Facultad de Matemática y Computación
Universidad de La Habana


# Algoritmos con estimación de distribuciones basados en cópulas y vines

Autor: **Yasser González Fernández**

Tutora: **Dra. Marta R. Soto Ortiz**

Trabajo de Diploma presentado en opción al título de
Licenciado en Ciencia de la Computación

Junio de 2011

# Agradecimientos

Quiero agradecer de manera especial a Marta Soto, por guiarme en mis primeros pasos en la investigación científica, compartir sus conocimientos acerca de los algoritmos con estimación de distribuciones y por su extraordinaria dedicación.

Agradezco a Alberto Ochoa y Juan Carlos Jiménez, por brindarme su ayuda cuando fue necesaria. A Omar Ochoa, por garantizar los recursos computacionales para los experimentos. A Yunay Hernández, por la revisión del documento. A Ernesto Moreno, por su ayuda para la obtención de los resultados en el problema de acoplamiento molecular. A Yanely Milanés y Adriel Álvarez, con quienes compartí preocupaciones durante el desarrollo de la tesis.

También quiero agradecer a mis profesores de la facultad, en especial a Yudivián Almeida y Ernesto Rodríguez. A mis amigos de la carrera, en especial a los que han sido más cercanos: Ariel Hernández, Julio Carlos Menéndez, Octavio Cuenca, Andy Venet y Abel Puentes.

Además, agradezco a Susana su apoyo, comprensión y estar junto a mí en estos momentos de mi vida. Finalmente, agradezco a mis padres y a mi abuela que han sido un apoyo fundamental durante todos mis años de estudios. Quisiera que ellos sientan como suyos estos resultados.

Yasser González Fernández
La Habana, junio de 2011

# Resumen

En este trabajo se introducen los modelos gráficos probabilísticos C-vines y D-vines en optimización evolutiva para modelar las distribuciones de búsqueda de los Algoritmos con Estimación de Distribuciones (EDA), dando lugar a los algoritmos CVEDA y DVEDA. El uso de estos modelos facilita el tratamiento de problemas con diferentes y complejos patrones de dependencia.

En la tesis se utilizan las cópulas bivariadas independencia, normal, t, Clayton, Clayton rotada, Gumbel y Gumbel rotada. Para el proceso de selección de las cópulas en los C-vines y D-vines, se diseña una estrategia que utiliza pruebas de hipótesis basadas en estadígrafos Cramér-von Mises. Además, se estudia el impacto en los algoritmos CVEDA y DVEDA del uso de diferentes estrategias de construcción de los vines. En particular, se propone una técnica de truncamiento del vine con el fin de construirlo de manera parcial, pero no de manera arbitraria, sino mediante un procedimiento de selección de modelos.

Para estudiar el comportamiento de los algoritmos CVEDA y DVEDA, se implementan dos paquetes para el ambiente estadístico R: `copulaedas` y `vines`. Se realiza un estudio empírico en varias funciones de prueba ampliamente conocidas y en un problema real complejo: el acoplamiento molecular. Los algoritmos CVEDA y DVEDA se comparan con dos EDA basados en las cópulas multivariadas independencia y normal. Los resultados obtenidos muestran la superioridad de los primeros.

# Índice general

# Índice de tablas

# Índice de figuras

# Introducción

Los Algoritmos con Estimación de Distribuciones (EDA, Estimation of Distribution Algorithms) (Larrañaga y Lozano, 2002) son una clase de algoritmos de optimización evolutivos que se caracterizan por el uso explícito de distribuciones de probabilidad. Estos algoritmos exploran el espacio de búsqueda mediante la simulación de un modelo probabilístico (distribución de búsqueda o simplemente modelo) estimado previamente a partir de las mejores soluciones de la población.

Aunque la idea esencial de los EDA es la estimación y simulación de distribuciones de probabilidad, existen diferencias fundamentales si el dominio es continuo o discreto (Bosman y Thierens, 2006). En esta tesis se trata con problemas con dominio continuo o real. La distribución más utilizada para modelar la distribución de búsqueda en problemas con variables reales es la normal, ya que existen métodos numéricos y algoritmos computacionales eficientes para su estimación y simulación. Entre los EDA más conocidos basados en la distribución normal multivariada u otras variantes derivadas de la misma se encuentran los siguientes: UMDA para dominio continuo (Larrañaga et al., 2000), que utiliza una descomposición en distribuciones normales univariadas; EMNA (Larrañaga et al., 2001), donde se utiliza una distribución normal multivariada; EGNA (Larrañaga et al., 2000), que utiliza redes Gaussianas y el uso de núcleos Gaussianos en los algoritmos propuestos por Bosman y Thierens (2000).

Independientemente de las bondades que ofrece la distribución normal multivariada, esta distribución es muy restrictiva debido a que solamente describe dependencias o interacciones lineales entre las variables y los marginales son todos normales. Asumir normalidad en los datos es raramente consistente con la evidencia empírica de los problemas reales y en ocasiones conlleva a la construcción de modelos probabilísticos incorrectos del espacio de búsqueda.

Una alternativa para evadir las restricciones de la distribución normal multivariada son

las funciones cópulas. Estas funciones permiten representar la estructura de dependencia del problema independientemente de las propiedades estadísticas de los marginales univariados. Esta propiedad de las cópulas brinda la posibilidad de construir distribuciones de búsqueda en las que se combinen marginales de diferentes distribuciones, mientras que la estructura de dependencia queda determinada por la cópula.

La utilización de cópulas en los EDA es muy reciente y es evidente que existe un interés creciente en la aplicación de la teoría de cópula en optimización. Varias son las propuestas realizadas en la literatura: GCEDA (Soto et al., 2007; Arderí, 2007), que utiliza una cópula normal multivariada; una extensión de MIMIC que utiliza cópulas presentada por Salinas-Gutiérrez et al. (2009); los algoritmos basados en la cópula Clayton presentados por Wang et al. (2010a,b) y el uso de cópulas empíricas bivariadas en (Cuesta-Infante et al., 2010). GCEDA es uno de los algoritmos pioneros en el uso de cópulas y constituye el antecedente más importante de la presente investigación.

Aunque el enfoque basado en cópulas multivariadas supera algunas de las restricciones del modelo normal multivariado, tiene varias desventajas. Entre estas desventajas se encuentran que la mayoría de las cópulas existentes son bivariadas y que el uso de una única cópula puede no ser apropiado cuando todas las variables no presentan el mismo tipo de dependencia. Para superar estas deficiencias se ha desarrollado un método que permite descomponer la distribución conjunta en cópulas bivariadas y marginales univariados. Esta descomposición se puede representar mediante un modelo gráfico probabilístico no dirigido llamado vine[1] (Cooke, 1997; Bedford y Cooke, 2001, 2002). Los vines son herramientas flexibles que permiten representar distribuciones de grandes dimensiones combinando una amplia variedad de dependencias y marginales de diferentes distribuciones. Un tipo especial de vines son los C-vines (canonical vines) y los D-vines (drawable vines). Desde el punto de vista gráfico, tanto un C-vine como un D-vine están formados por una cascada de árboles y a cada uno de ellos le corresponde una manera específica de descomponer la distribución.

A partir de las reflexiones enunciadas, se formula el siguiente problema científico: se conjetura que las deficiencias de los EDA basados en la distribución normal multivariada y la cópula normal multivariada pueden aliviarse con la utilización de modelos gráficos basados en vines. El problema científico consiste en reunir evidencias razonables que soporten dicha conjetura.

El objetivo general de esta tesis es desarrollar la clase Algoritmos con Estimación de

[1] La palabra vine no se ha traducido del inglés para evitar introducir un nuevo término en la literatura.

Distribuciones basados en Vines (VEDA, Vine Estimation of Distribution Algorithms). Los objetivos específicos son los siguientes:

1. Crear dos algoritmos pertenecientes a la clase VEDA: CVEDA utilizando C-vines y DVEDA con D-vines, y evaluar la factibilidad de realizar una implementación experimental en el ambiente estadístico R.

2. Evaluar el impacto de utilizar en los modelos diferentes tipos de cópulas.

3. Evaluar el impacto de diferentes estrategias de construcción de los vines: construcción parcial o total y selección de la estructura de los C-vines y D-vines de manera aleatoria o de acuerdo a la fuerza de las interacciones entre las variables.

4. Explorar empíricamente el desempeño de los algoritmos creados en funciones de prueba con características conocidas y en un problema real: el acoplamiento molecular.

La novedad científica de este trabajo consiste en que se introduce en la optimización evolutiva los C-vines y D-vines como modelos de las distribuciones de búsqueda. El uso de este enfoque es una oportunidad para abordar problemas de grandes dimensiones con diferentes y complejos patrones de dependencias. Cabe señalar en este punto que los primeros resultados obtenidos en el marco de esta investigación se publicaron en (Soto y González-Fernández, 2010). Salinas-Gutiérrez et al. (2010) introducen el algoritmo D-vine EDA, basado en D-vines. Sin embargo, un análisis crítico sobre este trabajo permite afirmar que D-vine EDA no supera la desventaja fundamental de los EDA basados en la distribución normal, ya que solamente ajusta la cópula normal en los dos primeros árboles y utiliza la cópula independencia en el resto.

El principal resultado de la tesis es la creación de los Algoritmos con Estimación de Distribuciones basados en Vines.

Para evaluar el comportamiento de CVEDA y DVEDA se realiza un estudio empírico en varias funciones de prueba y en un problema del área de la Bioinformática conocido como acoplamiento molecular. CVEDA y DVEDA se comparan con otros dos EDA basados en cópulas. Los resultados empíricos muestran que los EDA basados en vines superan a los otros dos.

Como parte del desarrollo de la tesis, se implementaron dos paquetes para el ambiente estadístico R (R Development Core Team, 2010): `copulaedas` (González-Fernández y Soto,

2011a), que contiene implementaciones de EDA basados en cópulas, y `vines` (González-Fernández y Soto, 2011b), que provee funcionalidades relacionadas con el modelado de dependencias multivariadas utilizando vines. Estos paquetes se encuentran disponibles en CRAN (Comprehensive R Archive Network). La investigación empírica llevada a cabo en esta tesis se realiza utilizando estos paquetes.

Algunos resultados de la tesis han sido publicados como reporte de investigación del Instituto de Cibernética, Matemática y Física (Soto y González-Fernández, 2010) y aceptados para su publicación como capítulo de un libro (Soto et al., 2011). Además, algunos resultados parciales se presentaron en varios eventos y jornadas científicas. Todo lo anterior se especifica en el apéndice C.

La estructura del resto de la tesis es la siguiente. En el capítulo 1 se realiza una revisión de los enfoques que han sido utilizados para el modelado de dependencias en los EDA con dominio continuo. El capítulo 2 presenta la principal propuesta de la tesis: los algoritmos basados en vines CVEDA y DVEDA. En los capítulos 3 y 4 se presentan los resultados y discusiones acerca del estudio empírico de los algoritmos en funciones de prueba y en el problema de acoplamiento molecular, respectivamente. Finalmente, se brindan las conclusiones, recomendaciones y líneas de trabajo futuro.

# Capítulo 1

# Modelado de dependencias en los EDA continuos

El uso explícito de distribuciones permite a los EDA describir las relaciones de dependencia existentes entre las variables y usar esta información en la optimización. Una distribución de búsqueda que represente adecuadamente la estructura de dependencia del problema permitirá realizar una búsqueda eficiente; un modelo incorrecto puede incluso comprometer la convergencia del algoritmo.

El principal objetivo de este capítulo es identificar las limitaciones fundamentales de los EDA basados en la distribución normal multivariada y la cópula normal multivariada, y motivar la creación de nuevos algoritmos que utilicen paradigmas que permitan construir distribuciones de búsqueda con diferentes tipos de dependencias.

## 1.1. Modelado de dependencias

Para facilitar la comprensión de los resultados presentados en la tesis, en esta sección se incluye una breve descripción de algunos conceptos relacionados con el modelado de dependencias entre variables aleatorias continuas.

### 1.1.1. Medidas de dependencia

El concepto de independencia entre variables aleatorias es fundamental en la teoría de probabilidades (Kurowicka y Cooke, 2006). Se dice que, dos variables aleatorias son

*independientes* si su función de distribución conjunta puede descomponerse en el producto de las funciones de distribución marginales (Nelsen, 2006).

Al decir que dos variables aleatorias no son independientes, no se brinda información acerca de su comportamiento conjunto. Para describir la fuerza y la naturaleza de la dependencia, se deben utilizar medidas de dependencia entre variables aleatorias. En esta sección se presentan dos medidas de dependencia bivariadas: el coeficiente de correlación de Pearson y el coeficiente tau de Kendall.

### Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson indica la fuerza de una dependencia lineal entre dos variables aleatorias. Este coeficiente es un escalar que toma valores en el intervalo $[-1,1]$.

**Definición 1 (Coeficiente de correlación de Pearson)** *De acuerdo a Kurowicka y Cooke (2006), el coeficiente de correlación de Pearson de dos variables aleatorias $X$, $Y$ se define como*

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}},$$

*donde $\mathrm{Cov}(X,Y)$ denota la covarianza entre las variables, $\mathrm{Var}(X)$ la varianza de $X$ y $\mathrm{Var}(Y)$ la varianza de $Y$.*

Dados $N$ pares de observaciones $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$ de un vector aleatorio $(X,Y)$, se puede calcular el coeficiente de correlación de Pearson de la muestra de la forma

$$\hat{\rho}(X,Y) = \frac{\sum_{i=1}^{N}(x_i-\bar{X})\sum_{i=1}^{N}(y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{X})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{Y})^2}},$$

donde $\bar{X} = \sum_{i=1}^{N} x_i$ y $\bar{Y} = \sum_{i=1}^{N} y_i$ (Kurowicka y Cooke, 2006).

Este coeficiente es utilizado frecuentemente en la práctica como una medida de dependencia a pesar de presentar algunas limitaciones: no es invariante ante transformaciones crecientes de las variables y su valor depende de las distribuciones marginales (Kurowicka y Cooke, 2006). Estas limitaciones son desfavorables si se quiere modelar separadamente la estructura de dependencia y las distribuciones marginales de una distribución multivariada.

Embrechts et al. (1999) brindan una lista de observaciones a tomar en cuenta al utilizar el coeficiente de correlación de Pearson.

### Coeficiente tau de Kendall

El término *coeficiente de correlación* generalmente se reserva para referirse a medidas de la dependencia lineal entre variables aleatorias como el coeficiente de correlación de Pearson. Se utiliza el término más general *coeficiente de asociación* para referirse a un grupo de medidas de dependencia entre las que se encuentra el coeficiente tau de Kendall (Nelsen, 2006).

**Definición 2 (Coeficiente tau de Kendall)** *Sean* $(X_1, Y_1)$, $(X_2, Y_2)$ *vectores aleatorios continuos independientes e igualmente distribuidos. De acuerdo a Nelsen (2006), el coeficiente tau de Kendall está dado por*

$$\tau(X,Y) = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \tag{1.1}$$

Dada una muestra con $N$ observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ de un vector aleatorio continuo $(X,Y)$. Se dice que, dos pares $(x_i, y_i)$, $(x_j, y_j)$ de estas observaciones son *concordantes* si $x_i < x_j$ y $y_i < y_j$ ó $x_i > x_j$ y $y_i > y_j$. De manera similar, se dice que son *discordantes* si $x_i < x_j$ y $y_i > y_j$ ó $x_i > x_j$ y $y_i < y_j$. Informalmente, dos variables aleatorias son concordantes si los valores grandes de una suelen estar asociados con los valores grandes de la otra, o los valores pequeños de una suelen estar asociados con los valores pequeños de la otra. Nótese que, en (1.1), $P[(X_1 - X_2)(Y_1 - Y_2) > 0]$ y $P[(X_1 - X_2)(Y_1 - Y_2) < 0]$ denotan la probabilidad de concordancia y discordancia, respectivamente. Finalmente, el coeficiente tau de Kendall para la muestra se calcula como

$$\hat{\tau}(X,Y) = \frac{c-d}{c+d},$$

donde $c$ y $d$ denotan el número de pares concordantes y discordantes en la muestra, respectivamente (Nelsen, 2006).

Este coeficiente es también un escalar que toma valores en el intervalo $[-1, 1]$. A diferencia del coeficiente de correlación de Pearson, el coeficiente tau de Kendall es uno de los llamados *coeficientes de correlación por rangos* y representa dependencias monótonas en-

tre las variables. En este caso, el valor del coeficiente es invariante frente a transformaciones continuas crecientes de las variables (Kurowicka y Cooke, 2006).

### 1.1.2. Cópulas

De acuerdo a Kurowicka y Cooke (2006), una *cópula* es una distribución multivariada con distribuciones marginales uniformes en $[0,1]$. Las cópulas permiten representar de manera separada la estructura de dependencia y las distribuciones marginales de una distribución multivariada. Los artículos introductorios (Genest y Favre, 2007; Trivedi y Zimmer, 2005) y las monografías (Joe, 1997; Nelsen, 2006) contienen extensa información acerca del modelado de dependencias con cópulas. Combinando diferentes distribuciones marginales con diferentes cópulas, es posible modelar distribuciones multivariadas con una amplia variedad de marginales y tipos de dependencia. Esta flexibilidad se basa en el siguiente teorema, considerado el principal resultado en la teoría de cópulas (Romano, 2002).

