UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA I

TESIS DOCTORAL

Análisis de sensibilidad a la evidencia en Redes Bayesianas Gaussianas

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADA POR

Paola Viviani García

Directores

Miguel Ángel Gómez Villegas
Paloma Maín Yaque

Madrid, 2014

Universidad Complutense de Madrid
Facultad de Ciencias Matemáticas
Departamento de Estadística e Investigación Operativa I

Análisis de sensibilidad a la evidencia en Redes Bayesianas Gaussianas.

Paola Viviani García

Madrid, 2014

Bajo la dirección de los doctores:
Miguel Angel Gómez Villegas y Paloma Maín Yaque

Agradecimientos

Quiero agradecer a mis tutores, Paloma y Miguel Angel, por sus horas dedicadas a este trabajo, primero en forma personal y luego a la distancia. A mi familia por el apoyo y creer en mí. A mis queridos amigos de España que tanto extraño, en especial a Karina por su amistad y compañía durante esta aventura. A Marco por su amor y comprensión.

Y sólo por existir, gracias a mi Samuelito, él es quien moviliza mi vida.

Prólogo

Este trabajo nace frente a un interés por estudiar sensibilidad a la evidencia en redes Bayesianas Gaussianas. En el camino, ante el intento por ajustar una red de este tipo a datos reales, se genera un segundo eje de trabajo que consiste en la revisión del supuesto de Normalidad de los datos, y en encontrar un procedimiento que permita que este supuesto se cumpla.

Las redes Bayesianas pertenecen a la familia de modelos gráficos probabilísticos, donde se combinan conceptos de Teoría de grafos y de Teoría de Probabilidad. Un modelo gráfico probabilístico construído para un grupo de variables de interés, busca representar la interacción entre las variables a través de un grafo, considerado como la parte cualitativa del modelo; y explicar la distribución condicionada de estas variables, por medio de distribuciones de probabilidad que determinan la parte cuantitativa.

Así, el Capítulo 1 se ha dedicado a introducir conceptos que se hacen fundamentales en esta línea de trabajo. Se comienza por revisar los conceptos básicos de Teoría de grafos, con especial énfasis en diferenciar entre grafos dirigidos, donde las interacciones dan un sentido de causalidad entre las variables, y grafos no dirigidos. Dentro de los grafos dirigidos, se debe distinguir a los grafos acíclicos, donde cada trayectoria dentro del grafo contiene nodos diferentes, incluyendo el nodo inicial y el final. Finalmente, se revisa el tipo de conexiones que se pueden presentar en grafos acíclicos dirigidos, ya que esto determinará de qué manera es posible transmitir información a lo largo de la red, en presencia de valores conocidos que se han denominado evidencia.

El Capítulo 1 también contiene conceptos básicos y ampliamente conocidos respecto a la Teoría de Probabilidad. Se ha preferido incluir esta sección como una manera de aunar conocimiento y notación que luego será utilizada a lo largo de este trabajo.

Con estos primeros conceptos ya claros, se puede entonces definir una red Bayesiana como un modelo gráfico probabilístico, donde a través de un grafo acíclico dirigido (DAG) se representan relaciones de dependencia directa entre las variables aleatorias.

Las redes Bayesianas son de gran interés como herramienta de modelado, ya que permiten conocer la estructura de interacciones entre variables, y son utlizadas en áreas tan diversas como la biología, la medicina y la economía, entre otras. De acuerdo al tipo de variables que se consideren para modelar, se puede distinguir entre redes Bayesianas discretas, redes Bayesianas Gaussianas y redes Bayesianas mixtas, que corresponde a una combinación de las dos anteriores. En el Capítulo 2 se revisa más en detalle esta definición y algunas propiedades relacionadas con redes Bayesianas.

Por ser el tema central de este trabajo, el Capítulo 3 se ha dedicado por completo a redes Bayesianas Gaussianas (RBG), esto es, redes Bayesianas donde la distribución de probabilidad conjunta tiene distribución Normal Multivariante.

Después de presentar su definición formal, se dedica una sección a la construcción de una RBG, donde los objetivos a cumplir son determinar su estructura, determinar los parámetros involucrados y definir el método de propagación de evidencia. Respecto a los dos primeros objetivos, es importante distinguir si la red se construye a partir de información dada por expertos (enfoque tradicional), o si se construye a partir de los datos (enfoque de aprendizaje). Respecto a la propagación de evidencia, se propone utilizar el método paso a paso introducido por Castillo et al. [5].

Una segunda Sección de este capítulo corresponde a uno de los objetivos centrales de análisis en este trabajo; el supuesto de Normalidad. Como ya se mencionó, es de gran interés modelar con redes Bayesianas situaciones extraídas de la realidad. Sin embargo, en el caso de redes Bayesianas Gaussianas, por definición, se debe cumplir que estos datos sigan una distribución Normal Multivariante. Este trabajo muestra cómo aplicar un test de normalidad multivariante para revisar el supuesto, y aplica una nueva e interesante aproximación respecto a la metodología a utilizar para reparar la falta de Normalidad cuando el test resulta rechazar.

En el caso univariante, este es un tema ya muy desarrollado, y la metodología habitual para obtener Normalidad en los datos es aplicar la transformaciones de Box-Cox. Sin embargo, en el caso multivariante esto no es posible, ya que al tratar cada variable de forma individual, se perdería la estructura de correlación de los datos. En el trabajo

de Liu et al. [29], se introduce la distribución *nonparanormal*, la cual es presentada en este trabajo como una alternativa para obtener Normalidad de la matriz de datos manteniendo su estructura de correlación, lo cual permite entonces modelar con una RBG.

Al ajustar una red Bayesiana, generalmente se define alguna de las variables como salida principal. Un buen modelo, busca que la distribución de probabilidad de esta variable de interés presente la menor variablidad posible, como una medida de certeza. A partir de esta idea, y de los conceptos vistos hasta aquí, surge la pregunta de cómo o en qué medida las distintas variables del modelo afectan en esta variabilidad, y ante la posibilidad de observar al menos algunas de estas variables, cuál o cuáles de ellas aportan de manera más significativa a una mayor certeza en la salida. Para responder a estas preguntas, en este trabajo se propone utilizar como medida de dispersión la entropía, y como medida de aporte de cada variable a la salida, la información mutua. Es por esta razón que el Capítulo 4 consiste en revisar definiciones y propiedades de la Teoría de la Información.

A partir de las definiciones, conceptos y propiedades vistas en los capítulos anteriores, en el Capítulo 5 se propone un procedimiento para analizar sensibilidad a la evidencia en redes Bayesianas Gaussianas. En modelos que reflejan situaciones de la realidad, no necesariamente la evidencia está preestablecida. Entonces, este procedimiento busca determinar qué variables son las más informativas en el modelo, de manera que, en la medida de lo posible, sean éstas las variables observadas. Esto permite obtener una mayor eficiencia del modelo, y una reducción de costes, ya que será posible priorizar la observación de variables e incluso dejar de observar aquellas variables que no aporten al resultado de interés.

El procedimiento propuesto permite, dentro de una RBG, obtener valores que generan una prioridad de las variables. Pero además, como extensión, se propone un nuevo procedimiento que utliza como medida de contribución, la información mutua normalizada. Este nuevo procedimiento permite por ejemplo, comparar el aporte de una misma variable en redes diferentes, ya que al ser un valor normalizado, las escalas de medición de los valores informativos se hacen comparables.

En el Capítulo 6 se recogen los procedimientos y metodologías propuestas en este trabajo y se aplican a una base de datos real. El principal énfasis está en el análisis del supuesto de Normalidad de los datos y el análisis de sensibilidad a la evidencia.

La aplicación consiste en modelar el Indice de Masa Corporal ($IMC : Peso/Talla^2$) de una muestra de personas que participaron de la Encuesta Nacional de Salud 2009-2010 en Chile. A partir de datos recolectados de tipo clínico, de laboratorio y de información suministrada por cada persona, interesa modelar cómo este grupo de variables interactúa para explicar el IMC. Se espera que estas variables no afecten de la misma manera a hombres y mujeres, por lo que se ajusta una red diferente para cada sexo, y se realizan comparaciones del análsis de sensibilidad con los procedimientos propuestos en el Capítulo anterior.

Finalmente, en el Capítulo 7, se recogen comentarios respecto al trabajo realizado y se proponen futuras líneas de investigación.

Todos los análisis y cálculos desarrollados en este trabajo, especialmente en ejemplos y aplicación final, fueron desarrollados utilizando el programa estadístico de distribución libre R.

Extended abstract with conclusions

Introduction

In general, a Bayesian network attempts to model some interesting phenomenon, considering the random variables which are involved in the problem and the existing dependency structure between them. So, the main objective consists in obtaining the conditioned probability distribution of the unknown variables, basically of some of them defined as target variable, based on the variables which are observable ones (evidential variables). Many times, the variables that may be observed are fixed a priori; however, other times, these may be defined during the modelling process of the network. This study is about Gaussian Bayesian networks, that is networks in which the joint probability distribution associated with the variables $X = (X_1, ..., X_n)$ has a Multivariate Normal distribution.

Considering these models, a methodology is suggested to deal with the analysis of sensitivity to evidence in Gaussian Bayesian networks. For many problems in which the aim is modelling real situations, the set of evidential variables is not defined previously; in fact, a common practice is trying to collect the greatest amount of information you can, a practice that always will have associated costs. After that, the main idea following usually consists in evaluating which of all the available variables are the ones that deliver more information to obtain a better result of the model.

Objectives

The first objective of this study is proposing a methodology, based on Information Theory, to analyze the sensitivity to evidence of Gaussian Bayesian networks. On the way, attempting to adjust a network of this kind to real data, a second working axis is created, which brings up new objectives, making it possible to present a methodology

able to revise the assumption of Normality of the data and finding a procedure which permits to meet this assumption. In addition, there is a transversal objective to the previously exposed, which consists in adjusting a Gaussian Bayesian network to real data.

Results and Conclusions

In this study we have proposed a methodology for analyzing evidence in Gaussian Bayesian networks, theory which is based on the Information Theory. The primary motivation for this proposal came from the idea that not all variables of a Gaussian Bayesian network have the same influence on the variable of our interest. In this context, it's important to underline that the objective of the procedure consists in identifying which one or ones, among the available variables have a stronger influence in diminishing the entropy of the objective, which means a reduction of uncertainty.

A second contribution of the study consists in the use of normalized mutual information to detect the informative value of each variable with respect to the objective. This extension of the procedure permits to prioritize variables in the already exposed way and, besides that, it makes it possible to compare the contribution of different alternative variables within the same network, or similar variables among different networks. In this sense, we show an example in which the aim is to know if the replacement of a node in the network means a higher or lower informative contribution than the original node.

Up to now, in general, the analysis that have been carried out regarding Gaussian Bayesian networks, have been of theoretical character. So, an important contribution of this study is having carried out an application to real data. However, this procedure meant facing some issues that were not considered within the theoretical approaches. The first issue was the revision of the assumption of Normality. When working with adjusted networks under the traditional approach, Normality is assumed and the analyses are carried out without problems. But when facing a real database, it becomes necessary to revise the compliance of the assumption of Normality. Therefore, in this study we looked for different methodologies and used a Multivariate Normality test, which permitted to respond to this assumption.

On the other hand, when the hypothesis of Normality was refused, a new problem arose: how to solve the lack of Normality in order to make it possible to adjust a Gaussian Bayesian network. In order to face that, we propose considering the Nonpa-

ranormal distribution for the data matrix. Using a semiparametric transformation, it is possible to obtain a data matrix which meets the assumption of Normality. This fact becomes very important due to the possibility of using Bayesian networks as a real tool for modelling reality, as very frequently the real data don't have a Multivariate Normal distribution.

Indice General

Prólogo **I**

Extended abstract with conclusions **V**

1. Conceptos preliminares **9**

1.1. Teoría de grafos 9

1.1.1. Clasificación de grafos 10

1.1.2. Grafos dirigidos 11

1.2. Independencia e independencia condicionada 14

1.2.1. Independencia 14

1.2.2. Independencia condicionada 15

1.2.3. Independencia e independencia condicionada en variables aleatorias 16

1.3. Propagación de la Evidencia 17

2. Redes Bayesianas **21**

2.1. Definición y Caracterización 21

2.1.1. Independencia condicionada 24

2.1.2. Propiedad de Markov 25

2.1.3. Tipos de redes Bayesianas 27

3. Redes Bayesianas Gaussianas (RBG) 35

3.1. Definición y caracterización 35

3.1.1. Definición 35

3.1.2. Construcción de RBG 36

3.1.3. Propiedades 39

3.1.4. Propagación de la evidencia 43

3.2. Supuesto de Normalidad en RBG 44

3.2.1. Revisión del supuesto de Normalidad 45

3.2.2. Caso de no Normalidad 46

4. Teoría de la Información 53

4.1. Entropía y entropía diferencial 53

4.1.1. Entropía conjunta y entropía condicionada 55

4.1.2. Entropía diferencial conjunta y entropía diferencial condicionada 56

4.2. Entropía relativa e información mutua 57

4.2.1. Relación entre la entropía y la información mutua 59

4.2.2. Propiedades de la entropía y la información mutua 60

4.3. Información mutua normalizada (IMN) 63

5. Sensibilidad a la evidencia en Redes Bayesianas Gaussianas 65

5.1. Introducción 65

5.2. Antecedentes 66

5.3. Análisis de sensibilidad a la evidencia basado en información mutua . . 68

5.3.1. Procedimiento propuesto 69
5.3.2. Teoría de la información en RBG 70
5.3.3. Ejemplo: Análisis de sensibilidad (AS) 74
5.4. Extensión del análisis de sensibilidad a la evidencia basado en información mutua normalizada 77
5.4.1. Nuevo procedimiento propuesto 78
5.4.2. Información mutua normalizada en RBG 79
5.4.3. Ejemplo: Análisis de sensibilidad basado en IMN 80
5.4.4. Extensión del Ejemplo 5 (EEjemplo 5) 82

6. Aplicación a datos reales **85**

6.1. Objetivos y base de datos 85
6.1.1. Objetivo 85
6.1.2. Encuesta nacional de salud (ENS) 86
6.1.3. Base de datos 88
6.2. Supuesto de Normalidad 89
6.2.1. Revisión del supuesto de Normalidad 89
6.2.2. Aplicación de la distribución Nonparanormal 91
6.3. Ajuste de la RBG 93
6.3.1. Proceso de aprendizaje 94
6.3.2. Parámetros estimados 102
6.4. Análisis de sensibilidad 107

7. Comentarios y futuras líneas de investigación **111**

Bibliografía 113

Indice de Tablas

3.1. Test de Normalidad Multivariante inicial Ejemplo 7 50

3.2. Test de Normalidad Multivariante Ejemplo 7 50

3.3. Test de Normalidad Univariante Ejemplo 7 51

5.1. Ejemplo: Análisis de sensibilidad (AS). Paso 2. 75

5.2. Ejemplo: AS. Sensibilidad de X_7 a la evidencia. 76

5.3. Ejemplo: AS. Paso 6, condicionado en X_5 76

5.4. Ejemplo: AS. Paso 6, condicionado en X_4 y X_5. 77

5.5. Ejemplo: Análisis de sensibilidad con IMN. Paso 2 80

5.6. Ejemplo: AS con IMN. Pasos 6 y 7, condicionado en X_5 81

5.7. Ejemplo: AS con IMN. Paso 6, condicionado en X_4 y X_5 82

5.8. EEjemplo 5: Comparación entre RBG_1 y RBG_2 84

6.1. Test de Normalidad Multivariante inicial, hombres 89

6.2. Test de Normalidad Univariante inicial, hombres 90

6.3. Test de Normalidad Multivariante inicial, mujeres 90

6.4. Test de Normalidad Univariante inicial, mujeres 90

6.5. Test de Normalidad Multivariante, hombres 91

6.6. Test de Normalidad Univariante, hombres 92
6.7. Test de Normalidad Multivariante, mujeres 92
6.8. Test de Normalidad Univariante, mujeres 92
6.9. Coeficientes de fuerza RBG hombres 95
6.10. Coeficientes de fuerza RBG mujeres 99
6.11. Aplicación. Paso 2. RBG hombres. 108
6.12. Aplicación. Paso 2. RBG mujeres. 109
6.13. Razón de Información mutua normalizada hombres vs. mujeres. 109

Indice de Figuras

1.1. Arista dirigida 10

1.2. Arista no dirigida 10

1.3. a) grafo dirigido. b) grafo no dirigido. 11

1.4. Subgrafo. 12

1.5. a) Grafo cíclico. b) Grafo acíclico. 13

1.6. a) Conexión serial b) Conexión divergente c) Conexión convergente. 14

1.7. D-separación en grafos dirigidos 19

2.1. Ejemplo 1: red Bayesiana. 23

2.2. Propiedad de Markov. 26

2.3. Ejemplo 2: red Bayesiana discreta. 28

2.4. Red Ejemplo 2 en Hugin. 29

2.5. Red Ejemplo 2 con evidencia en Hugin. 29

2.6. Ejemplo 3: red Bayesiana Gaussiana. 32

2.7. Ejemplo 4: red Bayesiana Gaussiana condicionada lineal. 34

3.1. DAG Ejemplo 5. 41

4.1. Relación entre la información mutua y la entropía 60

5.1. Entropía y función valor. 68

5.2. DAG EEjemplo 5. 82

6.1. Instrumentos de Medición ENS 2009-2010 (1) 87

6.2. Instrumentos de Medición ENS 2009-2010 (2) 87

6.3. RBG inicial ENS hombres 94

6.4. Fuerza de enlaces en RBG hombres 96

6.5. RBG final ENS hombres 96

6.6. Normal Q-Q Plot red hombres 97

6.7. Histogram of residuals red hombres 97

6.8. RBG inicial ENS mujeres 98

6.9. Fuerza de enlaces en RBG mujeres 100

6.10. RBG final ENS mujeres 100

6.11. Normal Q-Q Plot red mujeres 101

6.12. Histogram of residuals red mujeres 101

Capítulo 1

Conceptos preliminares

Los modelos gráficos probabilísticos combinan resultados de la teoría de grafos y de la probabilidad; la primera, considerada como el aspecto cualitativo del modelo, permite representar la estructura de dependencia (o independencia) de un conjunto de variables aleatorias, y la segunda, el aspecto cuantitativo, define numéricamente estas relaciones.

Las redes Bayesianas pertenecen a esta familia de modelos gráficos probabilísticos, también conocidos como redes probabilísticas, por lo que es pertinente comenzar por introducir algunos conceptos básicos de Teoría de grafos y de Probabilidad, para luego relacionarlos en una sección dedicada a causalidad.

1.1. Teoría de grafos

En el contexto de redes probabilísticas, un grafo representa interacciones entre un conjunto finito de variables aleatorias, digamos $V = \{V_1, V_2, .., V_n\}$, donde cada variable V_i corresponde a un vértice o nodo del grafo y las interacciones entre ellas, quedan definidas por aristas o arcos.

Definición 1.1 *Un* ***grafo*** G *se define como un par* (V,E)*, donde* $V = \{V_1, V_2, .., V_n\}$ *es un conjunto de elementos denominados* vértices *o* nodos *y* $E \subseteq V \times V$ *es un conjunto de* aristas *o* arcos*. Cada elemento de* E*, se representa por* E_{ij}*, indicando que corresponde a la arista que une* V_i *con* V_j*, para* $i \neq j$*.*

1.1.1. Clasificación de grafos

Una primera clasificación de grafo se puede realizar de acuerdo al tipo de aristas que contiene. Así, se debe distinguir entre:

Arista dirigida

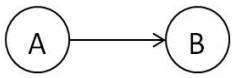

Figura 1.1: Arista dirigida

La presencia de una arista dirigida E_{ij} entre los vértices V_i y V_j, significa que el nodo V_i se relaciona con el nodo V_j, pero no al revés; se refiere a una relación *Causa-Efecto* y se escribe como $V_i \rightarrow V_j$. En la Figura 1.1 se muestra un ejemplo de dos vértices, A y B, conectados a través de una arista dirigida.

Arista no dirigida

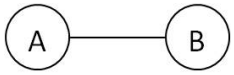

Figura 1.2: Arista no dirigida

Aquí, tanto la arista E_{ij} como la E_{ji} están contenidas en E, es decir, los nodos V_i y V_j se relacionan en forma recíproca, por lo que se dice que están *Asociados*. La Figura 1.2 muestra una conexión no dirigida entre los nodos A y B.

De acuerdo a esto, si todas las aristas contenidas en E son no dirigidas, corresponde a un ***grafo no dirigido***. En forma análoga, si todas las aristas son dirigidas, es un ***grafo dirigido***. Por último, si el grafo contiene los dos tipos de aristas, se trata de un ***grafo de cadena o mixto***. En la Figura 1.3, se muestran ejemplos de los dos primeros casos.

Las redes Bayesianas, tema principal de esta memoria, se representan con un caso particular de grafos dirigidos, por lo que de aquí en adelante se hará referencia sólo a este tipo grafos.

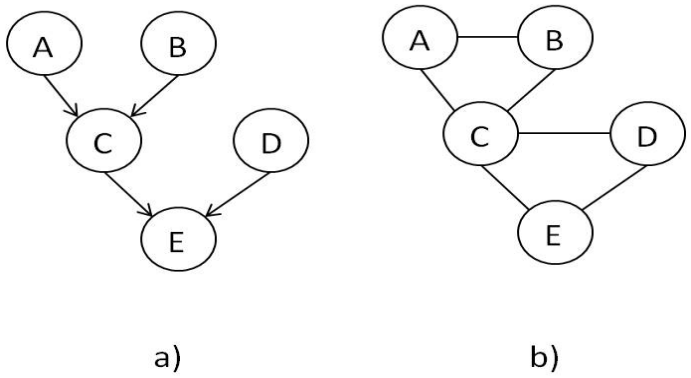

Figura 1.3: a) grafo dirigido. b) grafo no dirigido.

1.1.2. Grafos dirigidos

En esta Sección, se revisan los principales conceptos y propiedas de los grafos dirigidos.

