik}}^{-\theta_c}\right)^{-N-(\frac{1}{\theta_c})} \prod_{l=1}^{N} \left[u_{Y_{lk}}^{-\theta_c - 1}(\theta_c(l-1)+1)\right] \end{aligned} \tag{3.26}$$

donde $u_{X_{\cdot j}} = (u_{X_{1j}} \ldots, u_{X_{Nj}})$, $u_{Y_{\cdot k}} = (u_{Y_{1k}} \ldots, u_{Y_{Nk}})$, siendo $u_{X_{ij}} = F\left(x_{ij}\right)$ y $u_{Y_{ik}} = F\left(y_{ij}\right)$, $i = 1, \ldots, N$, $j = 1, 2, \ldots, n_x$ y $k = 1, 2, \ldots, n_y$.

Así, se tiene que el conjunto de parámetros del modelo es $\Theta = (\mu_X, \mu_Y, \sigma^2, p, \theta_c)$, donde, $\mu_X = (\mu_{X_1}, \ldots, \mu_{X_N})$, $\mu_Y = (\mu_{Y_1}, \ldots, \mu_{Y_N})$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_N^2)$ y $p = (p_1, \ldots, p_N)$, siendo p_i la probabilidad inicial de cada hipótesis nula, y θ_c el parámetro de la cópula.

Para la distribución a priori de $\pi(\Theta, \tau)$, se considera, al igual que en la sección anterior, distribuciones a priori independientes para todos los parámetros, excepto para las variables latentes τ_i cuya distribución depende de p_i, esto es, se considera

la distribución a priori de θ_c, el parámetro de la cópula, independiente de la distribución a priori de los parámetros de las marginales $(\mu_X, \mu_Y, \sigma^2, p)$ y para estos, se consideran también distribuciones a priori independientes. Igualmente, se considera una distribución a priori uniforme en el rango $[a, b]$ para el parámetro de la cópula y también se consideran distribuciones a priori uniformes para las medias bajo las dos condiciones de tratamiento, μ_{X_i}, μ_{Y_i}, $i = 1, \ldots, N$, en los rangos $[a_{X_i}, b_{X_i}]$ y $[a_{Y_i}, b_{Y_i}]$, respectivamente, y la densidad de Jeffreys para las varianzas σ_i^2, $i = 1, \ldots, N$. Asimismo para la probabilidad inicial de cada hipótesis nula, p_i, $i = 1, \ldots, N$, se considera una distribución $Beta(\alpha_i, \beta_i)$.

Finalmente, los parámetros τ_i, $i = 1, \ldots, N$, se incluyen en el modelo como procedentes de una distribución de Bernoulli dependiente de p_i, $\tau_i | p_i \sim Bernoulli(1 - p_i)$. Por tanto, la distribución a priori conjunta es la definida en (3.6) y la distribución a posteriori la definida en (3.7), donde $c_X(u_{X_{.j}}; \theta_c)$ y $c_Y(u_{Y_{.k}}; \theta_c)$ son ahora las cópulas de Clayton definidas en (3.26).

Como en la sección anterior, la distribución a posteriori es compleja y no tiene una forma conocida, por lo que también se aplican los métodos MCMC para realizar la inferencia bayesiana, utilizando el mismo algoritmo Metropolis-Hastings-within-Gibbs (cuadro 1), descrito en la subsección (3.1.3), con la única diferencia de que, en este caso, se utilizan las cópulas de Clayton, definidas en (3.26), en lugar de las cópulas Gaussianas.

Una vez obtenida una muestra de la cadena de Markov Monte Carlo en equilibrio $\left\{\mu_X^{(l)}, \mu_Y^{(l)}, \sigma^{2(l)}, p^{(l)}, \theta_c^{(l)}, \tau^{(l)}\right\}_{l=1}^{M}$, donde $\mu_X = (\mu_{X_1}, \ldots, \mu_{X_N})$, $\mu_Y = (\mu_{Y_1}, \ldots, \mu_{Y_N})$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_N^2)$, $p = (p_1, \ldots, p_N)$ y $\tau = (\tau_1, \ldots, \tau_N)$, mediante el algoritmo Metropolis-Hasting-within-Gibbs, se puede aproximar, igualmente, la probabilidad a posteriori de la hipótesis nula por:

$$\hat{Pr}\left(H_{0i} | X, Y\right) = 1 - \hat{Pr}\left(H_{1i} | X, Y\right) = 1 - \frac{1}{M} \sum_{l=1}^{M} I(\tau_i^{(l)} = 1)$$

para $i = 1, 2, \ldots, N$, y decidir, en función de estas probabilidades, las hipótesis que se aceptan y las que se rechazan, para lo que se utilizará el mismo criterio que en la sección anterior, es decir, rechazar las hipótesis nulas tales que $\hat{Pr}\left(H_{0i} | X, Y\right) \leq 0{,}5$,

aceptando el resto. También se puede obtener la estimación de los parámetros del modelo, μ_{X_i}, μ_{Y_i}, σ_i^2 y p_i, $i = 1, \ldots, N$, a partir de las estimaciones de las medias marginales como en (3.20), (3.21), (3.22) y (3.23), respectivamente, y el estimador del parámetro de la cópula de Clayton, θ_c como:

$$\hat{\theta_c} = E\left[\theta_c | \, X, Y\right] \approx \frac{1}{M} \sum_{l=1}^{M} \theta_c^{(l)}$$

En la subsección siguiente, se aplica el procedimiento a datos simulados.

3.2.1. Simulación. Resultados

Se aplicó el procedimiento, modelando la dependencia con cópulas de Clayton, a los mismos conjuntos de datos utilizados en la subsección 3.1.4 con cópulas Gaussinas, generados de la distribución Gaussiana multivariante $N_{50}\left(\mu_X, \Sigma\right)$ para la primera condición de tratamiento, X, y de la distribución Gaussiana multivariante $N_{50}\left(\mu_Y, \Sigma\right)$ para la segunda condición de tratamiento, Y, considerando la misma matriz de varianzas-covarianzas Σ, con un coeficiente de correlación $\rho = 0{,}8$ y los mismos rangos de valores para las componentes de los vectores μ_X, μ_Y y σ^2. Los tres conjuntos de datos se generaron con 50 hipótesis, con un 80 % de hipótesis nulas verdaderas para el primer conjunto, un 50 % para el segundo y un 20 % para el tercero y, para los tres conjuntos de datos, con $n_x = 7$ observaciones para la primera condición de tratamiento X y $n_y = 13$ para la segunda Y.

De acuerdo con Žežula (2009), cuando se trabaja con distribuciones no Elípticas, es preferible no utilizar el coeficiente de correlación de Pearson, como alternativa se puede utilizar el coeficiente de correlación de rangos τ de Kendall o el coeficiente ρ_s de Spearman.

Bajo normalidad, estos coeficientes están relacionados de la forma siguiente:

$$\tau = (2/(\pi))arcsen\rho \Leftrightarrow \rho = sen(\frac{\pi}{2}\tau) \tag{3.27}$$

Para la cópula de Clayton, el valor de τ de Kendall se puede obtener en función del parámetro de la cópula, θ_c, mediante la siguiente expresión:

$$\tau(\theta_c) = \frac{\theta_c}{\theta_c + 2} \tag{3.28}$$

Se eligió para el parámetro de la cópula, θ_c, la distribución uniforme en el intervalo $(1{,}3; 4)$, como distribución a priori y como distribución generadora de candidatos para el algorítmo MCMC, puesto que, teniendo en cuenta las relaciones (3.27) y (3.28), este intervalo se corresponde con $(0{,}58; 0{,}87)$ para ρ. Para la probabilidad inicial de cada hipótesis nula p_i, $i = 1, \ldots, N$, se consideró también como distribución a priori la distribución $Beta(\alpha, \beta)$, como en la subsección 3.1.4.

Para realizar un análisis de sensibilidad, se consideraron igualmente los siguientes valores para los parámetros α y β de la distribución a priori de p_i, $(0{,}5; 1)$, $(1; 1)$, $(1; 0{,}5)$, y $(2; 0{,}5)$.

Posteriormente, se ejecutó el algoritmo con 15000 iteraciones, descartando las primeras 7500 de la salida del MCMC. Finalmente se obtuvo la estimación de la probabilidad a posteriori de cada hipótesis nula, rechazando aquellas con esta probabilidad menor o igual que 0,5.

En las tablas (3.4), (3.5) y (3.6) se presentan los resultados obtenidos de los tres conjunto de datos simulados con un 80 %, un 50 % y un 20 % de hipótesis nulas verdaderas, respectivamente.

Tabla 3.4: Resultados para el modelo con cópulas de Clayton y para el conjunto de datos correspondientes al 80 % de hipótesis nulas verdaderas, para diferentes valores de los parámetros (α, β) de la distribución a priori de p_i.