**Teorema 1 (Sklar, 1959)** *Sea una función de distribución $n$-variada $F$, con funciones de distribución marginales continuas $F_1,\ldots,F_n$. Existe una única cópula $C$ tal que*

$$F(x_1,\ldots,x_n) = C(F_1(x_1),\ldots,F_n(x_n)). \tag{1.2}$$

*Además, si $C$ es una cópula y $F_1,\ldots,F_n$ son funciones de distribución univariadas continuas, la función $F$ definida en (1.2) es una función de distribución con funciones de distribución marginales $F_1,\ldots,F_n$ (Sklar, 1959, 1973).*

A continuación se describe un grupo de cópulas multivariadas y bivariadas a las que se hará referencia durante el desarrollo de la tesis.

#### Cópulas multivariadas

En esta sección se presenta la definición de dos cópulas multivariadas. Estas cópulas se incluyen también en la próxima sección, con sus expresiones particulares para el caso bivariado.

**Cópula independencia multivariada.** La cópula independencia multivariada caracteriza la independencia de un grupo de variables aleatorias continuas. Su función de distribución está dada por

$$C_{\mathrm{I}}(u_1,\ldots,u_n) = u_1 \cdots u_n, \tag{1.3}$$

y se obtiene a partir de la descomposición de la función de distribución conjunta como el producto de las funciones de distribución marginales.

**Cópula normal multivariada.** El teorema de Sklar provee un método para la construcción de cópulas basado en el método de inversión (Nelsen, 2006). Dada una distribución $n$-variada $F$ con distribuciones marginales inversibles $F_1,\ldots,F_n$, se obtiene una cópula como

$$C(u_1,\ldots,u_n) = F(F_1^{-1}(x_1),\ldots,F_n^{-1}(x_n)).$$

Una familia de cópulas construida utilizando este método es la cópula Gaussiana o normal multivariada (Song, 2000) cuya función de distribución está dada por

$$C_{\mathrm{N}}(u_1,\ldots,u_n;R) = \Phi_R(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_n)), \tag{1.4}$$

donde $\Phi^{-1}$ denota la inversa de la función de distribución normal estándar y $\Phi_R$ la función de distribución normal multivariada estándar, con matriz de correlación $R$.

### Cópulas bivariadas

Las cópulas bivariadas descritas en esta sección se utilizan como bloques constructivos de los modelos descritos en la sección 2.1.1. Este grupo de cópulas permite modelar diferentes estructuras de dependencia en las colas de las distribuciones (Aas et al., 2009; Brechmann, 2010). Para ilustrar estas características, en el apéndice A se incluyen gráficos de dispersión de muestras generadas a partir de estas cópulas con diferentes valores de sus parámetros.

**Cópula independencia.** De acuerdo a la definición multivariada de esta cópula, dada en (1.3), la función de distribución de la cópula independencia bivariada está dada por

$$C_{\mathrm{I}}(u,v) = uv. \tag{1.5}$$

**Cópula normal.** De manera semejante a la cópula independencia bivariada, la función de distribución de esta cópula se obtiene como un caso particular de la expresión multivariada dada en (1.4). En este caso, la matriz de correlación se reduce a un escalar y la función de distribución se define como

$$C_{\mathrm{N}}(u,v;\rho) = \Phi_\rho(\Phi^{-1}(u),\Phi^{-1}(v)), \tag{1.6}$$

donde $\Phi^{-1}$ denota la inversa de la función de distribución normal estándar y $\Phi_\rho$ la función de distribución normal bivariada estándar con coeficiente de correlación $\rho \in (-1,1)$. Cuando $\rho \to 1$ ó $\rho \to -1$, esta cópula representa una dependencia perfecta positiva o negativa respectivamente, mientras que para $\rho = 0$ se obtiene la cópula independencia bivariada.

**Cópula t.** La cópula t se construye de manera semejante a la cópula normal. Su función de distribución para el caso bivariado está dada por

$$C_{\mathrm{t}}(u,v;\rho,\nu) = t_{\rho,\nu}(t_\nu^{-1}(u),t_\nu^{-1}(v)), \tag{1.7}$$

donde $t_\nu^{-1}$ denota la inversa de la función de distribución t estándar con $\nu > 0$ grados de libertad y $t_{\rho,\nu}$ la función de distribución t bivariada estándar con coeficiente de correlación $\rho \in (-1,1)$ y $\nu > 0$ grados de libertad. Al igual que la cópula normal, esta cópula representa una dependencia perfecta positiva o negativa cuando $\rho \to 1$ ó $\rho \to -1$, respectivamente. Cuando $\rho = 0$ se obtiene la cópula independencia. El parámetro $\nu$ controla la dependencia en las colas. Al igual que la distribución t bivariada tiende a la distribución normal bivariada con el aumento de los grados de libertad, la cópula t se acerca a la cópula normal a medida que $\nu \to \infty$.

**Cópula Clayton.** La función de distribución de la cópula Clayton tiene la forma

$$C_{\mathrm{C}}(u,v;\delta) = \left(u^{-\delta} + v^{-\delta} - 1\right)^{-1/\delta}, \tag{1.8}$$

donde $\delta > 0$ es un parámetro que controla la dependencia. Las variables son independientes cuando $\delta \to 0$, mientras que cuando $\delta \to \infty$ la cópula representa una dependencia positiva perfecta. Esta cópula exhibe dependencias fuertes en la cola inferior, pero relativamente débiles en la cola superior.

**Cópula Clayton rotada.** La cópula Clayton con función de distribución definida en (1.8) solamente captura dependencias positivas. Por esta razón, siguiendo la transformación utilizada por Brechmann (2010), se considera una rotación de $90^{\text{o}}$ de esta cópula. Se dice que un vector aleatorio $(U,V) \in [0,1]^2$ distribuye como una cópula Clayton rotada con parámetro $\delta < 0$ si el vector $(U, 1-V)$ distribuye como una cópula Clayton con parámetro $-\delta$. La función de distribución de la cópula Clayton rotada está dada por

$$C_{\text{RC}}(u,v;\delta) = u - C_{\text{C}}(u,1-v;-\delta),$$

donde $C_{\text{C}}$ denota la función de distribución de la cópula Clayton dada en (1.8). Esta cópula mantiene las características de la dependencia en las colas de la cópula Clayton y representa la independencia cuando $\delta \to 0$ pero, a diferencia de la cópula Clayton, representa una dependencia negativa perfecta cuando $\delta \to -\infty$.

**Cópula Gumbel.** La función de distribución de la cópula Gumbel se define como

$$C_{\text{G}}(u,v;\delta) = \exp\left(-\left((-\log u)^{\delta} + (-\log v)^{\delta}\right)^{1/\delta}\right), \tag{1.9}$$

donde $\delta \geq 1$ es un parámetro que controla la dependencia. Esta cópula se transforma en la cópula independencia cuando $\delta = 1$ y cuando $\delta \to \infty$ representa una dependencia positiva perfecta. En contraste con la cópula Clayton, esta cópula exhibe dependencias fuertes en la cola superior, pero relativamente débiles en la cola inferior.

**Cópula Gumbel rotada.** Al igual que la cópula Clayton, la cópula Gumbel solamente representa dependencias positivas. Por esta razón, al igual que Brechmann (2010), se considera una rotación de $90^{\text{o}}$ de la cópula Gumbel. La función de distribución de la cópula Gumbel rotada está dada por

$$C_{\text{RG}}(u,v;\delta) = u - C_{\text{G}}(u,1-v;-\delta),$$

donde $C_{\text{G}}$ denota la función de distribución de la cópula Gumbel dada en (1.9) y $\delta \leq -1$ el parámetro que controla la dependencia. Esta cópula mantiene las características de la dependencia en las colas de la cópula Gumbel, se transforma en la cópula independencia para $\delta = -1$ y representa una dependencia negativa perfecta cuando $\delta \to -\infty$.

### Estimación de los parámetros

Existen varios métodos para estimar los parámetros de una cópula a partir de una muestra (Genest y Favre, 2007; Genest et al., 1995; Joe, 1997). Los métodos que realizan la estimación mediante máxima verosimilitud son aplicables a un gran número de cópulas. Otros métodos calculan el parámetro de la cópula a partir de una medida de asociación calculada en la muestra. En este último caso, la cópula debe satisfacer ciertas restricciones.

Los métodos que realizan la estimación mediante máxima verosimilitud generalmente requieren el uso de algoritmos de optimización para encontrar el valor de los parámetros. Entre estos métodos se encuentran diferencias en cuanto a:

- si se estiman los parámetros de las distribuciones marginales conjuntamente con los de la cópula o de manera separada y
- si se asume una distribución paramétrica de las distribuciones marginales o se realiza una estimación empírica.

En el caso de las cópulas bivariadas con un parámetro real, es posible expresar el coeficiente tau de Kendall en términos del parámetro de la cópula (Nelsen, 2006). Estas expresiones se pueden invertir para obtener el valor del parámetro de la cópula a partir del valor del coeficiente calculado en la muestra (sección 1.1.1). En la tabla 1.1 se presentan las expresiones del parámetro de las cópulas bivariadas con un parámetro real descritas en la sección 1.1.2 en función del coeficiente tau de Kendall.

**Tabla 1.1:** Expresiones del parámetro de un grupo de cópulas bivariadas en función del coeficiente tau de Kendall.

| Cópula | Parámetro de la cópula en función del coeficiente tau de Kendall |
|---|---|
| Normal | $\operatorname{sen}\left(\frac{\pi}{2}\tau\right)$ |
| Clayton | $2\tau/(1-\tau)$ |
| Clayton rotada | $2\tau/(1+\tau)$ |
| Gumbel | $1/(1-\tau)$ |
| Gumbel rotada | $-1/(1+\tau)$ |

### 1.1.3. Selección de modelos

En esta sección se describen métodos para determinar si un modelo es una buena representación de las características presentes en una muestra. Estos métodos también permiten seleccionar el modelo que brinda una mejor representación si se cuenta con varios modelos candidatos.

#### Prueba de bondad de ajuste para cópulas

Las pruebas de bondad de ajuste para cópulas permiten valorar cómo una cópula describe la estructura de dependencia de una muestra. Esta valoración se realiza en términos de una prueba de hipótesis. A continuación se describe, para el caso bivariado, una prueba de bondad de ajuste para cópulas propuesta por Genest y Rémillard (2008) que ha mostrado buenos resultados en varios estudios comparativos (Berg, 2009; Genest et al., 2009).

Sean $N$ observaciones $(u_1, v_1), \ldots, (u_N, v_N)$ de un vector aleatorio $(U, V) \in [0,1]^2$ con función de distribución conjunta $C$. La información sobre las dependencias en estas observaciones se representa mediante la cópula empírica (Deheuvels, 1979), dada por la función de distribución empírica de la muestra,

$$C_{\mathrm{E}}(u,v) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{u_i \leq u, v_i \leq v}.$$

La prueba de hipótesis se basa en una distancia entre la cópula empírica y una estimación $C_{\hat{\theta}}$ de la cópula $C$ a partir de la muestra, bajo la hipótesis nula de que $C$ pertenece a una familia de cópulas $C_\theta$ con parámetro $\theta$. Esta distancia se mide mediante un estadígrafo Cramér-von Mises definido por

$$S_N = N \int_{[0,1]^2} \left( C_{\mathrm{E}}(u,v) - C_{\hat{\theta}}(u,v) \right)^2 dC_{\mathrm{E}}(u,v) = \sum_{i=1}^{N} \left( C_{\mathrm{E}}(u_i,v_i) - C_{\hat{\theta}}(u_i,v_i) \right)^2. \tag{1.10}$$

Como la distribución de $S_N$ no se conoce en la práctica y depende de la cópula candidata y el valor de sus parámetros, los p-valores (p-values) deben ser calculados siguiendo un procedimiento de remuestreo (bootstrap) (Genest et al., 2009). Una alternativa más eficiente es utilizar el procedimiento propuesto por Kojadinovic y Yan (2011), pero aún siguiendo este procedimiento el cálculo de los p-valores es un proceso costoso computacionalmente

(Brechmann, 2010).

### Prueba de independencia

Una prueba de independencia puede considerarse como un caso particular de la prueba de bondad de ajuste para cópulas descrita en la sección 1.1.3. En este caso, la hipótesis nula es que la cópula desconocida $C$ es la cópula independencia. Esta prueba de independencia fue inicialmente sugerida por Deheuvels (1981) y posteriormente estudiada por Genest et al. (2007); Genest y Rémillard (2004).

En el caso bivariado, se utiliza el estadígrafo definido en (1.10), sustituyendo $C_{\hat{\theta}}(u,v)$ por $C_{\mathrm{I}}(u,v)$. Los p-valores pueden ser calculados con el mismo procedimiento de remuestreo utilizado para las pruebas de bondad de ajuste para cópulas. En este caso, el procedimiento es más eficiente debido a que no es necesario estimar los parámetros de la cópula en cada repetición.

### Criterios AIC y BIC

En esta sección, se presentan dos métodos de selección de modelos que no se expresan en términos de pruebas de hipótesis. Estos métodos son el *criterio de información de Akaike* (AIC) y el *criterio de información bayesiano* (BIC).

**Definición 3 (AIC)** *Dado un grupo de observaciones $\mathbf{x}_i$, $i = 1,\ldots,n$, el criterio de información de Akaike para un modelo se define como*

$$\mathrm{AIC} = -2\sum_{i=1}^{n} \log f(\mathbf{x}_i;\hat{\theta}) + 2k$$

*donde $f$ denota la función de verosimilitud del modelo y $\hat{\theta}$ el vector de los $k$ parámetros del modelo estimado mediante máxima verosimilitud. El primer término es una medida de la bondad del ajuste y el segundo penaliza la complejidad del modelo en términos del número de parámetros (Akaike, 1974).*

BIC incluye el número de observaciones de la muestra como parte del término de penalización, favoreciendo la selección de modelos con un menor número de parámetros (Schepsmeier, 2010).

**Definición 4 (BIC)** *Dado un grupo de observaciones* $\mathbf{x}_i$*,* $i = 1, \ldots, n$*, el criterio de información bayesiano para un modelo se define como*

$$\mathrm{BIC} = -2 \sum_{i=1}^{n} \log f(\mathbf{x}_i; \hat{\theta}) + \log(n)k$$

*donde* $f$ *denota la función de verosimilitud del modelo y* $\hat{\theta}$ *el vector de los* $k$ *parámetros del modelo estimado mediante máxima verosimilitud (Schwarz, 1978).*

Usualmente, estos criterios se emplean en la comparación de modelos anidados (Brechmann, 2010). La selección se realiza escogiendo el modelo para el cual se obtuvo el menor valor del criterio utilizado.

## 1.2. Algoritmos con estimación de distribuciones continuos

En esta sección se describe el esquema del funcionamiento de los EDA y se realiza una breve revisión de los algoritmos con dominio continuo propuestos en la literatura. Además, se explican detalladamente los algoritmos UMDA y GCEDA.

### 1.2.1. Esquema del funcionamiento de los EDA

Un EDA comienza con la generación de un grupo de soluciones de manera aleatoria. Estas soluciones constituyen la población inicial. Cada individuo de esta población se evalúa de acuerdo a la función objetivo y posteriormente se selecciona un grupo de estos individuos para formar la población seleccionada. Generalmente, este grupo de individuos son los mejores con respecto a la función objetivo del problema de optimización. El siguiente paso consiste en la estimación de un modelo probabilístico de la población seleccionada. Mediante la simulación del modelo estimado se crea una nueva población. De esta manera, se ejecuta un procedimiento iterativo de evaluación, selección, estimación y simulación de un modelo probabilístico hasta cumplir alguna condición de parada. El procedimiento descrito anteriormente se muestra en el algoritmo 1.1.

La estimación y la simulación del modelo probabilístico son los pasos esenciales en los EDA. De acuerdo al modelo probabilístico utilizado, es posible modelar diferentes estructuras de dependencia e interacciones entre las variables de la función objetivo. Aunque la idea

**Algoritmo 1.1** Esquema del funcionamiento de los EDA.

*Inicialización:* Generar los individuos de la población inicial aleatoriamente.

*Evaluación:* Evaluar los individuos de la población de acuerdo a la función objetivo.

*Selección:* Crear la población seleccionada con los mejores individuos de la población.

*Estimación*: Estimar un modelo probabilístico a partir de la población seleccionada.

*Simulación:* Simular una nueva población a partir del modelo estimado.

*Terminación:* Si la condición de parada no se satisface, ir al paso *Evaluación*.

esencial de los EDA es la estimación y simulación de distribuciones, existen diferencias fundamentales si el dominio es continuo o discreto. En esta tesis se tratan problemas con dominio continuo o real.

### 1.2.2. Breve revisión de los EDA continuos

Los modelos probabilísticos basados en la distribución normal multivariada y descomposiciones de la misma han dominado el modelado de las distribuciones de búsqueda de los EDA para espacios continuos (Bosman y Thierens, 2006; Kern et al., 2003). Algunas excepciones a esta situación son: la subdivisión del espacio de búsqueda utilizando árboles de decisión (Ocenasek y Schwarz, 2002), la aplicación de análisis de componentes principales (Cho y Zhang, 2001, 2004), el uso de la distribución de Cauchy (Pŏsík, 2009), el modelado probabilístico basado en histogramas (Tsutsui et al., 2001) y el modelado con cópulas (Soto et al., 2007; Arderí, 2007). Por su relevancia para la tesis, a continuación se describen algunos EDA basados en la distribución normal y en cópulas.