Sea *G=(V,E)* un grafo dirigido, formado por un conjunto de elementos $V = \{V_1, V_2, .., V_n\}$, conectados por un conjunto de aristas dirigidas E, entonces se pueden introducir las siguientes definiciones:

- Si $V_i \rightarrow V_j$ se dice que el nodo V_i es *padre* del nodo V_j, y que el nodo V_j es hijo del nodo V_i.
- El conjunto de todos los padres del nodo V_j se escribe como $pa(V_j)$.
- La *familia* del nodo V_j corresponde al conjunto que contiene la unión del nodo con sus padres, y se denota por $fa(V_j) = V_j \cup pa(V_j)$.
- Al conjunto de todos los nodos con al menos un posible camino hasta el nodo V_i, se denomina ancestros o ascendientes de V_i, y se escribe $as(V_i)$.
- El conjunto de todos los nodos que no son ascendientes o ancestros de V_i se escribe como $na(V_i)$.
- Al conjunto de todos los nodos a los que se puede llegar por al menos un camino desde V_i, se denomina descendientes de V_i y se escribe $de(V_i)$.
- El conjunto de todos los nodos que no son descendientes de V_i se escribe como $nd(V_i)$.

Como ejemplo, en el grafo a) de la Figura 1.3 se tiene que E es hijo de C, $pa(C) = \{A, B\}$ y que $fa(C) = \{A, B, C\}$.

Definición 1.2 *Sea $G = (V, E)$ un grafo dirigido, entonces $G' = (V', E')$ es **subgrafo** de G si cumple que $V' \subseteq V$, $E' \subseteq E$ y G' es un grafo. Es decir, G' es un subconjunto de nodos, con sus enlaces, de G.*

En la Figura 1.4 se muestra un posible subgrafo del grafo a) de la Figura 1.3.

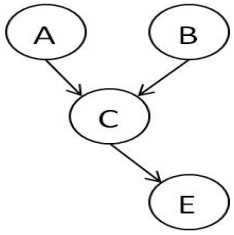

Figura 1.4: Subgrafo.

Definición 1.3 *Sea $G = (V, E)$ un grafo dirigido, se llama **trayectoria** en G a una sucesión de nodos $V_1V_2...V_{p-1}V_p$, tales que $E_{i\ i+1} \in E$, para cada $i = 1, 2, ..., p-1$.*
*Así, se dice que la trayectoria **conecta** o **une** al nodo V_1 con el nodo V_p y que el número de arcos recorridos, $p-1$, es la **longitud** de la trayectoria.*
*Un **camino** se define como una trayectoria sin nodos ni arcos repetidos.*
*Cuando dicho camino contiene nodos diferentes, excepto el inicial y final, se habla de un **ciclo**. Luego, se dice que un grafo es **acíclico** si no tiene ciclos.*

En la Figura 1.5 se muestran ejemplos simples de grafos cíclicos y no cíclicos.

Los grafos que se utilizan para describir las redes Bayesianas son grafos acíclicos dirigidos, en adelante, DAGs (Directed Acyclic Graphs).

Otras defininiciones interesantes en Teoría de grafos son:

Definición 1.4 *Un grafo es **conexo** si cada par de nodos está conectado por un camino; es decir, desde cualquier nodo inicial, existe al menos un camino que permite ir a todos los demás.*

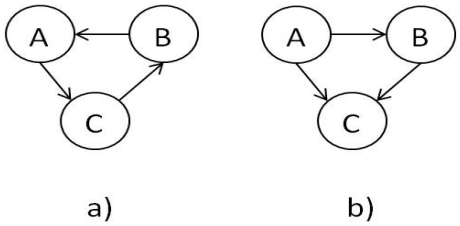

Figura 1.5: a) Grafo cíclico. b) Grafo acíclico.

Definición 1.5 *Un grafo* ***moral*** G^m *se obtiene a partir de un DAG* $G = (V, E)$*; primero, uniendo con aristas no dirigidas a todos los nodos adyacentes no conectados con hijos comunes, y luego, cambiando todos los enlaces dirigidos por conexiones no dirigidas.*

Este último tipo de grafos es relevante en el aspecto computacional de las redes Bayesianas.

Definición 1.6 *El* ***esqueleto*** *de un grafo se obtiene a partir de un DAG* $G = (V, E)$ *reemplazando todos los enlaces dirigidos por conexiones no dirigidas.*

Definición 1.7 *Es necesario distinguir entre tres tipos de conexiones que pueden presentarse en un grafo acíclico dirigido:*

- ***Conexión serial*** *Corresponde a una dependencia consecutiva entre los nodos. Así, en la Figura 1.6 a), se observa que el nodo A es* padre *del nodo B, quien a su vez, es* padre *del nodo C.*

- ***Conexión divergente*** *En este caso, de un nodo* padre*, se desprenden dos o más nodos* hijos *no conectados entre sí. En la Figura 1.6 b) se muestra un ejemplo.*

- ***Conexión convergente*** *Se refiere a una conexión en la cual dos o más nodos* padres*, no conectados entre sí, tienen asociado un mismo nodo* hijo*, como se muestra en la Figura 1.6 c). En este caso se suelen denominar a los nodos A y B como esposos.*

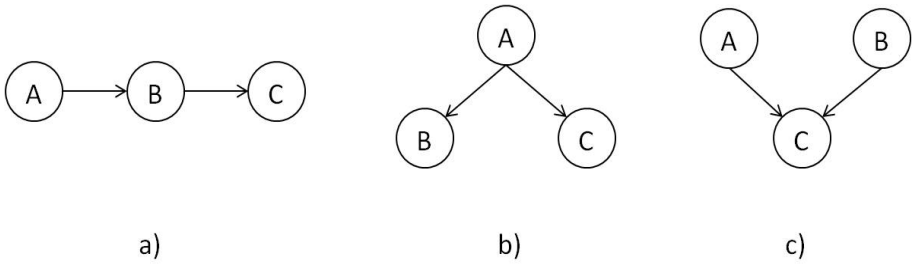

Figura 1.6: a) Conexión serial b) Conexión divergente c) Conexión convergente.

1.2. Independencia e independencia condicionada

El siguiente paso, es asociar los conceptos generales de grafos ya vistos, a modelos de redes probabilísticas, donde los nodos pasan a ser variables aleatorias y la estructura de dependencia se describe por medio de distribuciones de probabilidad condicionadas. Sin embargo, para esto, es necesario primero revisar conceptos de independencia e independencia condicionada.

1.2.1. Independencia

Si al conocer la ocurrencia de un suceso, no se modifica la probabilidad de ocurrencia del otro, entonces se dice que los sucesosos son independientes.

Definición 1.8 *El suceso A es **independiente** de B si*

$$p(A|B) = p(A)$$

Esta definición es simétrica, es decir, si A es independiente de B, también es cierto que B es independiente de A, luego:

$$p(B|A) = p(B)$$

En forma análoga, si A y B no son independientes (son dependientes), entonces:

$$p(A|B) \neq p(A) \quad \text{ó} \quad p(B|A) \neq p(B)$$

Aplicando la regla de multiplicación y la definición de independencia 1.8, se obtiene que si A es independiente de B entonces,

$$p(A \cap B) = p(A|B)p(B) = p(A)p(B)$$

1.2.2. Independencia condicionada

Sean tres sucesos A, B y C, entonces se dice que A es condicionalmente independiente de B dado C si,

$$p(A|B \cap C) = p(A|C)$$

En forma análoga a la independencia, la independencia condicionada también es simétrica, es decir, si A es condicionalmente independente de B dado C, entonces B es condicionalmente independente de A dado C:

$$p(B|A \cap C) = p(B|C)$$

Por último, en forma similar al caso anterior, si se aplica la regla de la cadena, entonces se tiene que si A es condicionalmente independente de B dado C:

$$p(A \cap B|C) = P(A|C)p(B|C)$$

Se puede ver que, si se considera $C = \emptyset$, entonces la independencia es un caso particular de la independencia condicionada.

A continuación, se aplican las definiciones anteriores a variables aleatorias.

1.2.3. Independencia e independencia condicionada en variables aleatorias

Como ya se ha mencionado, en una red Bayesiana los nodos representan variables aleatorias. Por esta razón, a continuación se presentan las definiciones y propiedades de independencia e independencia condicionada aplicado a variables aleatorias (Dawid [10]).

Definición 1.9 *Sean X e Y variables aleatorias; se dice que X e Y son **independientes**, y se denota como $X \perp\!\!\!\perp Y$, si y solo si*

$$p(x, y) = p(x)p(y)$$

Es decir, se dicen independientes si y sólo si, su función de probabilidad conjunta se puede obtener como el producto de sus funciones de probabilidad marginales.

Intuitivamente, si $X \perp\!\!\!\perp Y$ entonces la información que pueda aportar Y no afecta en la incertidumbre de X; es decir, de forma equivalente se puede escribir

$$p(x|y) = p(x)$$

A partir de la definición, se obtiene el siguiente teorema.

Teorema 1.1 *Si $X \perp\!\!\!\perp Y$, entonces $Y \perp\!\!\!\perp X$.*

A continuación, se incorpora una tercera variable aleatoria, Z, para ver la independencia condicional.

Definición 1.10 *Sean las variables aleatorias X, Y y Z,; se dice que X e Y son **condicionalmente independiente** dado Z, y se denota como $X \perp\!\!\!\perp Y \mid Z$, si y solo si*

$$p(x, y|z) = p(x|z)p(y|z)$$

De manera análoga al caso de la independencia, en el caso de la independencia condicionada, se puede escribir de forma equivalente,

$$p(x|y,z) = p(x|z)$$

Propiedades de la Independencia condicionada

Se presentan las principales propiedades de independencia condicionada, lo que permite obtener nuevas relaciones entre variables que podrían ser de utilidad de acuerdo a la información disponible.

- *Simetría:* si $X \perp\!\!\!\perp Y \mid Z$, entonces $Y \perp\!\!\!\perp X \mid Z$.
- *Descomposición:* si $X \perp\!\!\!\perp (Y \cup W) \mid Z$ entonces $X \perp\!\!\!\perp Y \mid Z$ y $X \perp\!\!\!\perp W \mid Z$.
- *Unión débil:* si $X \perp\!\!\!\perp (Y \cup W) \mid Z$, entonces $X \perp\!\!\!\perp Y \mid (W \cup Z)$ y $X \perp\!\!\!\perp W \mid (Y \cup Z)$.
- *Contracción:* si $X \perp\!\!\!\perp Y \mid Z$ y $X \perp\!\!\!\perp W \mid (Y \cup Z)$, entonces $X \perp\!\!\!\perp (Y \cup W) \mid Z$.
- *Intersección:* si $X \perp\!\!\!\perp Y \mid (Z \cup W)$ y $X \perp\!\!\!\perp W \mid (Y \cup Z)$, entonces $X \perp\!\!\!\perp (Y \cup W) \mid Z$.

La independencia y la independencia condicionada tiene una considerable importancia para este trabajo de Tesis, pues implica que no es rentable obtener información sobre variables independientes, ya que dicha información es irrelevante.

1.3. Propagación de la Evidencia

En esta sección se revisarán algunas reglas importantes respecto a la transmisión de la *evidencia* en grafos acíclicos dirigidos, las cuales permitirán el posterior trabajo y comprensión de las redes Bayesianas.

Evidencia

En un grafo, o específicamente en un grafo acíclico dirigido para el interés de este trabajo, cada uno de los nodos representa una variable aleatoria, y los arcos representan la estructura de dependencia entre ellas. Así, cada variable puede tomar valores

categóricos o continuos, dentro del correspondiente espacio muestral asociado.

Cuando se conoce el valor exacto que toma una de las variables, dicho valor será considerado como **evidencia** y es de interés conocer de qué manera, el introducir esta información en la red, afecta a la incertidumbre del resto de las variables.

d-separación en grafos dirigidos

Para estudiar cómo se traspasa la información de la *evidencia* a lo largo de una red, es necesario conocer los criterios de separación que se cumplen en el grafo; esto hace referencia a las relaciones de independencia o dependencia condicionada entre las variables, que dependerán del tipo de conexión que se considere (Definición 1.7).

- **Conexión serial**
  En este caso el nodo A influye en B, quien a su vez, influye en C. Luego, tener evidencia sobre el estado de A o de C, transmitirá información en la trayectoria a través de B. Sin embargo, si hay evidencia acerca de B, la conexión queda bloqueada, y A y C se hacen independientes; también se dice que A y C están d-separados dado B. En la Figura 1.7 a) se muestra esta situación, donde la evidencia queda representada por el nodo sombreado.

  En resumen, en una conexión serial, se transmite información de la evidencia, excepto que dicha información esté contenida en el nodo intermedio.

- **Conexión divergente**
  Ahora , la información puede pasar a los *hijos* del nodo A, excepto si A es conocido; luego, si la evidencia está contenida en el nodo A, como se muestra en la Figura 1.7 b), B y C están d-separados dado A.

  En una conexión divergente la información fluye a través de la red, excepto si la evidencia se encuentra en el nodo padre, ya que queda bloqueada la comunicación entre los nodos hijos.

- **Conexión convergente**
  Una conexión convergente funciona radicalmente diferente a las dos anteriores; si no hay información sobre el nodo C, entonces los nodos A y B son independientes, de hecho, si la evidencia se encuentra en uno de ellos, digamos A, dicha información no tendrá influencia en B. Sin embargo, si hay evidencia en el efecto C, entonces las causas A y B se hacen dependientes, como muestra la Figura 1.7 c), y se dice que A y B están d-conectados.

  En una conexión convergente la información puede ser transmitida a través de la red sólo si se tiene la evidencia sobre el nodo hijo, o un descendiente de éste.

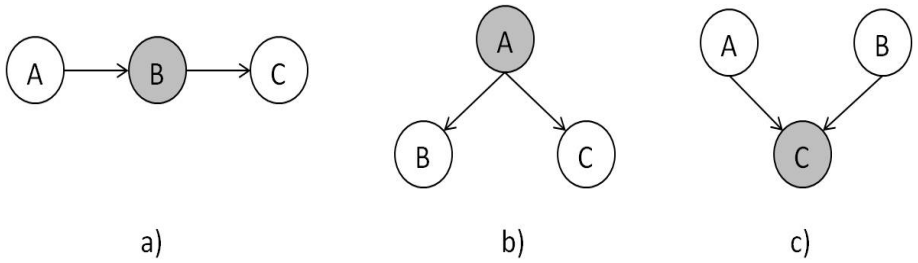

Figura 1.7: a) A y C d-separados dado B. b) B y C d-separados dado A. c) A y B d-conectados dado C.

Capítulo 2

Redes Bayesianas

Una red Bayesiana es un modelo gráfico probabilístico que representa, a través de un gráfico acíclico dirigido (DAG), la estructura de dependencia de un conjunto de variables aleatorias. Se ha convertido en una herramienta de gran utilidad y popularidad para razonar bajo incertidumbre, ya que su representación gráfica permite una mejor e intuitiva comprensión del fenómeno que se modela.

Formalmente, una red Bayesiana consiste en un modelo formado por dos partes; una cualitativa y otra cuantitativa. La dimensión cualitativa de la red está dada por el DAG, cuyos nodos representan variables aleatorias y los arcos representan relaciones de dependencia directa entre las variables. La parte cuantitativa de la red, asociada al DAG, especifica las distribuciones de probabilidad condicionadas de cada nodo dado sus padres.

2.1. Definición y Caracterización

Una **red Bayesiana** está formada por el par $(\mathcal{G}, \mathcal{P})$, donde $\mathcal{G}$ es una grafo acíclico dirigido (DAG) formado por un nodo para cada variable aleatoria de $\mathbf{X} = \{X_1, ..., X_n\}$ y arcos que representan la estructura de dependencia probabilística entre ellas, $\mathcal{P} = \{p(x_1|pa(x_1)), ..., p(x_n|pa(x_n))\}$ es un conjunto de n distribuciones de probabilidad condicionadas, y $pa(x_i)$ es el conjunto de *padres* del nodo X_i en $\mathcal{G}$.

Una característica importante de las redes Bayesianas, es que aplicando la regla de la cadena y las independencias condicionadas, se puede determinar la distribución de probabilidad conjunta a partir de $\mathcal{P}$ mediante la siguiente factorización (demostración disponible en Jensen [23]):

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i|pa(X_i)). \tag{2.1}$$

Es interesante observar que en la ecuación 2.1 se tiene una distribución de probabilidad condicionada por cada variable aleatoria, es decir, cada nodo X_i es una variable condicionada por sus padres. Luego, esta definición de red Bayesiana, establece una relación directa entre la parte cualitativa y la parte cuantitativa de la red, ya que es el grafo el que permite determinar las distribuciones de probabilidades condicionadas que se consideren en la factorización de la distribución de probabilidad conjunta. Es decir, cada factorización de la distribución de probabilidad conjunta de una red Bayesiana, está asociada a un determinado DAG.

Para revisar y comentar algunas propiedades presentes en redes Bayesianas, a continuación se presenta un ejemplo introducido por Ben-Gal [2].

Ejemplo 1 *Este ejemplo, tiene una estructura similar al ampliamente conocido ejemplo "Terremoto"de Pearl [33]. En la Figura 2.1 se muestra el DAG y las tablas de probabilidad condicionadas correspondientes porque las variables que intervienen son discretas. Considera la situación de una persona que podría estar con dolor de espalda, representado por la variable* Dolor *(D). Este dolor, puede ser causado por una* Lesión *(L), la cual a su vez, podría ser producida por una* Actividad *(A) física mal hecha, o porque en la oficina en que trabaja se han comprado nuevas e incómodas* Sillas *(S) de trabajo. Si la lesión se presentara por esta segunda razón, es de esperar que su* Compañero *(C) de trabajo, desarrolle también una lesión similar. Todas las variables aleatorias consideradas son discretas de tipo binario, y para cada caso se tiene los posibles estados excluyentes verdadero (V) o falso (F).*

A continuación se revisarán algunas propiedades con este ejemplo. Lo primero es observar que las variables *Silla* y *Actividad*, son *padres* de *Lesión*, por otro lado *Dolor* es *hijo* de *Lesión*, y por último, los *ancestros* de *Dolor* son las variables *Silla*, *Actividad* y *Lesión*.

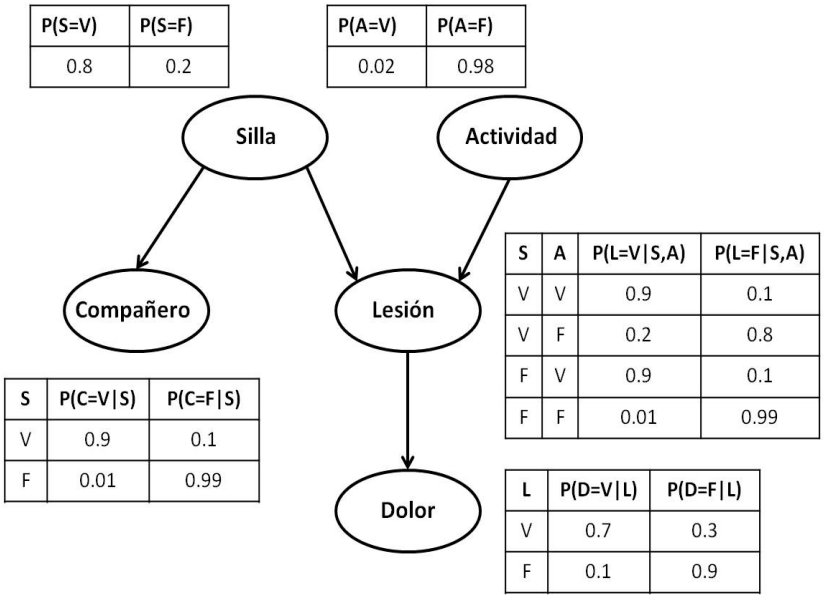

Figura 2.1: Ejemplo 1: red Bayesiana.

Es interesante reconocer al menos dos posibles enfoques como objetivo del modelo con una red Bayesiana. Usualmente, al modelar un fenómeno con una red Bayesiana, se identifica una variable de interés, digamos X_i, y se busca conocer la distribución de probabilidad condicionada de dicha variable dada la evidencia, proceso que se conoce como inferencia en la red.

Respecto a la interpretación de los resultados, si se tiene evidencia en variables que son *hijos* de X_i entonces se está contribuyendo a considerar X_i como una variable de *Diagnóstico*; por otro lado, si la evidencia se incorpora en *padres* de X_i, la inferencia es a favor de realizar una *Predicción* en el nodo X_i. Así, mirando el ejemplo, si se considera $X_i = A$ y se observa evidencia en L y D, se podrá interpretar la probabilidad condicionada que se obtenga para A como la capacidad de la variable *Actividad* para diagnosticar una lesión, y por lo tanto, el dolor de espalda. En cambio, si $X_i = D$ y la evidencia se incorpora en S y A, a través de la inferencia se podrá pronosticar tener dolor de espalda.

2.1.1. Independencia condicionada

Considerando que una red Bayesiana es un caso particular de grafo dirigido, entonces es natural asumir que se cumplen las propiedades de d-separación vistas en la Sección 1.3, las que determinan estructuras de independencia (y dependencia) condicionada.

A continuación, se describirán algunas situaciones de dependencia (o independencia) condicionada presentes en el Ejemplo 1, de acuerdo al tipo de convergencia.

- Las variables *Silla* y *Actividad* son independientes, pero si se conoce el verdadero estado de *Lesión* y se incorpora como evidencia, entonces ellas pasan a ser condicionalmente dependientes. Esto tiene sentido intuitivo, ya que inicialmente, sin saber si existe o no una lesión, las variables Silla y Actividad no tienen ninguna relación entre ellas; pero esta situación cambia al saber que por ejemplo sí hay lesión, ya que ellas, o al menos una de ellas, podría ser responsable de dicha lesión (conexión convergente).

- Si se introduce evidencia acerca del estado de la variable *Silla*, las variables *Compañero* y *Lesion* pasan a ser condicionalmente independientes. Por ejemplo, si se sabe que las sillas nuevas no producen ningún efecto físico, entonces las variables *Compañero* y *Lesión* ya no tienen relación ninguna (conexión divergente).

- Por último, si se observa la variable *Lesión*, *Dolor* será condicionalmente independiente de cada uno de sus ancestros, *Silla* y *Actividad*. Esto tiene sentido porque son éstas variables las que logran explicar *Dolor* pero a través de *Lesión*, por lo tanto, si ya se conoce *Lesión*, los ancestros no agregan nueva información (conexión serial).

Como ya se dijo, estas estructuras de independencia condicionada permiten escribir la distribución de probabilidad conjunta de una red Bayesiana, de acuerdo a la factorización presentada en 2.1. Ahora se va a revisar la implicación de este hecho con el Ejemplo 1.