$(\alpha; \beta)$	**(0,5; 1)**		**(1; 1)**		**(1; 0,5)**		**(2; 0,5)**		
$\widehat{\theta_c}$	1,94		1,99		1,94		1,99		
$\widehat{FDR}$	0.28		0,385		0,258		0,27		
	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Total
Verdaderas	0	39	10	29	39	0	39	0	39
Falsas	0	11	0	11	1	10	6	5	11
Total	0	50	10	40	40	10	45	5	50

Como puede verse en las tres tablas, en cuanto a la estimación de θ_c, el parámetro de la cópula de Clayton, el procedimiento resulta ser robusto con respecto a las distribuciones a priori de p_i, como sucedió en el modelo con cópulas Gaussianas, puesto que la estimación de θ_c es similar para los diferentes valores elegidos de

los parámetros, con una estimación de θ_c en torno a 2 que, teniendo en cuenta las relaciones (3.27) y (3.28), correspondería a una estimación de ρ en torno a 0,7, próximo también al valor con el que se generaron los datos. Sin embargo, en cuanto

Tabla 3.5: Resultados para el modelo con cópulas de Clayton y para el conjunto de datos correspondientes al 50 % de hipótesis nulas verdaderas, para diferentes valores de los parámetros (α, β) de la distribución a priori de p_i.

$(\alpha;\beta)$	(0,5; 1)		(1; 1)		(1; 0,5)		(2; 0,5)		
$\widehat{\theta_c}$	1,89		1,93		2,01		1,97		
$\widehat{FDR}$	0,226		0,29		0,166		0,177		
	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Total
Verdaderas	0	25	6	19	25	0	25	0	25
Falsas	0	25	1	24	8	17	11	14	25
Total	0	50	7	43	33	17	36	14	50

al número de hipótesis nulas rechazadas, el procedimiento no es robusto respecto a la elección de los parámetros de la distribución a priori p_i, obteniendo mejores resultados con las distribuciones a priori sesgadas a la derecha, incluso en el caso de los datos con un bajo porcentaje de hipótesis nulas verdaderas, como ocurría en el modelo con cópulas Gaussianas.

Tabla 3.6: Resultados para el modelo con cópulas de Clayton y para el conjunto de datos correspondientes al 20 % de hipótesis nulas verdaderas, para diferentes valores de los parámetros (α, β) de la distribución a priori de p_i.

$(\alpha;\beta)$	(0,5; 1)		(1; 1)		(1; 0,5)		(2; 0,5)		
$\widehat{\theta_c}$	2,59		2,19		2,29		2,27		
$\widehat{FDR}$	0,203		0,26		0,22		0,194		
	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Aceptadas	Rechazadas	Total
Verdaderas	0	12	4	8	12	0	12	0	12
Falsas	0	38	1	37	15	23	24	14	38
Total	0	50	5	45	27	23	36	14	50

Por otra parte, en el modelo con cópulas de Clayton y distribuciones a priori sesgadas a la derecha, a diferencia de lo que ocurría en el modelo con cópulas Gaussianas, se obtiene un número de hipótesis nulas rechazadas inferior al número de

hipótesis nulas falsas. Además, los valores estimados del FDR son significativamente más altos a los obtenidos con el modelo con cópulas Gaussianas. Por lo que el procedimiento con el modelo con cópulas de Clayton, para los ejemplos simulados, presenta una potencia menor que con el modelo con cópulas Gaussianas, como era de esperar de acuerdo a como se habían generado los datos.

3.2.2. Aplicación a Datos Reales

En esta subsección se aplica el procedimiento, con el modelo con cópulas de Clayton, a los mismos datos reales utilizados en las subsección (3.1.5). El objetivo es identificar los genes con expresión diferencial, para lo que se contrasta si existen diferencias significativas, para cada gen, en el nivel medio de expresión entre las muestras de tejidos procedentes de pacientes celíacos y las muestras procedentes de los controles. Los datos se modelaron mediante distribuciones marginales normales y, para representar la dependencia, mediante cópulas de Clayton.

En cuanto a las distribuciones a priori, se consideraron para μ_X, μ_Y y σ_i^2 las mismas que en la subsección 3.1.5, es decir, distribuciones uniformes para la media del nivel de expresión de cada gen, tanto en pacientes con enfermedad celíaca como en controles y la densidad de Jeffreys para la varianza de cada gen. Para el parámetro de la cópula, θ_c, se consideró una distribución uniforme en el intervalo $(0{,}01; 5)$, debido a que θ_c debe ser mayor que cero y a que este intervalo se corresponde, teniendo en cuenta las relaciones (3.27) y (3.28), con el intervalo $(0{,}008; 0{,}9)$ para ρ. Esta distribución también se utilizó en el algoritmo como distribución generadora de candidatos para θ_c.

Debido a que el modelo mediante el cual se identificaban mejor las hipótesis nulas ciertas y falsas en el análisis con datos simulados, correspondía al de las distribuciones a priori de p_i sesgadas a la derecha y principalmente, para el caso con cópulas de Clayton, al modelo con la distribución a priori $Beta(1; 0{,}5)$, para el análisis de datos reales se asume esta distribución para p_i, $i = 1, \ldots, 38$.

Se ejecutó el algoritmo mediante 40000 iteraciones descartando las 20000 itera-

ciones iniciales.

Se obtuvo $\hat{\theta}_c = 0{,}1515$ y se identificaron 22 genes con expresión diferencial, entre ellos, todos los genes identificados en Pascual et al. (2016): $C1orf106$, $C2orf74$, $CCR4$, $CCR6$, $FASLG$, $ICOSLG$, $IL18R1$, $IL18RAP$, $IL23A$, $IL6$, $JAK2$, $PLEK$, $TAGAP$, $TNFSF18$ y $UBE2L3$. Además identifica los siguientes genes no identificados por dichos autores: $FBX048$, $HCFC1$, $IL2$, $PTPN2$, $STAT3$, $TNFAIP3$ y $TYK2$.

3.2.3. Selección de Modelos

En esta subsección se compara el modelo utilizando cópulas Gaussianas y distribuciones marginales normales y el modelo con cópulas de Clayton y distribuciones marginales normales, aplicados ambos modelos a los datos simulados y a los datos sobre celiaquía de Pascual et al. (2016).

Existen distintos criterios para la selección de modelos, algunos de los más utilizados son el Akaike Information Criterium (AIC), el Bayesian Criterium (BIC) y el Deviance Information Criterion (DIC). El DIC se puede considerar como una generalización del AIC y del BIC, siendo especialmente útil en la selección de modelos bayesianos en los que la distribución a posteriori de los parámetros del modelo se obtiene mediante los métodos MCMC, puesto que este criterio puede ser estimado de manera sencilla mediante una muestra MCMC, como puede verse en Silva and Lopes (2008) y en Gómez et al. (2017), de modo que en este trabajo se utiliza el DIC para comparar los modelos utilizados para los datos simulados y para los datos sobre celiaquía de Pascual et al. (2016). Cuando se comparan dos o más modelos, el modelo con menor valor del DIC se considera el de mejor ajuste (Spiegelhalter et al., 2002).

El valor DIC puede ser escrito como una función del logaritmo de la verosimilitud de la forma siguiente:

$$DIC = -4E_{\Theta,\tau}\left[logL\left(\Theta,\tau|X,Y\right)|X,Y\right] + 2logL\left(E_{\Theta,\tau}\left[\Theta,\tau|X,Y\right]|X,Y\right) \quad (3.29)$$

donde la función de verosimilitud es la definida en (3.2), con las cópulas (3.4)

y (3.3) para el modelo con cópulas Gaussianas, siendo el conjunto de parámeros $\Theta = (\mu_X, \mu_Y, \sigma^2, p, \rho)$, y con las cópulas (3.26) para el modelo con cópulas de Clayton, siendo el conjunto de parámetros $\Theta = (\mu_X, \mu_Y, \sigma^2, p, \theta_c)$.

Como se ha comentado anteriormente, el valor del DIC se puede estimar a partir de una muestra MCMC. Así, dada una muestra de la cadena de Markov Monte Carlo (MCMC) de tamaño M de los parámetros de la distribución a posteriori para el modelo con cópulas Gaussianas,

$$\left\{\mu_X^{(l)}, \mu_Y^{(l)}, \sigma^{2(l)}, p^{(l)}, \rho^{(l)}, \tau^{(l)}\right\}_{l=1}^{M}$$

y una muestra de tamaño M de los parámetros de la distribución a posteriori para el modelo con cópulas de Clayton,

$$\left\{\mu_X^{(l)}, \mu_Y^{(l)}, \sigma^{2(l)}, p^{(l)}, \theta_c^{(l)}, \tau^{(l)}\right\}_{l=1}^{M}$$

el valor del DIC se puede estimar, para ambos modelos, como sigue:

$$DIC = -\frac{4}{M}\sum_{l=1}^{M} logL\left(\Theta^{(l)}, \tau^{(l)}|X, Y\right) + 2logL\left(\frac{1}{M}\sum_{l=1}^{M}\Theta^{(l)}, \frac{1}{M}\sum_{l=1}^{M}\tau^{(l)}|X, Y\right) \quad (3.30)$$

como puede verse en Gómez et al. (2017) y Nguyen et al. (2019), que también utilizan el DIC para comparar modelos con cópulas.

De esta manera, se obtuvo el valor del DIC, con los datos simulados, para el modelo con cópulas Gaussianas y para el modelo con cópulas de Clayton utilizando, en ambos modelos, las distribuciones a priori para p_i sesgadas a la derecha $Beta(1; 0{,}5)$ y $Beta(2; 0{,}5)$ y para los tres conjuntos de datos simulados con un 80 %, un 50 % y un 20 % de hipótesis nulas verdaderas, respectivamente. En la tabla (3.7) se muestran estos resultados.

Tabla 3.7: Valores del DIC para los modelos con cópulas Gaussianas y cópula de Clayton, con los datos correspondientes a los diferentes porcentajes de hipótesis nulas verdaderas y para las distribuciones a priori de p_i segadas a la derecha.