#### EDA basados en la distribución normal

El algoritmo más simple basado en la distribución normal asume la independencia entre las variables, lo cual corresponde a descomponer la densidad conjunta en el producto de densidades univariadas con distribución normal. El Algoritmo con Distribución Marginal Univariada con dominio continuo (UMDA, Univariate Marginal Distribution Algorithm) presentado por Larrañaga et al. (2000) se basa en este modelo.

Otros EDA basados en la distribución normal consideran correlaciones entre las variables y modelan la distribución de búsqueda utilizando una distribución normal multivariada o

una descomposición de la misma en probabilidades condicionales. El algoritmo EMNA (Larrañaga et al., 2001) utiliza como modelo probabilístico una distribución normal multivariada. En cada generación de este algoritmo se calculan los estimadores de máxima verosimilitud para el vector de medias y la matriz de covarianza. Ochoa (2010) introduce el algoritmo SEDAmn, que utiliza estimadores de la covarianza regularizados por encogimiento en lugar de los de máxima verosimilitud.

Con el objetivo de reducir el número de parámetros requeridos para definir la distribución conjunta, en (Larrañaga et al., 2000) se propone utilizar una red Gaussiana que descompone la distribución conjunta en probabilidades condicionales. Este modelo representa las relaciones de dependencia entre las variables mediante un grafo dirigido acíclico. En esta misma línea, otro algoritmo que construye redes Gaussianas, pero esta vez basado en pruebas de independencia es CMMHC (Madera, 2009).

Cuando la estructura de dependencia de los datos es compleja, resulta difícil para la red Gaussiana capturar las correlaciones. Como una alternativa para solucionar esta situación, se han aplicado mezclas de distribuciones. Bosman y Thierens (2001); Gallagher et al. (1999) proponen mezclas de distribuciones normales multivariadas. Como un caso extremo de mezcla de distribuciones, Bosman y Thierens (2000) utilizan núcleos Gaussianos donde existe una función de densidad para cada individuo de la población seleccionada. Los núcleos Gaussianos tienen propiedades interesantes pero es difícil interpretar la estructura de dependencia descrita por el modelo (Bosman y Thierens, 2006).

En la distribución normal multivariada, todos los marginales son normales y la estructura de dependencia entre las variables aleatorias está dada por el coeficiente de correlación de Pearson. En muchos problemas asumir que los datos siguen una distribución normal no es consistente con la evidencia empírica y puede conllevar a la construcción de distribuciones de búsqueda incorrectas. Aunque en ocasiones los términos dependencia y correlación se utilizan indistintamente, la correlación es un tipo particular de dependencia que caracteriza la existencia de una dependencia lineal. Debido a esta situación, los EDA basados en la distribución normal multivariada presentan limitaciones para capturar dependencias no lineales entre las variables.

El uso de un enfoque basado en cópulas para modelar las distribuciones de búsqueda de los EDA ofrece una alternativa para evadir las limitaciones anteriores. Las cópulas permiten representar separadamente las propiedades estadísticas de las variables y su estructura de dependencia. De esta manera, es posible construir distribuciones de búsqueda con diferen-

tes tipos de distribuciones marginales. Esta propiedad es útil para modelar características de las distribuciones marginales que no son representadas apropiadamente utilizando una distribución normal, como son: la multimodalidad y la asimetría de la muestra de cada variable en la población inicial con respecto al valor de cada variable en la configuración del óptimo (Arderí, 2007). En la siguiente sección se describen de manera general algunos EDA basados en cópulas propuestos en la literatura.

### Algoritmos basados en cópulas

El uso de cópulas en los EDA comienza con la introducción en (Arderí, 2007; Soto et al., 2007) del Algoritmo con Estimación de Distribuciones basado en Cópula Gaussiana (GCEDA, Gaussian Copula Estimation of Distribution Algorithm). En GCEDA los marginales pueden ser de diferentes distribuciones, mientras la estructura de dependencia se describe por la cópula normal multivariada. En la sección 1.2.3 se describe este algoritmo detalladamente.

Barba (2007) introduce un algoritmo llamado COPULEDA que modela la distribución de búsqueda utilizando una cópula normal multivariada, por lo que este algoritmo es equivalente a GCEDA. Las distribuciones marginales se modelan utilizando la distribución empírica.

Salinas-Gutiérrez et al. (2009) presentan una extensión del algoritmo MIMIC con dominio continuo (Larrañaga et al., 1999) que utiliza cópulas normal o Frank. Al ser una extensión de MIMIC, la estructura de dependencia aprendida es una cadena. Para determinar la estructura de la cadena, se encuentra la permutación de las variables que minimice la entropía relativa entre la función de densidad empírica y la aproximación basada en cópulas. En esta propuesta no se brinda un criterio para la selección de la cópula, sino que debe ser fijada antes de la ejecución del algoritmo y no se brinda la posibilidad de combinar cópulas de diferentes familias. Las distribuciones marginales se modelan con distribuciones normal o Beta. En el caso de la distribución Beta, se aplica una transformación lineal para extender los valores fuera del intervalo $[0, 1]$.

Wang et al. han propuesto varios algoritmos basados en cópulas. En (Wang et al., 2009) se presenta un algoritmo que modela la estructura de dependencia utilizando una cópula normal bivariada y las distribuciones marginales con la distribución normal. Este algoritmo solamente puede optimizar funciones de dos variables. En (Wang et al., 2010a,b) se presenta un algoritmo que modela la estructura de dependencia utilizando una cópula Clayton multivariada. En el primero los marginales se modelan con la distribución empírica, mientras

que en el segundo con la distribución normal. El parámetro de la cópula Clayton no se estima, sino que toma un valor constante. Ello es equivalente a asumir la misma estructura de dependencia durante toda la evolución. Como trabajo futuro los autores plantean el estudio de la selección del valor del parámetro de la cópula.

Cuesta-Infante et al. (2010) introducen dos EDA basados en cópulas. El primero de ellos utiliza cópulas empíricas bivariadas y marginales empíricos, ambos suavizados mediante interpolación lineal. Las cópulas empíricas bivariadas se combinan utilizando una construcción en forma de cadena. Esta construcción no resulta en una cópula de dimensiones superiores ni en una distribución de probabilidad de la población seleccionada. El segundo algoritmo modela la estructura de dependencia utilizando una de las cópulas arquimedianas: Frank, HRT o Clayton. La cópula y el parámetro de la misma se fijan antes de la ejecución del algoritmo. Se realiza una comparación entre estos algoritmos en un amplio grupo de funciones de prueba.

### 1.2.3. UMDA y GCEDA

En esta sección se describen más detalladamente dos EDA presentados en la sección 1.2.2 que son relevantes a los objetivos de la tesis, debido a que están basados en cópulas y se utilizan para comparar los algoritmos propuestos en este trabajo.

Una consecuencia del teorema de Sklar es que un grupo de variables aleatorias son independientes si la estructura de dependencia entre ellas está dada por la cópula independencia multivariada (sección 1.1.2). El algoritmo UMDA está directamente relacionado con esta cópula debido a que asume la independencia entre las variables.

Otra cópula importante es la normal multivariada (sección 1.1.2) que está asociada a la distribución normal multivariada mediante el teorema de Sklar. Esta cópula permite representar una estructura de dependencia normal y utilizar diferentes tipos de distribuciones marginales. El algoritmo GCEDA se basa en esta cópula.

El cálculo de la matriz de correlación en el paso de estimación de GCEDA es diferente de acuerdo a las distribuciones marginales utilizadas. Si se modelan los marginales utilizando la distribución normal, se puede calcular la matriz de correlación directamente de la población seleccionada, ya que el modelo resultante corresponde a la distribución normal multivariada. Si se utilizan otras distribuciones marginales, la matriz de correlación se construye a partir de la inversión del coeficiente tau de Kendall (sección 1.1.2) para cada par de variables. Si esta construcción no resulta en una matriz definida positiva, se aplica la corrección propuesta por

Rousseeuw y Molenberghs (1993).

La generación de un nuevo individuo de dimensión $n$ durante la simulación del algoritmo GCEDA se divide en dos partes: la simulación de un vector $(u_1, \ldots, u_n)$ a partir de la normal multivariada mediante el algoritmo 1.2 y la obtención de cada componente $x_i$ del nuevo individuo a partir de la componente $u_i$ utilizando el método de inversión (Devroye, 1986).

**Algoritmo 1.2** Simulación de la cópula normal multivariada.

1. Encontrar la descomposición de Cholesky $A$ de la matriz de correlación empírica $\hat{R}$.
2. Simular $n$ variables independientes con distribución normal estándar, resultando $\mathbf{z} = (z_1, \ldots, z_n)$.
3. Asignar $\mathbf{x} = A\mathbf{z}$.
4. Determinar $u_i = \Phi(x_i)$ con $i = 1, \ldots, n$ y $\Phi$ la función de distribución normal estándar.
5. El vector $(u_1, \ldots, u_n)$ es una simulación de la cópula normal multivariada con matriz de correlación empírica $\hat{R}$.

Como se ha mencionado anteriormente, una ventaja fundamental de estos algoritmos basados en cópulas es la posibilidad de tratar con diferentes distribuciones marginales. Esta propiedad se utiliza en el capítulo 4 para satisfacer las restricciones de dominio de las variables del problema.

## 1.3. Conclusiones del capítulo

Aunque la cópula normal multivariada supera algunas restricciones de la distribución normal multivariada, el enfoque basado en cópulas multivariadas tiene algunas desventajas: existe un número reducido de cópulas con extensiones multivariadas; además, generalmente poseen un único parámetro para describir las interacciones entre todos los pares de variables. De aquí la necesidad de buscar otras alternativas que permitan crear EDA capaces de modelar diferentes tipos de dependencias y combinarlas adecuadamente.

# Capítulo 2

# VEDA — Algoritmos con estimación de distribuciones basados en vines

Este capítulo presenta la principal propuesta de la tesis. El objetivo fundamental es crear dos algoritmos basados en vines: CVEDA utilizando C-vines y DVEDA con D-vines.

## 2.1. De las cópulas multivariadas a los vines

Las cópulas permiten construir distribuciones de búsqueda más flexibles que evaden las limitaciones impuestas por la distribución normal multivariada, debido a la posibilidad que brindan de separar las distribuciones multivariadas en la información marginal y la estructura de dependencia (sección 1.2.3).

No obstante, el enfoque basado en cópulas multivariadas presenta algunos inconvenientes. Primeramente, el número de cópulas disponibles para modelar la estructura de dependencia entre más de dos variables es limitado. De hecho, la mayoría de las cópulas paramétricas disponibles son bivariadas. Además, el enfoque basado en una cópula multivariada puede no resultar apropiado cuando los pares de variables presentan estructuras de dependencia diferentes. Finalmente, en muchos casos en que sí se cuenta con una extensión multivariada de la cópula, se utiliza un solo parámetro para describir la estructura de dependencia entre todos los pares de variables, lo cual es una limitación importante si las relaciones entre los pares de variables son diferentes.

Las construcciones con cópula bivariadas y los vines constituyen una alternativa al enfoque basado en una cópula multivariada. Estos modelos permiten extender las cópulas

bivariadas a dimensiones superiores, utilizando solamente cópulas bivariadas y densidades univariadas como bloques constructivos. El creciente interés en estos modelos se debe a la gran flexibilidad que brindan para modelar una amplia variedad de dependencias combinando cópulas bivariadas de familias diferentes. Cuando se modela la estructura de dependencias utilizando vines, no es necesario asumir una estructura de dependencia entre las variables, ya que los procedimientos de estimación pueden seleccionar las cópulas bivariadas que ajusten los datos apropiadamente.

### 2.1.1. Construcciones con cópulas bivariadas

La definición de cópulas multivariadas generalmente se reconoce como un problema difícil. Existen muchas cópulas bivariadas, pero el número de cópulas multivariadas es limitado (Aas et al., 2009). Las construcciones con cópulas bivariadas son una forma flexible de extender las cópulas bivariadas a dimensiones superiores. Este método ha mostrado un buen comportamiento en comparación con otros métodos para el modelado de dependencias multivariadas (Berg y Aas, 2007).

El desarrollo de estas construcciones se basa en la descomposición de la densidad multivariada de un vector aleatorio continuo $\mathbf{X} = (X_1, \ldots, X_n)$ en el producto de densidades condicionales, siguiendo la regla de la cadena

$$f(x_1, \ldots, x_n) = f(x_n) f(x_{n-1} \mid x_n) f(x_{n-2} \mid x_{n-1}, x_n) \cdots f(x_1 \mid x_2, \ldots, x_n). \tag{2.1}$$

Utilizando la definición de densidad condicional y la expresión de la densidad multivariada que se obtiene derivando (1.2), cada factor en (2.1) puede descomponerse en el producto de una cópula bivariada y una densidad condicional de la forma

$$f(x \mid \mathbf{v}) = c_{xv_j|\mathbf{v}_{-j}} \left( F\left(x \mid \mathbf{v}_{-j}\right), F\left(v_j \mid \mathbf{v}_{-j}\right) \right) f\left(x \mid \mathbf{v}_{-j}\right), \tag{2.2}$$

donde $\mathbf{v}$ es un vector con $m$ elementos, $v_j$ es una componente del vector $\mathbf{v}$ seleccionada arbitrariamente y $\mathbf{v}_{-j}$ denota el vector con $m-1$ elementos resultante de excluir $v_j$ de $\mathbf{v}$. Nótese que se asume que la cópula bivariada es independiente de las variables en el condicionante de sus argumentos. Hobæk Haff et al. (2010) muestran que esta simplificación es una buena aproximación de la descomposición correcta.

Combinando (2.1) y (2.2), se obtienen descomposiciones de la densidad multivariada en términos de densidades marginales univariadas y cópulas bivariadas que pueden pertenecer

a diferentes familias. Estas descomposiciones se denominan *construcciones con cópulas bivariadas* (Aas et al., 2009).

Los argumentos de la cópula bivariada en (2.2) son distribuciones condicionales de la forma $F(x \mid \mathbf{v})$. Estas expresiones se calculan utilizando

$$F\left(x \mid \mathbf{v}\right) = \frac{\partial C_{xv_j|\mathbf{v}_{-j}}\left(F\left(x \mid \mathbf{v}_{-j}\right), F\left(v_j \mid \mathbf{v}_{-j}\right)\right)}{\partial F\left(v_j \mid \mathbf{v}_{-j}\right)},$$

donde $C_{xv_j|\mathbf{v}_{-j}}$ es la función de distribución de una cópula bivariada (Joe, 1996). Para facilitar el cálculo de $F\left(x \mid \mathbf{v}\right)$, se denota por $h\left(x, v; \theta\right)$ la expresión anterior con una sola variable condicionante y distribuciones marginales uniformes

$$h(x, v, \theta) = F(x \mid v) = \frac{\partial C_{xv}\left(x, v; \theta\right)}{\partial v},$$

donde $\theta$ denota el vector de los parámetros de la cópula $C_{xv}$. Siguiendo esta notación, es posible expresar las distribuciones condicionales de cualquier orden como evaluaciones de funciones $h$ anidadas. Además, se denota por $h^{-1}(u, v; \theta)$ la inversa de la función $h$ con respecto a la primera variable $x$, o equivalentemente, la inversa de la función de distribución condicional. En el apéndice B se incluye la definición de las funciones $h$ y $h^{-1}$ de las cópulas bivariadas descritas en la sección 1.1.2.

Como la componente $v_j$ del vector $\mathbf{v}$ en (2.2) se puede seleccionar arbitrariamente, para una misma densidad multivariada existen varias construcciones con cópulas bivariadas. En la siguiente sección se presenta un modelo gráfico que organiza estas construcciones.

### 2.1.2. Vines

Con el objetivo de organizar las construcciones con cópulas bivariadas, Bedford y Cooke (2001, 2002) introducen un modelo gráfico llamado vine regular. Este modelo gráfico es un caso particular de un modelo más general presentado por Cooke (1997).

De acuerdo a Kurowicka y Cooke (2006), un *vine regular* es un conjunto de árboles anidados, donde las aristas del árbol $j$ son los nodos del árbol $j+1$ y dos aristas en el árbol $j$ están unidas por una arista en el árbol $j+1$ si comparten un nodo en el árbol $j$. Las aristas representan las cópulas bivariadas en la descomposición.

Dos casos particulares de los vines regulares son el C-vine y el D-vine, descritos en la siguiente definición (Kurowicka y Cooke, 2006).