Si se quisiera calcular la distribución de probabilidad conjunta del Ejemplo, aplicando la regla de la cadena, pero sin considerar las independencias condicionadas, se tendría que:

$$P(S,A,C,L,D) = \underbrace{P(S)}_{1}\underbrace{P(A \mid S)}_{2}\underbrace{P(C \mid A,S)}_{4}\underbrace{P(L \mid C,A,S)}_{8}\underbrace{P(D \mid L,C,A,S)}_{16}$$

donde en cada probabilidad requerida, se ha especificado el número de parámetros involucrados. Es decir, con sólo 5 variables, este modelo tiene 31 parámetros (este cálculo se puede obtener como $2^5 - 1 = 31$).

Ahora, aplicando la factorización que caracteriza a una red Bayesiana, en que se considera su estructura de independencia condicionada, este mismo cálculo se puede realizar de la siguiente manera:

$$P(S,A,C,L,D) = \underbrace{P(S)}_{1}\underbrace{P(A)}_{1}\underbrace{P(C \mid S)}_{2}\underbrace{P(L \mid A,S)}_{4}\underbrace{P(D \mid L)}_{2}$$

es decir, esta factorización reduce el número de parámetros a un total de 10, lo que implica importantes ventajas en el momento de realizar procesos de estimación y/o aprendizaje de la red.

2.1.2. Propiedad de Markov

Sea una red Bayesiana formada por el par $(\mathcal{G}, \mathcal{P})$, donde $\mathcal{G}$ es un grafo acíclico dirigido (DAG), en que cada nodo representa una variable aleatoria X_i de $\mathbf{X} = \{X_1, ..., X_n\}$ y $\mathcal{P}$ una distribución de probabilidad conjunta sobre $\mathbf{X}$; entonces la red cumple la *propiedad de Markov* si y sólo si, cada nodo X_i es condicionalmente independiente de sus no descendientes, $nd(x_i)$, dado sus padres, $pa(x_i)$ (Korb and Nicholson [25]). Es decir,

$$P(x_i \mid pa(x_i), nd(x_i)) = P(x_i \mid pa(x_i)) \tag{2.2}$$

En la siguiente Figura 2.2 se muestra la propiedad de Markov.

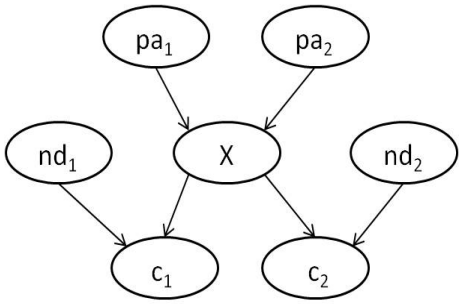

Figura 2.2: Propiedad de Markov.

Esta propiedad, permite una definición de red Bayesiana equivalente a la propiedad de factorización de la distribución de probabilidad conjunta recogida en 2.1, a través de los siguientes teoremas.

Teorema 2.1 *Todo par* $(\mathcal{G}, \mathcal{P})$ *que cumple la propiedad de Markov, constituye una red Bayesiana.*

Demostración

Aplicando la regla de la cadena, la distribución de probabilidad conjunta de $\mathbf{X} = \{X_1, ..., X_n\}$ puede ser factorizada por:

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(x_i | x_1, ...x_{i-1}).$$

Considerando que los nodos están en orden ancestral, entonces el conjunto $\{X_1, ...X_{i-1}\}$ contiene a todos los padres de X_i y a los nodos no descendientes de X_i. Luego, aplicando la propiedad de Markov,

$$P(x_i \mid x_1, ..., x_{i-1}) = P(x_i \mid pa(x_i)).$$

y entonces, la distribución de probabilidad conjunta de $\mathbf{X}$ puede ser factorizada según la definición de red Bayesiana recogida en 2.1.

Teorema 2.2 *Toda red Bayesiana formada por el par* $(\mathcal{G}, \mathcal{P})$*, cumple la propiedad de Markov.*

La demostración consiste en tomar la distribución de probabilidad conjunta de $\mathbf{X}$ y escribir la factorización 2.1 que se cumple por ser red Bayesiana. Luego, si se toma cualquier subconjunto de $nd(x_i)$, es evidente que cumple la propiedad de Markov.

En resumen, si se consideran estos dos teoremas recíprocos, se puede concluir que toda red Bayesiana definida por el par $(\mathcal{G}, \mathcal{P})$, cumple dos propiedades equivalentes: la factorización presentada en 2.1 y la propiedad de Markov presentada en 2.2.

2.1.3. Tipos de redes Bayesianas

La variable aleatoria $\mathbf{X}$ considerada en la red Bayesiana, puede ser de tipo discreta y/o continua, lo que determinará familias paramétricas específicas en 2.1. Así, se pueden distinguir tres tipos principales de redes Bayesianas:

Redes Bayesianas discretas

En este caso, se considera que todas las variables aleatorias de $\mathbf{X} = \{X_1, ..., X_n\}$ son discretas, es decir, cada variable sólo puede tomar un número finito de posibles estados. Además, la distribución de probabilidad de cada variable, condicionada a sus padres, es multinomial, por lo que dicha distribución queda especificada por tablas de probabilidades para las posibles combinaciones de estados entre las variables involucradas.

Ejemplo 2 *La red Bayesiana que se presenta a continuación, cuyo DAG y correspondientes tablas de probabilidad condicionadas se muestran en la Figura 2, consiste en un modelo de diagnóstico médico. Comienza por considerar la presencia (o ausencia) de* Metástasis *(M), la que podría causar un* Tumor cerebral *(T) e* Incremento en los niveles de calcio *(I). Estas dos variables, Tumor e Incremento de calcio, podrían causar un estado de* Coma *(C), y el tumor cerebral puede procucir además fuertes* Jaquecas *(J). Todas las variables consideradas son discretas dicotómicas que tienen como posible resultado Verdadero (V) en caso de presencia del síntoma o enfermedad, y Falso (F) en caso de ausencia.*

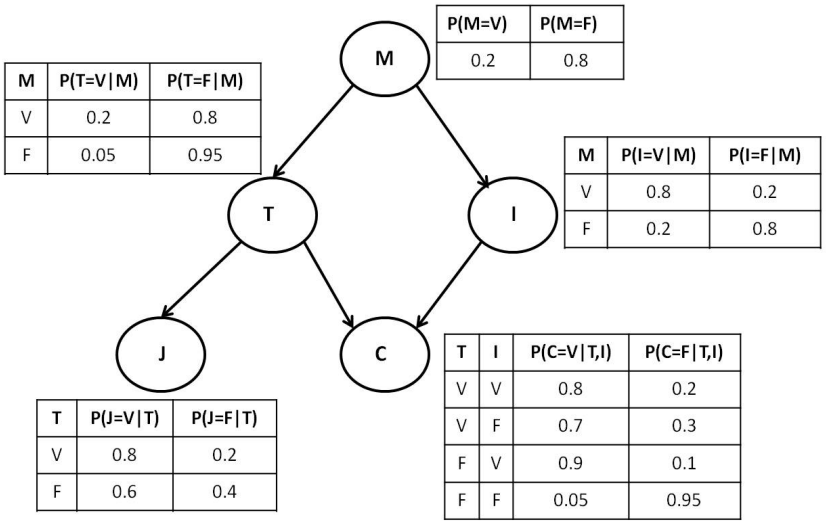

Figura 2.3: Ejemplo 2: red Bayesiana discreta.

Luego, considerando la factorización presentada en 2.1, se puede calcular la probabilidad conjunta como:

$$P(M,T,I,J,C) = P(M)P(T \mid M)P(I \mid M)P(J \mid T)P(C \mid T,I)$$

Una vez conocida la distribución de probabilidad conjunta, es posible conocer la distribución de cualquier conjunto de variables, incorporando evidencia acerca de una o más variables de la red. Así por ejemplo, si se quiere saber la probabilidad de Metástasis, para un paciente que ha estado en Coma y que no presenta Jaqueca, entonces hay que calcular:

$$\begin{aligned} P(M \mid J = F, C = V) &= \frac{P(M, J = F, C = V)}{P(J = F, C = V)} \\ &= \frac{\sum_{I,T} P(M, I, T, J = F, C = V)}{\sum_{M,I,T} P(M, I, T, J = F, C = V)} \end{aligned} \tag{2.3}$$

Este cálculo, a pesar de ser una red con sólo 5 variables, no es simple, requiere un importante esfuerzo computacional. Para poder obtener conclusiones respecto a la incorporación de evidencia, se implementó el ejemplo en el programa Hugin (Hugin

Lite 7.4 [22]). La Figura 2.4 muestra la red inicial, sin incorporar evidencia, y luego, en la Figura 2.5 se puede observar las probabilidades condicionadas en la evidencia propuesta en 2.3.

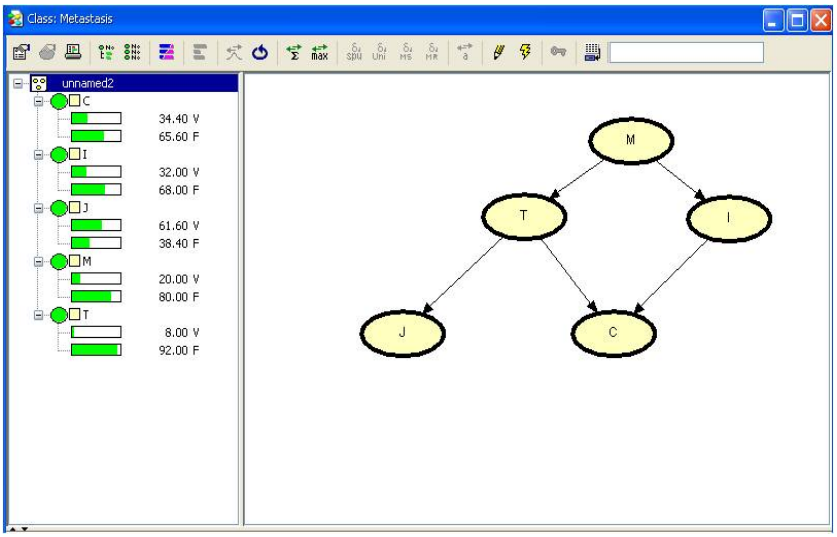

Figura 2.4: Red Ejemplo 2 en Hugin.

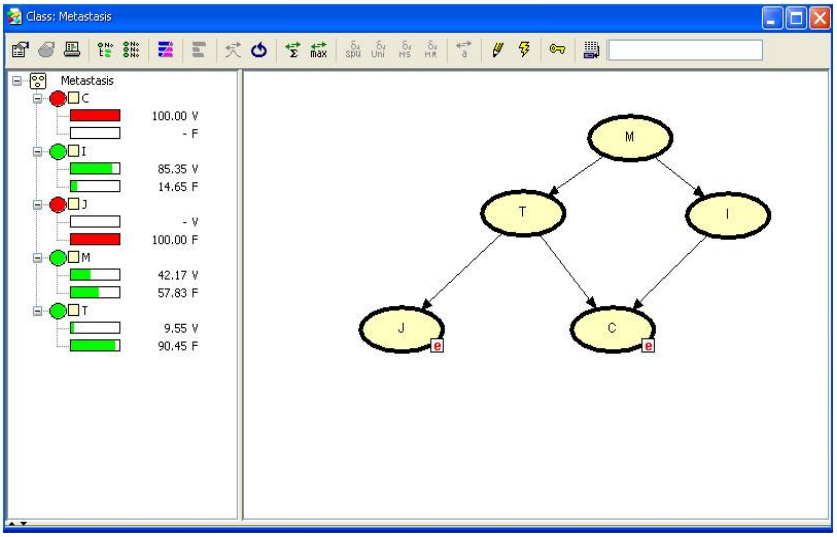

Figura 2.5: Red Ejemplo 2 con evidencia en Hugin.

Luego, la probabilidad de tener metástasis, una vez observado que el paciente ha estado en coma y no ha tenido jaqueca, es:

$$P(M = V \mid J = F, C = V) \;\; = \;\; 0.42 \tag{2.4}$$

Es decir, la probabilidad de metástasis aumentó de 0.2 inicial a 0.42 al conocerse esta evidencia. También es interesante observar qué pasó con la variable Incremento de calcio; inicialmente (ver Figura 2.4) la probabilidad marginal de presentar Incremento era de 0.32, sin embargo, conocida la evidencia, ésta aumenta a 0.85, lo cual tiene mucho sentido al observar la red, ya que la ausencia de jaqueca hace que sea razonable pensar que el estado de coma se deba al Incremento de Calcio más que a un Tumor cerebral, de hecho la probabilidad de Tumor aumentó muy poco al conocerse la presencia de estado de Coma.

Redes Bayesianas Gaussianas (RBG)

Al considerar redes Bayesianas con variables de tipo continuas, el modelo más comúnmente utilizado son las redes Bayesianas Gaussianas, donde se asume que las variables tienen distribución Normal por tanto la relación entre ellas es lineal. Sin embargo, esta no es la única alternativa; por ejemplo, en el trabajo de Main y Navarro [30] se propone un modelo más general considerando que la distribución conjunta pertenece a la familia Potencial Exponencial Multivariante (MEP)introducida por Gómez et al. [16]. Otra variedad a estos modelos se puede encontrar en el trabajo publicado por Driver and Morrell [13] donde se propone aproximar las densidades condicionadas usando distribuciones de sumas Gaussianas ponderadas.

En redes Bayesianas Gaussianas, la distribución de probabilidad conjunta asociada a las variables $\mathbf{X} = (X_1, X_2, ..., X_n)$ tiene distribución Normal Multivariante $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con función de densidad conjunta dada por,

$$f(\mathbf{x}) = (2\pi)^{-n/2} \left|\boldsymbol{\Sigma}\right|^{-1/2} exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right\} \tag{2.5}$$

donde $\boldsymbol{\mu}$ es el vector de medias con dimensión $n \times 1$, $\boldsymbol{\Sigma}$ la matriz definida positiva de covarianzas.

De acuerdo a las propiedades de la distribución Normal y a la factorización presentada en 2.1, la densidad conjunta de una red Bayesiana Gaussiana puede ser escrita como el producto de densidades de probabilidad condicionadas, donde cada una de ellas corresponde a una distribución Normal tal que:

$$f(x_i \mid pa(x_i)) \sim N\left(\mu_i + \sum_{j=1}^{i-1} \beta_{ji}(x_j - \mu_j), v_i\right), \tag{2.6}$$

donde β_{ji} es el coeficiente de regresión de X_j en la regresión de X_i condicionada a sus padres y $v_i = \Sigma_i - \Sigma_{ipa(X_i)}\Sigma^{-1}_{pa(X_i)}\Sigma^T_{ipa(X_i)}$ es la varianza condicionada de X_i dado sus padres, con Σ_i la varianza marginal de X_i, $\Sigma_{ipa(X_i)}$ el vector de covarianzas entre X_i y las variables del conjunto $pa(X_i)$, y $\Sigma_{pa(X_i)}$ la matriz de covarianzas de los $pa(X_i)$.

El coeficiente de regresión β_{ji} cuantifica el grado de relación entre X_i y X_j, luego, si $\beta_{ji} = 0$ $(j < i)$, entonces X_j no es padre de X_i, es decir, no existe una arista dirigida entre ellas.

Ejemplo 3 *A continuación se presenta una RBG estudiada por Castillo and Kjaerulff [6]. El objetivo es evaluar los daños en estructuras de hormigón armado en edificios. Para esto, se determina una variable de interés, X_{24}, la que indica el daño observado en una viga. Un ingeniero civil identifica 16 variables (de X_1 a X_{16}) como las principales variables que influyen en el daño de la estructura. Además, el ingeniero también define 7 variables intermedias no observables (de X_{17} a X_{23}) que indican algunos estados parciales de la estructura. Todas las variables fueron medidas en una escala construída en directa relación con la variable objetivo, es decir, entre más alto el valor, mayor posibilidad de daño. En Castillo and Kjaerulff [6] se encuentra en detalle la definición de cada variable. En la Figura 2.6 se muestra el DAG correspondiente a la parte cualitativa de la red. Para definir la parte cuantitativa de acuerdo a 3.1, se considera que todas las medias de todas las variables son iguales a cero, que los coeficientes de regresión β_{ji} se definen como se muestra en la Figura 2.6 y que las varianzas condicionadas están dadas por:*

$$v_i = \begin{cases} 10^{-4}, & \textit{si } X_i \textit{ no es observable} \\ 1, & \textit{en otro caso} \end{cases}$$

En esta Memoria, se centrará el trabajo en el estudio de redes Bayesianas Gaussianas, por lo que se dedicará una sección posterior a revisar propiedades y técnicas

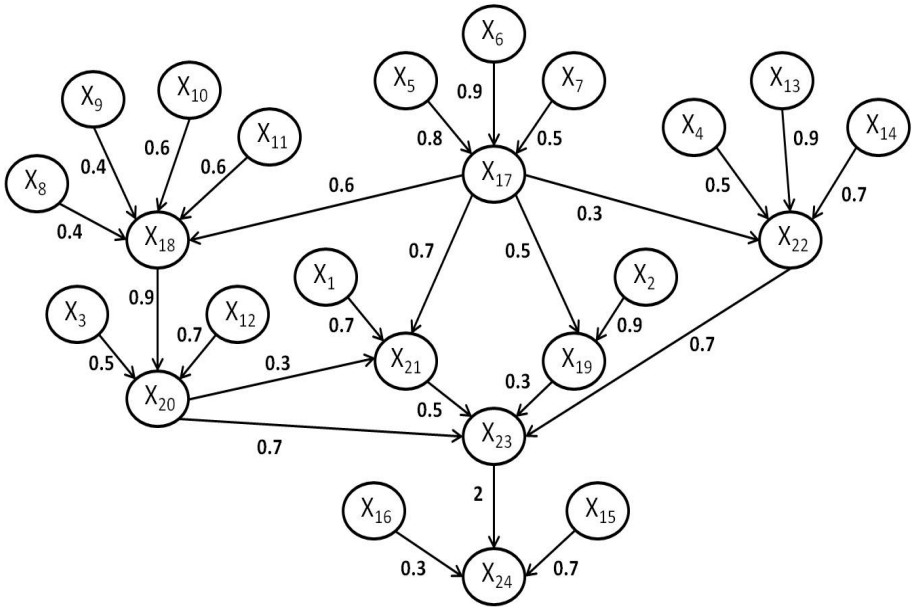

Figura 2.6: Ejemplo 3: red Bayesiana Gaussiana.

de este tipo de redes; principalmente cómo definir los parámetros de la distribución Normal Multivariante $N(\boldsymbol{\mu}, \Sigma)$ y cómo propagar la evidencia a lo largo de la red.

Redes Bayesianas mixtas

En redes Bayesianas mixtas, o también conocidas como redes Bayesianas híbridas, el conjunto de variables considerado $\mathbf{X} = (X_1, X_2, ..., X_n)$ contiene variables de tipo continuo y discreto. Este tipo de redes presentan una alta complejidad cuando se quiere realizar inferencia en la red, es decir, al propagar información.

Un caso particular de este tipo de redes son las redes Gaussianas condicionadas lineales; las primeras en que se desarrollaron métodos exactos de propagación (Lauritzen [28]). Así, siguiendo la notación presentada en Kjaerulff and Madsen [24] y Cobb et al. [8], se considera que $\mathbf{X}$ se particiona en dos, $\mathbf{X} = \mathbf{Y} \cup \mathbf{Z}$, donde $\mathbf{Y}$ representa las variables discretas y $\mathbf{Z}$ a las continuas. Luego, cada nodo en el DAG $\mathcal{G}$ representa una variable discreta con un conjunto finito de estados (exclusivos y excluyentes), o una variable continua con distribución condicionada lineal Gaussiana en sus padres discre-

tos. Si la variable continua tiene padres continuos, entonces la media de la distribucion condicionada Gaussiana dependerá linealmente del estado de sus padres continuos.

Luego, en un modelo condicionado Gaussiano, la distribución de una variable continua, dado sus padres discretos, será una distribución Gaussiana Multivariante:

$$\mathbf{Z} \mid \mathbf{Y} = \mathbf{y} \sim N(\boldsymbol{\mu}(\mathbf{y}), \Sigma(\mathbf{y})) \tag{2.7}$$

con Σ definida positiva.

Una importante restricción en este tipo de redes es que toda variable aleatoria discreta, sólo puede tener padres discretos, no se permiten padres continuos para hijos discretos.

Definición 2.1 *Una variable aleatoria mixta* $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ *tiene* ***distribución condicionada Gaussiana*** *si la distribución conjunta tiene densidad:*

$$f(\mathbf{x}) = f(\mathbf{y}, \mathbf{z}) = \chi(\mathbf{y}) exp\{g(\mathbf{y}) + h(\mathbf{y})^T \mathbf{z} - \mathbf{z}^T K(\mathbf{y}) \mathbf{z}/2\}$$

donde $\chi(\mathbf{y}) \in \{0, 1\}$ *indica si* f *es positiva en* $\mathbf{y}$*,* g *es una función real,* h *es un vector y* K *una matriz.*

Es usual, para distinguir entre variables discretas y continuas, representar el DAG de una red Bayesiana mixta utilizando doble-óvalos para las variables continuas.

Ejemplo 4 *Se presenta un ejemplo muy sencillo introducido por Kjaerulff and Madsen [24] donde* $\mathbf{X} = \{X_1, X_2, X_3\}$ *con* $\mathbf{Y} = \{X_1\}$ *es una variable aleatoria discreta binaria (toma valores V o F) y* $\mathbf{Z} = \{X_2, X_3\}$ *son variables continuas. El correspondiente DAG se muestra en la Figura 2.7.*

Luego, es necesario especificar la parte cuantitativa de la red, es decir, las distribuciones de probabilidad condicionadas del modelo:

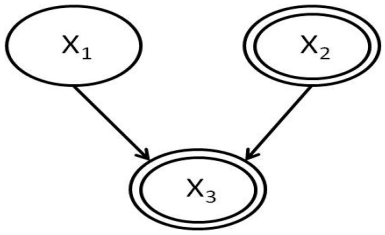

Figura 2.7: Ejemplo 4: red Bayesiana Gaussiana condicionada lineal.

- *La variable X_3 sigue una distribución Gaussiana condicionada lineal tal que:*

$$f(x_3 \mid x_1 = V, x_2) \sim N(-5 + (-2 * x_2), 1.1))$$
$$f(x_3 \mid x_1 = F, x_2) \sim N(5 + (2 * x_2), 1.2))$$

- *La variable $X_2 \sim N(0, 1)$.*
- $P(x_1) = (0.75, 0.25)$

Calcular la distribución conjunta en este tipo de redes, y luego hacer inferencia, requiere cálculos computacionales más complejos. Una primera propuesta para realizar propagación en este caso, fue planteada por Lauritzen [28], basando los cálculos computaciones en un esquema de árbol de unión.