	Modelo	**Cópula Gaussiana**		**Cópula de Clayton**	
	$(\alpha;\beta)$	**(1; 0,5)**	**(2; 0,5)**	**(1; 0,5)**	**(2; 0,5)**
% **de hipótesis nulas verdaderas**	80 %	8548.326	8540.206	8961.517	8973.411
	50 %	8682.981	8696.98	9033.87	9034.77
	20 %	8565,80	8543,045	9114,768	9130,447

Como puede verse en la tabla (3.7), las estimaciones más bajas del DIC corresponden al modelo con cópulas Gaussianas, no habiendo diferencias importantes, para cada modelo, entre el valor del DIC correspondiente a los parámetros $(1; 0{,}5)$ y $(2; 0{,}5)$ de la distribución a priori de p_i. Por lo que el modelo con cópulas Gaussianas resulta ser el más adecuado para los datos simulados, como se esperaba según se habían generado los datos, siendo también con este modelo con el que se obtenían, mediante el procedimiento que se propone de contraste de hipótesis múltiples, un número de hipótesis nulas rechazadas muy próximo al número de hipótesis nulas falsas, con valores de FDR en niveles aceptables y más bajos que los obtenidos con el modelo con cópulas de Clayton, por lo que nuestro procedimiento funciona bien cuando la dependencia se modela mediante la función cópula más adecuada.

Igualmente, se estimó el valor del DIC para los datos sobre celiaquía de Pascual et al. (2016), obteniendo, para el modelo con cópulas Gaussianas $DIC = 1487{,}213$ y para el modelo con cópulas de Clayton $DIC = 1504{,}445$. Puesto que el menor valor del DIC corresponde al modelo con cópulas Gaussianas y distribuciones marginales normales, este modelo resulta ser el más apropiado para los datos asociados a la enfermedad celíaca de Pascual et al. (2016) y, por tanto, los resultados obtenidos mediante el procedimiento de contrastes de hipótesis múltiples de la subsección 3.1.5 serían los más adecuados para estos datos.

Capítulo 4

Conclusiones y Extensiones

4.1. Conclusiones

En esta tesis se propone un procedimiento bayesiano para contrastar simultáneamente un elevado número de hipótesis, cada una de ellas referente a una variable medida en dos situaciones distintas de tratamiento independientes y bajo el supuesto de dependencia entre las variables. El procedimiento bayesiano tiene la ventaja, frente al enfoque frecuentista, de utilizar toda la información disponible, tanto objetiva como subjetiva. La dependencia se modela mediante funciones cópulas, utilizando los datos completos en lugar de los estadísticos, por lo que el proceso de modelado resulta ser más complejo y puede presentar problemas computacionales.

El procedimiento se aplica cuando se realiza, para cada variable, un contraste de igualdad de medias. Para modelar los datos se consideran distribuciones marginales normales, debido a que es habitual utilizar la distribución normal para modelar datos procedentes de experimentos con microarrays de ADN en el campo de la genómica, considerando, para cada variable, varianzas iguales en las dos condiciones de tratamiento y distintas para todas las variables. Asimismo, se considera la misma estructura de dependencia para las dos condiciones de tratamientos y se modela, en primer lugar, mediante la cópula Gaussiana de la familia Elíptica, por utilizar la correlación entre pares de variables para modelar la dependencia de la mismas forma que la distribución normal multivariante, considerando la matriz de correlación

uniforme con la finalidad de reducir la alta dimensión del espacio paramétrico. En segundo lugar se consideró, para modelar la dependencia, la cópula de Clayton de la familia Arquimediana, por permitir modelar distribuciones multivariantes con dependencia asimétrica, mediante una única función univariante, reduciendo también de esta manera el elevado número de parámetros.

En este contexto y a partir de los resultados obtenidos en todos los ejemplos simulados, podemos concluir que, en cuanto a la estimación del parámetro de dependencia, el procedimiento es robusto frente a la elección de los parámetros de la distribución a priori de la probabilidad inicial de cada hipótesis nula, puesto que, en todos los ejemplos y con ambas cópulas, se obtienen valores muy similares y próximos al valor del parámetro con el que se generaron los datos.

Sin embargo, en cuanto al número de hipótesis nulas rechazadas y aceptadas, el procedimiento no es robusto respecto a la elección de dichos parámetros, siendo este número muy distinto según los valores utilizados para los mismos, obteniendo un número de hipótesis nulas rechazadas y aceptadas más próximo al número de hipótesis nulas falsas y verdaderas, cuando se utilizan valores para los parámetros que dan lugar a distribuciones a priori beta sesgadas a la derecha.

Por otro lado, mediante el criterio de selección de modelos DIC y con las distribuciones a priori sesgadas a la derecha, se obtuvo que el modelo con cópulas Gaussianas resulta ser el más adecuado para los datos simulados, como se esperaba según se habían generado los datos, siendo también con este modelo con el que se obtienen, mediante el procedimiento que se propone de contraste de hipótesis múltiples, un número de hipótesis nulas rechazadas más próximo al número de hipótesis nulas falsas que el que se obtiene con el modelo construido utilizando cópulas de Clayton. Además, con el modelo con cópulas Gaussianas se rechaza un número de hipótesis nulas ligeramente superior al número de hipótesis nulas falsas, a la vez que el FDR estimado se mantiene en niveles aceptables y significativamente más bajos que con el modelo con cópulas de Clayton, por lo que el procedimiento propuesto funciona bien, con una elevada potencia, puesto que rechaza un alto número de hipótesis nulas falsas, cuando la dependencia se modela mediante la

función cópula más adecuada.

Cabe destacar también que utilizando el modelo con cópulas Gaussianas, en los ejemplos analizados y cuando se utilizan distribuciones a priori sesgadas a la derecha, el valor de FDR más alto fue de 0,079, superior al 0,05 habitual. No obstante, en el contexto de los experimentos con microarrays de ADN, se puede estar dispuesto a tolerar un mayor número de falsos positivos con el objetivo de obtener el mayor número posible de genes que potencialmente se expresan de manera diferencial, para realizar con ellos posteriormente estudios más de detallados.

Por otra parte, el procedimiento que se propone es flexible en la medida en que puede utilizarse con otras matrices de correlación, o con otras funciones cópulas para modelar la dependencia, así como con otras funciones de distribuciones marginales.

Finalmente, la literatura en la que se utilizan las funciones cópulas, en el contexto de los contrastes de hipótesis múltiples, para modelar la dependencia entre un elevado número de variables es escasa, en la que solo hemos encontrado trabajos que utilizan estas funciones para modelar la dependencia entre un bajo número de variables y con un enfoque frecuentista, por lo que el procedimiento que se propone en esta tesis resulta fundamental, especialmente en el campo de la genómica.

Parte del contenido de esta tesis ha sido publicada en Maria et al. (2020) y presentada en:

- Conferencia: FuzzMAD 2019. 13/12/2019. Madrid, España.
- Seminario: Ciência em Foco. 15-17/06/2020. Nampula, Mozambique.

4.2. Futuras Investigaciones

A partir de la línea de investigación desarrollada en esta tesis, surgen algunas posibilidades de investigaciones futuras. A continuación se describen las de mayor interés.

En esta memoria se ha asumido, por simplicidad, varianzas iguales para cada variable en las dos condiciones de tratamiento, por lo que una extensión natural sería

considerar distintas varianzas en las dos condiciones de tratamiento. Además, en el contexto de los experimentos con microarrays de ADN, se puede esperar que exista dependencia entre la media y la varianza (Baldi and Long, 2001), por tanto, si para cada variable, bajo la hipótesis alternativa, se suponen diferencias en las medias de las dos condiciones de tratamiento, se puede suponer también que existan diferencias entre las correspondientes varianzas.

Con el objetivo de simplificar el modelo, se han considerado distribuciones a priori independientes para todos los parámetros del modelo, por lo que otra posible extensión, teniendo en cuenta lo comentado en el párrafo anterior, podría ser considerar, para cada variable, una distribución a priori para la media dependiente de la varianza (Baldi and Long, 2001; Ausín et al., 2011).

Para el modelo con cópulas Gaussianas, se ha considerado la matriz de correlación uniforme con el objetivo de reducir la dimensión de los parámetros, donde $\rho \in \left[\frac{-1}{N-1}, 1\right]$, excluyendo las correlaciones altas negativas cuando N, el número de hipótesis nulas, es elevado, por lo que para una futura investigación y manteniendo el objetivo de reducir la dimensión de los parámetros, se podrían considerar otras matrices de correlación más realistas como, por ejemplo, la *Serial correlaction structure* propuesta por Žežula (2009), donde $\rho \in [-1, 1]$, incluyendo así las correlaciones negativas.

Otra posible extensión de la tesis es considerar una mixtura de dos o más cópulas, permitiendo capturar de este modo las variadas estructuras de dependencia en un mismo conjunto de datos, pues en los datos procedentes de experimentos con microarrays de ADN, se espera que haya grupos de genes con diferentes estructuras de dependencia. En este sentido, también podría utilizarse previamente el análisis cluster para agrupar los genes que presentan correlaciones similares y, posteriormente, modelar en cada grupo la correspondiente estructura de dependencia mediante la función cópula más adecuada.