**Definición 5 (C-vine y D-vine)** *Un vine regular en $n$ variables se denomina*

- C-vine*, si cada árbol $T_j$ tiene un único nodo con grado $n-j$. El nodo con grado $n-1$ en el árbol $T_1$ se denomina raíz.*
- D-vine*, si cada nodo en el árbol $T_1$ tiene a lo sumo grado 2.*

La figura 2.1 muestra ejemplos de un C-vine y un D-vine con cuatro variables. Cada uno de estos modelos brinda una forma específica de descomponer la densidad multivariada utilizando construcciones con cópulas bivariadas. Para un vector $\mathbf{x}$ con $n$ variables, la densidad de un C-vine está dada por

$$f(\mathbf{x}) = \prod_{k=1}^{n} f(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j,j+i|i,\ldots,j-1} \left( F\left(x_j | x_1, \ldots, x_{j-1}\right), F\left(x_{j+i} | x_1, \ldots, x_{j-1}\right) \right)$$

y la densidad de un D-vine por

$$f(\mathbf{x}) = \prod_{k=1}^{n} f(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,i+j|i+1,\ldots,i+j-1} \left( F\left(x_i | x_{i+1}, \ldots, x_{i+j-1}\right), F\left(x_{i+j} | x_{i+1}, \ldots, x_{i+j-1}\right) \right),$$

donde el índice $j$ identifica los árboles e $i$ las aristas de cada árbol.

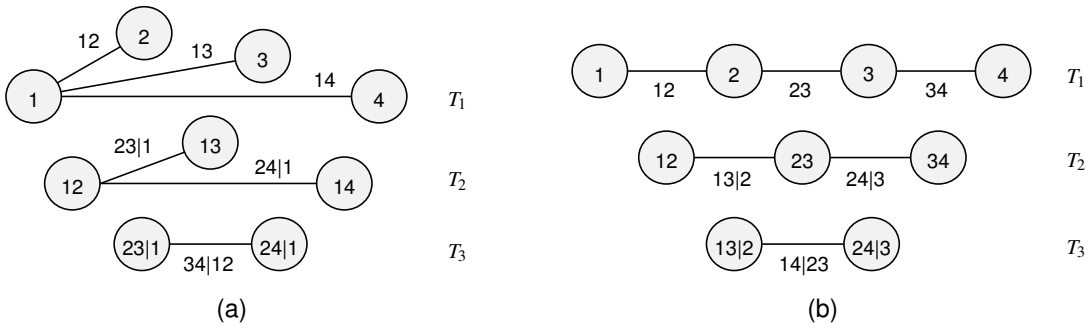


**Figura 2.1:** Un C-vine (a) y un D-vine (b) con cuatro variables. En un C-vine, cada árbol tiene un único nodo conectado al resto. En un D-vine, ningún nodo está conectado a más de dos.

## 2.2. EDA basados en vines

Los Algoritmos con Estimación de Distribuciones basados en Vines (VEDA, Vine Estimation of Distribution Algorithms) son una clase de EDA que utiliza los vines como modelos

de las distribuciones de búsqueda. CVEDA y DVEDA son instancias de VEDA basados en C-vines y D-vines, respectivamente (Soto y González-Fernández, 2010). A continuación se describen los pasos de estimación y simulación de estos dos algoritmos.

### 2.2.1. Estimación

Los métodos de construcción de C-vines y D-vines han sido desarrollados por Aas et al. (2009). La estimación de C-vines y D-vines debe considerar dos tareas fundamentales: 1) selección de la estructura de los C-vines y D-vines y 2) selección de cada cópula bivariada en la descomposición y estimación de sus parámetros. A continuación se describen estos pasos de acuerdo a las implementaciones de los algoritmos CVEDA y DVEDA.

1. Selección de la estructura de los C-vines y D-vines.

   Una vez que se ha decidido una estructura general (C-vine o D-vine), se debe seleccionar qué pares de variables serán modeladas explícitamente con cópulas. Un C-vine resulta apropiado si una variable gobierna las interacciones entre las variables. Un D-vine permite una selección más flexible de las dependencias.

   Para determinar la estructura de los vines se utilizan heurísticas que intentan representar las dependencias más fuertes en el primer árbol. La fuerza de las dependencias representadas en el árbol se mide como la suma de un peso asignado a cada arista. Este peso es el valor absoluto del coeficiente tau de Kendall (sección 1.1.1) entre el par de variables en la población seleccionada.

   - La selección de la estructura de un C-vine se realiza mediante la elección del nodo raíz. Este nodo se determina de manera iterativa, considerando cada uno de los nodos como nodo raíz y calculando la fuerza de las dependencias representadas en el árbol. El nodo raíz será el que maximice la suma de los pesos de las aristas del árbol resultante.
   - En un D-vine la estructura está completamente determinada por la cadena formada por el primer árbol. El método de selección de la estructura consiste en encontrar la secuencia de variables más larga (en términos de los pesos de las aristas) en la cual cada variable aparece solamente una vez. De acuerdo a Brechmann (2010), este problema puede transformarse en el problema del viajante añadiendo un nodo ficticio conectado con aristas con peso cero al resto de los

nodos. En este trabajo, se construye una solución aproximada del problema del viajante siguiendo la heurística de la inserción menos costosa descrita en (Rosenkrantz et al., 1977).

Es importante destacar que la flexibilidad que permiten los vines para modelar dependencias viene con un costo asociado. Ello implica que para una distribución con $n$ variables, hay que construir $n-1$ árboles y ajustar cópulas en $n(n-1)/2$ aristas. En el contexto de los EDA, donde en cada generación se debe construir un vine, es muy importante implementar estrategias que permitan disminuir el costo computacional de la construcción del mismo.

Una primera estrategia es construir el vine parcialmente hasta un árbol determinado de manera arbitraria. Esta opción tiene la desventaja de que no se está teniendo en cuenta información que puede ser decisiva para que el algoritmo converja en problemas con fuertes interacciones entre las variables.

Brechmann (2010) propone la técnica de truncamiento con el fin de construir el vine de manera parcial pero no de manera arbitraria, sino mediante un procedimiento de selección de modelos basado en AIC o BIC (sección 1.1.3). La idea general de esta técnica es la siguiente: el árbol $j+1$ se expande si el valor del criterio de información para el vine con los primeros $j+1$ árboles es menor que el valor obtenido para el vine con los primeros $j$ árboles; en otro caso, el vine es truncado en el árbol $j$, lo cual significa asignar la cópula independencia a las aristas del árbol $j+1$ y todos los árboles siguientes.

2. Selección de cada cópula bivariada en la descomposición y estimación de sus parámetros mediante el siguiente procedimiento iterativo:

   a) Determinar las cópulas en el árbol 1 a partir de la muestra.

   b) Obtener las observaciones del árbol 2 mediante la evaluación de las distribuciones condicionales utilizando las funciones $h$ de las cópulas en el árbol 1.

   c) Determinar las cópulas en el árbol 2 a partir de las observaciones obtenidas en el paso b).

   d) Repetir los pasos b) y c) para el resto de los árboles.

En CVEDA y DVEDA, para determinar cada cópula bivariada se procede de la siguiente manera. Primeramente, se aplica una prueba de independencia. Se selecciona la cópula independencia si no hay evidencia suficiente en contra de la hipótesis de independencia a un nivel de significación dado. En otro caso, se estiman los parámetros de las cópulas candidatas y se selecciona la cópula $C_{\hat{\theta}}$ que minimice la distancia entre $C_{\hat{\theta}}$ y la cópula empírica. Como medida de distancia se utiliza el estadígrafo Cramér-von Mises $S_N$ definido en (1.10). La prueba de independencia utilizada se basa también en el estadígrafo $S_N$, sustituyendo $C_{\hat{\theta}}$ por $C_I$ en (1.10).

En este trabajo se utilizan las siguientes cópulas bivariadas: normal, t, Clayton, Clayton rotada, Gumbel y Gumbel rotada. Estas cópulas permiten modelar un amplio rango de estructuras de dependencia en las colas de las distribuciones: las cópulas Clayton y Clayton rotada permiten modelar dependencias fuertes en la cola inferior pero no en la superior; las cópulas Gumbel y Gumbel rotada permiten modelar dependencias fuertes en la cola superior pero no en la inferior; la cópula t permite modelar dependencias fuertes en ambas colas, mientras que la cópula normal en ninguna de las dos (sección 1.1.2 y apéndice A).

Los parámetros de las cópulas normal, Clayton, Clayton rotada, Gumbel y Gumbel rotada se estiman utilizando la inversión del coeficiente tau de Kendall (sección 1.1.2). En el caso de la cópula t, se sigue el procedimiento sugerido por Demarta y McNeil (2005), donde el coeficiente de correlación $\rho$ se estima como en la cópula normal, y los grados de libertad $\nu$ se estiman mediante máxima verosimilitud (sección 1.1.2) con $\rho$ constante. Para $\nu$ se considera un límite superior de 30, debido a que para este valor la cópula t se hace indistinguible de la cópula normal (Fantazzini, 2010).

### 2.2.2. Simulación

La simulación en los algoritmos CVEDA y DVEDA se divide en dos partes: la simulación de una muestra de variables uniformes con la estructura de dependencia representada por el vine y la transformación de esta muestra para obtener la nueva población.

La simulación de un C-vine o un D-vine con $n$ variables comienza con la generación de igual número de observaciones uniformes independientes $w_i \in (0,1)$. A partir de estas observaciones, se obtienen $u_1, u_2, \ldots, u_n$ con la estructura de dependencia representada por

el vine de acuerdo a

$$\begin{aligned} u_1 &= w_1, \\ u_2 &= F^{-1}_{2|1}(w_2|u_1), \\ &\vdots \\ u_n &= F^{-1}_{n|1,2,\ldots,n-1}(w_n|u_1,\ldots,u_{n-1}). \end{aligned}$$

El cálculo de la inversa de las distribuciones condicionales se realiza, de manera recursiva, utilizando las definiciones de las funciones $h$ y $h^{-1}$ (sección 1.1.2 y apéndice B). En cada paso, el número de variables en el condicionante decrece en uno.

Los individuos de la nueva población se obtienen a partir de la muestra de variables uniformes, $u_1, u_2, \ldots, u_n$, utilizando el método de inversión (Devroye, 1986).

## 2.3. Conclusiones del capítulo

Los vines son modelos que permiten representar diferentes tipos de dependencias y combinarlas adecuadamente. Estas características ofrecen nuevas vías para tratar con diferentes fuentes de complejidad que surgen en la optimización. Consecuentemente, los EDA basados en vines son más flexibles que sus predecesores basados en cópulas multivariadas ya que utilizan modelos capaces de describir una amplia variedad de patrones de dependencia.

# Capítulo 3

# Estudio empírico de CVEDA y DVEDA

En este capítulo se investiga empíricamente el comportamiento de los algoritmos CVEDA y DVEDA en un grupo de problemas de prueba con características conocidas. Los objetivos principales del capítulo son 1) evaluar el impacto de utilizar en los modelos diferentes tipos de cópulas y 2) valorar el efecto de aplicar diferentes estrategias de construcción de los vines: la construcción parcial o total y la selección de la estructura de los C-vines y D-vines de manera arbitraria o de acuerdo a la fuerza de las interacciones entre las variables.

## 3.1. Implementación de los algoritmos

Durante el desarrollo de la tesis se crearon dos paquetes para el ambiente estadístico R (R Development Core Team, 2010): `copulaedas` (González-Fernández y Soto, 2011a), que contiene implementaciones de EDA basados en cópulas, y `vines` (González-Fernández y Soto, 2011b), que provee funcionalidades relacionadas con el modelado de dependencias multivariadas utilizando vines. La investigación empírica llevada a cabo en esta tesis se realiza utilizando estos paquetes.

El paquete `vines` contiene definiciones de clases tipo S4 (Chambers, 2008) para vines (C-vines y D-vines) y métodos para la estimación, pruebas de bondad de ajuste, evaluación de la densidad, evaluación de la función de distribución y simulación de estos modelos. Este paquete depende de los paquetes: `copula` (Kojadinovic y Yan, 2010; Yan, 2007), `ADGofTest` (Gil, 2009), `cubature` (Johnson y Narasimhan, 2009) y `TSP` (Hahsler y Hornik, 2007).

El paquete `copulaedas` contiene implementaciones de varias clases de EDA basados

en la teoría de cópulas. En este paquete los EDA se implementan utilizando clases tipo S4 con funciones genéricas para los pasos del algoritmo: inicialización, selección, estimación, simulación, remplazo de la población, optimización local, terminación y reporte de progreso. El paquete también incluye la definición de un grupo de problemas de prueba y funciones de utilidad para estudiar el comportamiento de los EDA. Este paquete depende de los paquetes: `copula` (Kojadinovic y Yan, 2010; Yan, 2007) y `vines` (González-Fernández y Soto, 2011b).

## 3.2. Diseño experimental

Las funciones Sphere, Griewank, Ackley y Summation Cancellation, estudiadas en (Arderí, 2007; Bengoetxea et al., 2002), se consideran como problemas de prueba en 10 dimensiones. A continuación se muestra la definición de estas funciones para un vector $\mathbf{x} = (x_1, \ldots, x_n)$.

$$f_{\text{Sphere}}(\mathbf{x}) = \sum_{i=1}^{n} x_i^2$$

$$f_{\text{Griewank}}(\mathbf{x}) = 1 + \sum_{i=1}^{n} \frac{x_i^2}{4000} - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right)$$

$$f_{\text{Ackley}}(\mathbf{x}) = -20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x^2}\right) - \exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right) + 20 + \exp(1)$$

$$f_{\text{Summation Cancellation}}(\mathbf{x}) = \frac{1}{10^{-5} + \sum_{i=1}^{n} |y_i|}, y_1 = x_1, y_i = y_{i-1} + x_i$$

Sphere, Griewank y Ackley son problemas de minimización, mientras que Summation Cancellation es un problema de maximización. Sphere, Griewank y Ackley alcanzan su óptimo global en $\mathbf{x} = (0, \ldots, 0)$ con evaluación cero. Summation Cancellation alcanza su óptimo global en $\mathbf{x} = (0, \ldots, 0)$ con evaluación $10^5$.

Este grupo de funciones presenta características diferentes en cuanto a la modalidad y las interacciones entre las variables. Sphere es una función unimodal que no presenta

dependencias entre las variables. Griewank y Ackley son funciones con muchos óptimos locales. Summation Cancellation presenta fuertes interacciones lineales multivariadas y el óptimo global se encuentra ubicado en un pico muy estrecho. Estas características se ilustran en la figura 3.1.

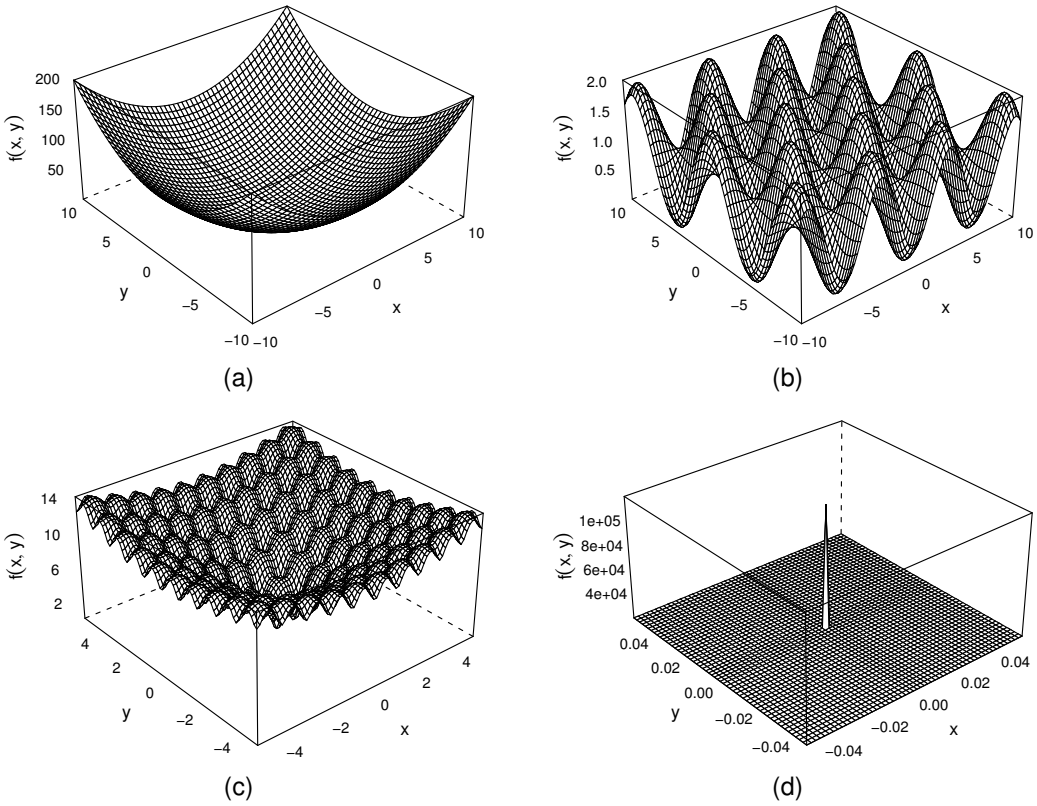


**Figura 3.1:** Gráficos de las funciones Sphere (a), Griewank (b), Ackley (c) y Summation Cancellation (d) en dos dimensiones.

CVEDA y DVEDA utilizan un nivel de significación al 10 % en la prueba de independencia para la selección de las cópulas bivariadas en el vine. Si se rechaza la hipótesis nula de independencia, se selecciona entre las cópulas candidatas normal, t, Clayton, Clayton rotada, Gumbel y Gumbel rotada.