En forma más reciente, Moral et al.[32] han propuesto representar la distribución de las variables de la red, discretas y continuas, usando una mixtura de exponenciales truncadas (MTE). Este modelo tiene la ventaja, respecto a las redes Gaussianas condicionadas lineales, de permitir que variables discretas tengan padres continuos; es decir, no pone restricciones respecto a cómo se relacionan las variables dentro de la red. Más detalles del modelo y sobre cómo realizar inferencia y aprendizaje en este tipo de redes, se pueden ver en Cobb et al. [8].

Capítulo 3

Redes Bayesianas Gaussianas (RBG)

Como ya se ha mencionado, en este trabajo se consideran redes Bayesianas Gaussianas, es decir redes donde la distribución de probabilidad conjunta asociada a las variables $\mathbf{X} = (X_1, X_2, ..., X_n)$ tiene distribución Normal Multivariante. Debido a la importancia que tiene este tipo de redes para este trabajo, se ha dedicado un capítulo completo a ellas. Se comienza por detallar su definición y propiedades, para luego dedicar una sección a la revisión del supuesto de Normalidad y de qué manera se puede hacer cumplir este supuesto.

3.1. Definición y caracterización

A continuación se revisará con más detalle la definición, construcción, propiedades y método de propagación de la evidencia, en redes Bayesianas Gaussianas.

3.1.1. Definición

Definición 3.1 *Una **red Bayesiana Gaussiana (RBG)** corresponde a un tipo de red Bayesiana donde la densidad de probabilidad conjunta asociada con* $\mathbf{X}$ *es una distribución Normal Multivariante* $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Comúnmente, al modelar una situación de incertidumbre utilizando una red Bayesiana, se considera una de las variables como variable respuesta de interés. Luego, a través de la inferencia en la red, se busca conocer principalmente la distribución marginal condicionada de esa variable. En ese sentido, las redes Bayesianas Gaussianas, gracias a las propiedades de la distribución Normal Multivariante, permiten obtener información relevante sin necesidad de una alta complejidad en los cálculos.

3.1.2. Construcción de RBG

Como ya se mencionó, en una RBG la densidad conjunta se puede escribir como el producto de las siguientes densidades condicionadas:

$$f(x_i \mid pa(x_i)) \sim N\left(\mu_i + \sum_{j=1}^{i-1} \beta_{ji}(x_j - \mu_j), v_i\right), \tag{3.1}$$

donde β_{ji} es el coeficiente de regresión de X_j en la regresión de X_i condicionada a sus padres y $v_i = \Sigma_i - \Sigma_{ipa(X_i)}\Sigma^{-1}_{pa(X_i)}\Sigma^T_{ipa(X_i)}$ es la varianza condicionada de X_i dado sus padres, con Σ_i la varianza de X_i, $\Sigma_{ipa(X_i)}$ el vector de covarianzas entre X_i y las variables del conjunto $pa(X_i)$, y $\Sigma_{pa(X_i)}$ la matriz de covarianzas de los $pa(X_i)$.

En relación a cálculos que se desarrollarán más adelante, es importante notar que la media condicionada $\mu_{x_i|pa(x_i)}$ depende de los valores que tomen los padres de X_i; en cambio la varianza condicionada no, ésta sólo depende de la estructura de covarianzas de X_i con sus padres.

Para construir una red Bayesiana, discreta o Gaussiana, se deben resolver las siguientes tres tareas:

1. **Determinar la estructura** Esta tarea consiste en determinar el DAG de la red, lo que se ha llamado la parte cualitativa. Básicamente consiste en determinar las relaciones de dependencia entre las variables involucradas en un estudio.

2. **Determinar los parámetros** Consiste en derminar la parte cuantitativa de la red, es decir, la función de densidad marginal o condicionada, de cada nodo de acuerdo a la estructura antes definida.

3. **Propagación de evidencia** Una vez que se ha definido la parte cualitativa y cuantitativa de la red, interesa conocer cómo cambian las distribuciones de probabilidad a lo largo de la red cuando los valores de algunas de las variables son conocidos.

Respecto a las tareas definidas en 1 y 2, se debe distinguir entre dos enfoques para la construcción de la red:

- Enfoque tradicional
- Enfoque de aprendizaje

A continuación se presenta cada uno de ellos, planteando en cada caso posibles metodologías para el caso particular que interesa en este trabajo: las redes Bayesianas Gaussianas.

Enfoque tradicional

En este caso, la construcción de la RBG se hace con un grupo de expertos; quienes definen las variables que se deben considerar, la forma en que ellas se relacionan y las distribuciones de probabilidad condicionada de cada una dado sus padres. Para realizar este trabajo, hay dos posibles caminos:

A partir de la distribución conjunta

Es necesario definir los parámetros involucrados en la distribución conjunta asociada al DAG, es decir, los parámetros de 2.5; estos son:

- $\boldsymbol{\mu}$: el vector de medias con dimensión $n \times 1$, donde cada μ_i es la media de la variable X_i para $i = 1, .., n$.
- $\boldsymbol{\Sigma}$: la matriz (definida positiva) de covarianzas de dimensión $n \times n$, donde σ_{ii} es la varianza de X_i y σij es la covarianza entre X_i y X_j, para $i, j = 1, .., n$.

A partir de las distribuciones condicionadas

Ahora, los parámetros que se necesitan definir están determinados por las distribuciones condicionadas de la expresión 3.1; es decir:

- $\boldsymbol{\mu}$: el vector de medias con dimensión $n \times 1$, donde cada μ_i es la media de la variable X_i para $i = 1, .., n$.
- los coeficientes de regresión β_{ji} de X_j en la regresión de X_i sobre $pa(X_i)$, para $i, j = 1, .., n$.

  Los dos conjuntos de parámetros anteriores permitirán definir las medias condicionadas $E(X_i \mid pa(X_i))$.
- las varianzas condicionadas v_i de cada X_i dado sus padres, con $i = 1, .., n$.

Enfoque de aprendizaje

El procedimiento de aprendizaje consiste en determinar tanto el DAG como las distribuciones de probabilidad, a partir de una base de datos.

Respecto a los algoritmos de aprendizaje de la estructura de una red Bayesiana se pueden distinguir tres grandes grupos:

- Algoritmos basados en restricciones: En este caso, la estructura de la red se determina aplicando test de independiencia condicionada de acuerdo a la propiedades de Markov presentes en toda red Bayesiana. Se busca construir un grafo que cumpla con las reglas de d-separación.
- Algoritmos basados en puntuaciones o medidas: Este algoritmo asigna una puntuación a toda posible red bayesiana que se pueda construir con los datos, e intenta maximizar dicha puntuación con mecanismos de búsqueda heurística.
- Algoritmos híbridos: Corresponde a una combinación de algoritmos basados en restricciones y algoritmos basados en puntuaciones. En general, los primeros se utilizan para limitar el espacio de búsqueda a través de test de independencia condicionada, y los segundos para encontrar, por medio de alguna medida, una red óptima en dicho espacio reducido.

En la Sección 6 de esta Tesis, se ajusta una RBG bajo el enfoque de aprendizaje, utilizando el algoritmo híbrido $max - min\ hill - climbing$ (mmhc) introducido por Tsamardinos et al [44]. El objetivo del algoritmo mmhc consiste en ajustar un grafo acíclico dirigido de una red Bayesiana a partir de un conjunto de datos. Corresponde a una combinación del algoritmo basado en restricciones *max-min parents and children* (mmpc) introducido por Tsamardinos et al. [43] y al algoritmo basado en puntuaciones $hill-climbling$ (hc) (detalles en Russell and Norvig [36]); en forma resumida, el primer algoritmo permite construir el esqueleto de la red, y el segundo establece la dirección de los enlaces.

3.1.3. Propiedades

Normalmente, dentro del enfoque tradicional, para los investigadores resulta más cómodo construir una RBG definiendo los parámetros de las distribuciones condicionadas (expresión 3.1). Para calcular luego la distribución conjunta, de acuerdo a 2.5, será necesario obtener la matriz de covarianzas $\mathbf{\Sigma}$.

Descomposición de la matriz $\mathbf{\Sigma}$

En el trabajo de Shachter y Kenley [38], además del algoritmo correspondiente, se muestra que la matriz $\mathbf{\Sigma}$, siendo definida positiva, se puede calcular como:

$$\mathbf{\Sigma} = [(\mathbf{I} - \mathbf{B})^{-1}]^T \mathbf{D} (\mathbf{I} - \mathbf{B})^{-1} \tag{3.2}$$

donde $\mathbf{D}$ es una matriz diagonal formada por las varianzas condicionadas v_i, $\mathbf{B}$ es una matriz triangular estrictamente superior de dimensión $n \times n$ con los coeficientes de regresión β_{ji} e $\mathbf{I}$ es la matriz identidad de dimensión $n \times n$.

Así, a partir de los parámetros definidos en las densidades condicionadas, es posible obtener la matriz $\mathbf{\Sigma}$ y por tanto, definir la distribución conjunta de la RBG.

En forma alternativa, se puede obtener la matriz de precisión $\mathbf{W} = \mathbf{\Sigma}^{-1}$ a partir de la siguiente fórmula recursiva presentada en el mismo trabajo de Shachter y Kenley [38]:

$$W(1) = \frac{1}{v_i}$$

$$W(i+1) = \begin{pmatrix} W(i) + \frac{\beta_{i+1}\beta_{i+1}^T}{v_{i+1}} & -\frac{\beta_{i+1}}{v_{i+1}} \\ -\frac{\beta_{i+1}^T}{v_{i+1}} & \frac{1}{v_{i+1}} \end{pmatrix} i = 1, ..., n.$$

donde $W(i)$ es una submatriz $i \times i$ y β_i es el vector columa $\{\beta_{ji}$ con $j < i\}$.

Independencia e Independencia condicionada

A continuación se revisan algunas propiedades conocidas de la distribución Normal Multivariante que permiten determinar la independencia y la independencia condicionada entre dos variables, a partir de la estructura de ceros de la matriz de covarianzas $\mathbf{\Sigma}$.

Proposición 3.1.1 *Sea* $\mathbf{X}$ *con distribución Normal Multivariante* $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ *tal que* $\mathbf{X}$ *se particiona como* $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}\}$ *con* $\boldsymbol{\mu}$ *y* $\mathbf{\Sigma}$ *dados por*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \quad y \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$$

entonces, se cumple:

- $\mathbf{X_1} \sim N(\boldsymbol{\mu_1}, \mathbf{\Sigma_{11}})$ y $\mathbf{X_2} \sim N(\boldsymbol{\mu_2}, \mathbf{\Sigma_{22}})$
- $\mathbf{X_1}$ *y* $\mathbf{X_2}$ *son independientes si y sólo si* $\mathbf{\Sigma_{12}} = \mathbf{\Sigma_{21}} = \mathbf{0}$

Como ya se dijo, cuando la matriz $\mathbf{\Sigma}$ es definida positiva, entonces su inversa existe y la matriz de precisión se puede obtener como $\mathbf{W} = \mathbf{\Sigma}^{-1}$. Luego, considerando la estructura de independencia introducida en la última Proposición, se puede obtener el siguiente resultado sobre independencia condicionada.

Proposición 3.1.2 *Sea* $\mathbf{X}$ *con distribución Normal Multivariante* $N(\boldsymbol{\mu}, \mathbf{\Sigma})$*, si* $\mathbf{\Sigma}$ *es una matriz invertible, entonces las variables* X_i *y* X_j *son condicionalmente independientes dado el resto de las variables de* $\mathbf{X}$*, si y sólo si,* $W_{ij} = 0$.

A continuación, se mostrarán estos cálculos y propiedades con un ejemplo.

Ejemplo 5 *Este problema, presentado por Gómez-Villegas et al. [18], trata sobre el tiempo de duración de una máquina trabajando. La máquina está formada por siete elementos, conectados como se muestra en el DAG de la Figura 3.1.*

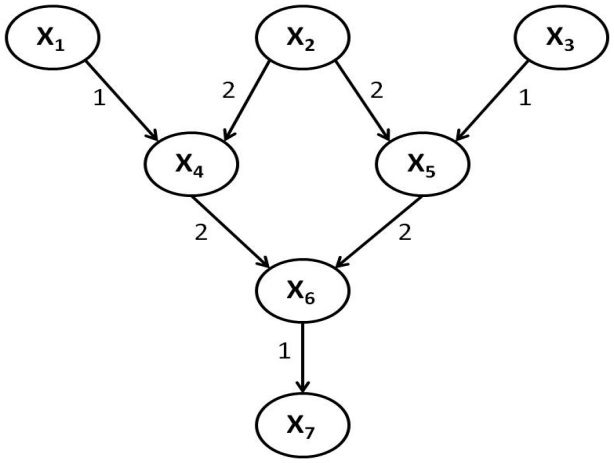

Figura 3.1: DAG Ejemplo 5.

El nodo de interés es X_7 y la parte cuantitativa de la red corresponde a una Normal Multivariante con los siguientes parámetros.

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 4 \\ 5 \\ 8 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix},$$

con $\boldsymbol{\mu}$ vector de medias, $\mathbf{D}$ matriz diagonal de varianzas v_i, $\mathbf{B}$ matriz triangular superior con los coeficientes de regresión β_{ji} para X_j padre de X_i.

Usando la expresión presentada en 3.2, se obtiene que la variable aleatoria $\mathbf{X} = \{X_1, X_2, X_3, X_4, X_5, X_6, X_7\} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con parámetros,

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 4 \\ 5 \\ 8 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 2 & 2 \\ 0 & 1 & 0 & 2 & 2 & 8 & 8 \\ 0 & 0 & 2 & 0 & 2 & 4 & 4 \\ 1 & 2 & 0 & 6 & 4 & 20 & 20 \\ 0 & 2 & 2 & 4 & 10 & 28 & 28 \\ 2 & 8 & 4 & 20 & 28 & 97 & 97 \\ 2 & 8 & 4 & 20 & 28 & 97 & 99 \end{pmatrix}$$

A continuación, se muestra la matriz de precisión obtenida a partir de $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$:

$$\mathbf{W} = \begin{pmatrix} 2 & 2 & 0 & -1 & 0 & 0 & 0 \\ 2 & 6 & 0.5 & -2 & -0.5 & 0 & 0 \\ 0 & 0.5 & 0.75 & 0 & -0.25 & 0 & 0 \\ -1 & -2 & 0 & 5 & 4 & -2 & 0 \\ 0 & -0.5 & -0.25 & 4 & 4.25 & -2 & 0 \\ 0 & 0 & 0 & -2 & -2 & 1.5 & -0.5 \\ 0 & 0 & 0 & 0 & 0 & -0.5 & 0.5 \end{pmatrix}$$

Alternativamente, se podría haber obtenido $\mathbf{W}$ por el método recursivo antes presentado y luego haber calculado $\boldsymbol{\Sigma} = \mathbf{W}^{-1}$.

Para aplicar los conceptos de independencia e independencia condicionada, a partir de las matrices $\boldsymbol{\Sigma}$ y $\mathbf{W}$ respectivamente, a continuación se van a revisar los resultados obtenidos en relación a la variable X_7:

- A partir de la matriz $\boldsymbol{\Sigma}$ y siguiendo la estructura del DAG de la Figura 3.1, se puede ver que la variable X_7 no es independiente de ninguna de las otras variables contenidas en $\mathbf{X}$, lo que se traduce en no observar ningún *cero* en la columna 7 (o fila 7) de la matriz de covarianzas.
- Sin embargo, siguiendo las propiedades de independencia condicionada, se sabe que X_7 es condicionalmente independiente de X_i dado el resto de las variables para $i = 1, ..., 5$, lo que se ve reflejado en los correspondientes *ceros* de la matriz de precisión $\mathbf{W}$.

3.1.4. Propagación de la evidencia

En problemas que consideran situaciones de la vida real, la información sobre el estado de una o más variables, conocidas como variables evidenciales, podría encontrarse disponible. En este caso, es de interés incorporar dicha información a la red y en base a ella, actualizar las distribuciones de probabilidad del resto de las variables; este proceso es conocido como propagación de evidencia.

Se han propuesto distintas metodologías para propagar evidencia en una red Bayesiana Gaussiana. En este trabajo, se considera el método paso a paso propuesto por Castillo et al. [5], que consiste en calcular la densidad condicionada de una distribución Normal cada vez que se introduce, una a una, cada evidencia en la variable evidencial. Luego, considerando el conjunto de variables no evidenciales $\mathbf{Y}$ y variables evidenciales $\mathbf{E} = \{X_e\}$, entonces, $\mathbf{X}$ se puede escribir como la partición $\mathbf{X} = (\mathbf{Y}, E)$ y la distribución condicionada de $\mathbf{Y}$ dado $E = e$ es una distribucion Normal Multivariante con parámetros:

$$\boldsymbol{\mu}^{\mathbf{Y}|\mathbf{E}=\mathbf{e}} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YE}}{\boldsymbol{\Sigma}_{\mathbf{EE}}}^{-1}(\mathbf{e} - \boldsymbol{\mu}_{\mathbf{E}}) \tag{3.3}$$

y

$$\boldsymbol{\Sigma}^{\mathbf{Y}|\mathbf{E}=\mathbf{e}} = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YE}}{\boldsymbol{\Sigma}_{\mathbf{EE}}}^{-1}\boldsymbol{\Sigma}_{\mathbf{EY}} \tag{3.4}$$

Si el interés está en la densidad marginal a posteriori de sólo una variable considerada como respuesta, $X_i \, \epsilon \, \mathbf{Y}$, y una variable evidencial E, entonces, al propagar la evidencia se obtiene:

$$X_i \mid E = e \sim N\left(\mu_i^{Y|E=e}, \sigma_{ii}^{Y|E=e}\right) = N\left(\mu_i + \frac{\sigma_{ie}}{\sigma_{ee}}(e - \mu_e), \sigma_{ii} - \frac{\sigma_{ie}^2}{\sigma_{ee}}\right)$$

con parámetros previos a la propagación, μ_i y μ_e las medias de X_i y E, σ_{ii} y σ_{ee} las varianzas de X_i y E, y σ_{ie} la covarianza entre X_i y E.

Para explicar esta metodología, se retoma el ejemplo 5 presentado por Gómez-Villegas et al. [18].

Ejemplo 6 *Se vuelve a considerar la RBG del Ejemplo anterior, representada por el DAG de la Figura 3.1, donde X_7 se definió como la variable respuesta. Ahora, se considerará que se conoce el valor de las variables X_1, X_2 y X_3:* $\mathbf{E} = \mathbf{e} = \{X_1 = 2, X_2 = 2, X_3 = 1\}$. *Luego, interesa conocer la distribución de probabilidad a posteriori de las variables no evidenciales, una vez que se realiza la correspondiente propagación de la evidencia contenida en* $\mathbf{E}$.

Incorporando las variables evidenciales una a una, de acuerdo a las expresiones 3.3 y 3.4, se obtiene que $\mathbf{Y}|\mathbf{E} = \mathbf{e} \sim N(y \mid \mu^{\mathbf{Y}|\mathbf{E}=e}, \Sigma^{\mathbf{Y}|\mathbf{E}=e})$ con,

$$\mu^{\mathbf{Y}|\mathbf{E}=\mathbf{e}} = \begin{pmatrix} 0 \\ 1 \\ -3 \\ 0 \end{pmatrix} \qquad \Sigma^{\mathbf{Y}|\mathbf{E}=\mathbf{e}} = \begin{pmatrix} 1 & 0 & 2 & 2 \\ 0 & 4 & 8 & 8 \\ 2 & 8 & 21 & 21 \\ 2 & 8 & 21 & 23 \end{pmatrix}$$

Es interesante notar que la distribución marginal de la variable respuesta X_7 en la red original era $X_7 \sim N(8, 99)$, lo que implica una considerable incertidumbre debido a una alta variabilidad. Sin embargo, después de propagar la evidencia, la distribución marginal actualizada para esta variable es $X_7 \sim N(0, 23)$, es decir, la incertidumbre disminuyó en forma importante al incorporar la información disponible.

A partir de este ejercicio, se genera la siguiente pregunta: ¿son estas variables y estos valores, ($\mathbf{E} = \{X_1 = 2, X_2 = 2, X_3 = 1\}$), la mejor opción para reducir la incertidumbre de X_7? El trabajo que se desarrolla en esta Tesis, contiene las herramientas necesarias para responder a esta pregunta.

3.2. Supuesto de Normalidad en RBG

Como ya se ha mencionado, se consideran redes Bayesianas Gaussianas, es decir redes donde la distribución de probabilidad conjunta asociada a las variables $\mathbf{X} = (X_1, X_2, ..., X_n)$ tiene distribución Normal Multivariante. Sin embargo, al trabajar con datos reales y construir una red bajo el enfoque de aprendizaje, dicho supuesto toma especial importancia, ya que debe ser revisado y se debe cerciorar que sea cumplido antes de modelar.

3.2.1. Revisión del supuesto de Normalidad

Para ajustar una RBG el primer supuesto es que la matriz de datos se distribuya mediante una distribución Normal Multivariante. Para revisar este supuesto, se utiliza el test *omnibus* de Doornick-Hansen [12] implementado en el paquete *asbio* de R [1].

Es importante señalar que probar Normalidad Multivariante es una tarea difícil de realizar, ya que cualquier gráfico diagnóstico tradicionalmente utilizado, sólo permite observar dos variables a la vez. Por otro lado, si a partir de una matriz de datos se prueba que cada una de sus variables no se distribuye como Normal Univariante en forma independiente, entonces la matriz no se distribuye mediante una Normal Multivariante. Sin embargo, si todas y cada una de las variables cumple con Normalidad Univariante, no necesariamente se cumple que la matriz completa siga una distribución Normal Multivariante. Esto hace necesario disponer de un test específico para resolver la hipótesis de que los datos tienen distribución Normal Multivariante.

El estadístico de Doornick-Hansen, en el caso univariante, se basa en el estadístico introducido por Bowman y Shenton [4], quienes a su vez proponen un estadístico basado en el cálculo de *kurtosis* y *asimetría* multivariante. Sin embargo, como se menciona en el trabajo de Doornick-Hansen [12], los valores de *kurtosis* y *asimetría* no cumplen con ser independientes aunque son incorrelados, y la *kurtosis* muestral se aproxima de forma muy lenta a la Normalidad. Para superar esta deficiencia, los autores proponen un test derivado del trabajo de Shenton y Bowman [40] quienes asignan a la *kurtosis* una distribución gamma, condicionando a que se cumpla que $kurtosis > asimetría + 1$.