Apéndice A

Algoritmo MCMC: Metropolis-Hastings-within-Gibbs

En este apéndice, se muestran los detalles del algoritmo MCMC propuesto. Como ya se ha mencionado a lo largo de la tesis, se utiliza el algoritmo Metropolis-Hastings-within-Gibbs, que combina las estrategias del muestreo de Gibbs, para distribuciones condicionadas a posteriori conocidas, y del Metropolis-Hastings, para distribuciones condicionadas a posteriori desconocidas, con el fin de obtener una muestra de la distribución a posteriori conjunta.

Algorithm MCMC

Require: initial values $(\Theta^{(0)}, \tau^{(0)}) = (\mu_X^{(0)}, \mu_Y^{(0)}, \sigma^{2(0)}, p^{(0)}, \rho^{(0)}, \tau^{(0)})$. Where $\tau^{(0)} = (\tau_1^{(0)}, ..., \tau_N^{(0)})$, $p^{(0)} = (p_1^{(0)}, .., p_N^{(0)})$, $\mu_X^{(0)} = (\mu_{X_1}^{(0)}, ..., \mu_{X_N}^{(0)})$, $\mu_Y^{(0)} = (\mu_{Y_1}^{(0)}, ..., \mu_{Y_N}^{(0)})$, $\sigma^{2(0)} = (\sigma_1^{2(0)}, .., \sigma_N^{2(0)})$

Procedure

1: Let the current state of the Markov chain be $(\Theta^{(l)}, \tau^{(l)}) = (\mu_X^{(l)}, \mu_Y^{(l)}, \sigma^{(l)}, p^{(l)}, \rho^{(l)}, \tau^{(l)})$
2: **for** $l \in 1 : M$ **do**
3: Update τ_i by sampling from $\tau_i^{(l+1)}$, for $i = 1, \ldots, N$ ▷ algorithm 1
4: Update p_i by sampling from $p_i^{(l+1)}$, for $i = 1, \ldots, N$ ▷ algorithm 2
5: Update μ_{X_i} by sampling from $\mu_{X_i}^{(l+1)}$, for $i = 1, \ldots, N$ ▷ algorithm 3
6: Update μ_{Y_i} by sampling from $\mu_{Y_i}^{(l+1)}$, for $i = 1, \ldots, N$ ▷ algorithm 4
7: Update σ_i^2 by sampling from $\sigma_i^{2(l+1)}$, for $i = 1, \ldots, N$ ▷ algorithm 5
8: Update ρ by sampling from $\rho^{(l+1)}$ ▷ algorithm 6
9: **end for**

End Procedure : Return $\left\{\left(\Theta^{(l)}, \tau^{(l)}\right) : l = 1, .., M\right\}$

Algorithm 1 MCMC for τ_i, $i = 1, \ldots, N$

Require: initial values $(\Theta^{(0)}, \tau^{(0)}) = (\mu_X^{(0)}, \mu_Y^{(0)}, \sigma^{2(0)}, p^{(0)}, \rho^{(0)}, \tau^{(0)})$. Where $\tau^{(0)} = (\tau_1^{(0)}, ..., \tau_N^{(0)})$, $p^{(0)} = (p_1^{(0)}, .., p_N^{(0)})$, $\mu_X^{(0)} = (\mu_{X_1}^{(0)}, ..., \mu_{X_N}^{(0)})$, $\mu_Y^{(0)} = (\mu_{Y_1}^{(0)}, ..., \mu_{Y_N}^{(0)})$, $\sigma^{2(0)} = (\sigma_1^{2(0)}, .., \sigma_N^{2(0)})$

Procedure

Update $\mu_Y^{(0)} = (\mu_{Y_1}^{(0)}, ..., \mu_{Y_N}^{(0)})$ ▷ **Ifelse** $\tau_i^{(0)} = 0$, $\mu_{Y_i}^{(0)} = \mu_{X_i}^{(0)}$, $\mu_{Y_i}^{(0)}$

Calculate copula $c_Y(u_Y; \rho^{(0)})$

Let the current state of the Markov chain be $(\Theta^{(l)}, \tau^{(l)}) = (\mu_X^{(l)}, \mu_Y^{(l)}, \sigma^{(l)}, p^{(l)}, \rho^{(l)}, \tau^{(l)})$

for $i \in 1 : N$ **do**

    Calculate $K_i = P_r(\tau_i^{l+1} = 0 | \mu_X^{(l)}, \mu_Y^{(l)}, \sigma^{2(l)}, p_i^{(l)}, \rho^{(l)}, \tau_{j<i}^{(l+1)}, \tau_{j>i}^{(l)})$ ▷ Equation (3.8)

    Generate a random uniform number $\mathcal{U}_i \in (0, 1)$

    if $\mathcal{U}_i \leq K_i$ **then**

        $\tau_i^{l+1} = 0$

        Update $\mu_{Y_i}^{(l)} = \mu_{X_i}^{(l)}$

    else

        $\tau_i^{l+1} = 1$

    end if

    Update $\mu_Y^{(l)} = (\mu_{Y_1}^{(l)}, ..., \mu_{Y_N}^{(l)})$ ▷ **Ifelse** $\tau_i^{(l+1)} = 0$, $\mu_{Y_i}^{(l)} = \mu_{X_i}^{(l)}$, $\mu_{Y_i}^{(l)}$

    Calculate copula $c_Y(u_Y; \rho^{(l)})$

end for

end procedure Return (τ_i^{l+1}), $i = 1, \ldots, N$

Algorithm 2 MCMC: GIBBS for p_i, $i = 1, \ldots, N$

Require: current values $(\Theta^{(0)}, \tau^{(l+1)}) = (\mu_X^{(0)}, \mu_Y^{(0)}, \sigma^{2(0)}, p^{(0)}, \rho^{(0)}, \tau^{(0)})$. Where $\tau^{(l+1)} = (\tau_1^{(1+1)}, ..., \tau_N^{(l+1)})$, $p^{(0)} = (p_1^{(0)}, .., p_N^{(0)})$, $\mu_X^{(0)} = (\mu_{X_1}^{(0)}, ..., \mu_{X_N}^{(0)})$, $\mu_Y^{(0)} = (\mu_{Y_1}^{(0)}, ..., \mu_{Y_N}^{(0)})$, $\sigma^{2(0)} = (\sigma_1^{2(0)}, .., \sigma_N^{2(0)})$

procedure

1: Let the current state of the Markov chain be $(\Theta^{(l)}, \tau^{(l+1)}) = (\mu_X^{(l)}, \mu_Y^{(l)}, \sigma^{(l)}, p^{(l)}, \rho^{(l)}, \tau^{(l+1)})$
2: **for** $i \in 1 : N$ **do**
3: Update p_i by sampling from $p_i^{l+1} \sim p(p_i \mid \tau_i^{(l+1)})$ ▷ Equation (3.9)
4: **end for**
5: **end procedure** Return (p_i^{l+1}), $i = 1, \ldots, N$

Algorithm 3 Single iteration of the Metropolis-Hasting for μ_{X_i}, $i = 1, \ldots, N$

Require: current values $\tau^{(l+1)} = (\tau_1^{(1+1)}, ..., \tau_N^{(l+1)})$, $\mu_X^{(l)} = (\mu_{X_1}^{(l)}, ..., \mu_{X_N}^{(l)})$, $\mu_Y^{(l)} = (\mu_{Y_1}^{(l)}, ..., \mu_{Y_N}^{(l)})$, $\sigma^{2(l)} = (\sigma_1^{2(l)}, .., \sigma_N^{2(l)})$, $\rho^{(l)}$

procedure

1: **for** $i \in 1 : N$ **do**
2: Sample candidate $\mu_{X_i}^{(c)}$ ▷ **Ifelse** $\tau_i^{(l+1)} = 0$, (3.11), from (3.13)
3: Calculate copula $c_X(u_X; \rho^{(l)})$ ▷ With $\mu_{X_i}^{(c)}$
4: Calculate copula $c_Y(u_Y; \rho^{(l)})$ ▷ **Ifelse** $\tau_i^{(l+1)} = 0$, $\mu_{Y_i}^{(l)} = \mu_{X_i}^{(c)}$, $\mu_{Y_i}^{(l)}$
5: Sample random uniform number $\mathcal{U}_i \in (0, 1)$
6: **if** $\mathcal{U}_i \leq \alpha(\mu_{X_i}^{(l)}, \mu_{X_i}^{(c)})$ **then** ▷ $\alpha(\mu_{X_i}^{(l)}, \mu_{X_i}^{(c)}) = min\{1, A_1\}$
7: $\mu_{X_i}^{(l+1)} = \mu_{X_i}^{(c)}$
8: **else**
9: $\mu_{X_i}^{(l+1)} = \mu_{X_i}^{(l)}$
10: **end if**
11: Update $\mu_Y^{(l)} = (\mu_{Y_1}^{(l)}, ..., \mu_{Y_N}^{(l)})$ ▷ **Ifelse** $\tau_i^{(l+1)} = 0$, $\mu_{Y_i}^{(l)} = \mu_{X_i}^{(l+1)}$, $\mu_{Y_i}^{(l)}$
12: Calculate copula $c_X(u_X; \rho^{(l)})$
13: Calculate copula $c_Y(u_Y; \rho^{(l)})$
14: **end for**
15: **end procedure** Return $\mu_{X_i}^{(l+1)}$, $i = (1, \ldots, N)$

$$A_1 = \frac{\prod_{j=1}^{n_x} c_X^c(u_{X.j};\rho) \prod_{k=1}^{n_y} c_Y^c(u_{Y.k};\rho) f_i(\mu_{X_i}^{(c)})}{\prod_{j=1}^{n_x} c_X(u_{X.j};\rho) \prod_{k=1}^{n_y} c_Y(u_{Y.k};\rho) f_i(\mu_{X_i}^{l})},$$

where $c_X^c(u_{X.j};\rho)$ and $c_Y^c(u_{Y.k};\rho)$ computed when $\mu_{X_i}^l = \mu_{X_i}^c$