Todos los algoritmos modelan las distribuciones marginales con la distribución normal. La condición de parada es alcanzar el óptimo global de la función objetivo, con una precisión menor que $10^{-6}$ o realizar un máximo de 500000 evaluaciones de la función objetivo. Los valores de las variables en la población inicial se generan en los intervalos $[-600, 600]$ para

Sphere y Griewank, $[-30, 30]$ para Ackley y $[-0.16, 0.16]$ para Summation Cancellation.

Los tamaños de población reportados en las tablas de las siguientes secciones corresponden al *tamaño de población crítico*: el tamaño de población mínimo requerido por el algoritmo para encontrar el óptimo global de la función objetivo en 30 de 30 ejecuciones independientes o un máximo de 2000 individuos si el algoritmo falló para poblaciones más pequeñas. El tamaño de población crítico se determina empíricamente utilizando el procedimiento de búsqueda dicotómica propuesto por Pelikan (2005). El número de evaluaciones de la función objetivo y la mejor evaluación de la misma corresponden al promedio y la desviación estándar de los resultados obtenidos en las 30 ejecuciones con el tamaño de población crítico.

## 3.3. Resultados y discusión

En esta sección se presentan los resultados obtenidos por CVEDA y DVEDA en las funciones de prueba y se discute sobre el comportamiento de los mismos.

### 3.3.1. Uso de diferentes tipos de cópulas

El objetivo de los experimentos presentados en esta sección es estudiar el efecto de utilizar diferentes tipos de cópulas en los vines en los algoritmos CVEDA y DVEDA.

Los resultados experimentales obtenidos con los algoritmos UMDA, CVEDA, DVEDA y GCEDA en las funciones Sphere, Griewank, Ackley y Summation Cancellation se muestran en las tablas 3.1, 3.2, 3.3 y 3.4, respectivamente. Los algoritmos CVEDA y DVEDA estiman los vines completamente (9 árboles) y utilizan las heurísticas para la selección de la estructura de los C-vines y D-vines que representan las dependencias más fuertes en el primer árbol (sección 2.2.1). Los resultados obtenidos se describen a continuación.

**En los problemas donde el algoritmo UMDA exhibe un buen comportamiento, la introducción de dependencias por los algoritmos CVEDA, DVEDA y GCEDA puede resultar en un deterioro del comportamiento de estos últimos.** Las funciones Sphere, Griewank y Ackley son problemas que pueden ser optimizados por el algoritmo UMDA. En estas funciones con débiles interacciones entre las variables, la información de los marginales parece ser suficiente para guiar la búsqueda hacia el óptimo global de la función. Los algoritmos

**Tabla 3.1:** Resultados de los algoritmos en la función Sphere.

| Algoritmo | Población | Éxito | Evaluaciones | Mejor evaluación |
|---|---|---|---|---|
| UMDA | 86 | $30/30$ | $3996.1 \pm 89.5$ | $6.9\text{E}-07 \pm 1.9\text{E}-07$ |
| $\text{CVEDA}_{9,\text{ greedy}}$ | 188 | $30/30$ | $8033.8 \pm 170.5$ | $6.8\text{E}-07 \pm 2.1\text{E}-07$ |
| $\text{DVEDA}_{9,\text{ greedy}}$ | 207 | $30/30$ | $8818.2 \pm 192.9$ | $7.0\text{E}-07 \pm 1.8\text{E}-07$ |
| GCEDA | 325 | $30/30$ | $13769.1 \pm 248.5$ | $6.6\text{E}-07 \pm 1.6\text{E}-07$ |

**Tabla 3.2:** Resultados de los algoritmos en la función Griewank.

| Algoritmo | Población | Éxito | Evaluaciones | Mejor evaluación |
|---|---|---|---|---|
| UMDA | 113 | $30/30$ | $5179.1 \pm 210.0$ | $7.2\text{E}-07 \pm 1.7\text{E}-07$ |
| $\text{CVEDA}_{9,\text{ greedy}}$ | 213 | $30/30$ | $9151.9 \pm 452.6$ | $6.5\text{E}-07 \pm 1.8\text{E}-07$ |
| $\text{DVEDA}_{9,\text{ greedy}}$ | 225 | $30/30$ | $9630.0 \pm 309.2$ | $6.9\text{E}-07 \pm 1.5\text{E}-07$ |
| GCEDA | 304 | $30/30$ | $12798.4 \pm 351.1$ | $6.6\text{E}-07 \pm 1.7\text{E}-07$ |

CVEDA, DVEDA y GCEDA tienen que estimar los parámetros de las cópulas, además de los parámetros de las distribuciones marginales. Por esta razón, estos algoritmos requieren poblaciones de mayor tamaño para estimar de manera fiable los parámetros de sus modelos y esto se refleja en la realización de un mayor número de evaluaciones de la función objetivo.

**Los algoritmos CVEDA y DVEDA pueden comportarse de manera robusta tanto en problemas con débiles o con fuertes interacciones entre las variables.** El algoritmo UMDA asume la independencia entre las variables. Por su parte, GCEDA asume una estructura de dependencia normal. Sin embargo, los algoritmos CVEDA y DVEDA no necesitan asumir una estructura de dependencia entre las variables. Los procedimientos de estimación de estos algoritmos asignan la cópula independencia si no existe evidencia suficiente de dependencia entre las variables y en otro caso, se selecciona entre el grupo de cópulas candidatas, la cópula que ajuste los datos apropiadamente.

**En los problemas fáciles para el algoritmo UMDA (Sphere, Griewank y Ackley) el algoritmo CVEDA tiende a comportarse mejor que DVEDA.** Esta observación está rela-

**Tabla 3.3:** Resultados de los algoritmos en la función Ackley.

| Algoritmo | Población | Éxito | Evaluaciones | Mejor evaluación |
|---|---|---|---|---|
| UMDA | 88 | $30/30$ | $5426.6 \pm 127.2$ | $8.2\text{E}-07 \pm 1.0\text{E}-07$ |
| $\text{CVEDA}_{9,\text{ greedy}}$ | 213 | $30/30$ | $11984.8 \pm 184.9$ | $7.9\text{E}-07 \pm 1.5\text{E}-07$ |
| $\text{DVEDA}_{9,\text{ greedy}}$ | 213 | $30/30$ | $11920.9 \pm 197.6$ | $7.9\text{E}-07 \pm 1.3\text{E}-07$ |
| GCEDA | 325 | $30/30$ | $18178.3 \pm 207.8$ | $8.0\text{E}-07 \pm 1.5\text{E}-07$ |

**Tabla 3.4:** Resultados de los algoritmos en la función Summation Cancellation.

| Algoritmo | Población | Éxito | Evaluaciones | Mejor evaluación |
|---|---|---|---|---|
| UMDA | 2000 | $0/30$ | $500000.0 \pm 0.0$ | $6.9\text{E}+02 \pm 5.0\text{E}+02$ |
| $\text{CVEDA}_{9,\text{ greedy}}$ | 625 | $30/30$ | $84958.3 \pm 786.0$ | $1.0\text{E}+05 \pm 1.1\text{E}-07$ |
| $\text{CVEDA}_{\text{N, 9, greedy}}$ | 319 | $30/30$ | $43373.3 \pm 539.5$ | $1.0\text{E}+05 \pm 1.3\text{E}-07$ |
| $\text{DVEDA}_{9,\text{ greedy}}$ | 1400 | $30/30$ | $161840.0 \pm 1352.5$ | $1.0\text{E}+05 \pm 9.3\text{E}-08$ |
| $\text{DVEDA}_{\text{N, 9, greedy}}$ | 488 | $30/30$ | $58494.9 \pm 457.3$ | $1.0\text{E}+05 \pm 1.3\text{E}-07$ |
| GCEDA | 325 | $30/30$ | $38848.3 \pm 327.6$ | $1.0\text{E}+05 \pm 1.2\text{E}-07$ |

cionada con el primer resultado. El modelo utilizado por DVEDA permite seleccionar más libremente las dependencias bivariadas representadas explícitamente, mientras que el modelo utilizado por CVEDA tiene una estructura más restrictiva. Estas características posibilitan que DVEDA encuentre relaciones de dependencia que CVEDA no puede encontrar, por lo que los modelos estimados por DVEDA tendrán un mayor número de cópulas que representan dependencias. Para estimar correctamente los parámetros de estas cópulas, se necesitan poblaciones de mayor tamaño y debido a esto DVEDA realiza un mayor número de evaluaciones.

**En la función Summation Cancellation, un problema difícil para UMDA, CVEDA se comporta mucho mejor que DVEDA.** La información sobre las fuertes interacciones lineales entre las variables de la función Summation Cancellation es esencial para encontrar el óptimo global. Esta situación se refleja en los malos resultados obtenidos por el algoritmo UMDA. Por otra parte, el algoritmo CVEDA obtiene mejores resultados que DVEDA. La ra-

zón de este comportamiento parece estar basada en que el modelo utilizado por CVEDA resulta más apropiado, a pesar de ser más restrictivo para la selección de las dependencias bivariadas modeladas explícitamente. La función Summation Cancellation alcanza su óptimo global cuando la sumatoria en el denominador de la fracción en su definición toma valor 0. El $i$-ésimo término de esta sumatoria es el resultado de la suma de los valores de las $i$ primeras variables de la función, por lo que los valores de las primeras variables tienen una mayor influencia en el valor de la sumatoria. Las poblaciones seleccionadas reflejan estas características presentando asociaciones más fuertes entre cada una de las primeras variables y las siguientes. Un C-vine permite representar apropiadamente esta estructura de dependencia debido a que es posible encontrar una variable que gobierne las interacciones de la muestra.

**Considerar otras cópulas además de la normal en los algoritmos CVEDA y DVEDA deteriora el comportamiento de los mismos en la función Summation Cancellation.** En los resultados mostrados en la tabla 3.4 resalta el mal comportamiento de los algoritmos CVEDA y DVEDA que utilizan todas las cópulas candidatas ($\text{CVEDA}_{9,\text{ greedy}}$ y $\text{DVEDA}_{9,\text{ greedy}}$) comparado con el comportamiento de GCEDA. Las interacciones lineales entre las variables de la función Summation Cancellation se corresponden con una estructura de dependencia normal, por lo que el algoritmo GCEDA tiene muy buen comportamiento. El deterioro de los algoritmos basados en vines parece causado por la selección de otras cópulas en lugar de la normal. Berg (2009) y Brechmann (2010) han señalado deficiencias en las pruebas de bondad de ajuste para cópulas cuando la dependencia entre las variables no es fuerte. Para buscar evidencias empíricas a favor de esta hipótesis, se realizaron experimentos incluyendo solamente la cópula normal como cópula candidata en los algoritmos CVEDA y DVEDA ($\text{CVEDA}_{\text{N, 9, greedy}}$ y $\text{CVEDA}_{\text{N, 9, greedy}}$). En este caso, el comportamiento de los algoritmos basados en vines se acerca al comportamiento de GCEDA, especialmente CVEDA.

### 3.3.2. Construcción parcial o total de los vines

En los experimentos realizados en la sección anterior los algoritmos CVEDA y DVEDA estiman en cada generación un vine con nueve árboles a partir de la población seleccionada. En este caso, el procedimiento de estimación debe seleccionar 45 cópulas bivariadas, lo cual implica realizar igual número de pruebas de independencia y, si se rechaza la hipótesis nula, una prueba de bondad de ajuste por cada cópula candidata. Debido a esta situación, se

hace necesario utilizar métodos que reduzcan el costo computacional de la estimación sin sacrificar las propiedades de los modelos.

El objetivo de los experimentos presentados en esta sección es comparar varios métodos para la construcción parcial de los vines que resultan en una reducción del costo computacional del procedimiento de estimación. Los primeros dos métodos construyen el vine parcialmente hasta tres o seis árboles (CVEDA$_{3,\text{ greedy}}$, CVEDA$_{6,\text{ greedy}}$, DVEDA$_{3,\text{ greedy}}$ y DVEDA$_{6,\text{ greedy}}$). Los métodos restantes determinan el número de árboles del vine basado en los criterios de información AIC o BIC (CVEDA$_{\text{AIC, greedy}}$, CVEDA$_{\text{BIC, greedy}}$, DVEDA$_{\text{AIC, greedy}}$ y DVEDA$_{\text{BIC, greedy}}$). Los experimentos realizados con las funciones Sphere y Summation Cancellation se muestran en las tablas 3.5 y 3.6. Los resultados obtenidos se describen a continuación.

**Tabla 3.5:** Resultados de los algoritmos CVEDA y DVEDA con diferentes métodos para la construcción parcial de los vines en la función Sphere.

| Algoritmo | Población | Éxito | Evaluaciones | Mejor evaluación |
|---|---|---|---|---|
| CVEDA$_{3,\text{ greedy}}$ | 175 | $30/30$ | $7536.6 \pm 151.9$ | $6.5\text{E}-07 \pm 2.2\text{E}-07$ |
| CVEDA$_{6,\text{ greedy}}$ | 191 | $30/30$ | $8174.8 \pm 176.6$ | $6.7\text{E}-07 \pm 1.9\text{E}-07$ |
| CVEDA$_{\text{AIC, greedy}}$ | 163 | $30/30$ | $7106.8 \pm 139.3$ | $6.6\text{E}-07 \pm 2.0\text{E}-07$ |
| CVEDA$_{\text{BIC, greedy}}$ | 113 | $30/30$ | $5017.2 \pm 134.6$ | $6.8\text{E}-07 \pm 1.6\text{E}-07$ |
| DVEDA$_{3,\text{ greedy}}$ | 191 | $30/30$ | $8149.3 \pm 161.2$ | $6.5\text{E}-07 \pm 1.8\text{E}-07$ |
| DVEDA$_{6,\text{ greedy}}$ | 207 | $30/30$ | $8818.2 \pm 128.6$ | $6.9\text{E}-07 \pm 1.8\text{E}-07$ |
| DVEDA$_{\text{AIC, greedy}}$ | 163 | $30/30$ | $6992.7 \pm 144.2$ | $6.5\text{E}-07 \pm 1.9\text{E}-07$ |
| DVEDA$_{\text{BIC, greedy}}$ | 138 | $30/30$ | $6026.0 \pm 127.2$ | $7.0\text{E}-07 \pm 2.2\text{E}-07$ |

**Los algoritmos CVEDA y DVEDA con la construcción de los vines hasta un número arbitrario de árboles no son algoritmos robustos.** La selección del número de árboles en el vine es dependiente del problema. El uso de un menor número de árboles en la función Sphere mejora el comportamiento de los algoritmos, mientras que en la función Summation Cancellation lo deteriora. Esta situación hace que los algoritmos CVEDA y DVEDA que utilizan este método de construcción parcial del vine no puedan comportarse de manera robusta. Por ejemplo, los algoritmos que construyen tres árboles tienen un buen comportamiento en la función Sphere pero no encuentran el óptimo de la función Summation Cancellation.

**Tabla 3.6:** Resultados de los algoritmos CVEDA y DVEDA con diferentes métodos para la construcción parcial de los vines en la función Summation Cancellation.

| Algoritmo | Población | Éxito | Evaluaciones | Mejor evaluación |
|---|---|---|---|---|
| $\text{CVEDA}_{\text{3, greedy}}$ | 2000 | $0/30$ | $500000.0 \pm 0.0$ | $2.6\text{E}+03 \pm 3.4\text{E}+03$ |
| $\text{CVEDA}_{\text{6, greedy}}$ | 2000 | $0/30$ | $500000.0 \pm 0.0$ | $3.7\text{E}+04 \pm 3.2\text{E}+04$ |
| $\text{CVEDA}_{\text{AIC, greedy}}$ | 650 | $30/30$ | $90003.3 \pm 1262.8$ | $1.0\text{E}+05 \pm 1.2\text{E}-07$ |
| $\text{CVEDA}_{\text{BIC, greedy}}$ | 800 | $30/30$ | $108506.6 \pm 1647.3$ | $1.0\text{E}+05 \pm 9.8\text{E}-08$ |
| $\text{DVEDA}_{\text{3, greedy}}$ | 2000 | $0/30$ | $500000.0 \pm 0.0$ | $8.4\text{E}+04 \pm 2.5\text{E}+04$ |
| $\text{DVEDA}_{\text{6, greedy}}$ | 2000 | $10/30$ | $412133.3 \pm 128711.1$ | $9.9\text{E}+04 \pm 1.7\text{E}+02$ |
| $\text{DVEDA}_{\text{AIC, greedy}}$ | 1300 | $30/30$ | $152750.0 \pm 1404.1$ | $1.0\text{E}+05 \pm 1.0\text{E}-07$ |
| $\text{DVEDA}_{\text{BIC, greedy}}$ | 2000 | $26/30$ | $285000.0 \pm 100221.0$ | $9.9\text{E}+04 \pm 6.9\text{E}-03$ |

**Los métodos basados en los criterios de información muestran un mejor comportamiento que la construcción del vine hasta un número arbitrario de árboles.** En ambas funciones los algoritmos basados en criterios de información se comportan mejor que todos los algoritmos que utilizan una selección arbitraria del número de árboles. Esto sucede así incluso entre algoritmos que utilizan diferentes modelos.