Para el caso multivariante, el test de Doornick-Hansen utiliza una transformación que permite, de forma aproximada, llevar una distribución multivariante a Normales estándar independientes. Luego calculan *kurtosis* y *asimetría* univariantes para cada variable y el test estadístico sigue la misma lógica del test univariante, pero aplicado a vectores de *kurtosis* y *asimetría*.

En el mismo trabajo de los autores, se muestra que el test *omnibus* de Doornick-Hansen propuesto tiene mayor potencia que el test de Shapiro-Wilks; razón por la que se ha escogido como test a utilizar en este trabajo.

3.2.2. Caso de no Normalidad

Frente a una matriz de datos para la cual no se ha pasado el test de Normalidad, surge el problema de no cumplir con los supuestos requeridos para poder ajustar una RBG si se busca trabajar bajo el enfoque de aprendizaje. Frente a esto, se hace necesario transformar la matriz de datos de tal manera que se satisfaga la suposición de Normalidad, sin afectar la estructura de correlación de los datos. Este último hecho es el motivo por el que no es posible aplicar la transformación de Box-Cox.

En el trabajo de Lui et al. [29] se introduce la distribución Nonparanormal como metodología de transformación de datos para el ajuste de grafos no dirigidos. En este trabajo, se propone aplicar a la matriz de datos la transformación Nonparanormal introducida por Lui et al. [29] e implementada en el paquete *huge* [45] de R.

Definición 3.2 *Se dice que el vector* $\mathbf{X} = (X_1, ..., X_n)$ *tiene distribución Nonparanormal (npn) si existen funciones* $\{f_j\}_{j=1}^n$ *tales que:*

$$Z \equiv f(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ con } f(\mathbf{x}) = (f_1(x_1), ..., f_n(x_n))$$

Luego, se tiene que $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f)$, donde las $f_j's$ son funciones monótonas y diferenciables.

Con esto, se tiene que la función de densidad conjunta de X es:

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} exp\{-\frac{1}{2}(f(\mathbf{x}) - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(f(\mathbf{x}) - \boldsymbol{\mu})\} \prod_{j=1}^n |f_j'(x_j)|$$

donde el producto corresponde al jacobiano.

Propiedades

- Si $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f)$ y cada f_j es diferenciable, entonces la estructura de independencia condicionada de los datos se mantiene a través de la matriz $\boldsymbol{\Sigma}^{-1}$.

- La estructura del grafo asociado a la matriz $\mathbf{\Sigma}^{-1}$ de las variables transformadas (ahora Normales) es idéntica a la estructura del grafo de las variables no Normales, esto se refleja en los ceros de $\mathbf{\Sigma}^{-1}$.

Estas dos propiedades, permiten estimar el grafo como si las variables fueran Normales.

Lema 1 *La distribución npn es una cópula Gaussiana cuando las $f_j's$ son monótonas y diferenciables.*

El Lema 1 se obtiene a partir del Teorema de Sklar (Sklar [41]) en el que se indica que cualquier función de distribución acumulada conjunta se puede escribir a partir de las funciones de distribuciones acumuladas marginales como:

$F(x_1, ..., x_n) = C\{F(x_1), ..., F(x_n)\}$ donde C es una función llamada cópula.

En el caso de la distribución npn, se tiene que:

$$F(x_1, ..., x_n) = \Phi_{\boldsymbol{\mu},\mathbf{\Sigma}}(\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_n(x_n)))$$

con $\Phi_{\boldsymbol{\mu},\mathbf{\Sigma}}$ la función de distribución de una Normal Multivariante, y Φ la función de distribución de una Normal estándar Univariante.

Luego, la cópula que se obtiene es de la forma:

$$C(u_1, ..., u_n) = \Phi_{\boldsymbol{\mu},\mathbf{\Sigma}}(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))$$

lo que corresponde a una cópula Gaussiana con parámetros $\boldsymbol{\mu}$ y $\mathbf{\Sigma}$.

Para que la densidad de X, con $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f)$, sea identificable, se debe cumplir que las funciones f_j preserven las medias y las varianzas, es decir:

$\mu_j = \mathrm{E}(Z_j) = \mathrm{E}(X_j)$ y $\sigma_j^2 \equiv \Sigma_{jj} = Var(Z_j) = Var(X_j)$.

Es interesante notar que estas condiciones sólo dependen de la diagonal de la matriz $\boldsymbol{\Sigma}$ y no de la matriz de covarianzas completa.

Sea $F_j(x)$ la función de distribución acumulada de X_j, considerando que la componente $f_j(X_j)$ tiene distribución Normal, entonces:

$$\begin{aligned} F_j(x) &= \mathrm{P}(X_j \leq x) \\ &= \mathrm{P}(Z_j \leq f_j(x)) \\ &= \Phi\left(\frac{f_j(x) - \mu_j}{\sigma_j}\right) \end{aligned}$$

luego,

$$f_j(x) = \mu_j + \sigma_j \ \Phi^{-1}(F_j(x)). \tag{3.5}$$

Método de estimación

A partir de la ecuación 3.5, es posible definir los parámetros y funciones de distribución marginal que es necesario estimar. A continuación se especifican dichos estimadores (Liu et al. [29] y Lafferty et al. [26]).

Sea $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(N)}$ una muestra aleatoria de tamaño N con $\mathbf{X}^{(i)} = (X_1^{(i)}, ..., X_n^{(i)})^T$ y sea $\widetilde{F}_j$ un estimador de F_j, entonces un posible candidato a estimador es la función de distribución marginal empírica:

$$\widehat{F}_j(t) = \frac{1}{N}\sum_{i=1}^{N} 1_{X_j^{(i)} \leq t}$$

Sin embargo, en el contexto de datos de alta dimensión (high dimensional data), los autores Liu et al. proponen utilizar el siguiente estimador *truncado* para acotar la variabilidad de $\widehat{F}_j$:

$$\widetilde{F}_j(x) = \begin{cases} \delta_N & si\ \widetilde{F}_j(x) < \delta_N \\ \widehat{F}_j(x) & si\ \delta_N \leq \widetilde{F}_j(x) \leq 1 - \delta_N \\ (1 - \delta_N) & si\ \widetilde{F}_j(x) > 1 - \delta_N \end{cases}$$

donde δ_N es el parametro de *truncamiento* con valor:

$$\delta_N = \frac{1}{4N^{1/4}\sqrt{\pi \log N}}$$

Una vez definido el estimador para las funciones acumuladas marginales, ya es posible obtener una estimación de las densidades marginales:

$$\widetilde{f}_j(x) = \widehat{\mu}_j + \widehat{\sigma}_j\ \Phi^{-1}(\widetilde{F}_j(x)).$$

donde $\widehat{\mu}_j$ y $\widehat{\sigma}_j$ corresponden a los estimadores muestreales de la media y la desviación estándar:

$$\widehat{\mu}_j \equiv \tfrac{1}{N}\textstyle\sum_{i=1}^{N} X_j^{(i)} \quad \text{y} \quad \widehat{\sigma}_j = \sqrt{\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left(X_j^{(i)} - \widehat{\mu}_j\right)^2}$$

A continuación se muestra un ejemplo donde el interés está primero en comprobar la Normalidad, y luego, al no cumplirse, aplicar la distribución Nonparanormal.

Ejemplo 7 *En este ejemplo, se consideran los parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ definidos en el Ejemplo 5, donde se modela el tiempo de duración de una maquina trabajando formada por siete elementos.*

Utilizando la función rmvt del paquete estadístico mvtnorm de R [15], se generan 200 observaciones de $\mathbf{X} = \{X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$ provenientes de una distribución t-Multivariante de parámetros,

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 2 & 2 \\ 0 & 1 & 0 & 2 & 2 & 8 & 8 \\ 0 & 0 & 2 & 0 & 2 & 4 & 4 \\ 1 & 2 & 0 & 6 & 4 & 20 & 20 \\ 0 & 2 & 2 & 4 & 10 & 28 & 28 \\ 2 & 8 & 4 & 20 & 28 & 97 & 97 \\ 2 & 8 & 4 & 20 & 28 & 97 & 99 \end{pmatrix}$$

Al aplicar el test *omnibus* de Doornick-Hansen [12], el supuesto de Normalidad resulta rechazado. En la Tabla 3.1 se muestra est resultado, donde E corresponde al valor calculado para el estadístico de Doornick-Hansen.

Tabla 3.1: Test de Normalidad Multivariante inicial Ejemplo 7

Var	E	df (2p)	$P(Chi > E)$
X	59.66	14	$1.345e-07$

Para corregir la falta de Normalidad, en este trabajo se propone aplicar la distribución Nonparanormal introducida por Lui et al. [29] mediante la función *huge.npn* del paquete *huge* de R. De esta manera, se obtiene una nueva matriz de datos, que contiene la estructura de correlación de la matriz original, y que se espera, permita aceptar el supuesto de Normalidad. En las siguientes Tablas 3.2 y 3.3 se puede observar que es efectivamente lo que ocurre.

Tabla 3.2: Test de Normalidad Multivariante Ejemplo 7

Var	E	df (2p)	$P(Chi > E)$
X	7.5702	14	0.9105

Tabla 3.3: Test de Normalidad Univariante Ejemplo 7

Var	E	df (2)	$P(Chi > E)$
X_1	0.147	2	0.9292
X_2	0.743	2	0.6898
X_3	0.912	2	0.6337
X_4	0.994	2	0.6073
X_5	1.586	2	0.4524
X_6	1.493	2	0.4739
X_7	1.695	2	0.4285

Finalmente, la propuesta de este trabajo consiste en considerar la matriz de datos **X** con distribución Nonparanormal y de acuerdo con esto obtener la matriz **Z** con distribución Normal. Luego, a esta matriz **Z** aplicar el proceso de ajuste. Esto permite encontrar la RBG considerando en forma correcta el supuesto de Normalidad requerido.

Capítulo 4

Teoría de la Información

La medida de la incertidumbre se realiza con la probabilidad, pero es necesario contar con un mecanismo que mida la sensibilidad y ese mecanismo se va a fijar a partir de la Teoría de la Información. Por esta razón a continuación se presentan las principales definiciones y propiedades para Entropía e Información Mutua.

4.1. Entropía y entropía diferencial

En esta sección, se comienza por presentar la Entropía de Shannon, introducida por Shannon en 1948 (Shannon [39]); como más adelante se considerarán redes Bayesianas cuyas variables tienen distribución Normal Multivariante, estos resultados, que originalmente en su mayoría fueron especificados para variables discretas, serán complementados por sus correspondientes expresiones para el caso de variables continuas(Cover and Thomas [9]).

Definición 4.1 *Sea X una v.a. discreta con probabilidad p(x), entonces la* ***entropía*** *es:*

$$\begin{aligned} H(X) &= -E_{p(x)}[log\ p(X)] \\ &= -\sum_{X} p(x)\ log\ p(x) \end{aligned}$$

Es decir, la *entropía* mide el grado de caos en la distribución de una variable; corresponde a una medida de incertidumbre.

En forma análoga al caso discreto, la entropía se define para variables continuas y se conoce como *entropía diferencial*.

Definición 4.2 *Sea X una variable aleatoria continua, con densidad $f(x)$, entonces la* ***entropía diferencial*** *es,*

$$\begin{aligned} h(X) &= -E_f[log\ f(X)] \\ &= -\int_S f(x)\ log\ f(x)\ dx \end{aligned}$$

donde S es el conjunto soporte de X (el conjunto en que se cumple que $f(x) > 0$).

Para el caso discreto, la entropía es una medida de incertidumbre o aleatoriedad de una variable aleatoria. Sin embargo, intuitivamente hablando, la incertidumbre de una variable continua es infinita (Ihara [21]). Luego, en el caso discreto, la entropía mide incertidumbre en forma *absoluta*, en cambio en el caso continuo, se refiere a una medida *relativa* y debe ser utilizada para hacer comparaciones entre dos o más variables, o en la misma variable bajo diferentes modelos.

Es importante hacer notar que la entropía y la entropía diferencial dependen sólo de la función de masa, o función de densidad según corresponda, de la variable aleatoria; no depende del valor específico que tome la variable. Además, se cumplirá siempre que en el caso discreto $H(X) \geq 0$, en cambio, para el caso continuo $h(X)$ puede tomar valores negativos.

Hasta aquí, se ha considerado la definición de entropía para sólo una variable aleatoria; a continuación, se presentan algunas extensiones para dos o más variables.

4.1.1. Entropía conjunta y entropía condicionada

Se comenzará por revisar el caso en que se consideren dos variables discretas, para luego hacer una generalización en el caso continuo.

Definición 4.3 *Sean dos variables aleatorias X e Y discretas, con su correspondiente función de probabilidad conjunta $p(x, y)$, entonces la* ***entropía conjunta*** *está definida por,*

$$\begin{aligned} H(X,Y) &= -E_{p(x,y)}[log\ p(X,Y)] \\ &= -\sum_X \sum_Y p(x,y)\ log\ p(x,y) \end{aligned}$$

A continuación se define la entropía condicionada de una variable aleatoria discreta dada la información de una segunda variable aleatoria discreta.

Definición 4.4 *Sea Y una v.a. discreta con función de probabilidad $p(y)$, entonces la* ***entropía condicionada*** *de X dado Y es:*

$$\begin{aligned} H(X \mid Y) &= -E_{p(x,y)}[log\ p(X \mid Y)] \\ &= -\sum_Y p(y)\ H(X \mid Y = y) \\ &= -\sum_Y p(y) \sum_X p(x \mid y)\ log\ p(x \mid y) \\ &= -\sum_Y \sum_X p(x,y)\ log\ p(x \mid y) \end{aligned}$$

A partir de la regla de la cadena, se obtiene la siguiente expresión que permite obtener una idea más intuitiva de lo que busca medir la entropía condicionada (demostración disponible en Cover and Thomas [9]):

$$H(X \mid Y) = H(X,Y) - H(Y)$$

Es decir, la entropía condicionada corresponde a la entropía conjunta de dos variables, menos la entropía de la variable que condiciona.

Se debe notar que $H(X \mid Y) \neq H(Y \mid X)$, sin embargo, sí se cumple que $H(X \mid Y) + H(Y) = H(Y \mid X) + H(X)$.

Para el caso de tres variables aleatorias discretas, se tiene que:

$$H(X, Y \mid Z) = H(X \mid Z) + H(Y \mid X, Z)$$

4.1.2. Entropía diferencial conjunta y entropía diferencial condicionada

Para el caso continuo, que corresponde al interés de este trabajo, se define a continuación la entropía diferencial conjunta y condicionada:

Definición 4.5 *La **entropía diferencial conjunta** de un conjunto de variables aleatorias $X_1, ..., X_n$, distribuídas de acuerdo a la densidad conjunta $f(x_1, ..., x_n)$, se define por*

$$h(X_1, ..., X_n) = -\int f(x_1, ..., x_n)\ log\ f(x_1, ..., x_n)\ dx_1 ... dx_n$$

Definición 4.6 *La **entropía diferencial condicionada** para dos variables aleatorias continuas X e Y, con función de densidad conjunta $f(x, y)$ es*

$$h(X \mid Y) = -\int f(x, y)\ log\ f(x \mid y)\ dxdy \tag{4.1}$$

Considerando que $f(x \mid y) = f(x, y)/f(y)$, entonces (4.1) tiene una expresión alternativa dada por

$$h(X \mid Y) = h(X,Y) - h(Y) \tag{4.2}$$

La entropía conjunta mide cuánta incertidumbre hay en un conjunto de variables aleatorias $X_1, ..., X_n$ consideradas en forma simultánea, y la entropía condicionada es una medida de cuánta incertidumbre queda en X una vez que se conoce Y.

Tanto en el caso discreto, como en el caso continuo, se cumple que al condicionar se reduce entropía; es decir

$$H(X \mid Y) \leq H(X)$$
$$h(X \mid Y) \leq h(X)$$

cumpliéndose la igualdad cuando X e Y son independientes.

4.2. Entropía relativa e información mutua

En esta sección se revisan dos conceptos relacionados con la entropía: la entropía relativa, también conocida como divergencia de Kullback-Leibler, y la información mutua.

La entropía relativa es una medida de discrepancia entre dos distribuciones de probabilidad, o dos densidades en el caso continuo.

Definición 4.7 *La **entropía relativa** o **divergencia de Kullback-Leibler**, entre dos distribuciones de probabilidad $p(x)$ y $q(x)$, se define por*

$$\begin{aligned} D(p \mid q) &= E_p\left[log\frac{p(X)}{q(X)}\right] \\ &= \sum_X p(x)\ log\frac{p(X)}{q(X)} \end{aligned}$$

A continuación, se presenta la definición de entropía relativa para el caso continuo, es decir, considerando dos densidades de variables aleatorias continuas.

Definición 4.8 *La* ***entropía relativa*** *o* ***divergencia de Kullback-Leibler*** *entre dos densidades $f(x)$ y $g(x)$, definidas para el mismo soporte S, está dada por,*

$$D(f(x) \mid g(x)) = \int_S f(x)\ log\frac{f(x)}{g(x)}\ dx$$

Por otro lado, la información mutua mide cuánta información comparten dos variables aleatorias; representa la reducción de incertidumbre de cualquiera de ellas si se conoce la otra.

Definición 4.9 *Sean X e Y dos v.a. discretas, con función de probabilidad conjunta $p(x, y)$ y funciones de probabilidad marginales $p(x)$ y $p(y)$, entonces la* ***información mutua*** *entre X e Y es:*

$$I(X;Y) = \sum_X \sum_Y p(x,y)\ log\frac{p(x,y)}{p(x)p(y)}$$

Luego, la información mutua $I(X;Y)$ corresponde a la divergencia de Kullback Leibler entre la distribución de probabilidad conjunta $p(x, y)$ y el producto de las distribuciones marginales $p(x)$ y $p(y)$, es decir

$$I(X;Y) = D(p(x,y) \mid p(x)p(y))$$

Nuevamente, si se trata de dos variables aleatorias continuas, las definiciones son análogas.

Definición 4.10 *La* ***información mutua*** *(IM) entre dos v.a. continuas con función de densidad conjunta $f(x, y)$ y densidades marginales $f(x)$ y $f(y)$, está definida por,*

$$I(X;Y) = \int f(x,y) log \frac{f(x,y)}{f(x)f(y)} dxdy$$

Y se cumple que,

$$I(X;Y) = D(f(x,y) \mid f(x)f(y))$$

La información mutua será siempre no negativa, ya sea entre variables discretas o continuas, es decir

$$I(X;Y) \geq 0$$

con igualdad si y sólo si X e Y son independientes.

4.2.1. Relación entre la entropía y la información mutua

Las relaciones que ahora se presentan, son válidas tanto para variables discretas como continuas, y sus respectivas demostraciones se encuentran disponibles en Cover and Thomas [9].

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

Además,

$$\begin{aligned} I(X;Y) &= I(Y;X) \\ I(X;X) &= H(X) \end{aligned}$$

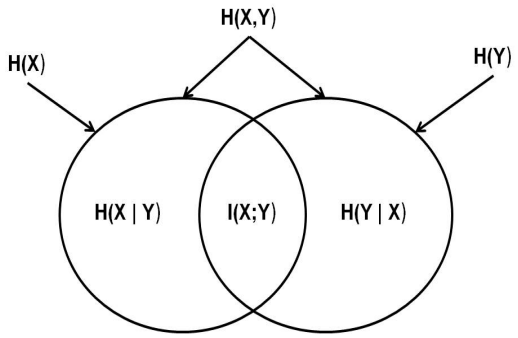

Figura 4.1: Relación entre la información mutua y la entropía

El siguiente diagrama (Figura 4.1), tomado de Cover and Thomas [9], muestra claramente la relación entre entropía e información mutua.

Como ya se mencionó, todas estas relaciones se cumplen también para variables de tipo continuo.

$$\begin{aligned} I(X;Y) &= h(X) - h(X \mid Y) \\ &= h(Y) - h(Y \mid X) \\ &= h(X) + h(Y) - h(X,Y) \end{aligned} \tag{4.3}$$

Cumpliéndose que,

$$\begin{aligned} I(X;Y) &= I(Y;X) \\ I(X;X) &= h(X) \end{aligned}$$

4.2.2. Propiedades de la entropía y la información mutua

En esta Sección se revisan algunas propiedades de gran utilidad para realizar cálculos de entropía e información mutua, debiendo distinguir algunas diferencias para el caso discreto y continuo. Todas las demostraciones se encuentran disponibles en Cover and Thomas [9].

Definición 4.11 *Sea un conjunto de variables aleatorias discretas $X_1, X_2, ..., X_n$, con su correspondiente función de probabilidad conjunta $p(x_1, x_2, ..., x_n)$, entonces la* ***regla de la cadena para la entropía*** *permite calcular la entropía conjunta como,*

$$H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, ..., X_1)$$

Y si se trata de variables continuas,

Definición 4.12 *Sea un conjunto de variables aleatorias continuas $X_1, X_2, ..., X_n$, con densidad conjunta $f(x_1, x_2, ..., x_n)$, entonces la entropía diferencial conjunta se puede calcular con la* ***regla de la cadena para entropía diferencial****:*

$$h(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} h(X_i \mid X_{i-1}, ..., X_1)$$

En el caso discreto y continuo, se cumple que:

$$\begin{aligned} H(X_1, X_2, ..., X_n) &\leq \sum_{i=1}^{n} H(X_i) \\ h(X_1, X_2, ..., X_n) &\leq \sum_{i=1}^{n} h(X_i), \end{aligned}$$

presentándose la igualdad si y sólo si las X_i son independientes.

La siguiente definición, corresponde a la información mutua condicionada; medida que indica la información que comparten las variables X e Y cuando Z es dada como evidencia.

Definición 4.13 *La* ***información mutua condicionada*** *entre dos v.a. discretas X e Y, dada una tercera v.a. discreta Z es,*

$$\begin{aligned} I(X;Y \mid Z) &= E_{p(x,y,z)} \left[log \, \frac{p\ (X;Y \mid Z)}{p\ (X \mid Z)\ p\ (Y \mid Z)} \right] \\ &= H(X \mid Z) - H(X \mid Y, Z) \end{aligned}$$

En el caso discreto, se cumple que $I(X;Y \mid Z) \geq 0$.