Algorithm 4 Single iteration of the Metropolis-Hasting for μ_{Y_i}, $i = 1, \ldots, N$

Require: current values $\tau^{(l+1)} = (\tau_1^{(1+1)}, ..., \tau_N^{(l+1)})$, $\mu_X^{(l+1)} = (\mu_{X_1}^{(l+1)}, ..., \mu_{X_N}^{(l+1)})$, $\mu_Y^{(l)} = (\mu_{Y_1}^{(l)}, ..., \mu_{Y_N}^{(l)})$, $\sigma^{2(l)} = (\sigma_1^{2(l)}, .., \sigma_N^{2(l)})$, $\rho^{(l)}$

procedure

for $i \in 1 : N$ **do**
  if $\tau_i^{(l+1)} = 0$ **then**
    $\mu_{Y_i}^{(l+1)} = \mu_{X_i}^{(l+1)}$
  else
    Sample candidate $\mu_{Y_i}^{(c)}$ ▷ candidate-generating density (3.15)
    Calculate copula $c_Y(u_Y;\rho^{(l)})$ ▷ With $\mu_Y^{(c)}$
    Sample random uniform number $\mathcal{U}_i \in (0,1)$
    if $\mathcal{U}_i \leq \alpha(\mu_{Y_i}^{(l)}, \mu_{Y_i}^{(c)})$ **then** ▷ $\alpha(\mu_{Y_i}^{(l)}, \mu_{Y_i}^{(c)}) = min\{1, A_2\}$
      $\mu_{Y_i}^{(l+1)} = \mu_{Y_i}^{(c)}$
    else
      $\mu_{Y_i}^{(l+1)} = \mu_{Y_i}^{(l)}$
    end if
    Calculate copula $c_Y(u_Y;\rho^{(l)})$
  end if
end for
end procedure Return $\mu_{Y_i}^{(l+1)}$, $i = (1, \ldots, N)$

$$A_2 = \frac{\prod_{k=1}^{n_y} c_Y^c(u_{Y.k};\rho) f_i(\mu_{X_i}^{(c)})}{\prod_{k=1}^{n_y} c_Y(u_{Y.k};\rho) f_i(\mu_{X_i}^{l})},$$ where $c_Y^c(u_{Y.k};\rho)$ computed when $\mu_{Y_i}^l = \mu_{Y_i}^c$

Algorithm 5 Single iteration of the Metropolis-Hasting for σ_i^2, $i = 1, \ldots, N$

Require: current values $\tau^{(l+1)} = (\tau_1^{(1+1)}, ..., \tau_N^{(l+1)})$, $\mu_X^{(l+1)} = (\mu_{X_1}^{(l+1)}, ..., \mu_{X_N}^{(l+1)})$, $\mu_Y^{(l+1)} = (\mu_{Y_1}^{(l+1)}, ..., \mu_{Y_N}^{(l+1)})$, $\sigma^{2(l)} = (\sigma_1^{2(l)}, .., \sigma_N^{2(l)})$, $\rho^{(l)}$

procedure

1: **for** $i \in 1 : N$ **do**
2: Sample candidate $\sigma_i^{2(c)}$ ▷ **Ifelse** $\tau_i^{(l+1)} = 0$,from (3.17), from (3.18)
3: Calculate copula $c_X(u_X; \rho^{(l)})$, $c_Y(u_Y; \rho^{(l)})$ ▷ With $\sigma_i^{2(c)}$
4: Sample random uniform number $\mathcal{U}_i \in (0, 1)$
5: **if** $\mathcal{U}_i \leq \alpha(\sigma_i^{2(l)}, \sigma_i^{2(c)})$ **then** ▷ $\alpha(\sigma_i^{2(l)}, \sigma_i^{2(c)}) = min\{1, A_3\}$
6: $\sigma_i^{2(l+1)} = \sigma_i^{2(c)}$
7: **else**
8: $\sigma_i^{2(l+1)} = \sigma_i^{2(l)}$
9: **end if**
10: Calculate copula $c_X(u_X; \rho^{(l)})$
11: Calculate copula $c_Y(u_Y; \rho^{(l)})$
12: **end for**
13: **end procedure** Return $\sigma_i^{2(l+1)}$, $i = (1, \ldots, N)$

$$A_3 = \frac{\prod_{j=1}^{n_x} c_X^c(u_{X_{\cdot j}};\rho) \prod_{k=1}^{n_y} c_Y^c(u_{Y_{\cdot k}};\rho) f_i(\sigma_i^{2(c)})}{\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}};\rho) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}};\rho) f_i(\sigma_i^{2(l)})},$$

where $c_X^c(u_{X_{\cdot j}}; \rho)$ and $c_Y^c(u_{Y_{\cdot k}}; \rho)$ computed when $\sigma_i^{2(l)} = \sigma_i^{2(c)}$

Algorithm 6 Single iteration of the Metropolis-Hasting for ρ

Require: current values $\mu_X^{(l+1)} = (\mu_{X_1}^{(l+1)}, ..., \mu_{X_N}^{(l+1)})$, $\mu_Y^{(l+1)} = (\mu_{Y_1}^{(l+1)}, ..., \mu_{Y_N}^{(l+1)})$, $\sigma^{2(l+1)} = (\sigma_1^{2(l+1)}, .., \sigma_N^{2(l+1)})$

procedure

1: Sample candidate $\rho^{(c)}$ ▷ from $\mathcal{U} \in (a, b)$, $a, b \in (0, 1)$

2: Sample random uniform number $\mathcal{U} \in (0, 1)$

3: **if** $\mathcal{U} \leq \alpha(\rho^{(l)}, \rho^{(c)})$ **then** ▷ $\alpha = min\{1, A_4\}$

4: $\rho^{(l+1)} = \rho^{(c)}$

5: **else**

6: $\rho^{(l+1)} = \rho^{(l)}$

7: **end if**

8: **end procedure** Return $\rho^{(l+1)}$

$$A_4 = \frac{\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; \rho^c) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; \rho^c)}{\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; \rho^l) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; \rho^l)}$$

Apéndice B

Distribuciones condicionadas a posteriori

En este apéndice, se describe la obtención de las distribuciones condicionadas a posteriori de forma más detallada, donde $\tau_{-i} = (\tau_1, \ldots, \tau_{i-1}, \tau_{i+1}, \ldots, \tau_N)$ y $\Theta_{-\theta_i}$ es el vector de parámetros $\Theta = (\mu_X, \mu_Y, \sigma^2, p, w)$ sin el parámetro indicado en el subíndice, siendo θ_i cualquier parámetro de ese vector, $\mu_X = (\mu_{X_1}, .., \mu_{X_N}), \mu_Y = (\mu_{Y_1}, ..., \mu_{Y_N}), \sigma^2 = (\sigma_1^2, ..., \sigma_N^2)$, $p = (p_1, .., p_N)$, $\tau = (\tau_1, .., \tau_N)$ y $c_X\left(u_{X_{.j}}; w\right)$ y $c_Y\left(u_{Y_{.k}}; w\right)$ son las cópulas Gaussianas (3.3) y (3.4), respectivamente, con $w = \rho$, o las cópulas de Clayton (3.26), con $w = \theta_c$.

Distribución condicionada a posteriori para $\tau_i = 0$, $i = 1, \ldots, N$, dadas las observaciones y el resto de parámetros.

$$
\begin{aligned}
&Pr\left(\tau_i = 0|\Theta, \tau_{-i}, X, Y\right) \\
&= \frac{\pi\left(\Theta, \tau_i = 0, \tau_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i}|X, Y\right)}{\sum_{v=0}^{1} \pi\left(\Theta, \tau_i = v, \tau_{-i}\right) L\left(\Theta, \tau_i = v, \tau_{-i}, |X, Y\right)} \\
&= \frac{\pi\left(\Theta\right) \pi\left(\tau_i = 0|p_i\right) \pi\left(\tau_{-i}|p_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i}|X, Y\right)}{\sum_{v=0}^{1} \pi\left(\Theta\right) \pi\left(\tau_i = v|p_i\right) \pi\left(\tau_{-i}|p_{-i}\right) L\left(\Theta, \tau_i = v, \tau_{-i}|X, Y\right)} \\
&= \frac{Pr(\tau_i = 0|p_i) Q_0}{Pr(\tau_i = 0|p_i) Q_0 + (1 - \pi(\tau_i = 0|p_i)) Q_1} \\
&= \frac{p_i \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; w\right) f_i\left(y_{ik}\middle|\mu_{X_i}, \sigma_i^2\right)}{p_i \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; w\right) f_i\left(y_{ik}\middle|\mu_{X_i}, \sigma_i^2\right) + (1 - p_i) \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; w\right) f_i\left(y_{ik}\middle|\mu_{Y_i}, \sigma_i^2\right)} \\
&= \frac{p_i \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; w\right) exp\left\{-\frac{1}{2{\sigma_i}^2} S_{\mu_{X_i}}\right\}}{p_i \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; w\right) exp\left\{-\frac{1}{2{\sigma_i}^2} S_{\mu_{X_i}}\right\} + (1 - p_i) \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; w\right) exp\left\{-\frac{1}{2{\sigma_i}^2} S_{\mu_{Y_i}}\right\}}
\end{aligned}
$$

donde $S_{\mu_{X_i}} = \left[n_y\left(\overline{y}_i. - \mu_{X_i}\right)^2\right]$ y $S_{\mu_{Y_i}} = \left[n_y\left(\overline{y}_i. - \mu_{Y_i}\right)^2\right]$,
$Q_0 = \prod_{j=1}^{n_x} c_X\left(u_{X.j}; w\right) f_i\left(x_{ij}|\mu_{X_i}, \sigma_i^2\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; w\right) f_i\left(y_{ik}|\mu_{X_i}, \sigma_i^2\right)$,
$Q_1 = \prod_{j=1}^{n_x} c_X\left(u_{X.j}; w\right) f_i\left(x_{ij}|\mu_{X_i}, \sigma_i^2\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y.k}; w\right) f_i\left(y_{ik}|\mu_{Y_i}, \sigma_i^2\right)$