**Entre los métodos basados en los criterios de información, el método basado en AIC muestra el mejor comportamiento.** En la función Sphere los algoritmos que utilizan BIC tienen un mejor comportamiento que los que utilizan AIC. En la función Summation Cancellation se presenta la situación contraria, resaltando la incapacidad del algoritmo $\text{DVEDA}_{\text{BIC, greedy}}$ para encontrar el óptimo de la función en las 30 ejecuciones. Ambas situaciones se deben a que, debido a la diferencia en el término de penalización de los parámetros, el criterio de selección basado en BIC prefiere modelos con un menor número de cópulas que la selección basada en AIC. Esta característica resulta una ventaja en la función Sphere pero compromete la convergencia del algoritmo en la función Summation Cancellation. Los algoritmos que utilizan AIC tienen un buen comportamiento en ambas funciones y el uso de este criterio no compromete la convergencia del algoritmo. En la construcción parcial utilizando AIC en la función Sphere el número de árboles de los vines nunca fue mayor que tres y cuatro en CVEDA y DVEDA, respectivamente. En la función Summation Cancellation los dos algoritmos llegaron a construir los nueve árboles.

### 3.3.3. Selección de la estructura de los C-vines y D-vines

Los experimentos presentados en esta sección tienen como objetivo valorar el efecto de la selección de la estructura de los C-vines y D-vines estimados en CVEDA y DVEDA, respectivamente.

Los experimentos se realizan en las funciones Sphere y Summation Cancellation. En cada caso se ejecuta el algoritmo CVEDA y DVEDA con construcción parcial del vine basada en criterios de información que mostró el mejor comportamiento en los resultados de la sección anterior. La selección de la estructura se realiza de manera aleatoria ($\text{CVEDA}_{\text{BIC, random}}$, $\text{DVEDA}_{\text{BIC, random}}$, $\text{CVEDA}_{\text{AIC, random}}$ y $\text{DVEDA}_{\text{AIC, random}}$) o utilizando las heurísticas para representar las dependencias más fuertes en el primer árbol ($\text{CVEDA}_{\text{BIC, greedy}}$, $\text{DVEDA}_{\text{BIC, greedy}}$, $\text{CVEDA}_{\text{AIC, greedy}}$ y $\text{DVEDA}_{\text{AIC, greedy}}$). Los resultados correspondientes al primer grupo de algoritmos se muestran en las tablas 3.5 y 3.6 presentadas en la sección anterior y los del segundo grupo en las tablas 3.7 y 3.8. El principal resultado obtenido se describe a continuación.

**Tabla 3.7:** Resultados de los algoritmos CVEDA y DVEDA con selección aleatoria de la estructura de los vines en la función Sphere.

| Algoritmo | Población | Éxito | Evaluaciones | Mejor evaluación |
|---|---|---|---|---|
| $\text{CVEDA}_{\text{BIC, random}}$ | 100 | 30/30 | $4523.3 \pm 100.6$ | $6.9\text{E}-07 \pm 1.8\text{E}-07$ |
| $\text{DVEDA}_{\text{BIC, random}}$ | 100 | 30/30 | $4526.6 \pm 114.2$ | $6.6\text{E}-07 \pm 1.6\text{E}-07$ |

**Tabla 3.8:** Resultados de los algoritmos CVEDA y DVEDA con selección aleatoria de la estructura de los vines en la función Summation Cancellation.

| Algoritmo | Población | Éxito | Evaluaciones | Mejor evaluación |
|---|---|---|---|---|
| $\text{CVEDA}_{\text{AIC, random}}$ | 775 | 30/30 | $110360.0 \pm 2020.9$ | $1.0\text{E}+05 \pm 1.1\text{E}-07$ |
| $\text{DVEDA}_{\text{AIC, random}}$ | 1500 | 30/30 | $255900.0 \pm 5205.7$ | $1.0\text{E}+05 \pm 1.2\text{E}-07$ |

**En los algoritmos CVEDA y DVEDA, es importante utilizar heurísticas para representar las dependencias más fuertes en el primer árbol de los vines.** En la función Sphere los algoritmos que utilizan una estructura aleatoria de los C-vine y D-vines realizan un menor número de evaluaciones. Al considerar una estructura arbitraria el número de cópulas distintas de la independencia en los vines es menor que al utilizar una selección de las dependencias más fuertes. Esta situación favorece la construcción de modelos más cercanos al UMDA, que tiene el mejor comportamiento en esta función. En Summation Cancellation se refleja la situación contraria: utilizar una estructura arbitraria provoca que dependencias importantes para realizar una búsqueda más eficiente no sean capturadas. Esta situación deteriora el comportamiento de los algoritmos. Estos resultados muestran que la selección de las dependencias modeladas explícitamente en los vines es importante para contar con algoritmos basados en vines que exhiban un comportamiento robusto.

## 3.4. Conclusiones del capítulo

Los resultados obtenidos en las funciones de prueba muestran que CVEDA y DVEDA son algoritmos robustos que se comportan relativamente bien, tanto en problemas con débiles interacciones (Sphere, Ackley y Griewank) como con fuertes interacciones (Summation Cancellation) entre las variables. El estudio realizado sobre algunas características intrínsecas de estos algoritmos permite afirmar que el uso de un modelo C-vine o D-vine influye en el comportamiento del algoritmo, así como también es importante la selección de la estructura de los modelos de acuerdo a la fuerza de las interacciones de las variables. Además, la construcción parcial de los vines basada en AIC es una alternativa a la construcción total del vine que permite disminuir el costo computacional de la estimación de los modelos sin comprometer la convergencia de los algoritmos.

# Capítulo 4

# Aplicación de CVEDA y DVEDA en el problema de acoplamiento molecular

En este capítulo se estudia la aplicación de CVEDA y DVEDA en el problema de acoplamiento molecular con el objetivo de comprobar si las características de estos algoritmos observadas en las funciones de prueba se manifiestan también en este problema real.

El acoplamiento molecular es un procedimiento que tiene como objetivo predecir la geometría de la unión entre dos moléculas. Una de estas moléculas, llamada *receptor*, es una proteína; mientras que la otra, llamada *ligando*, es una molécula pequeña que se une al sitio activo de la proteína. Este procedimiento tiene un papel muy importante en el diseño racional de fármacos. En la actualidad, este problema constituye un área activa de investigación ya que los algoritmos para explorar las posibles conformaciones y las funciones para medir la calidad de las mismas presentan limitaciones significativas (Warren et al., 2006).

En las simulaciones desarrolladas en esta sección el receptor se trata como un cuerpo rígido, mientras que el ligando es flexible. Cada solución representa solamente el ligando.

El ligando se representa como un vector con valores reales que determina su posición, orientación y ángulos de torsión. Las tres primeras componentes del vector representan la posición del ligando en el espacio tridimensional. Los valores de estas componentes están limitados por las dimensiones de un ortoedro colocado sobre el sitio activo del receptor. Las dimensiones del ortoedro se calculan añadiendo cinco angstroms (Å) a los valores mínimos y máximos, en cada dimensión, de las coordenadas del ligando en su conformación experimental. Las siguientes tres componentes del vector representan la orientación del ligando y toman valores en los intervalos $[0, 2\pi]$, $[-\pi/2, \pi/2]$ y $[-\pi, \pi]$, respectivamente. Finalmente,

cada ángulo de torsión flexible del ligando se representa con una variable adicional que toma valores en $[-\pi, \pi]$.

Se utiliza la función objetivo semiempírica descrita por Huey et al. (2007) e implementada en AutoDock 4.2. Esta función determina la energía del acoplamiento entre ligando y el receptor, la cual se calcula mediante la suma de las interacciones entre los átomos del ligando y del receptor (energía intermolecular) y las interacciones entre los átomos del ligando (energía intramolecular).

La estructura tridimensional propuesta por el algoritmo de acoplamiento se compara con la estructura tridimensional conocida experimentalmente para evaluar su calidad. Esta comparación se realiza mediante el cálculo de la desviación cuadrática media,

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(dx_i^2 + dy_i^2 + dz_i^2)}{n}},$$

donde $n$ denota el número de átomos del ligando y $dx_i$, $dy_i$, $dz_i$ denotan la diferencia entre las coordenadas experimentales del átomo $i$ del ligando y las obtenidas mediante el procedimiento de acoplamiento molecular. Si se obtienen valores de RMSD menores o cercanos a 2 Å, se considera que el acoplamiento fue exitoso. Por otra parte, valores cercanos a 3 Å indican un acoplamiento parcial.

## 4.1. Diseño experimental

En la tabla 4.1 se presentan las características de los cuatro compuestos utilizados en el estudio del comportamiento de los algoritmos. La información sobre los compuestos fue tomada de PDB (Protein Data Bank) (Berman et al., 2000).

Se realiza una comparación entre UMDA, GCEDA y los dos algoritmos basados en vines propuestos en la tesis. CVEDA y DVEDA utilizan un nivel de significación al 1 % en la prueba de independencia para la selección de las cópulas bivariadas en el vine. Si se rechaza la hipótesis nula de independencia, se ajusta una cópula normal. Ambos algoritmos utilizan la construcción parcial de los vines basada en AIC que mostró un buen comportamiento en los experimentos presentados en el capítulo anterior.

La capacidad de los algoritmos basados en cópulas para tratar con diferentes distribuciones marginales se utiliza para satisfacer las restricciones del dominio de las variables del problema. En estos experimentos, todos los algoritmos utilizan distribuciones marginales

**Tabla 4.1:** Compuestos utilizados en el problema de acoplamiento molecular. Los hidrógenos apolares no se incluyen en el número de átomos.

| PDB | Número de átomos | Número de torsiones | Dimensiones del ortoedro (Å) |
|---|---|---|---|
| 1adb | 56 | 15 | $28 \times 20 \times 22$ |
| 1bmm | 43 | 10 | $17 \times 19 \times 22$ |
| 1cjw | 71 | 26 | $26 \times 22 \times 30$ |
| 2z5u | 73 | 20 | $28 \times 32 \times 24$ |

normal truncada (Johnson et al., 1994). Esta distribución corresponde a una variable con distribución normal acotada en un intervalo $[a,b]$. La función de densidad de la distribución normal truncada con media $\mu$ y desviación estándar $\sigma \geq 0$ está dada por

$$f(x;\mu,\sigma,a,b) = \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$$

y la función de distribución por

$$F(x;\mu,\sigma,a,b) = \frac{\Phi(\frac{x-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})},$$

donde $\Phi$ y $\phi$ denotan las funciones de distribución y de densidad de la distribución normal estándar, respectivamente.

Como no se conoce el óptimo de la función objetivo, los algoritmos utilizan como condición de parada que la población pierda diversidad. Los algoritmos se detienen si la desviación estándar de los valores de energía de la población es menor que 0.01.

Para lograr una comparación satisfactoria de los algoritmos, se realizaron 30 ejecuciones independientes con diferentes tamaños de población entre 200 y 2000 con un incremento de 200. Se seleccionó el tamaño de población para el cual se obtuvieron los menores valores de la media de la energía con el menor número de evaluaciones de la función objetivo. Los resultados reportados en la próxima sección corresponden a las 30 ejecuciones con el tamaño de población seleccionado.

## 4.2. Resultados y discusión

En esta sección se reporta el comportamiento de los algoritmos UMDA, CVEDA, DVEDA y GCEDA en el problema de acoplamiento molecular. Los resultados de UMDA y GCEDA corresponden a los reportados por Milanés y Álvarez (2011).

En la tabla 4.2 se muestran los resultados de los algoritmos en cuanto a la media y la desviación estándar de la menor energía obtenida en una ejecución, el número de evaluaciones de la función objetivo y los valores de RMSD correspondientes a las soluciones con la menor energía. La figura 4.1 ilustra la relación entre la menor energía alcanzada y el número de evaluaciones de la función objetivo durante la ejecución de los algoritmos.

**Tabla 4.2:** Resultados de los algoritmos en el problema de acoplamiento molecular.

| PDB | Algoritmo | Población | Evaluaciones | Mejor energía | RMSD |
|---|---|---|---|---|---|
| 1adb | UMDA | 1800 | $157933.3 \pm 11286.0$ | $-16.55 \pm 2.09$ | $3.17 \pm 0.76$ |
| | CVEDA | 1400 | $114258.1 \pm 17585.9$ | $-17.01 \pm 2.18$ | $2.63 \pm 0.98$ |
| | DVEDA | 1400 | $112133.3 \pm 13920.2$ | $-17.61 \pm 1.79$ | $2.50 \pm 1.03$ |
| | GCEDA | 1400 | $108200.0 \pm 20162.1$ | $-18.69 \pm 0.84$ | $1.44 \pm 0.81$ |
| 1bmm | UMDA | 600 | $48633.3 \pm 6099.6$ | $-9.42 \pm 1.37$ | $4.58 \pm 0.40$ |
| | CVEDA | 800 | $65766.6 \pm 10500.1$ | $-9.11 \pm 1.39$ | $4.88 \pm 0.49$ |
| | DVEDA | 800 | $62266.6 \pm 9776.5$ | $-9.41 \pm 1.33$ | $4.77 \pm 0.71$ |
| | GCEDA | 1000 | $72166.6 \pm 18632.7$ | $-8.31 \pm 1.10$ | $4.99 \pm 0.90$ |
| 1cjw | UMDA | 1200 | $180000.0 \pm 13341.6$ | $-12.19 \pm 1.46$ | $6.24 \pm 1.04$ |
| | CVEDA | 1200 | $180233.3 \pm 13103.5$ | $-12.85 \pm 1.69$ | $6.35 \pm 1.33$ |
| | DVEDA | 1200 | $165400.0 \pm 18872.7$ | $-14.40 \pm 2.40$ | $5.20 \pm 1.34$ |
| | GCEDA | 1600 | $201033.3 \pm 37696.1$ | $-15.55 \pm 2.02$ | $4.97 \pm 1.26$ |
| 2z5u | UMDA | 1400 | $171900.0 \pm 11442.0$ | $-29.14 \pm 1.97$ | $0.61 \pm 0.18$ |
| | CVEDA | 1400 | $157600.0 \pm 11391.4$ | $-29.58 \pm 1.23$ | $0.58 \pm 0.12$ |
| | DVEDA | 1200 | $125266.6 \pm 11965.2$ | $-30.16 \pm 1.28$ | $0.52 \pm 0.12$ |
| | GCEDA | 1600 | $140966.6 \pm 17835.4$ | $-29.43 \pm 0.56$ | $0.51 \pm 0.05$ |

De manera general, se observa que UMDA solamente obtiene mejores resultados que el resto de algoritmos en el compuesto 1bmm, y que DVEDA y GCEDA tienen un comportamiento ligeramente superior que CVEDA.

Los algoritmos basados en vines tienen un comportamiento superior al algoritmo con el

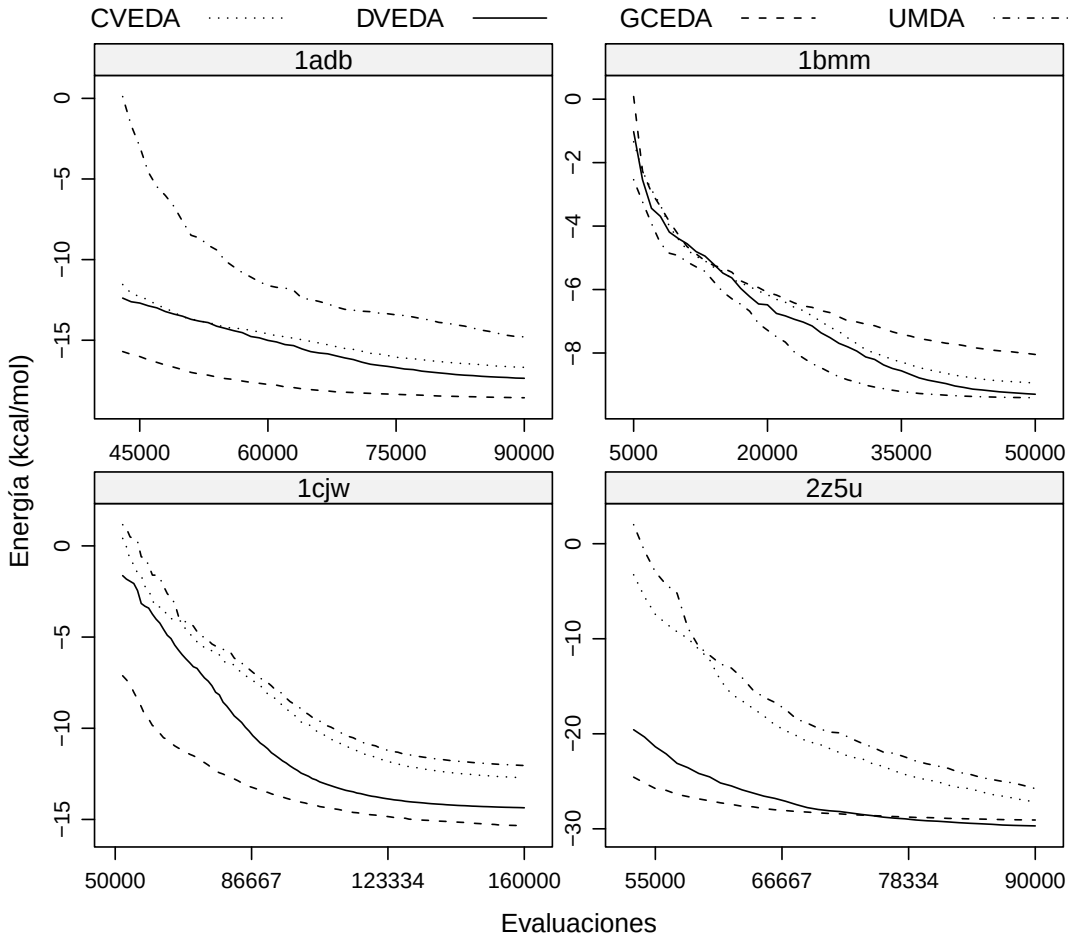


**Figura 4.1:** Media de la mejor energía alcanzada por los algoritmos en el problema de acoplamiento molecular, en función del número de evaluaciones.

peor desempeño en cada compuesto. En la figura 4.1 se ilustra claramente el comportamiento robusto de los algoritmos CVEDA y DVEDA, reportándose durante toda la ejecución resultados que se encuentran entre los obtenidos por el algoritmo UMDA, que asume la independencia de las variables, y el algoritmo GCEDA, que asume una estructura de dependencia normal multivariada. CVEDA presenta dificultades en la optimización de los compuestos 1cjw y 2z5u, donde DVEDA muestra un comportamiento notablemente superior. En

el compuesto 2z5u DVEDA es el algoritmo que obtiene los mejores resultados.