Para el caso continuo la definición es análoga,

Definición 4.14 *La* ***información mutua condicionada*** *entre dos v.a. continuas X e Y, dada una tercera v.a. continua Z es,*

$$I(X;Y \mid Z) = h(X \mid Z) - h(X \mid Y,Z) \tag{4.4}$$

A continuación se presentan otras maneras de escribir 4.4 que serán de gran utilidad para realizar cálculos posteriores. Estas equivalencias son válidas para el caso discreto y continuo.

$$\begin{aligned} I(X;Y \mid Z) &= h(X \mid Z) + h(Y \mid Z) - h(X,Y \mid Z) \\ &= I(X;Y) - I(X;Y;Z) \\ &= h(X,Z) + h(Y,Z) - h(X,Y,Z) - h(Z) \end{aligned} \tag{4.5}$$

Por último, en las siguientes definiciones se presenta la regla de la cadena para información mutua, tanto para el caso discreto como continuo.

Definición 4.15 *Sea un conjunto de variables aleatorias discretas $X_1, X_2, ..., X_n$ con distribución de probabilidad conjunta $p(x_1, x_2, ..., x_n)$, y sea otra v.a. discreta Y, entonces la* ***regla de la cadena para información mutua en el caso discreto*** *se define por,*

$$I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, X_{i-2}, ..., X_1)$$

Y para variables aleatorias continuas,

Definición 4.16 *Sea un conjunto de variables aleatorias continuas $X_1, X_2, ..., X_n$ con densidad conjunta $f(x_1, x_2, ..., x_n)$, y sea otra v.a. continua Y, entonces por la* ***regla de la cadena para información mutua en el caso continuo*** *se cumple que,*

$$I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, X_{i-2}, ..., X_1)$$

4.3. Información mutua normalizada (IMN)

Las definiciones que se presentan en esta Sección, originalmente fueron introducidas para variables aleatorias discretas, sin embargo, por el interés de este trabajo, a continuación serán presentadas con la notación correspondiente a variables continuas.

En el artículo de Strehl and Ghosh [42] se propone la siguiente definición para información mutua normalizada.

Definición 4.17 *Sean dos variables aleatorias X e Y, entonces la* ***información mutua normalizada*** *es,*

$$IN(X;Y) = \frac{I(X;Y)}{\sqrt{h(X)h(Y)}}, \tag{4.6}$$

donde h(X) *y* h(Y) *son las correspondientes entropías diferenciales.*

Luego, IN es una medida que puede tomar valores que van desde *cero*, cuando X e Y son independientes, a *uno* en el límite en que $X = Y$ (en este caso, se puede ver en Cover and Thomas [9] que $I(X;X) = h(X)$).

Esta definición permite una mejor interpretación de resultados y comparar valores de información mutua correspondientes a diferentes conjuntos de variables (modelos).

Es importante notar que la entropía relativa, como ya se dijo, puede tomar valores negativos; luego, si las dos entropías relativas involucradas son negativas, el denominador será un número real y el cálculo se puede realizar sin problemas, sin embargo

esto no se cumple si sólo una de ella es negativa, ya que el resultado será un número complejo.

El siguiente resultado es una extensión de la definición anterior, propuesta por Richiardi [35].

Definición 4.18 *Sean tres variables aleatorias X, Y y Z, entonces la* ***información mutua condicionada normalizada*** *viene dada por,*

$$IN(X;Y \mid Z) = \frac{I(X;Y \mid Z)}{\sqrt{h(X \mid Z)h(Y \mid Z)}}, \tag{4.7}$$

con $IN(X;Y \mid Z)$ entre *cero* (cuando X es independiente de Y dado Z) y *uno* (cuando $X = Y$ dado Z). En forma similar a lo presentado para $IN(X;Y)$, se debe tener cuidado con las restricciones necesarias para obtener un número real.

Finalmente, se presenta una medida propuesta por Besson et al. [3]. que también contribuye a la interpretación de resultados.

Definición 4.19 *Sean tres variables aleatorias X, Y y Z, se define como* ***diferencia normalizada*** *a la expresión:*

$$\Delta I_{XYZ} = \frac{[IN(X;Y) - IN(X;Y \mid Z)]}{IN(X;Y)} \tag{4.8}$$

Capítulo 5

Sensibilidad a la evidencia en Redes Bayesianas Gaussianas

5.1. Introducción

En general, una red Bayesiana busca modelar algún fenómeno de interés, considerando las variables aleatorias involucradas en el problema y la estructura de dependencia existente entre ellas. Así, el objetivo principal consiste en obtener la distribución de probabilidad condicionada de las variables que no son conocidas, principalmente de alguna de ellas definida como variable respuesta, en base a las variables que sí lo son (variables evidenciales). Muchas veces, las variables observables están fijas a priori; sin embargo, otras veces, éstas pueden ir siendo definidas durante el mismo proceso de modelado de la red.

En este contexto, el análisis de sensibilidad estudia de qué manera, la información que se introduce en la red, produce efectos o cambios fundamentales en la distribución condicionada de la variable respuesta. Dicha información puede variar al producirse cambios en los parámetros considerados en la distribución de probabilidad condicionada, o en los valores específicos que se asigne a las variables evidenciales. Bajo este enfoque de sensibilidad, es posible encontrar bastante literatura, de la última década principalmente, donde se presentan diversas técnicas para realizar análisis de sensibilidad en redes Bayesianas. La gran mayoría de estos trabajos se refieren a redes Bayesianas discretas; por ejemplo, Malhas y Aghbari [31] introducen un procedimiento basado en incrementos de la información mutua que permite descubrir nuevos patrones de interés; Chan y Darwiche [7] proponen una medida de distancia entre la distribución

original y una nueva distribución donde se han modificado los parámetros; y Laskey [27] mide sensibilidad calculando las derivadas parciales de la probabilidad condicionada respecto a los parámetros considerados. En redes Bayesianas Gaussianas, Castillo y Kjaerulff [6] desarrollan análisis de sensibilidad usando derivadas parciales y propagación simbólica; y en los trabajos de Gómez-Villegas et al. [17] y Gómez-Villegas et al. [18] consideran la divergencia de Kulback-Leibler como medida de sensibilidad.

En este trabajo, se propone una metodología para tratar el análisis de sensibilidad desde un aspecto diferente. Como ya se mencionó, en muchos problemas en que se busca modelizar situaciones reales, el conjunto de variables evidenciales no está previamente definido; de hecho, es una práctica usual el intentar recoger la mayor cantidad de información posible, lo cual siempre tendrá un coste asociado. Luego, la idea principal que a continuación se desarrolla, y que ya se encuentra en proceso de publicación (Gómez-Villegas et al. [19]), consiste en evaluar cuál o cuáles, de todas las variables disponibles, son las más informativas para obtener un mejor resultado en el modelo. Así, por ejemplo, si se intenta obtener un diagnóstico médico, en base a antecedentes previos del paciente y a exámenes clínicos que se realicen, nace la pregunta natural sobre qué exámenes serán de mayor utilidad, y por lo tanto prioritarios, para obtener un diagnóstico más preciso. Responder esta pregunta en forma correcta, permite hacer el modelo más eficiente, en términos de obtener el mejor resultado al menor coste.

Para cumplir con el objetivo recién propuesto, es necesario dejar claro que se considera el siguiente supuesto: un mejor resultado será aquél en que se obtiene la distribución de probabilidad condicionada de la variable respuesta con menor incertidumbre, es decir, con menor entropía. Luego, en esta sección, se introducen las herramientas necesarias, considerando medidas basadas en teoría de la información, para priorizar la información disponible de manera que se reduzca lo más posible la incertidumbre de la variable respuesta.

5.2. Antecedentes

En esta Sección, se presenta una metodología presentada por Kjaerulff y Madsen [24] para redes Bayesianas discretas. Este proceso, denominado *análisis de valor de la información (value of information analysis)*, es un proceso paso a paso que permite cuantificar cambios en la distribución de la variable respuesta, al incorporar a la red una variable adicional como evidencia.

Considerando que la entropía puede ser usada como una medida de incertidumbre de la distribución de la variable respuesta, la metodología se basa en utilizar esta medida para determinar un orden de prioridad al incorporar nuevas evidencias.

A continuación se introduce un resumen del procedimiento propuesto. Sea X la variable de interés en una red Bayesiana discreta, y sea Y el conjunto de todas las restantes variables discretas involucradas en la red (inicialmente no observadas), entonces:

Procedimiento

1. Calcular la función valor, definida por $V(X) = -H(X)$
2. Calcular $I(X;Y)$ para todo Y no observado.
3. Incorporar a la red como evidencia (e) la variable Y con la que se obtuvo una mayor $I(X;Y)$ en 2.
4. Calcular el aumento que se ha producido en $V(X) = -(H(X) - I(X;Y=y)) = -H(X \mid Y=y)$
5. Volver a calcular $I(X;Y)$ para todo $Y \neq e$ incorporado en 3.

Notas:

$*$ Si la variable con mayor $I(X;Y)$ no es observable, considerar la segunda mayor.

$*$ Las variables con $I(X;Y) = 0$ no necesitan ser observadas porque no aportan información a la variable respuesta.

Hay que notar que se propone como medida de decisión, la función valor $V(X) = -H(X)$, es decir, el valor negativo de la entropía, en vez de la entropía directamente. Para justificar esta elección, Kjaerulff y Madsen [24] proponen un ejemplo. Si se considera que la variable de interés X es de tipo binario, con posibles estados *verdadero* y *falso*, entonces la distribución de X corresponde a una distribución de Bernoulli de parámetro (p). Luego, $Prob(X = verdadero, X = falso) = (p, 1-p)$ y la entropía máxima se alcanzará en $p = 0.5$. Sin embargo, los autores enfocan el análisis en el sentido de obtener la mayor información posible para la distribución de la variable

respuesta, más que en disminuir la entropía (aunque en ambos casos se llega al mismo objetivo); por esta razón, eligen trabajar con la función valor, la que en el ejemplo será mínima cuando $p = 0.5$ y máxima en los extremos. En la Figura 5.1 se pueden comparar ambas medidas para el ejemplo recién dado.

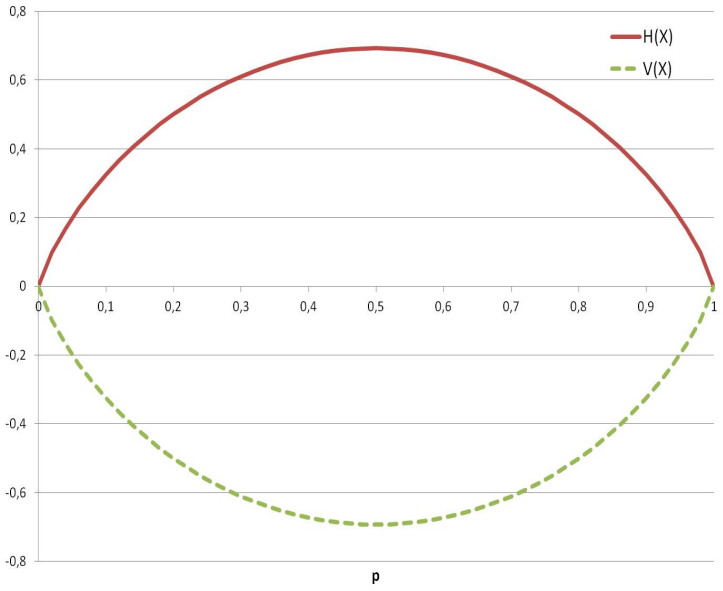

Figura 5.1: Entropía y función valor.

Así, esta metodología permite generar un orden de prioridad para incorporar variables evidenciales a una red Bayesiana discreta. En las siguientes secciones, se propone una generalización para redes Bayesianas Gaussianas, incorporando luego, nuevas medidas que permitirán realizar comparaciones entre distintos modelos.

5.3. Análisis de sensibilidad a la evidencia basado en información mutua

Como ya se ha comentado, el principal objetivo de la metodología que a continuación se propone, consiste en identificar aquellas variables que proporcionen la mayor información mutua con la variable respuesta; estas potenciales variables evidenciales serán las que permitan, al ser observadas, reducir la incertidumbre (entropía) de la distribución condicionada de la variable respuesta.

5.3.1. Procedimiento propuesto

Sea el conjunto de variables no evidenciales $\mathbf{Y}$ (inicialmente se puede considerar $\mathbf{Y} = \mathbf{X}$ porque $E = \phi$) y la variable respuesta $X_i \, \epsilon \, \mathbf{Y}$; se hará referencia a $\mathbf{Y}_{-i}$ como al conjunto de variables no evidenciales sin considerar la variable respuesta X_i, es decir, $\mathbf{Y}_{-i} = \mathbf{Y} \backslash X_i$. Luego, el procedimiento propuesto (Gómez-Villegas et al. [19]) involucra los siguientes pasos:

1. Calcular la entropía para la variable respuesta $h(X_i)$.
2. Calcular la información mutua entre la variable respuesta X_i y cada una de las variables no evidenciales $\mathbf{Y}$ diferentes de X_i, es decir, $I(X_i; \mathbf{Y}_{-\mathbf{i}})$.
3. Elegir como variable evidencial la X_k de $\mathbf{Y}_{-i}$ que proporcione la mayor información mutua con X_i.
4. Determinar la disminución de entropía en $h(X_i \mid \mathbf{E}) - I(X_i; X_k \mid \mathbf{E}) = h(X_i \mid \mathbf{E}, X_k)$.
5. Considerar que $X_k \, \epsilon \, \mathbf{E}$.
6. Calcular la información mutua condicionada entre la variable respuesta y el nuevo conjunto de variables no evidenciales $\mathbf{Y}_{-\mathbf{i}}$: $I(X_i; \mathbf{Y}_{-\mathbf{i}} \mid \mathbf{E})$.
7. Volver al paso 3.

Notas:

- Detener el proceso cuando la incertidumbre de X_i sea suficientemente pequeña o cuando no queden variables no evidenciales disponibles con información mutua mayor a cero.
- Si la variable con la información mutua más alta no está disponible, entonces se debe considerar como evidencia aquella variable que contenga la segunda información mutua más alta (y así sucesivamente).

- Las variables con información mutua igual a cero, o cercano a cero, no deben ser consideradas como evidencia, ya que ellas no van a aportar información a la distribución de la variable respuesta.

En la siguiente Sección, se definen los principales conceptos de teoría de la información para el caso específico en que las variables aleatorias consideradas tengan distribución Normal. De esta forma, se busca contar con las herramientas necesarias para aplicar este procedimiento a RBG.

5.3.2. Teoría de la información en RBG

Considerando que este trabajo se basa en redes Bayesianas Gaussianas, se hace necesario presentar los valores para entropía diferencial cuando las variables aleatorias utilizadas tienen distribución Normal Univariante y distribución Normal Multivariante.

Entropía diferencial para la distribución Normal

Sea la variable aleatoria X con distribución Normal de media μ y varianza σ^2, es decir, $X \sim N(\mu, \sigma^2)$, entonces:

$$h(X) = \frac{1}{2}\, ln\, (2\pi e\sigma^2) \tag{5.1}$$

Demostración

Si $X \sim N(\mu, \sigma^2)$ entonces $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, luego

$$
\begin{aligned}
h(X) &= -\int_S \phi(x)\ ln\ \phi(x) dx \\
&= -\int_S \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\right) dx \\
&= -\ln\frac{1}{\sqrt{2\pi\sigma^2}}\ \int_S \phi(x) + \frac{1}{2\sigma^2}\ \int_S \phi(x)\ (x-\mu)^2\ dx \\
&= -\ln\frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2\sigma^2}\left[\int_S \phi(x)\ x^2\ dx - 2\mu\int_S \phi(x)\ x\ dx + \mu^2\right] \\
&= -\ln\frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2\sigma^2}\left[E(X^2) - 2\mu\ E(X) + \mu^2\right]
\end{aligned}
$$

Considerando que:

$$E(X) = \mu$$

y que,

$$V(X) = E(X^2) - E^2(X) \ \text{ por lo tanto } \quad E(X^2) = \sigma^2 + \mu^2,$$

entonces,

$$
\begin{aligned}
h(X) &= -\ln\frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2\sigma^2}\sigma^2 \\
&= \frac{1}{2}\ ln\ (2\pi e\sigma^2)
\end{aligned}
$$

Entropía diferencial para la distribución Normal Multivariante

Sea el vector aleatorio $X_1, X_2, ..., X_n$ con distribución Normal Multivariante, de vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas Σ, es decir, $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, entonces:

$$h(X_1, X_2, ..., X_n) = \frac{1}{2}\ ln\ (2\pi e)^n |\Sigma|, \tag{5.2}$$

donde $|\Sigma|$ es el determinante de Σ.

Demostración

Si $\mathbf{X} = (X_1, ..., X_n) \sim N(\boldsymbol{\mu}, \Sigma)$ entonces,

$$\phi(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n \ |\Sigma|^{1/2}} e^{\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \ \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Luego, siguiendo un procedimiento análogo al caso anterior,

$$\begin{aligned} h(\mathbf{x}) &= -\int_S \phi(\mathbf{x}) \ ln \ \phi(\mathbf{x}) \ d\mathbf{x} \\ &= -\ln \frac{1}{(\sqrt{2\pi})^n |\Sigma|^{1/2}} \ \int_S \phi(\mathbf{x}) + \frac{1}{2} \ \int_S \phi(\mathbf{x}) \ (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \ d\mathbf{x} \\ &= -\ln \frac{1}{(\sqrt{2\pi})^n |\Sigma|^{1/2}} + \frac{1}{2} \ E\left[(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right] \end{aligned}$$

Así, es necesario calcular,

$$\begin{aligned} E\left[(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right] &= E\left[\sum_i \sum_j (x_i - \mu_i) \Sigma_{ij}^{-1} (x_j - \mu_j)\right] \\ &= \sum_i \sum_j E\left[(x_i - \mu_i)(x_j - \mu_j)\right] \Sigma_{ij}^{-1} \\ &= \sum_i \sum_j \left[E(x_i x_j) - \mu_j E(x_i) - \mu_i E(x_j) + \mu_i \mu_j\right] \Sigma_{ij}^{-1} \\ &= \sum_i \sum_j \Sigma_{ij} \Sigma_{ij}^{-1} \\ &= \sum_i I_{ii} \\ &= n \end{aligned}$$

Esta misma demostración se puede obtener considerando que $(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})$ corresponde a una forma cuadrática Normal con distribución Chi-cuadrado, con n gra-

dos de libertad, y por tanto su media es n.

Luego,

$$\begin{aligned} h(\mathbf{x}) &= -\ln \frac{1}{(\sqrt{2\pi})^n |\Sigma|^{1/2}} + \frac{1}{2} n \\ &= \frac{1}{2}\, ln\ (2\pi e)^n |\Sigma| \end{aligned}$$

Con estas definiciones, ya es posible realizar el cálculo de la información mutua e información mutua condicionada, para el caso Normal, siguiendo las notaciones presentadas en 4.3 y 4.5.

En particular, a continuación se incluye como ejemplo el valor de información mutua para el caso Normal Bivariante.

Información mutua para la distribución Normal Bivariante

Sea el vector aleatorio (X, Y) con distribución Normal Bivariante, de vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\Sigma_{2\times 2}$, entonces:

$$I(X;Y) = -\frac{1}{2}\, ln\ (1 - \rho^2_{X,Y}) \tag{5.3}$$

donde $\rho^2_{X,Y}$ es el coeficiente de correlación entre X e Y.

Demostración

Utilizando la expresión 4.3 y los resultados obtenidos en 5.1 y 5.2, entonces:

$$
\begin{aligned}
I(X;Y) &= h(X) + h(Y) - h(X,Y) \\
&= \frac{1}{2}\, ln\, (2\pi e\sigma_X^2) + \frac{1}{2}\, ln\, (2\pi e\sigma_Y^2) - \frac{1}{2}\, ln\, (2\pi e)^2|\Sigma| \\
&= \frac{1}{2}\ln\, \left(\frac{2\pi e\ 2\pi e}{(2\pi e)^2}\right) + \frac{1}{2}\ln\, \left(\frac{\sigma_X^2\ \sigma_Y^2}{|\Sigma|}\right)
\end{aligned}
$$

Considerando que $\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$, entonces:

$$
\begin{aligned}
I(X;Y) &= \frac{1}{2}\, ln \left(\frac{\sigma_X^2\ \sigma_Y^2}{\sigma_X^2\ \sigma_Y^2 - \sigma_{XY}^2}\right) \\
&= -\frac{1}{2}\, ln \left(1 - \frac{\sigma_{XY}^2}{\sigma_X^2\ \sigma_Y^2}\right) \\
&= -\frac{1}{2}\, ln \left(1 - \left(\frac{\sigma_{XY}}{\sigma_X\ \sigma_Y}\right)^2\right) \\
&= -\frac{1}{2}\, ln\, (1 - \rho_{X,Y}^2)
\end{aligned}
$$

Luego, para el caso Normal Bivariante, la información mutua queda en función del coeficiente de correlación, como era de esperar.

5.3.3. Ejemplo: Análisis de sensibilidad (AS)

Para mostrar el comportamiento del procedimiento propuesto y los resultados que se pueden obtener con él, se considera nuevamente el Ejemplo 5, correspondiente a una RBG, cuyo DAG se muestra en la Figura 3.1. Se ha definido la variable X_7 como la respuesta de interés y se trabajará bajo el supuesto de que inicialmente no hay variables evidenciales determinadas.

A partir de (4.3) y utilizando la definiciones de entropía para el caso Normal (5.1) y Normal Multivariante (5.2), se calculan los pasos 1 y 2 especificados en la Sección 5.3.1. Así, se obtiene la entropía diferencial de la variable respuesta X_7, $h(X_7) = 3.7165$, y los valores de información mutua para cada una de las variables no evidenciales con la

Tabla 5.1: Ejemplo: Análisis de sensibilidad (AS). Paso 2.

Y_{-i}	$h(Y_{-i})$	$h(X_7, Y_{-i})$	$I(X_7; Y_{-i})$
X_1	1.4189	5.1148	0.0206
X_2	1.4989	4.6155	0.5199
X_3	1.7655	5.4399	0.0421
X_4	2.3148	5.4718	0.5595
X_5	2.5702	5.5018	0.7849
X_6	3.7063	5.4718	1.9509

variable respuesta. Los cálculos se realizaron con el software estadístico de distribución libre R [34] y los resultados obtenidos se muestran en la Tabla 5.1.