Distribución condicionada a posteriori para p_i, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros.

$$
\begin{aligned}
\pi(p_i|X, Y, \Theta_{-pi}, \tau) &= \frac{\pi\left(\Theta\right) \pi\left(\tau|p\right) L\left(\Theta, \tau|X, Y\right)}{\int_0^1 \pi\left(\Theta\right) \pi\left(\tau|p\right) L\left(\Theta, \tau|X, Y\right) dp_i} \\
&= \frac{\pi(p_i)\pi(\tau_i|p_i)}{\int_0^1 \pi(p_i)\pi(\tau_i|p_i) dp_i} \\
&\propto p_i^{\tau_i}(1 - p_i)^{1-\tau_i} p_i^{\alpha_i - 1}(1 - p_i)^{\beta_i - 1} \\
&\propto p_i^{\tau_i + \alpha_i - 1}(1 - p_i)^{(1 - \tau_i + \beta_i)}
\end{aligned}
$$

$$\pi(p_i|X, Y, \Theta_{-pi}, \tau) \sim Beta(\alpha_i + 1 - \tau_i, \beta_i + \tau_i)$$

Distribución condicionada a posteriori de μ_{X_i}, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros, cuando $\tau_i = 0$.

$$
\begin{aligned}
&\pi(\mu_{X_i}|X, Y, \Theta_{-\mu_{X_i}}, \tau_i = 0, \tau_{-i}) \\
&= \frac{\pi\left(\Theta, \tau_i = 0, \tau_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i}|X, Y\right)}{\displaystyle\int_{\mu_{X_i}} \pi\left(\Theta, \tau_i = 0, \tau_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i}|X, Y\right) d\mu_{X_i}} \\
&= \frac{\pi\left(\mu_{X_i}\right) \pi\left(\Theta_{-\mu_{X_i}}\right) \pi\left(\tau_i = 0|p_i\right) \pi\left(\tau_{-i}|p_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i}|X, Y\right)}{\displaystyle\int_{\mu_{X_i}} \pi\left(\mu_{X_i}\right) \pi\left(\Theta_{-\mu_{X_i}}\right) \pi\left(\tau_i = 0|p_i\right) \pi\left(\tau_{-i}|p_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i}|X, Y\right) d\mu_{X_i}} \\
&= \frac{L\left(\Theta, \tau_i = 0, \tau_{-i}|X, Y\right)}{\displaystyle\int_{\mu_{X_i}} L\left(\Theta, \tau_i = 0, \tau_{-i}|X, Y\right) d\mu_{X_i}} \\
&= \frac{\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij}\middle|\mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik}\middle|\mu_{X_i}, \sigma_i^2\right)}{\displaystyle\int_{\mu_{X_i}} \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij}\middle|\mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik}\middle|\mu_{X_i}, \sigma_i^2\right) d\mu_{X_i}} \\
&= \frac{\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) exp\left\{-\frac{1}{2{\sigma_i}^2}\left[\sum_j \left(x_{ij} - \mu_{X_i}\right)^2 + \sum_k \left(y_{ik} - \mu_{X_i}\right)^2\right]\right\}}{\displaystyle\int_{\mu_{X_i}} \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) exp\left\{-\frac{1}{2{\sigma_i}^2}\left[\sum_j \left(x_{ij} - \mu_{X_i}\right)^2 + \sum_k \left(y_{ik} - \mu_{X_i}\right)^2\right]\right\} d\mu_{X_i}} \\
&= \frac{\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i(\mu_{X_i})}{E_{f_i(\mu_{X_i})}\left[\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w)\right]}
\end{aligned}
$$

donde

$$f_i(\mu_{X_i}) \sim N\left(\frac{n_x \overline{x}_{i\cdot} + n_y \overline{y}_{i\cdot}}{n_x + n_y}, \frac{\sigma_i}{\sqrt{n_x + n_y}}\right)$$

Distribución condicionada a posteriori de μ_{X_i}, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros, cuando $\tau_i = 1$.

$$
\begin{aligned}
&\pi(\mu_{X_i}|X, Y, \Theta_{-\mu_{X_i}}, \tau_i = 1, \tau_{-i}) \\
&= \frac{\pi\left(\Theta, \tau_i = 1, \tau_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right)}{\displaystyle\int_{\mu_{X_i}} \pi\left(\Theta, \tau_i = 1, \tau_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right) d\mu_{X_i}} \\
&= \frac{\pi\left(\mu_{X_i}\right) \pi\left(\Theta_{-\mu_{X_i}}\right) \pi\left(\tau_i = 1|p_i\right) \pi\left(\tau_{-i}|p_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right)}{\displaystyle\int_{\mu_{X_i}} \pi\left(\mu_{X_i}\right) \pi\left(\Theta_{-\mu_{X_i}}\right) \pi\left(\tau_i = 1|p_i\right) \pi\left(\tau_{-i}|p_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right) d\mu_{X_i}} \\
&= \frac{L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right)}{\displaystyle\int_{\mu_{X_i}} L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right) d\mu_{X_i}} \\
&= \frac{\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij}\middle|\mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik}\middle|\mu_{Y_i}, \sigma_i^2\right)}{\displaystyle\int_{\mu_{X_i}} \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij}\middle|\mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik}\middle|\mu_{Y_i}, \sigma_i^2\right) d\mu_{X_i}} \\
&= \frac{\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) f_i\left(x_{ij}\middle|\mu_{X_i}, \sigma_i^2\right)}{\displaystyle\int_{\mu_{X_i}} \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) f_i\left(x_{ij}\middle|\mu_{X_i}, \sigma_i^2\right) d\mu_{X_i}} \\
&= \frac{\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) exp\left\{-\frac{1}{2{\sigma_i}^2} \sum_j \left(x_{ij} - \mu_{X_i}\right)^2\right\}}{\displaystyle\int_{\mu_{X_i}} \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) exp\left\{-\frac{1}{2{\sigma_i}^2} \sum_j \left(x_{ij} - \mu_{X_i}\right)^2\right\} d\mu_{X_i}} \\
&= \frac{\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) exp\left\{-\frac{n_x(\overline{x}_{i\cdot} - \mu_{X_i})^2}{2{\sigma_i}^2}\right\}}{\displaystyle\int_{\mu_{X_i}} \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) exp\left\{-\frac{n_x(\overline{x}_{i\cdot} - \mu_{X_i})^2}{2{\sigma_i}^2}\right\} d\mu_{X_i}} \\
&= \frac{\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) f_i(\mu_{X_i})}{E_{f_i(\mu_{X_i})}\left[\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w)\right]}
\end{aligned}
$$

donde

$$f_i(\mu_{X_i}) \sim N\left(\overline{x}_{i\cdot}, \frac{\sigma_i}{\sqrt{n_x}}\right)$$

Distribución condicionada a posteriori para μ_{Y_i}, cuando $\tau_i = 1$, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros.