En la figura 4.2 se comparan los algoritmos CVEDA y DVEDA en los compuestos 1bmm y 2z5u en cuanto a la proporción entre el número de cópulas normal y el total de aristas en el vine estimado en cada generación. En el compuesto 1bmm los vines tienen un total de 15 árboles y 120 aristas, mientras que en el 2z5u tienen 25 árboles y 325 aristas. En ambos casos la proporción aumenta con el número de generaciones. DVEDA ajusta un número mayor de cópulas normal que CVEDA, lo cual es particularmente significativo en el compuesto 2z5u. Aunque en la construcción del C-vine, CVEDA intenta representar explícitamente las dependencias más fuertes, las restricciones del modelo evitan que algunas de ellas sean capturadas. Un C-vine resulta apropiado cuando una variable gobierna las interacciones entre las variables.

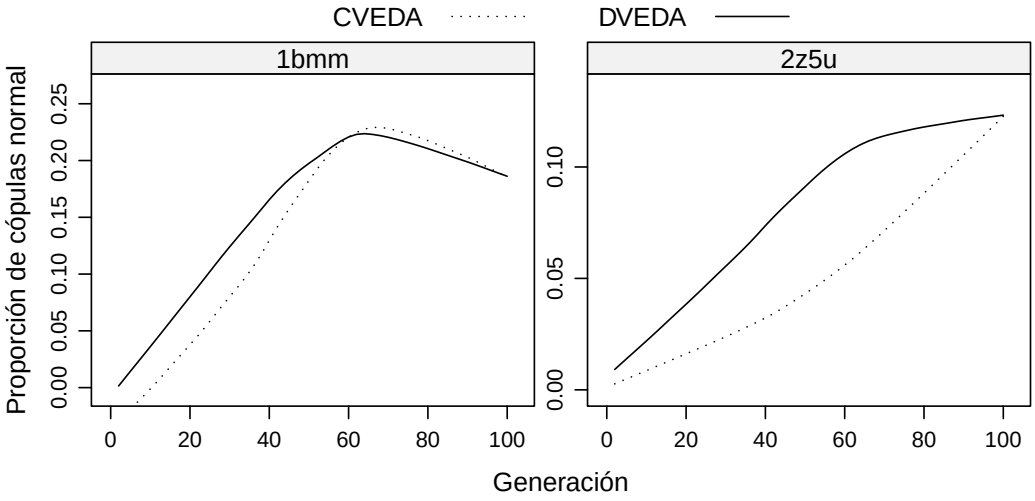


**Figura 4.2:** Comparación de los algoritmos CVEDA y DVEDA en los complejos 1bmm y 2z5u en cuanto a la media de la proporción entre el número de cópulas normal y el total de aristas del vine en cada generación.

Es importante notar que, gracias al uso del método de truncamiento, el número de pruebas estadísticas necesarias para ajustar los vines se reduce drásticamente. En el compuesto 2z5u, el número de árboles en los vines en ninguna generación fue mayor que siete y nueve en CVEDA y DVEDA, respectivamente. En el compuesto 1bmm se expandieron a lo sumo nueve árboles en los dos algoritmos.

## 4.3. Conclusiones del capítulo

Los resultados empíricos obtenidos en el problema de acoplamiento molecular ratifican el comportamiento robusto de los algoritmos CVEDA y DVEDA observado en los experimentos con las funciones de prueba. Estas características hacen que los EDA basados en vines puedan ser considerados herramientas prometedoras para crear nuevos programas de acoplamiento molecular.

# Conclusiones

En esta tesis se ha reportado una investigación sobre la conveniencia del uso de vines en los EDA. Las conclusiones fundamentales son las siguientes:

1. Los vines son herramientas que permiten construir mejores distribuciones de búsqueda ofreciendo nuevas vías para tratar con diferentes fuentes de complejidad que surgen en la optimización.

2. Los EDA basados en vines son más flexibles que sus predecesores basados en cópulas multivariadas, como son UMDA y GCEDA, ya que están basados en modelos capaces de describir una amplia variedad de patrones de dependencia.

3. CVEDA y DVEDA son algoritmos robustos dotados de mecanismos que les permiten comportarse relativamente bien tanto en problemas con débiles como con fuertes interacciones entre las variables.

4. La investigación empírica del comportamiento de los algoritmos CVEDA y DVEDA en un conjunto de funciones de prueba muestra que la construcción parcial de los vines basada en AIC es una alternativa a la construcción total que permite disminuir el costo computacional de la estimación de dichos modelos, y que es importante realizar una selección de la estructura de los vines de acuerdo a la fuerza de las interacciones entre las variables.

5. Los resultados empíricos en un conjunto de complejos moleculares muestran que los EDA basados en vines son herramientas prometedoras para crear nuevos programas de acoplamiento molecular.

Los principales resultados son:

1. Se crea una nueva clase de EDA llamada Algoritmos con Estimación de Distribuciones basados en Vines. Dos algoritmos pertenecientes a esta clase son: CVEDA, que utiliza C-vines, y DVEDA con D-vines.

2. Se implementan dos paquetes para el ambiente estadístico R. El primero provee funcionalidades relacionadas con el modelado de dependencias multivariadas utilizando vines. El segundo contiene implementaciones de EDA basados en la teoría de cópulas, entre los que se encuentran los algoritmos CVEDA y DVEDA.

# Recomendaciones y trabajo futuro

Se proponen las siguientes recomendaciones:

- Realizar un estudio de los métodos para la selección de las cópulas bivariadas en los algoritmos CVEDA y DVEDA teniendo en cuenta las deficiencias en la selección de la cópula normal en la función Summation Cancellation.
- Extender el estudio empírico de los algoritmos a un mayor grupo de funciones de prueba caracterizadas de acuerdo a la complejidad de las mismas.

Las siguientes líneas de investigación garantizan la continuidad de este trabajo científico:

- Considerar la extensión de los algoritmos propuestos al dominio discreto.
- Extender la clase Algoritmos con Estimación de Distribuciones basados en Vines con un nuevo algoritmo que utilice vines regulares generales, en lugar de las descomposiciones específicas correspondientes a los C-vines y los D-vines.

# Bibliografía


Aas, K., Czado, C., Frigessi, A. y Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44:182–198.

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

Arderí, R. J. (Junio de 2007). *Algoritmo con Estimación de Distribuciones con cópula gaussiana*. Trabajo de diploma, Universidad de La Habana.

Barba, S. E. (Diciembre de 2007). *Una propuesta para Algoritmos de Estimación de Distribución no paramétricos*. Tesis de maestría, Centro de Investigación en Matemáticas, A.C.

Bedford, T. y Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32:245–268.

Bedford, T. y Cooke, R. M. (2002). Vines — a new graphical model for dependent random variables. *The Annals of Statistics*, 30:1031–1068.

Bengoetxea, E., Miquélez, T., Lozano, J. A. y Larrañaga, P. (2002). Experimental results in function optimization with EDAs in continuous domain. En Larrañaga, P. y Lozano, J. A., editores, *Estimation of Distribution Algorithms. A new tool for Evolutionary Computation*, páginas 181–194. Kluwer Academic Publisher.

Berg, D. (2009). Copula goodness-of-fit testing: An overview and power comparison. *The European Journal of Finance*, 15:675–701.

Berg, D. y Aas, K. (2007). Models for construction of multivariate dependence. Reporte de investigación SAMBA/23/07, Norwegian Computing Center, NR.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. y Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28:235–242.

Bosman, P. A. N. y Thierens, D. (2000). Continuous iterated density estimation evolutionary algorithms within the idea framework. En *Proceedings of the Optimization by Building and Using Probabilistic Models Workshop at the Genetic and Evolutionary Computation Conference (GECCO 2000)*, páginas 197–200.

Bosman, P. A. N. y Thierens, D. (2001). Advancing continous IDEAs with mixture distributions and factorization selection metrics. En *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2001)*, páginas 208–212.

Bosman, P. A. N. y Thierens, D. (2006). Numerical optimization with real-valued Estimation of Distribution Algorithms. En Pelikan, M., Sastry, K. y Cantú-Paz, E., editores, *Scalable optimization via probabilistic modeling. From algorithms to applications*. Springer-Verlag.

Brechmann, E. C. (2010). *Truncated and simplified regular vines and their applications*. Trabajo de diploma, Technische Universität München.

Brent, R. P. (1973). *Algorithms for minimization without derivatives*. Prentice-Hall.

Chambers, J. (2008). *Software for data analysis: Programming with R*. Springer-Verlag.

Cho, D.-Y. y Zhang, B.-T. (2001). Continuous Estimation of Distribution Algorithms with probabilistic principal component analysis. En *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2001)*, páginas 521–526.

Cho, D.-Y. y Zhang, B.-T. (2004). Evolutionary continuous optimization by distribution estimation with variational bayesian independent component analyzers mixture model. En *Parallel Problem Solving from Nature — PPSN VIII*, páginas 212–221. Springer-Verlag.

Cooke, R. M. (1997). Markov and entropy properties of tree- and vine-dependent variables. En *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*.

Cuesta-Infante, A., Santana, R., Hidalgo, J. I., Bielza, C. y Larrañaga, P. (2010). Bivariate empirical and $n$-variate Archimedean copulas in Estimation of Distribution Algorithms. En *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2010)*.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance. *Bulletin de la Classe des Sciences, V. Série, Académie Royale de Belgique*, 65:274–292.

Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis*, 11:102–113.

Demarta, S. y McNeil, A. J. (2005). The t copula and related copulas. *International Statistical Review*, 73:111–129.

Devroye, L. (1986). *Non-uniform random variate generation*. Springer-Verlag.

Embrechts, P., McNeil, A. J. y Straumann, D. (1999). Correlation: Pitfalls and alternatives. *Risk*, 5:69–71.

Fantazzini, D. (2010). Three-stage semi-parametric estimation of t-copulas: Asymptotics, finite-sample properties and computational aspects. *Computational Statistics and Data Analysis*, 54:2562–2579.

Gallagher, M., Fran, M. y Downs, T. (1999). Real-valued evolutionary optimization using a flexible probabilty density estimator. En *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 1999)*, páginas 840–846.

Genest, C. y Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368.

Genest, C., Ghoudi, K. y Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552.

Genest, C., Quessy, J. F. y Rémillard, B. (2007). Asymptotic local efficiency of Cramér-von Mises tests for multivariate independence. *The Annals of Statistics*, 35:166–191.

Genest, C. y Rémillard, B. (2004). Tests of independence or randomness based on the empirical copula process. *Test*, 13:335–369.

Genest, C. y Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques*, 44:1096–1127.

Genest, C., Rémillard, B. y Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44:199–213.

Gil, C. J. (2009). *ADGofTest: Anderson-Darling GoF test*. URL `http://CRAN.R-project.org/package=ADGofTest`. Paquete para R versión 0.1.

González-Fernández, Y. y Soto, M. (2011a). *copulaedas: Estimation of Distribution Algorithms based on copula theory*. URL `http://CRAN.R-project.org/package=copulaedas`. Paquete para R versión 1.0.1.

González-Fernández, Y. y Soto, M. (2011b). *vines: Multivariate dependence modeling with vines*. URL `http://CRAN.R-project.org/package=vines`. Paquete para R versión 1.0.1.

Hahsler, M. y Hornik, K. (2007). TSP — Infrastructure for the traveling salesperson problem. *Journal of Statistical Software*, 23:1–21.

Hobæk Haff, I., Aas, K. y Frigessi, A. (2010). On the simplified pair-copula construction — simply useful or too simplistic? *Journal of Multivariate Analysis*, 101:1296–1310.

Huey, R., Morris, G. M., Olson, A. J. y Goodsell, D. S. (2007). A semiempirical free energy force field with charge-based desolvation. *Journal Computational Chemistry*, 28:1145–1152.

Joe, H. (1996). Families of $m$-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. En Rüschendorf, L., Schweizer, B. y Taylor, M. D., editores, *Distributions with fixed marginals and related topics*, páginas 120–141.

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

Johnson, N. L., Kotz, S. y Balakrishnan, N. (1994). *Continuous univariate distributions*, tomo 1. John Wiley & Sons, segunda edición.

Johnson, S. G. y Narasimhan, B. (2009). *cubature: Adaptive multivariate integration over hypercubes*. URL `http://CRAN.R-project.org/package=cubature`. Paquete para R versión 1.0.

Kern, S., Müller, S. D., Hansen, N., Büche, D., Ocenasek, J. y Koumoutsakos, P. (2003). Learning probability distributions in continuous evolutionary algorithms – A comparative review. *Natural Computing*, 3:77–112.

Kojadinovic, I. y Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34:1–20.

Kojadinovic, I. y Yan, J. (2011). A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems. *Statistics and Computing*, 21:17–30.

Kurowicka, D. y Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling*. John Wiley & Sons.

Larrañaga, P., Etxeberria, R., Lozano, J. A. y Peña, J. M. (1999). Optimization by learning and simulation of bayesian and gaussian networks. Reporte de investigación KZZA-IK-4-99, University of the Basque Country.

Larrañaga, P., Etxeberria, R., Lozano, J. A. y Peña, J. M. (2000). Optimization in continous domains by learning and simulation of gaussian networks. En *Proceedings of the Workshop in Optimization by Building and Using Probabilistic Models in the Genetic and Evolutionary Computation Conference (GECCO 2000)*, páginas 201–204.

Larrañaga, P. y Lozano, J. A., editores (2002). *Estimation of Distribution Algorithms. A new tool for Evolutionary Computation*. Kluwer Academic Publisher.

Larrañaga, P., Lozano, J. A. y Bengoetxea, E. (2001). Estimation of Distribution Algorithms based on multivariate Normal and Gaussian networks. Reporte de investigación EHU-KZAA-IK-1-01, University of the Basque Country.

Madera, J. (2009). *Hacia una generación eficiente de Algoritmos Evolutivos con Estimación de Distribuciones: Pruebas de (in)dependencia+paralelismo*. Tesis de doctorado, Instituto de Cibernética, Matemática y Física.

Milanés, Y. y Álvarez, A. (Junio de 2011). *Acoplamiento molecular utilizando algoritmos con estimación de distribuciones basados en copulas*. Trabajo de diploma, Universidad de La Habana.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer-Verlag, segunda edición.

Ocenasek, J. y Schwarz, J. (2002). Estimation of Distribution Algorithm for mixed continous-discrete optimization. En *Proceedings of the Second Euro-International Symposium on Computational Intelligence*, páginas 227–232.

Ochoa, A. (2010). Opportunities for expensive optimization with Estimation of Distribution Algorithms. En Tenne, Y. y Goh, C.-K., editores, *Computational Intelligence in Expensive Optimization Problems*. Springer-Verlag.

Pelikan, M. (2005). *Hierarchical Bayesian Optimization Algorithm. Toward a new generation of Evolutionary Algorithms*. Springer-Verlag.

Pŏsík, P. (2009). BBOB-benchmarking a simple Estimation of Distribution Algorithm with Cauchy distribution. En *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2009)*, páginas 2309–2314.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Romano, C. (2002). Calibrating and simulating copula functions: An application to the italian stock market. Reporte de investigación 12, Centro Interdipartimale sul Diritto e l'Economia dei Mercati.

Rosenkrantz, D. J., Stearns, R. E. y Philip, M. L., II (1977). An analysis of several heuristics for the Traveling Salesman Problem. *SIAM Journal on Computing*, 6:563–581.

Rousseeuw, P. y Molenberghs, G. (1993). Transformation of nonpositive semidefinite correlation matrices. *Communications in Statistics: Theory and Methods*, 22:965–984.

Salinas-Gutiérrez, R., Hernández-Aguirre, A. y Villa-Diharce, E. (2009). Using copulas in Estimation of Distribution Algorithms. En *Proceedings of the Eighth Mexican International Conference on Artificial Intelligence (MICAI 2009)*.