Al presentar el Ejemplo 5, se determinó la red y se obtuvo la distribución conjunta de $\mathbf{X}$, correspondiente a una distribución Normal Multivariante con parámetros $\boldsymbol{\mu}$ y $\sum$. A partir de estos parámetros, se obtiene que la distribución marginal de la variable X_7 corresponde a una distribución Normal $X_7 \sim N(8, 99)$. Luego, es posible observar que la distribución contiene asociada una alta variabilidad, lo que se ve reflejado en una alta entropía también. Por lo tanto, es de interés buscar qué variables, hasta ahora no observables, permitirían disminuir este nivel de incertidumbre si fueran incorporadas como información a la red.

Los resultados presentados en la Tabla 5.1, permiten obtener un orden de prioridad para disminuir la entropía de X_7. Es evidente que la variable más informativa para X_7 es la variable X_6 y luego, la variable X_5.

Para mostrar cómo la elección de la variable a observar afecta en la disminución de la incertidumbre de la variable respuesta X_7, se calculó la varianza condicionada de X_7 y su entropía diferencial condicionada, considerando el caso supuesto de que cada una de las variables evidenciales fuera observada. Para realizar estos cálculos, que se presentan en la Tabla 5.2, se utilizaron las expresiones 3.4 y 4.2. Es importante notar que estas medidas no dependen del valor que tome la variable observada, si no sólo, del conjunto de variables evidenciales que se considere.

Como ya se dijo, la distribución inicial de X_7 era $N(8, 99)$ con una entropía diferencial de $h(X_7) = 3.7165$; luego, es evidente que la variable más informativa para X_7 es X_6. De hecho, si se pudiera incorporar esta variable como evidencia, la varianza de X_7 disminuiría a 2 y su entropía a 1.76; es decir, sería suficiente observar X_6 para que el problema quedara resuelto con una alta precisión. Por otro lado, es interesante observar

Tabla 5.2: Ejemplo: AS. Sensibilidad de X_7 a la evidencia.

Y_{-i}	$Var(X_7\|Y_{-i})$	$h(X_7\|Y_{-i})$
X_1	95	3.6959
X_2	35	3.1966
X_3	91	3.6743
X_4	32.3	3.1564
X_5	20.6	2.9316
X_6	2	1.7655

que las dos variables que fueron incorporadas como evidencia cuando se presentó este ejemplo, X_1 y X_3, en realidad no contribuyen individualmente en forma significativa a la reducción de la incertidumbre en X_7.

Para seguir con el procedimiento propuesto, se va a suponer que la variable X_6 no está disponible, pero que sí se cuenta con información para la variable X_5, que es la que produce el siguiente mayor efecto. Si se incorpora esta variable a la red como evidencia (paso 3), entonces la nueva distribución de $X_7|X_5$ es N(8,20.6) y del paso 4 se obtiene una entropía diferencial condicionada $h(X_7|X_5) = 2.9316$, lo que implica una importante reducción de incertidumbre.

Ahora, de acuerdo con los pasos 5 y 6 respectivamente, $X_5 \ \epsilon \ \mathbf{E}$ y se calcula la información mutua condicionada del resto de las variables Y_{-i} con X_7. Para realizar estos cálculos, es posible utilizar las definiciones 4.2 y 4.5 para entropía e información mutua condicionada; en forma equivalente, se puede calcular la distribución Normal Multivariante condicionada por el resto de las variables al propagar X_5 a través de (3.3) y (3.4), es decir, la distribución de $(X_1, X_2, X_3, X_4, X_6, X_7 \mid X_5)$, y luego hacer los cálculos en forma análoga a los realizados para la Tabla 5.1. Los resultados para el paso 6 se muestran en la Tabla 5.3.

Tabla 5.3: Ejemplo: AS. Paso 6, condicionado en X_5

Y_{-i}	$h(Y_{-i}\|X_5)$	$h(X_7, Y_{-i}\|X_5)$	$I(X_7; Y_{-i}\|X_5)$
X_1	1.4189	4.2426	0.1079
X_2	1.1635	3.7814	0.3137
X_3	1.6539	4.5451	0.0404
X_4	2.1597	4.1279	0.9633
X_6	2.8805	4.6460	1.1661

Es claro, de la Tabla 5.3, que sigue siendo la variable X_6 la que permite disminuir en forma más significativa la incertidumbre de X_7, pero como se ha considerado que esta variable no puede ser observada, se incorporará como evidencia la segunda variable más informativa para X_7, la variable X_4. Así, volviendo al paso 3, después de propagar la variable evidencial X_4, se obtiene que la nueva $Var(X_7 \mid X_5, X_4) = 3$, lo que significa que se ha podido obtener un resultado con una alta precisión.

Finalmente, la entropía diferencial condicionada que se ha obtenido para $X_7 \mid X_5, X_4$ es $h(X_7 \mid X_5, X_4) = 1.968$, y en la Tabla 5.4 se muestran los nuevos valores de entropía e información mutua de las variables que van quedando como no evidenciales. Se puede observar que la entropía diferencial de la variable respuesta X_7 ha disminuído notoriamente y que ya no hay variables no evidenciales que aporten información relevante, ni siquiera X_6; de hecho, las variables X_1, X_2 y X_3 tienen valores de información mutua con X_7 de *cero*, lo que significa que ellas son independientes de X_7 dado X_4 y X_5, como se muestra en el DAG.

Tabla 5.4: Ejemplo: AS. Paso 6, condicionado en X_4 y X_5.

Y_{-i}	$h(Y_{-i}\|X_5, X_4)$	$h(X_7, Y_{-i}\|X_5, X_4)$	$I(X_7; Y_{-i}\|X_5, X_4)$
X_1	1.2883	3.2565	0
X_2	0.7643	2.7325	0
X_3	1.6047	3.5729	0
X_6	1.4189	3.1844	0.2027

5.4. Extensión del análisis de sensibilidad a la evidencia basado en información mutua normalizada

En esta Sección se presenta una extensión al procedimiento propuesto en 5.3.1 incorporando medidas basadas en información mutua normalizada introducidas en 4.3. Estas medidas tienen la ventaja de tomar valores entre 0 y 1, independientemente de la unidad de la variable aleatoria que se considere, lo que permite realizar comparaciones entre distintas variables dentro de un mismo modelo, e incluso entre modelos diferentes.

5.4.1. Nuevo procedimiento propuesto

Este procedimiento se basa en el procedimiento presentado en 5.3.1; el cambio consiste en reemplazar los valores de información mutua por medidas normalizadas, además se incorpora la diferencia normalizada.

Sea el conjunto de variables no evidenciales $\mathbf{Y}$ (inicialmente se puede considerar $\mathbf{Y} = \mathbf{X}$ porque $E = \phi$) y la variable respuesta $X_i \epsilon \mathbf{Y}$; considérese $\mathbf{Y}_{-i}$ como el conjunto de variables no evidenciales sin utilizar la variable respuesta X_i, es decir, $\mathbf{Y}_{-i} = \mathbf{Y} \backslash X_i$. Entonces, este nuevo procedimiento consiste en:

1. Calcular la entropía para la variable respuesta $h(X_i)$.
2. Calcular la información mutua normalizada entre la variable respuesta X_i y cada una de las variables no evidenciales $\mathbf{Y}$ diferentes de X_i, es decir, $IN(X_i; \mathbf{Y_{-i}})$
3. Elegir como variable evidencial la X_k de $\mathbf{Y}_{-i}$ que proporcione la mayor información mutua normalizada con X_i.
4. Mirar la disminución de entropía en $h(X_i \mid \mathbf{E}) - I(X_i; X_k \mid \mathbf{E}) = h(X_i \mid \mathbf{E}, X_k)$.
5. Considerar que $X_k \epsilon \mathbf{E}$.
6. Calcular la información mutua condicionada normalizada entre la variable respuesta y el nuevo conjunto de variables no evidenciales $\mathbf{Y_{-i}}$: $IN(X_i; \mathbf{Y_{-i}} \mid \mathbf{E})$.
7. Calcular la diferencia normalizada $\Delta I_{X_i Y_{-i} E}$.
8. Volver al paso 3.

Las notas para este procedimiento son análogas al procedimiento anterior.

- Detener el proceso cuando la incertidumbre de X_i sea suficientemente pequeña o cuando no queden variables no evidenciales disponibles con información mutua normalizada mayor que cero.
- Si la variable con la información mutua normalizada más alta no está disponible, entonces se debe considerar como evidencia aquella variable que contenga la segunda información mutua normalizada más alta (y así sucesivamente).

- Las variables con información mutua normalizada igual a cero, o cercano a cero, no deben ser consideradas como evidencia, ya que ellas no van a aportar información a la distribución de la variable respuesta.

Antes de aplicar este procedimiento a un ejemplo, es necesario revisar algunas restricciones importantes que se deben cumplir al trabajar con redes Bayesianas Gaussianas.

5.4.2. Información mutua normalizada en RBG

No es necesario incorporar nuevas definiciones respecto a información mutua normalizada en el caso de redes Bayesianas Gaussianas, ya que las expresiones para la entropía en el caso Normal dadas en 5.1 y 5.2, son suficientes para aplicar el nuevo procedimiento propuesto.

Sin embargo, sí es importante considerar la siguiente restricción; como ya se mencionó al introducir los conceptos de información mutua normalizada en 4.3, se debe procurar no obtener un número complejo en el denominador, lo cual sucede si sólo una de las entropías diferenciales consideradas es negativa. En el caso particular en que las variables aleatorias tengan distribución Normal, las entropías diferenciales definidas en 5.1 y 5.2, serán negativas si:

$$\sigma^2 < \frac{1}{2\pi e} \text{ i.e. } (\sim 0.0585) \text{ para la distribución Normal Univariante } N(\mu, \sigma^2)$$

y

$$\mid \Sigma \mid < \frac{1}{(2\pi e)^n} \text{ para la distribución Normal Multivariante } N(\boldsymbol{\mu}, \Sigma)$$

Esto significa que si se están comparando dos variables aleatorias X e Y, y sólo una de ellas tiene una varianza muy pequeña (menor que 0.0585 si tiene distribución Normal univariante, 0.0585^2 para el caso bivariante, y así sucesivamente), entonces en el denominador se obtendrá un número complejo; en este caso, es mejor no utilizar la definición propuesta en 4.6. Sin embargo, en la mayoría de los casos esta situación no sucederá, ya que usualmente los problemas que interesa modelar comienzan conteniendo variables con alta varianza, no muy diferentes entre ellas, y lo que se busca

justamente es reducir esta incertidumbre.

5.4.3. Ejemplo: Análisis de sensibilidad basado en IMN

Con el objetivo de poder comparar procedimientos y resultados, se trabaja nuevamente con la RBG introducida en el Ejemplo 5, donde X_7 se define como la variable de interés.

Para aplicar el nuevo procedimiento basado en la información mutua normalizada, lo primero que se debe hacer es comprobar la restricción presentada en 5.4.2 para obtener un número real en $\sqrt{h(X)h(Y)}$. De la matriz de covarianzas Σ, se puede ver que todas las varianzas individuales son mayores a 0.0585, por lo que no se obtendrán entropías diferenciales univariantes negativas. Respecto al caso Normal Multivariante, se calculó el determinante de Σ, obteniendo un valor de $\mid \Sigma \mid = 16 > 1/(2\pi e)^7$, luego en este caso tampoco se observan entropías diferenciales menores de *cero*. Así, es posible aplicar el nuevo procedimiento propuesto a esta RBG.

Se comienza por realizar los pasos 1 y 2, usando 5.1 y 5.2, respectivamente. La entropía diferencial inicial no cambia, es decir, $h(X_7) = 3.7165$; los valores de información mutua normalizada de las variables no evidenciales con la variable respuesta, se muestran en la Tabla 5.5.

Tabla 5.5: Ejemplo: Análisis de sensibilidad con IMN. Paso 2

Y_{-i}	$h(Y_{-i})$	$IN(X_7; Y_{-i})$
X_1	1.4189	0.0089
X_2	1.4989	0.2263
X_3	1.7655	0.0164
X_4	2.3148	0.1907
X_5	2.5702	0.2539
X_6	3.7063	0.5257

De la Tabla 5.5 se puede observar que los resultados obtenidos por este procedimiento son consistentes con los obtenidos anteriormente; la variable X_6 es la más informativa para X_7, seguida por la variable X_5. Para hacer comparables los resultados, a continuación se vuelve a suponer que la variable X_6 no se encuentra disponible y se incorpora la variable X_5 como evidencia a la red; los pasos 3 y 4 coinciden con los

cálculos anteriores, es decir, se obtiene $Var(X_7 \mid X_5) = 20.6$ y la entropía diferencial $h(X_7 \mid X_5) = 2.9316$. En la Tabla 5.6 se muestran los resultados para los pasos 6 y 7, calculados con 4.7 y 4.8 respectivamente.

Tabla 5.6: Ejemplo: AS con IMN. Pasos 6 y 7, condicionado en X_5

Y_{-i}	$h(Y_{-i}\|X_5)$	$IN(X_7; Y_{-i}\|X_5)$	$\Delta I_{X_7 Y_{-i} X_5}$
X_1	1.4189	0.0529	-4.9438
X_2	1.1635	0.1698	0.2496
X_3	1.6539	0.0183	-0.1158
X_4	2.1597	0.3828	-1.0073
X_6	2.8805	0.4013	0.2365

Al igual que en el caso anterior, de la Tabla 5.6 se puede concluir que la variable X_6 sigue siendo la más informativa para X_7, ahora seguida por la variable X_4.

Se ha incorporado un elemento nuevo en la Tabla 5.6, la diferencia $\Delta I_{X_7 Y_{-i} X_5}$; este valor se puede interpretar como una medida de la pérdida relativa de influencia de cada variable en la variable respuesta, después de haber incorporado evidencia a la red. Así, un valor negativo indica un aumento de la importancia de la variable, en términos de ser más informativa para X_7, y un valor positivo, lo contrario. Por ejemplo, es interesante notar que inicialmente la variable X_6 era significativamente la más informativa para X_7; sin embargo, después de incorporar X_5 como evidencia, la contribución de X_4 aumenta mientras que la de X_6 disminuye, obteniéndose, en este paso, una reducción de entropía similar con ambas variables.

Es interesante también, para entender mejor cada medida, analizar lo que sucede con la variable X_1. Esta variable presenta un alto incremento en su importancia relativa, sin embargo, esto no implica que sea una variable que aporte demasiado a la reducción de entropía de la variable respuesta; de hecho, se puede ver que sigue siendo poco informativa para X_7, lo que deja en evidencia que estos valores deben ser analizados siempre en forma conjunta, mirarlos en forma independiente puede llevar a tomar decisiones equivocadas.

Continuando con el procedimiento, corresponde volver al paso 3, después de comprobar las restricciones necesarias, se calcula la reducción en la entropía diferencial de X_7 condicionada ahora por X_5 y X_4, y se incorpora la variable X_4 como evidencia a la red. Los resultados para el paso 6 se muestran en la Tabla 5.7.

Tabla 5.7: Ejemplo: AS con IMN. Paso 6, condicionado en X_4 y X_5

Y_{-i}	$h(Y_{-i}\|X_5, X_4)$	$IN(X_7; Y_{-i}\|X_5, X_4)$
X_1	1.2882	0
X_2	0.7643	0
X_3	1.6047	0
X_6	1.4189	0.1213

Finalmente, se mostrará la importancia de trabajar con medidas normalizadas incorporando una nueva red Bayesiana al ejemplo.

5.4.4. Extensión del Ejemplo 5 (EEjemplo 5)

Se considera el Ejemplo 5 ya presentado, sobre una red que busca modelar la duración en tiempo de funcionamiento de una máquina. Sin embargo, esta vez se supondrá que la máquina está constituída por dos nuevos elementos, los que reemplazan a dos elementos anteriores. Estos cambios definen un nuevo DAG, como se muestra en la Figura 5.2.

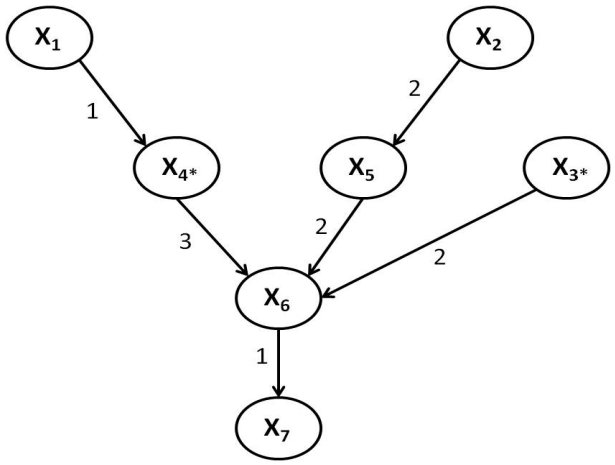

Figura 5.2: DAG EEjemplo 5.

La variable respuesta de interés sigue siendo X_7, y la distribución de probabilidad conjunta de $\mathbf{X} = \{X_1, X_2, X_3*, X_4*, X_5, X_6, X_7\}$ tendrá nuevos parámetros asociados

a las nuevas variables incorporadas (X_3* y X_4*). Luego, se tiene que el vector $\mathbf{X}$ tiene distribución Normal Multivariante con los siguientes parámetros definidos por expertos:

$$\boldsymbol{\mu^*} = \begin{pmatrix} 1 \\ 3 \\ 3 \\ 6 \\ 4 \\ 5 \\ 8 \end{pmatrix} \mathbf{B}^* = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{D}^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

donde $\boldsymbol{\mu^*}$ es el vector de medias de dimensión n, $\mathbf{D}^*$ es una matriz diagonal con varianzas condicionadas v_i^*, y $\mathbf{B}^*$ es una matriz triangular superior estricta con coeficientes de regresión β_{ji}^*, donde X_j es *padre* de X_i.

De manera análoga al caso anterior, se calcula $\mathbf{\Sigma}^*$ como $\mathbf{\Sigma} = [(\mathbf{I}-\mathbf{B})^{-1}]^T\mathbf{D}(\mathbf{I}-\mathbf{B})$. Luego, se obtiene que $\mathbf{X}$ tiene distribución Normal Multivariante con parámetros:

$$\mu^* = \begin{pmatrix} 1 \\ 3 \\ 3 \\ 6 \\ 4 \\ 5 \\ 8 \end{pmatrix} \quad \Sigma^* = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 3 & 3 \\ 0 & 1 & 0 & 0 & 2 & 4 & 4 \\ 0 & 0 & 2 & 0 & 0 & 4 & 4 \\ 1 & 0 & 0 & 4 & 0 & 12 & 12 \\ 0 & 2 & 0 & 0 & 7 & 14 & 14 \\ 3 & 4 & 4 & 12 & 14 & 74 & 74 \\ 3 & 4 & 4 & 12 & 14 & 74 & 75 \end{pmatrix}$$

A continuación, se utiliza la notación h_k, I_k y IN_k para hacer referencia a la entropía diferencial, la información mutua y la información mutua normalizada, respectivamente, de la k-ésima RBG, con $k = 1$ para el ejemplo introducido en el ejemplo 5 y $k = 2$ para este Ejemplo.

Es necesario comenzar por revisar las restricciones presentadas en 5.4.2, para obtener un número real en el cálculo de la información mutua normalizada. Para este ejemplo, a partir de Σ^*, se obtiene que las desviaciones individuales son mayores que 0.0585 y que $\mid \Sigma^* \mid = 36 > 1/(2\pi e)^7$, lo que permite seguir con el procedimiento.

Para la RBG del DAG presentado en la Figura 3.1, ya se calculó la entropía diferencial de la variable respuesta, obteniéndose que $h_1(X_7) = 3.7165$. Ahora, para el DAG

de la Figura 5.2, se obtiene que $h_2(X_7) = 3.5777$. En la Tabla 5.8 que se presenta a continuación, se muestran los valores de la información mutua normalizada de las dos RBG consideradas en esta extensión del ejemplo.

Tabla 5.8: EEjemplo 5: Comparación entre RBG_1 y RBG_2

Y_{-i}	$IN_1(X_7; Y_{-i})$	$IN_2(X_7; Y_{-i})$
X_1	0.0089	0.0284
X_2	0.2263	0.0532
X_3	0.0164	
X_3*		0.0224
X_4	0.1907	
X_4*		0.1189
X_5	0.2539	0.0799
X_6	0.5257	0.6039

A partir de la Tabla 5.8, se puede comparar la influencia relativa que tienen las variables comunes en ambos modelos. Estas comparaciones se pueden realizar gracias a que se está considerando la información mutua normalizada, lo que constituye una importante ventaja al utilizar esta medida. Así, se pueden destacar algunas conclusiones de interés; por ejemplo, el incorporar las variables X_3* y X_4*, reemplazando a las antiguas variables X_3 y X_4, hace disminuir en forma significativa la importancia, sobre X_7, de las variables X_2 y X_5. Es más, si se mantiene el supuesto de que la variable X_6 no es observable (sigue siendo la más informativa en ambos modelos), en el segundo modelo correspondería incorporar como evidencia la variable X_4* en vez de X_5, ya que ésta permitiría obtener una menor incertidumbre en X_7.

Capítulo 6

Aplicación a datos reales

En este Capítulo se presenta un problema real con el cual se busca aplicar los procedimientos propuestos en este trabajo. Se comienza por hacer una referencia al problema, indicando las fuentes de información, para luego explicar el contenido de la base de datos. A continuación, se discute el ajuste de la RBG, incluyendo especialmente el análisis de la hipótesis de Normalidad que se debe cumplir. Finalmente, sobre la RBG ajustada, se realiza un análisis de sensibilidad a la evidencia.

6.1. Objetivos y base de datos

El objetivo es realizar una aplicación de los procedimientos vistos en este trabajo en una base de datos real, definiendo una variable de interés. A continuación se describe la base de datos elegida y la variable seleccionada.

6.1.1. Objetivo

Para realizar una aplicación a los conceptos y propuestas vistas en este trabajo, se ha escogido trabajar con datos obtenidos a través de la Encuesta Nacional de Salud (ENS) 2009-2010 de Chile. Como objetivo del estudio, se ha planteado la idea de estudiar, a partir de la información disponible, qué variables se relacionan, y de qué forma, con el Indice de Masa Corporal (IMC) en adultos.

El IMC se calcula como $Peso/Talla^2$ y es considerada como una medida del estado

nutricional en los adultos. Este es un índice de gran importancia ya que se ha visto que el sobrepeso y la obesidad se relacionan con muerte prematura, enfermedades cardiovasculares, hipertensión arterial, artrosis, algunos tipos de cáncer y la diabetes [20].

6.1.2. Encuesta nacional de salud (ENS)

La Encuesta Nacional de Salud para el periodo 2009-2010 [14], fue realizada por el Departamento de Salud Pública de la Universidad Católica de Chile y el Ministerio de Salud fue el encargado del financiamiento, de la coordinación técnica y ejecutó el rol de contraparte técnica.