$$
\begin{aligned}
&\pi(\mu_{Y_i}|X, Y, \Theta_{-\mu_{Y_i}}, \tau_i = 1, \tau_{-i}) \\
&= \frac{\pi\left(\Theta, \tau_i = 1, \tau_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right)}{\displaystyle\int_{\mu_{Y_i}} \pi\left(\Theta, \tau_i = 1, \tau_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right) d\mu_{Y_i}} \\
&= \frac{\pi\left(\mu_{Y_i}\right) \pi\left(\Theta_{-\mu_{Y_i}}\right) \pi\left(\tau_i = 1|p_i\right) \pi\left(\tau_{-i}|p_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right)}{\displaystyle\int_{\mu_{Y_i}} \pi\left(\mu_{Y_i}\right) \pi\left(\Theta_{-\mu_{Y_i}}\right) \pi\left(\tau_i = 1|p_i\right) \pi\left(\tau_{-i}|p_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right) d\mu_{Y_i}} \\
&= \frac{L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right)}{\displaystyle\int_{\mu_{Y_i}} L\left(\Theta, \tau_i = 1, \tau_{-i}|X, Y\right) d\mu_{Y_i}} \\
&= \frac{\displaystyle\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij}\middle|\mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik}\middle|\mu_{Y_i}, \sigma_i^2\right)}{\displaystyle\int_{\mu_{Y_i}} \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij}\middle|\mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik}\middle|\mu_{Y_i}, \sigma_i^2\right) d\mu_{Y_i}} \\
&= \frac{\displaystyle\prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(y_{ik}\middle|\mu_{Y_i}, \sigma_i^2\right)}{\displaystyle\int_{\mu_{Y_i}} \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(y_{ik}\middle|\mu_{Y_i}, \sigma_i^2\right) d\mu_{Y_i}} \\
&= \frac{\displaystyle\prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) exp\left\{-\frac{1}{2{\sigma_i}^2} \sum_k \left(y_{ik} - \mu_{Y_i}\right)^2\right\}}{\displaystyle\int_{\mu_{Y_i}} \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) exp\left\{-\frac{1}{2{\sigma_i}^2} \sum_k \left(y_{ik} - \mu_{Y_i}\right)^2\right\} d\mu_{Y_i}} \\
&= \frac{\displaystyle\prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) exp\left\{-\frac{n_y(\overline{y}_{i\cdot} - \mu_{Y_i})^2}{2{\sigma_i}^2}\right\}}{\displaystyle\int_{\mu_{Y_i}} \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) exp\left\{-\frac{n_y(\overline{y}_{i\cdot} - \mu_{Y_i})^2}{2{\sigma_i}^2}\right\} d\mu_{Y_i}} \\
&= \frac{\displaystyle\prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i(\mu_{Y_i})}{E_{f_i(\mu_{Y_i})}\left[\displaystyle\prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w)\right]}
\end{aligned}
$$

donde

$$f_i(\mu_{Y_i}) \sim N\left(\overline{y}_{i\cdot}, \frac{\sigma_i}{\sqrt{n_y}}\right)$$

Distribución condicionada a posteriori para ${\sigma_i}^2$, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros, cuando $\tau_i = 0$

$$
\begin{aligned}
&\pi\left({\sigma_i}^2 \middle| X, Y, \Theta_{-{\sigma_i}^2}, \tau_i = 0, \tau_{-i}\right) \\
&= \frac{\pi\left(\Theta, \tau_i = 0, \tau_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i} | X, Y\right)}{\int_{\sigma_i^2} \pi\left(\Theta, \tau_i = 0, \tau_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i} | X, Y\right) d\sigma_i^2} \\
&= \frac{\pi\left(\sigma_i^2\right) \pi\left(\Theta_{-\sigma_i^2}\right) \pi\left(\tau_i = 0 | p_i\right) \pi\left(\tau_{-i} | p_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i} | X, Y\right)}{\int_{\sigma_i^2} \pi\left(\sigma_i^2\right) \pi\left(\Theta_{-\mu_{Y_i}}\right) \pi\left(\tau_i = 0 | p_i\right) \pi\left(\tau_{-i} | p_{-i}\right) L\left(\Theta, \tau_i = 0, \tau_{-i} | X, Y\right) d\sigma_i^2} \\
&= \frac{\pi\left(\sigma_i^2\right) L\left(\Theta, \tau_i = 0, \tau_{-i} | X, Y\right)}{\int_{\sigma_i^2} \pi\left(\sigma_i^2\right) L\left(\Theta, \tau_i = 0, \tau_{-i} | X, Y\right) d\sigma_i^2} \\
&= \frac{\pi\left(\sigma_i^2\right) \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij} \middle| \mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik} \middle| \mu_{X_i}, \sigma_i^2\right)}{\int_{\sigma_i^2} \pi\left(\sigma_i^2\right) \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij} \middle| \mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik} \middle| \mu_{X_i}, \sigma_i^2\right) d\sigma_i^2} \\
&= \frac{\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) (\frac{1}{\sigma_i^2})^{\frac{n_x+n_y}{2}+1} exp\left\{-\frac{1}{2{\sigma_i}^2} A\right\}}{\int_{\sigma_i^2} \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) (\frac{1}{\sigma_i^2})^{\frac{n_x+n_y}{2}+1} exp\left\{-\frac{1}{2{\sigma_i}^2} A\right\} d\sigma_i^2} \\
&= \frac{\prod_{j=1}^{n_x} c_X\left(u_{X_{\cdot j}}; w\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}}; w\right) f_i\left(\sigma_i^2\right)}{E_{f_i\left(\sigma_i^2\right)}\left[\prod_{j=1}^{n_x} c_X\left(u_{X_{\cdot j}}; w\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}}; w\right)\right]}
\end{aligned}
$$

donde

$$
f_i\left(\sigma_i^2\right) \sim \text{Gamma Inversa}\left(\frac{n_x + n_y}{2}, \frac{A}{2}\right) \text{ con A} = \sum_j \left(x_{ij} - \mu_{X_i}\right)^2 + \sum_k \left(y_{ik} - \mu_{X_i}\right)^2
$$

Distribución condicionada a posteriori para ${\sigma_i}^2$, $i = 1, 2, \ldots, N$, dadas las observaciones y el resto de parámetros, cuando $\tau_i = 1$

$$
\begin{aligned}
&\pi\left({\sigma_i}^2 \middle| X, Y, \Theta_{-{\sigma_i}^2}, \tau_i = 1, \tau_{-i}\right) \\
&= \frac{\pi\left(\Theta, \tau_i = 1, \tau_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i} | X, Y\right)}{\int_{\sigma_i^2} \pi\left(\Theta, \tau_i = 1, \tau_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i} | X, Y\right) d\sigma_i^2} \\
&= \frac{\pi\left(\sigma_i^2\right) \pi\left(\Theta_{-\sigma_i^2}\right) \pi\left(\tau_i = 1 | p_i\right) \pi\left(\tau_{-i} | p_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i} | X, Y\right)}{\int_{\sigma_i^2} \pi\left(\sigma_i^2\right) \pi\left(\Theta_{-\mu_{Y_i}}\right) \pi\left(\tau_i = 1 | p_i\right) \pi\left(\tau_{-i} | p_{-i}\right) L\left(\Theta, \tau_i = 1, \tau_{-i} | X, Y\right) d\sigma_i^2} \\
&= \frac{\pi\left(\sigma_i^2\right) L\left(\Theta, \tau_i = 1, \tau_{-i} | X, Y\right)}{\int_{\sigma_i^2} \pi\left(\sigma_i^2\right) L\left(\Theta, \tau_i = 1, \tau_{-i} | X, Y\right) d\sigma_i^2} \\
&= \frac{\pi\left(\sigma_i^2\right) \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij} \middle| \mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik} \middle| \mu_{Y_i}, \sigma_i^2\right)}{\int_{\sigma_i^2} \pi\left(\sigma_i^2\right) \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) f_i\left(x_{ij} \middle| \mu_{X_i}, \sigma_i^2\right) f_i\left(y_{ik} \middle| \mu_{Y_i}, \sigma_i^2\right) d\sigma_i^2} \\
&= \frac{\prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) (\frac{1}{\sigma_i^2})^{\frac{n_x+n_y}{2}+1} exp\left\{-\frac{1}{2{\sigma_i}^2} B\right\}}{\int_{\sigma_i^2} \prod_{j=1}^{n_x} c_X(u_{X_{\cdot j}}; w) \prod_{k=1}^{n_y} c_Y(u_{Y_{\cdot k}}; w) (\frac{1}{\sigma_i^2})^{\frac{n_x+n_y}{2}+1} exp\left\{-\frac{1}{2{\sigma_i}^2} B\right\} d\sigma_i^2} \\
&= \frac{\prod_{j=1}^{n_x} c_X\left(u_{X_{\cdot j}}; w\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}}; w\right) f_i\left(\sigma_i^2\right)}{E_{f_i(\sigma_i^2)}\left[\prod_{j=1}^{n_x} c_X\left(u_{X_{\cdot j}}; w\right) \prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}}; w\right)\right]}
\end{aligned}
$$

donde

$$
f_i\left(\sigma_i^2\right) \sim \text{Gamma Inversa}\left(\frac{n_x + n_y}{2}, \frac{B}{2}\right) \text{con B} = \sum_k \left(x_{ij} - \mu_{Xi}\right)^2 + \sum_k \left(y_{ik} - \mu_{Y_i}\right)^2
$$

Distribución condicionada a posteriori para w, dadas las observaciones y el resto de parámetros

$$
\begin{aligned}
&\pi\left(w|X,Y,\Theta_{-w},\tau_i=v,\tau_{-i}\right)\\
&=\frac{\pi\left(\Theta,\tau_i=v,\tau_{-i}\right)L\left(\Theta,\tau_i=v,\tau_{-i}|X,Y\right)}{\int_w \pi\left(\Theta,\tau_i=v,\tau_{-i}\right)L\left(\Theta,\tau_i=v,\tau_{-i}|X,Y\right)dw}\\
&=\frac{\pi\left(w\right)\pi\left(\Theta_{-w}\right)\pi\left(\tau_i=v|p_i\right)\pi\left(\tau_{-i}|p_{-i}\right)L\left(\Theta,\tau_i=0,\tau_{-i}|X,Y\right)}{\int_w \pi\left(w\right)\pi\left(\Theta_{-w}\right)\pi\left(\tau_i=v|p_i\right)\pi\left(\tau_{-i}|p_{-i}\right)L\left(\Theta,\tau_i=v,\tau_{-i}|X,Y\right)dw}\\
&=\frac{\pi\left(w\right)L\left(\Theta,\tau_i=v,\tau_{-i}|X,Y\right)}{\int_w \pi\left(w\right)L\left(\Theta,\tau_i=v,\tau_{-i}|X,Y\right)dw}\\
&=\frac{\prod_{j=1}^{n_x} c_X\left(u_{X_{\cdot j}};w\right)\prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}};w\right)}{\int_w \prod_{j=1}^{n_x} c_X\left(u_{X_{\cdot j}};w\right)\prod_{k=1}^{n_y} c_Y\left(u_{Y_{\cdot k}};w\right)dw}, \qquad v\in\{0,1\}
\end{aligned}
$$

Bibliografía

Ausín, M. C., Gómez-Villegas, M. A., González-Pérez, B., Rodríguez-Bernal, M. T., Salazar, I., and Sanz, L. (2011). Bayesian analysis of multiple hypothesis testing with applications to microarray experiments. *Communications in Statistics-Theory and Methods*, 40(13):2276–2291.