Salinas-Gutiérrez, R., Hernández-Aguirre, A. y Villa-Diharce, E. (Julio de 2010). D-vine EDA: A new estimation of distribution algorithm based on regular vines. En *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2010)*.

Schepsmeier, U. (2010). *Maximum likelihood estimation of C-vine pair-copula constructions based on bivariate copulas from different families*. Trabajo de diploma, Technische Universität München.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.

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

Song, P. X. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27:305–320.

Soto, M. y González-Fernández, Y. (Mayo de 2010). Vine Estimation of Distribution Algorithms. Reporte de investigación ICIMAF 2010-561, Instituto de Cibernética, Matemática y Física. ISSN 0138-8916.

Soto, M., Ochoa, A. y Arderí, R. J. (Junio de 2007). El Algoritmo con Estimación de Distribuciones basado en cópula gausiana. Reporte de investigación ICIMAF 2007-406, Instituto de Cibernética, Matemática y Física. ISSN 0138-8916.

Soto, M., Ochoa, A., González-Fernández, Y., Milanés, Y., Álvarez, A. y Moreno, E. (2011). Vine Estimation of Distribution Algorithms with application to Molecular Docking. En Santana, R. y Shakya, S., editores, *Markov networks in Evolutionary Computation*. Springer-Verlag.

Trivedi, P. K. y Zimmer, D. M. (2005). Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1:1–111.

Tsutsui, S., Pelikan, M. y Goldberg, D. E. (2001). Evolutionary algorithm using marginal histogram in continuous domain. En *Proceedings of the Optimization by Building and Using Probabilistic Models Workshop at the Genetic and Evolutionary Computation Conference (GECCO 2001)*, páginas 230–233.

Wang, L.-F., Wang, Y., Zeng, J.-C. y Hong, Y. (2010a). An Estimation of Distribution Algorithm based on Clayton copula and empirical margins. En Li, K., Li, X., Ma, S. y Irwin, G. W., editores, *Life System Modeling and Intelligent Computing*, páginas 82–88. Springer-Verlag.

Wang, L.-F., Zeng, J.-C. y Hong, Y. (2009). Estimation of Distribution Algorithm based on copula theory. En *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2009)*.

Wang, L.-F., Zeng, J.-C., Hong, Y. y Guo, X. (2010b). Copula Estimation of Distribution Algorithm sampling from Clayton copula. *Journal of Computational Information Systems*, 6:2431–2440.

Warren, G. L., Andrews, C. W., Capelli, A. M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E. y Head, M. S. (2006). A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry*, 49:5912–5931.

Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21:1–21.

# Apéndice A

# Gráficos de dispersión de las cópulas

En este apéndice se incluyen gráficos de dispersión de muestras con 500 observaciones generadas a partir de las cópulas bivariadas descritas en la sección 1.1.2. Los valores de los parámetros de las cópulas se obtienen, utilizando las expresiones dadas en la tabla 1.1, a partir de los valores -0.5, 0.25 y 0.75 del coeficiente tau de Kendall.

## Cópula independencia

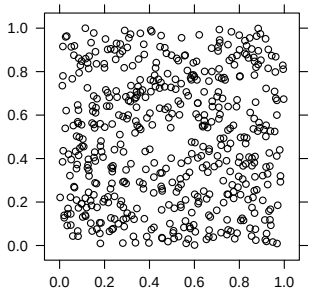


**Figura A.1:** Gráfico de dispersión de una muestra generada a partir de la cópula independencia bivariada.

# Cópula normal

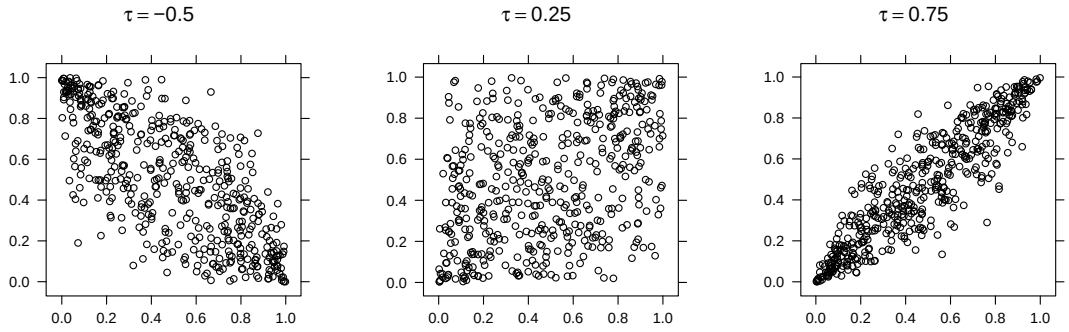


**Figura A.2:** Gráficos de dispersión de tres muestras generadas a partir de la cópula normal bivariada con diferentes valores de su parámetro (correspondientes a los valores -0.5, 0.25 y 0.75 del coeficiente tau de Kendall).

# Cópula t

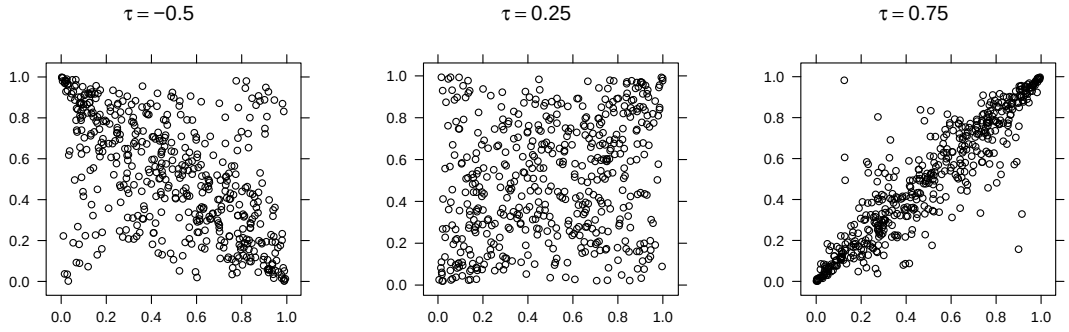


**Figura A.3:** Gráficos de dispersión de tres muestras generadas a partir de la cópula t bivariada con diferentes valores del parámetro de correlación (correspondientes a los valores -0.5, 0.25 y 0.75 del coeficiente tau de Kendall) y dos grados de libertad.

# Cópulas Clayton y Clayton rotada

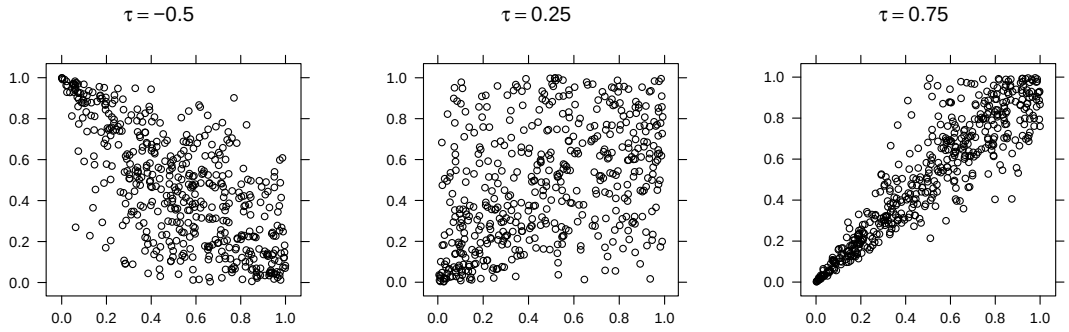


**Figura A.4:** Gráficos de dispersión de tres muestras generadas a partir de las cópulas Clayton (centro y derecha) y Clayton rotada (izquierda) con diferentes valores de sus parámetros (correspondientes a los valores -0.5, 0.25 y 0.75 del coeficiente tau de Kendall).

# Cópulas Gumbel y Gumbel rotada

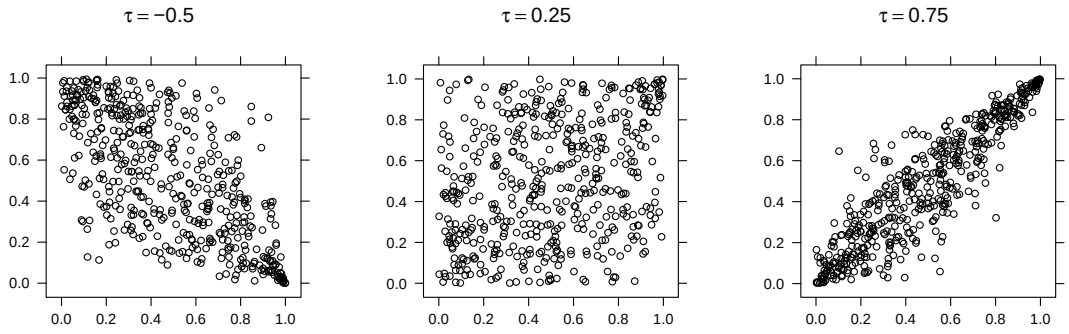


**Figura A.5:** Gráficos de dispersión de tres muestras generadas a partir de las cópulas Gumbel (centro y derecha) y Gumbel rotada (izquierda) con diferentes valores de sus parámetros (correspondientes a los valores -0.5, 0.25 y 0.75 del coeficiente tau de Kendall).

# Apéndice B

# Funciones $h$ y $h^{-1}$ de las cópulas

En el desarrollo de la tesis se utilizaron las cópulas bivariadas independencia, normal, t, Clayton, Clayton rotada, Gumbel y Gumbel rotada (sección 1.1.2). Este apéndice contiene la definición de las funciones $h$ y $h^{-1}$ (sección 2.1.1) para este grupo de cópulas. Aas et al. (2009) presentan la derivación de estas expresiones para las cópulas normal, t, Clayton y Gumbel.

## Cópula independencia

A partir de la función de distribución de la cópula independencia bivariada dada en (1.5), se obtiene que $h_{\mathrm{I}}(x,v) = x$ y $h_{\mathrm{I}}^{-1}(u,v) = u$.

## Cópula normal

De acuerdo a la función de distribución de la cópula normal bivariada dada en (1.6), las funciones $h$ y $h^{-1}$ para esta cópula son

$$h_{\mathrm{N}}(x,v;\rho) = \Phi\left(\frac{\Phi^{-1}(x) - \rho\,\Phi^{-1}(v)}{\sqrt{1-\rho^2}}\right),$$

$$h_{\mathrm{N}}^{-1}(u,v;\rho) = \Phi\left(\Phi^{-1}(u)\sqrt{1-\rho^2} + \rho\,\Phi^{-1}(v)\right).$$

## Cópula t

A partir de la función de distribución de la cópula t bivariada dada en (1.7), se obtiene que las funciones $h$ y $h^{-1}$ para esta cópula son

$$h_{\mathrm{t}}\left(x,v;\rho,\nu\right)=t_{\nu+1}\left(\frac{t_{\nu}^{-1}\left(x\right)-\rho\,t_{\nu}^{-1}\left(v\right)}{\sqrt{\frac{\left(\nu+\left(t_{\nu}^{-1}(v)\right)^{2}\right)\left(1-\rho^{2}\right)}{\nu+1}}}\right),$$

$$h_{\mathrm{t}}^{-1}\left(u,v;\rho,\nu\right)=t_{\nu}\left(t_{\nu+1}^{-1}\left(u\right)\sqrt{\frac{\left(\nu+\left(t_{\nu}^{-1}\left(v\right)\right)^{2}\right)\left(1-\rho^{2}\right)}{\nu+1}}+\rho\,t_{\nu}^{-1}\left(v\right)\right).$$

## Cópula Clayton

De acuerdo a la función de distribución de la cópula Clayton bivariada dada en (1.8), las funciones $h$ y $h^{-1}$ para esta cópula son

$$h_{\mathrm{C}}\left(x,v;\delta\right)=v^{-\delta-1}\left(x^{-\delta}+v^{-\delta}-1\right)^{-1-1/\delta},$$

$$h_{\mathrm{C}}^{-1}\left(u,v;\delta\right)=\left(\left(uv^{\delta+1}\right)^{-\delta/(\delta+1)}+1-v^{-\delta}\right)^{-1/\delta}.$$

## Cópula Clayton rotada

Las funciones $h$ para la cópula Clayton rotada, dadas por $h_{\mathrm{RC}}(x,v;\delta)=h_{\mathrm{C}}(x,1-v;-\delta)$ y $h_{\mathrm{RC}}^{-1}(u,v;\delta)=h_{\mathrm{C}}^{-1}(u,1-v;-\delta)$, se expresan en términos de las funciones $h$ de la versión no rotada.

## Cópula Gumbel

La función $h$ se obtiene a partir de la función de distribución dada en (1.8),

$$h_{\mathrm{G}}(x,v;\delta) = C_{\mathrm{G}}(x,v;\delta)\,\frac{1}{v}\,(-\log v)^{\delta-1}\left[(-\log x)^{\delta}+(-\log v)^{\delta}\right]^{1/\delta-1}.$$

La inversa de $h_{\mathrm{G}}$ respecto al primer argumento no se puede escribir en forma cerrada por lo que se evalúa numéricamente utilizando el método propuesto por Brent (1973).

## Cópula Gumbel rotada

Al igual que para la cópula Clayton rotada, las funciones $h$ para la cópula Gumbel rotada se expresan en términos de las funciones $h$ de la versión no rotada y están dadas por $h_{\mathrm{RG}}(x,v;\delta) = h_{\mathrm{G}}(x,1-v;-\delta)$ y $h^{-1}_{\mathrm{RG}}(u,v;\delta) = h^{-1}_{\mathrm{G}}(u,1-v;-\delta)$.

# Apéndice C

# Publicación de los resultados

Los resultados relacionados con el estudio del comportamiento de los algoritmos en las funciones de prueba se encuentran publicados en el siguiente reporte de investigación:

> M. Soto y Y. González-Fernández (Mayo de 2010). Vine Estimation of Distribution Algorithms. Reporte de investigación ICIMAF 2010-561, Instituto de Cibernética, Matemática y Física. ISSN 0138-8916.

Los resultados obtenidos en el problema de acoplamiento molecular fueron aceptados para su publicación en el siguiente capítulo:

> M. Soto, A. Ochoa, Y. González-Fernández, Y. Milanés, A. Álvarez y E. Moreno (2011). Vine Estimation of Distribution Algorithms with application to Molecular Docking. En R. Santana y S. Shakya, editores, *Markov networks in Evolutionary Computation*. Springer-Verlag.

Algunos resultados parciales se presentaron en los eventos nacionales e internacionales:

- M. Soto, Y. González-Fernández, Y. Garcés y D. Carrera. Evolutionary Estimation of Distribution Algorithms Based on Copulas. *En* $9^{th}$ *International Conference on Operations Research (ICOR 2010)*. La Habana, Cuba.
- M. Soto, A. Ochoa, D. Carrera, Y. González-Fernández y Y. Garcés. On the Use of the Concept of Copula in Evolutionary Optimization. En *XX Encuentro de Estadísticos Cuba-México*. La Habana, Cuba.

- M. Soto, A. Ochoa y Y. González-Fernández. Estimation of Distribution Algorithms based on Vine-Copula. En *Taller de Cómputo de Alto Desempeño (HPC 2010)*. México, D.F., México.

- M. Soto, E. Moreno, A. Ochoa, Y. Milanés, A. Álvarez y Y. González-Fernández. Estimation of Distribution Algorithms for the Molecular Docking Problem. En $9^{th}$ *International Workshop on Operations Research (IWOR 2011)*. La Habana, Cuba.

- M. Soto y Y. González-Fernández. Copula-Vine Modeling in Evolutionary Optimization. En $9^{th}$ *International Workshop on Operations Research (IWOR 2011)*. La Habana, Cuba.

Algunos resultados parciales se presentaron en las jornadas científicas:

- Y. Garcés, Y. González-Fernández y D. Carrera. Algoritmo con Estimación de Distribuciones basado en cópula. En *Jornada Científica Estudiantil de la Facultad de Matemática y Computación de la Universidad de La Habana 2010*. La Habana, Cuba. Este trabajo obtuvo los premios *Primera Mención de Tercer Año* y *Mejor Exposición de la Jornada* de la especialidad de Matemática.

- M. Soto y Y. González-Fernández. Estimation of Distribution Algorithms based on Vine-Copula. En *Jornada Científica del Instituto de Cibernética, Matemática y Física 2010*. La Habana, Cuba.

- Y. Milanés, A. Álvarez y Y. González-Fernández. Estimation of Distribution Algorithms for the Molecular Docking Problem. En *Jornada Científica Juvenil del Instituto de Cibernética, Matemática y Física 2011*. La Habana, Cuba.

- Y. González-Fernández y Y. Garcés. Copula-Vine Modeling in Evolutionary Optimization. En *Jornada Científica Juvenil del Instituto de Cibernética, Matemática y Física 2011*. La Habana, Cuba.

- Y. González-Fernández. Modelado con cópulas y vines en Optimización Evolutiva. En *Jornada Científica Estudiantil de la Facultad de Matemática y Computación de la Universidad de La Habana 2011*. La Habana, Cuba. Este trabajo obtuvo el premio al *Mejor Trabajo Escrito de Matemática Aplicada de Quinto Año* de la especialidad de Computación.