Ausín, M. C. and Lopes, H. F. (2010). Time-varying joint distribution through copulas. *Computational Statistics & Data Analysis*, 54(11):2383–2399.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Berry, D. A. and Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82(1):215–227.

Bodnar, T. and Dickhaus, T. (2014). False discovery rate control under Archimedean copula. *Electronic Journal of Statistics*, 8(2):2207–2241.

Box, G. E. and Tiao, G. C. (2011). *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons.

Broët, P., Richardson, S., and Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology*, 9(4):671–683.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174.

Chaudhuri, J. D. (2005). Genes arrayed out for you: the amazing world of microarrays. *Medical Science Monitor*, 11(2):RA52–RA62.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons.

Chi, Z. (2011). Effects of statistical dependence on multiple testing under a hidden Markov model. *The Annals of Statistics*, 39(1):439–473.

Chib, S. and Greenberg, E. (1995). Markov Chain Monte Carlo Simulation Methods in Econometrics.

Clarke, S. and Hall, P. (2009). Robustness of multiple testing procedures against dependence. *The Annals of Statistics*, pages 332–358.

Clayton, D. G. (1978a). 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.

Clayton, D. G. (1978b). 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.

Clemen, R. T. and Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224.

Costa Dias, A. d. (2004). *Copula inference for finance and insurance*. PhD thesis, ETH Zurich.

Dahl, D. B. and Newton, M. A. (2007). Multiple hypothesis testing by clustering treatment effects. *Journal of the American Statistical Association*, 102(478):517–526.

Daniels, M. J. and Pourahmadi, M. (2009). Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis*, 100(10):2352–2363.

Díaz, G. (2014). A note on the multivariate Archimedean dependence structure in small wind generation sites. *Wind Energy*, 17(8):1287–1295.

Dickhaus, T. and Gierl, J. (2012). Simultaneous test procedures in terms of p-value copulae. Technical report, SFB 649 Discussion Paper.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2):363–375.

Do, K.-A., Müller, P., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *J. R. Stat. Soc. Ser. C-Appl. Stat.*, 54(3):627–644.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, pages 71–103.

Duncan, D. B. (1955). Multiple range and multiple F tests. *Biometrics*, 11(1):1–42.

Duncan, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics*, 7(2):171–222.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121.

Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(suppl 1):S105–S110.

Efron, B. (2007). Correlation and Large-Scale Simultaneous Significance Testing. *As of November*, 9.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160.

Embrechts, P., Lindskog, F., and McNeil, A. (2001). Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*, 14.

Embrechts, P., Lindskog, F., McNeil, A., and Rachev, S. (2003). Handbook of heavy tailed distributions in finance. *Modelling Dependence with Copulas and Applications to Risk Management. Handbooks in Finance: Book*, 1:329–385.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1:176–223.

Feller, W. (1966). *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons: New York-London-Sidney.

Fisher, R. A. (1935). The design of experiments, London.

François, O., Ancelet, S., and Guillot, G. (2006). Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, 174(2):805–816.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Gelman, A. and Meng, X.-L. (1991). A note on bivariate distributions that are conditionally normal. *Am. Stat.*, 45(2):125–126.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368.

Genest, C. and MacKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283.

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

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc. B-Stat. Methodol.*, 64(3):499–517.

Genovese, C. and Wasserman, L. (2003). Bayesian and Frequentist Multiple Testing. In Proceedings of the Seventh Valencia International Meeting, June 2-6, 2002, Bayesian Statistics 7, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, pages 1035–1061.

Ghosal, S. and Roy, A. (2012). Predicting false discovery proportion under dependence. *Journal of the American Statistical Association*.

Gómez, M., Ausín, M. C., and Domínguez, M. C. (2017). Seasonal copula models

for the analysis of glacier discharge at King George Island, Antarctica. *Stochastic environmental research and risk assessment*, 31(5):1107–1121.

Gómez-Villegas, M. A., Salazar, I., and Sanz, L. (2014). A Bayesian decision procedure for testing multiple hypotheses in DNA microarray experiments. *Statistical applications in genetics and molecular biology*, 13(1):49–65.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Jeffreys, H. (1967). Theory of Probability (1939). Clarendon.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Kowalski, C. J. (1973). Non-normal bivariate distributions with normal marginals. *Am. Stat.*, 27(3):103–106.

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

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

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événements. *Mém. de math. et phys. présentés à l'Acad. roy. des sci*, 6:621–656.

Lewis, C. and Thayer, D. T. (2004). A loss function related to the FDR for random effects multiple comparisons. *Journal of statistical planning and inference*, 125(1-2):49–58.

Liu, J., Peissig, P., Zhang, C., Burnside, E., McCarty, C., and Page, D. (2012). Graphical-model based multiple testing under dependence, with applications to genome-wide association studies. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2012, page 511. NIH Public Access.

Liu, J., Zhang, C., Burnside, E. S., and Page, D. (2014a). Learning Heterogeneous Hidden Markov Random Fields. In *AISTATS*, pages 576–584.

Liu, J., Zhang, C., Burnside, E. S., and Page, D. (2014b). Multiple Testing under Dependence via Semiparametric Graphical Models. In *ICML*, pages 955–963.

Liu, J., Zhang, C., and Page, D. (2016). Multiple testing under dependence via graphical models. *The Annals of Applied Statistics*, 10(3):1699–1724.

Maria, E. C. J., Salazar, I., Sanz, L., and Gómez-Villegas, M. A. (2020). Using Copula to Model Dependence When Testing Multiple Hypotheses in DNA Microarray Experiments: A Bayesian Approximation. *Mathematics*, 8, 1514.

Marín, J. M. and Rodríguez-Bernal, M. T. (2012). Multiple hypothesis testing and clustering with mixtures of non-central t-distributions applied in microarray data analysis. *Computational Statistics & Data Analysis*, 56(6):1898–1907.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468):990–1001.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Nguyen, H., Ausín, M. C., and Galeano, P. (2019). Parallel Bayesian Inference for

High-Dimensional Dynamic Factor Copulas. *Journal of Financial Econometrics*, 17(1):118–151.

Niemeyer, C. M. and Blohm, D. (1999). DNA microarrays. *Angewandte Chemie International Edition*, 38(19):2865–2869.

Nikoloulopoulos, A. K. and Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Statistics in Medicine*, 27(30):6393–6406.

Pascual, V., Medrano, L., López-Palacios, N., Bodas, A., Dema, B., Fernández-Arquero, M., González-Pérez, B., Salazar, I., and Núñez, C. (2016). Different gene expression signatures in children and adults with celiac disease. *PloS one*, 11(2).

Patz, R. J. and Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of educational and behavioral Statistics*, 24(2):146–178.

Qiu, X., Klebanov, L., and Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Rayaprolu, S. and Chi, Z. (2014). Multiple Testing under Dependence with Approximate Conditional Likelihood. *arXiv preprint arXiv:1412.7778*.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Salazar, I. (2011). *Aproximación bayesiana a los contrastes de hipótesis múltiples con aplicaciones a los microarrays*. Madrid: E-Prints Complutense.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448.

Sarkar, S. K. (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *The Annals of Statistics*, pages 394–415.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Schweizer, B. and Sklar, A. (1983). Probabilistic Metric Spaces, North-Holland Series in Probability and Applied Mathematics.

Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *The annals of statistics*, 9(4):879–885.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.

Silva, R. d. S. and Lopes, H. F. (2008). Copula, marginal distributions and model selection: a Bayesian note. *Statistics and Computing*, 18(3):313–320.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

Smith, M. S. (2011). Bayesian approaches to copula modelling. *arXiv preprint arXiv:1112.4204*.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035.

Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):393–424.

Van de Rijn, M. and Gilks, C. (2004). Applications of microarrays to histopathology. *Histopathology*, 44(2):97–108.

Waller, R. A. and Duncan, D. B. (1969). A Bayes rule for the symmetric multiple comparisons problem. *Journal of the American Statistical Association*, 64(328):1484–1503.

Yuan, M. and Kendziorski, C. (2006). A unified approach for simultaneous gene clustering and differential expression identification. *Biometrics*, 62(4):1089–1098.

Žežula, I. (2009). On multivariate Gaussian copulas. *Journal of Statistical Planning and Inference*, 139(11):3942–3946.

Zhao, H., Chan, K.-L., Cheng, L.-M., and Yan, H. (2008). Multivariate hierarchical Bayesian model for differential gene expression analysis in microarray experiments. *BMC bioinformatics*, 9(Suppl 1):S9.