Respecto a los objetivos de la ENS, los principales son:

- Medir la prevalencia de 42 problemas de salud del adulto en población general que vive en Chile,
- Describir estos problemas de acuerdo a las variables sexo, edad, nivel socioeconómico, factor rural, zona geográfica y regiones,
- Constituir una seroteca (colección de muestras biológicas en condiciones de temperatura controlada) que queda disponible para estudios futuros.

En referencia a la metodología utilizada, la ENS corresponde a un estudio de prevalencia, aplicada sobre una muestra aleatoria (estratificada, trietápica, por conglomerados) de la población general de 15 o más años que vive en Chile. El tamaño muestral fue de 5416 casos, considerando como exclusiones a las mujeres embarazadas y personas violentas. La representatividad obtenida a partir de este diseño es nacional, por macrozonas y regional.

Se utilizaron tres instrumentos de medición: Encuesta, mediciones biofisiológicas y exámenes de laboratorio. En las siguientes Figuras 6.1 y 6.2 se muestra el detalle de las mediciones realizadas en relación a los 42 problemas de salud estudiados. Estos instrumentos fueron aplicados y medidos sobre el terreno entre el 17 de octubre del 2009 y el 6 de agosto del 2010.

		INSTRUMENTO DE MEDICIÓN		
Nº	TEMA DE SALUD	ENCUESTA	MEDICION CLÍNICA	TEST LABORATORIO
1	Hipertensión arterial	X	X	X
2	Dislipidemia	X		X
3	Estado nutricional	X	X	
4	Diabetes	X		X
5	Exposición a tabaco	X		
6	Consumo de alcohol y problemas relacionados	X		X
7	Consumo de sal			X
8	Consumo de alimentos protectores	X		
9	Actividad física	X	X	
10	Síndrome metabólico	X	X	X
11	Daño hepático crónico	X		X
12	Riesgo cardiovascular	X	X	X
13	Enfermedad cardiovascular	X		
14	Síntomas respiratorios crónicos	X		
15	Síntomas músculo-esqueléticos	X		
16	Patología biliar	X		X
17	Síntomas digestivos	X		
18	Síntomas depresivos	X		
19	Patología tiroidea	X		X
20	Deterioro cognitivo del adulto mayor	X		
21	Visión	X		

Figura 6.1: Instrumentos de Medición ENS 2009-2010 (1)

22	Audición	X		
23	Salud dental	X		
24	Trastornos del sueño	X		
25	Daño renal crónico			X
26	Cáncer de mama	X		
27	Cáncer cérvicouterino	X		
28	Calidad de vida relacionada con salud	X		
29	Discapacidad	X		
30	Determinantes sociales y psicológicos de la salud	X		
31	Salud sexual y reproductiva	X		
32	Consumo de medicamentos y productos naturales	X		
33	Percepción del modelo de atención primaria	X		
34	Uso de medicinas alternativas	X		
35	Déficit de vitamina B12 y de ácido fólico			X
36	Virus de hepatitis B y C			X
37	Virus de inmunodeficiencia humano	X		X
38	Enfermedad de Chagas			X
39	Virus HTLV I-II			X
40	Grupo sanguíneo y Rh			X
41	Enfermedad celíaca	X		X
42	Riesgo de fracturas y caídas	X		

Figura 6.2: Instrumentos de Medición ENS 2009-2010 (2)

6.1.3. Base de datos

Para modelar el Indice de Masa Corporal, como se planteó en los objetivos, se ha decidido considerar las siguientes variables:

- **Edad** Edad de la persona en el momento de ser aplicada la encuesta.
- **GluBasal** Glucosa basal en ayuno.
- **CREASANG** Creatinina en sangre
- **POTASIOR** Electrolito K (orina)
- **SODIOR** Electrolito Na (orina)
- **HDL** Colesterol HDL
- **LDL** Colesterol LDL
- **TGD** Triglicéridos
- **PAS** Presión arterial sistólica
- **PAD** Presión arterial diastólica
- **GRSAL24** Consumo diario de sal (gramos)
- **GRSDIAFYV** Consumo diario de frutas y verduras (gramos)
- **ANOSCUR** Número de años cursados
- **IMC** Indice de masa corporal ($Peso/Talla^2$). Variable de interés del modelo.

Para modelar, se considera una RBG para mujeres y otra para hombres, ya que la manera en que las variables inciden en el Indice de Masa Corporal en cada caso podrían ser diferentes.

Sobre la base de datos completa de la ENS 2009-2010, se consideró sólo aquellos casos con información completa en las variables antes mencionadas. Este proceso condujo a obtener una base de datos correspondiente a 2438 participantes, de los cuales 1018 son hombres.

Por simplicidad, se ha decidido tomar una muestra aleatoria de un 20 % para ajustar la RBG. Así, se trabaja con 488 encuestados, de los cuales 207 son hombres.

6.2. Supuesto de Normalidad

En esta Sección se revisa el cumplimiento de Normalidad, y se prepara la base de datos, aplicando la transformación de la distribución Nonparanormal, para que dicho supuesto sea cumplido.

6.2.1. Revisión del supuesto de Normalidad

Para ajustar una RBG el primer supuesto es que la matriz de datos sigue una distribución Normal Multivariante. Para revisar este supuesto, se utiliza el test de Doornick-Hansen [12] implementado en el paquete *asbio* de R [1].

A continuación, se presenta la revisión del supuesto de Normalidad de las bases de datos, primero para hombres y luego para mujeres.

Base de datos ENS para hombres

Al comprobar el supuesto de Normalidad Multivariante (y Univariante), por medio del test de Doornick-Hansen, se obtienen los resultados que se muestran en las Tablas 6.1 y 6.2.

Tabla 6.1: Test de Normalidad Multivariante inicial, hombres

Var	E	df (2p)	$P(Chi > E)$
X	591.761	28	$7.086e-107$

Base de datos ENS para mujeres

Al comprobar el supuesto de Normalidad Multivariante para la base de mujeres, utilizando nuevamente el test de Doornick-Hansen, se obtiene los resultados que se muestran en las Tablas 6.3 y 6.4 .

Tabla 6.2: Test de Normalidad Univariante inicial, hombres

Var	E	df (2)	$P(Chi > E)$
Edad	57.43	2	$3.375e-13$
GluBasal	32.90	2	$7.166e-08$
CreaSang	46.19	2	$9.319e-11$
PotasioR	23.59	2	$7.547e-06$
SodioR	32.54	2	$8.569e-08$
Hdl	48.78	2	$2.557e-11$
Ldl	110.66	2	$9.347e-25$
Tgd	21.14	2	$2.567e-05$
Pas	19.91	2	$4.743e-05$
Pad	56.38	2	$5.708e-13$
GrSal24	52.67	2	$3.653e-12$
GrsDiaFyV	35.92	2	$1.582e-08$
AnosCur	35.72	2	$1.751e-08$
Imc	17.91	2	$1.293e-04$

Tabla 6.3: Test de Normalidad Multivariante inicial, mujeres

Var	E	df (2p)	$P(Chi > E)$
X	747.675	28	$2.043e-139$

Tabla 6.4: Test de Normalidad Univariante inicial, mujeres

Var	E	df (2)	$P(Chi > E)$
Edad	60.99	2	$5.677e-14$
GluBasal	21.29	2	$2.376e-05$
CreaSang	73.02	2	$1.391e-16$
PotasioR	7.03	2	$2.974e-02$
SodioR	52.78	2	$3.461e-12$
Hdl	92.08	2	$1.011e-20$
Ldl	68.15	2	$1.587e-15$
Tgd	43.39	2	$3.767e-10$
Pas	44.29	2	$2.416e-10$
Pad	102.18	2	$6.485e-23$
GrSal24	23.08	2	$9.712e-06$
GrsDiaFyV	64.37	2	$1.049e-14$
AnosCur	60.18	2	$8.555e-14$
Imc	34.81	2	$2.764e-08$

Como es posible observar, en ambos casos la matriz de datos no sigue una distribución Normal Multivariante, se rechaza la hipótesis nula de Normalidad con un $p-valor$ menor a 0.0001. De hecho, el test aplicado en forma individual a cada una de las variables consideradas, es rechazado en todos los casos.

6.2.2. Aplicación de la distribución Nonparanormal

Dado que el supuesto de Normalidad se debe cumplir para ajustar una RBG, se aplica a los datos la transformación Nonparanormal introducida por Liu et al. [29] y presentada en la Sección 3.2.

Para encontrar la matriz **Z** definida por 3.2 se utiliza la función *huge.npn* que se encuentra implementada en el paquete *High-dimensional Undirected Graph Estimation* para R (huge [45]).

Así, a partir de la matriz de datos con las variables consideradas de la ENS para hombres y para mujeres, se obtienen nuevas matrices con datos transformados, tal que en ambos casos se espera se acepte la hipótesis nula de Normalidad Multivariante.

Luego, se vuelve a revisar el supuesto de Normalidad, pero esta vez de la nueva matriz **Z**; para esto, nuevamente se considera el test de Normalidad Multivariante de Doornick and Hansen [12].

A continuación se muestra el análisis de Normalidad para la matriz transformada, en primer lugar para hombres (Tablas 6.5 y 6.6) y luego para mujeres (Tablas 6.7 y 6.8).

Base de datos de hombres

Tabla 6.5: Test de Normalidad Multivariante, hombres

Var	E	df (2p)	$P(Chi > E)$
X	30.873	28	0.3228

Tabla 6.6: Test de Normalidad Univariante, hombres

Var	E	df (2)	$P(Chi > E)$
Edad	1.201	2	0.548
GluBasal	0.675	2	0.713
CreaSang	0.455	2	0.796
PotasioR	6.288	2	0.044
SodioR	3.273	2	0.195
Hdl	0.126	2	0.938
Ldl	0.284	2	0.867
Tgd	1.769	2	0.413
Pas	1.044	2	0.593
Pad	2.280	2	0.319
GrSal24	5.557	2	0.062
GrsDiaFyV	0.270	2	0.874
AnosCur	4.482	2	0.106
Imc	3.226	2	0.199

Tabla 6.7: Test de Normalidad Multivariante, mujeres

Var	E	df (2p)	$P(Chi > E)$
X	31.795	28	0.2829

Tabla 6.8: Test de Normalidad Univariante, mujeres

Var	E	df (2)	$P(Chi > E)$
Edad	8.949	2	0.011
GluBasal	0.043	2	0.979
CreaSang	1.714	2	0.424
PotasioR	6.758	2	0.034
SodioR	2.144	2	0.342
Hdl	1.206	2	0.547
Ldl	0.232	2	0.890
Tgd	2.151	2	0.341
Pas	1.465	2	0.481
Pad	1.585	2	0.453
GrSal24	1.166	2	0.558
GrsDiaFyV	2.248	2	0.325
AnosCur	1.559	2	0.459
Imc	0.574	2	0.751

De esta manera, en ambos casos la matriz de datos ya cumple con el supuesto de Normalidad Multivariante necesario para ajustar una RBG. Luego, con el supuesto de Normalidad ya controlado, es posible ajustar una RBG, de forma independiente, para hombres y mujeres.

6.3. Ajuste de la RBG

Para realizar el aprendizaje de la estructura de la red y las estimaciones de los parámetros, se consideran las siguientes variables:

- **Edad** Edad de la persona al momento de ser aplicada la encuesta.
- **GluBasal** Glucosa basal en ayuno.
- **CREASANG** Creatinina en sangre
- **POTASIOR** Electrolito K (orina)
- **SODIOR** Electrolito Na (orina)
- **HDL** Colesterol HDL
- **LDL** Colesterol LDL
- **TGD** Triglicéridos
- **PAS** Presión arterial sistólica
- **PAD** Presión arterial diastólica
- **GRSAL24** Consumo diario de sal (gramos)
- **GRSDIAFYV** Consumo diario de frutas y verduras (gramos)
- **ANOSCUR** Número de años cursados
- **IMC** Indice de masa corporal ($Peso/Talla^2$). Resultado del modelo.

La variable de interés es IMC.

6.3.1. Proceso de aprendizaje

Para ajustar una RBG a estos datos, se ha utilizado el paquete *bnlearn* de R creado por Scutari [37].

Como ya se mencionó en la Sección 3.1.2, para ajustar la red se trabaja con el algoritmo híbrido $max-min\ hill-climbing$ (mmhc) presentado por Tsamardinos et al [44].

RBG para Hombres

La Figura 6.3 muestra la RBG ajustada para hombres de acuerdo al procedimiento $max - min\ hill - climbing$ ya descrito.

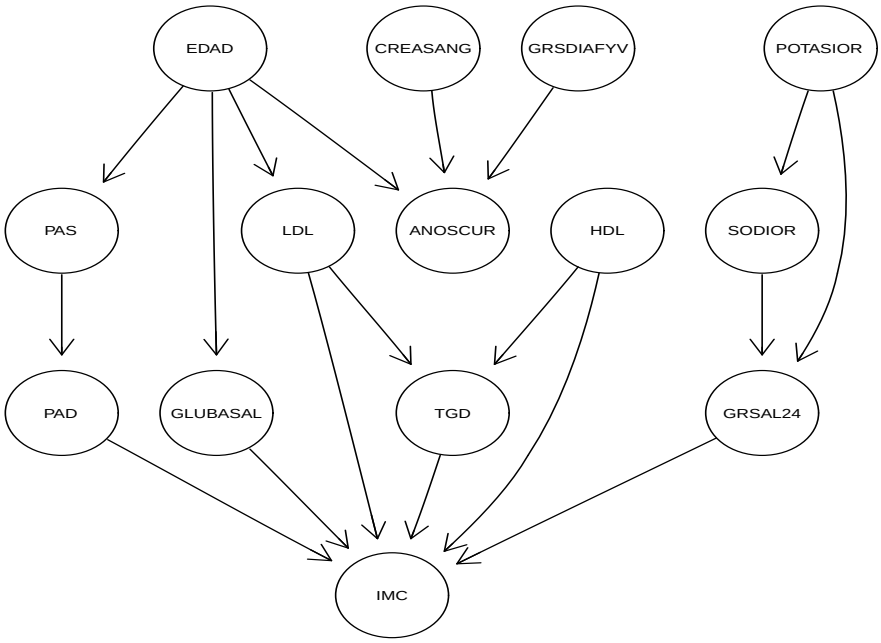

Figura 6.3: RBG inicial ENS hombres

El programa *bnlearn* cuenta con la función *arc.strength*, que permite conocer la importancia o $fuerza$ de cada enlace de la red ajustada. Cada uno de estos coeficientes de fuerza, corresponde al aumento o disminución que se produciría al eliminar dicho

enlace de la red en la medida total de la red. Es decir, enlaces con coeficientes positivos, indican que la eliminación del enlace produce una mejoría en la red, por lo tanto son enlaces no significativos. Por otro lado, mientras más alto sea, en valor negativo, el coeficiente, implica que el enlace tiene una mayor importancia. En la Tabla 6.9 se muestra los coeficientes obtenidos de la red mostrada en la Figura 6.3.

Tabla 6.9: Coeficientes de fuerza RBG hombres

	desde	hacia	strength
1	PAS	PAD	-90.2
2	$EDAD$	PAS	-28.3
3	HDL	TGD	-24.1
4	TGD	IMC	-1.8
5	$EDAD$	$ANOSCUR$	-25.4
6	$POTASIOR$	$GRSAL24$	-42.4
7	$SODIOR$	$GRSAL24$	-38.9
8	$EDAD$	$GLUBASAL$	-12.6
9	PAD	IMC	-5.4
10	$POTASIOR$	$SODIOR$	-10.7
11	$CREASANG$	$ANOSCUR$	-8.6
12	$EDAD$	LDL	-7.6
13	LDL	TGD	-7.4
14	$GRSDIAFYV$	$ANOSCUR$	-1.8
15	HDL	IMC	-2.1
16	$GRSAL24$	IMC	-2.1
17	LDL	IMC	-0.6
18	$GLUBASAL$	IMC	-0.3

Por defecto, el programa considera un umbral igual a *cero* para distinguir enlaces significativos, es por eso que todos los coeficientes de la Tabla 6.9 son negativos. Sin embargo, se puede observar que el coste de eliminar los enlaces 17 y 18 es muy bajo en relación a los restantes enlaces. En la Figura 6.4, obtenida directamente con el programa *bnlearn*, se observa por el grosor de la lineas, la fuerza de cada enlace. Se han dejado con linea de puntos los enlaces más débiles.

Con el objetivo de obtener un modelo plausible, y en consideración al bajo coste que tiene borrar los enlaces 17 y 18, se decide eliminar estos de la red. En la Figura 6.5 se muestra la red ajustada definitiva con la que se continuará trabajando.

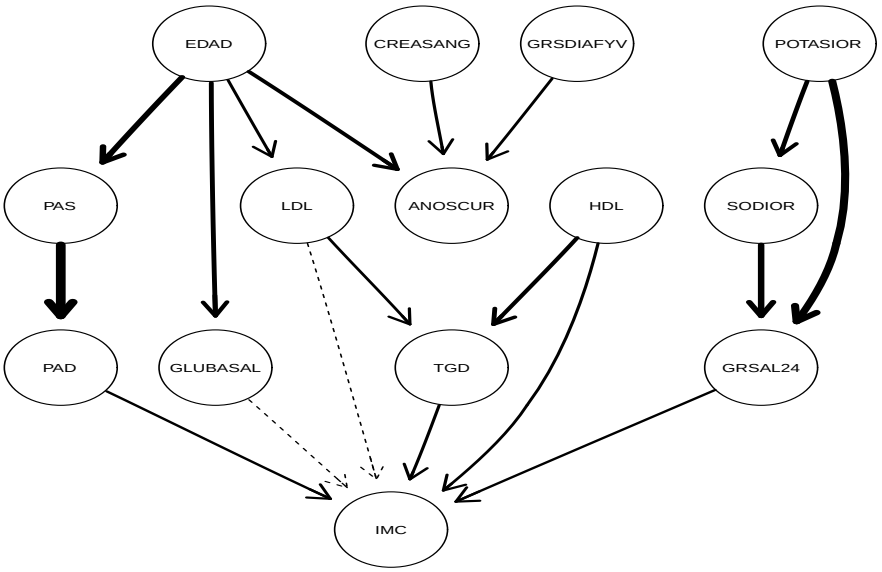

Figura 6.4: Fuerza de enlaces en RBG hombres

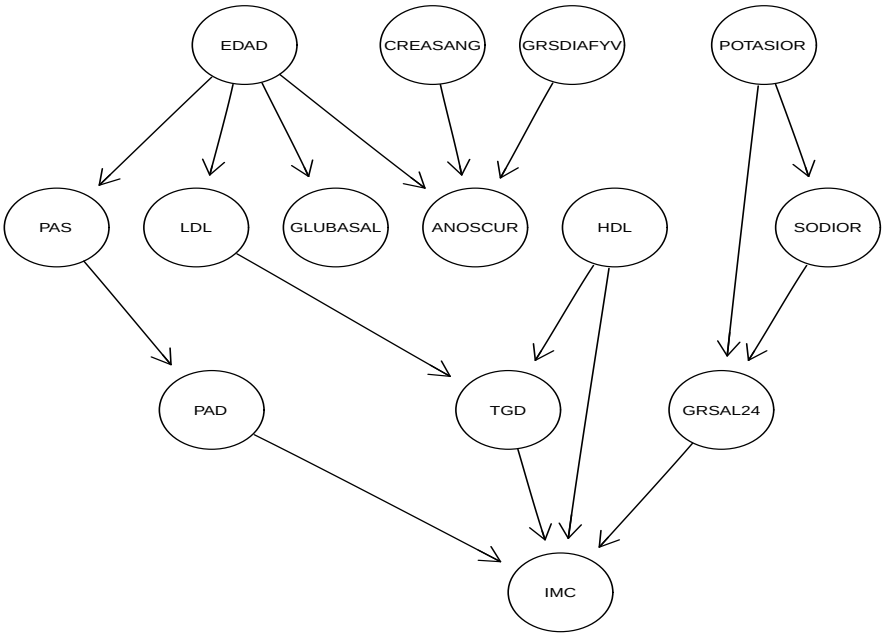

Figura 6.5: RBG final ENS hombres

A continuación se presentan dos gráficos que corresponden a la verificación de los supuestos del modelo. Las Figuras 6.6 y 6.7 muestran para cada variable, el gráfico de probabilidad Normal y el gráfico de residuos, respectivamente.

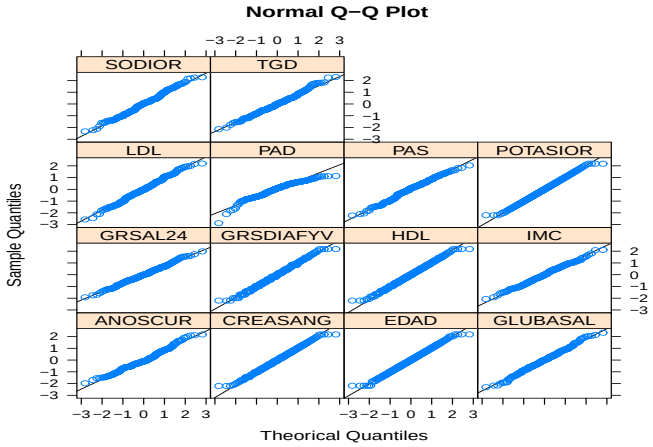

Figura 6.6: Normal Q-Q Plot red hombres

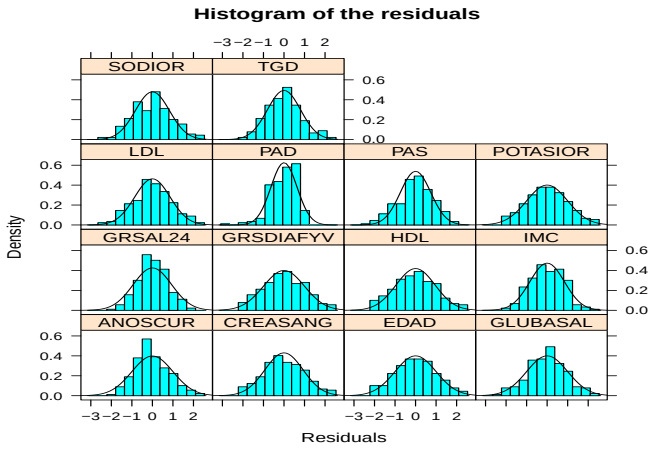

Figura 6.7: Histogram of residuals red hombres

RBG para mujeres

Una primera aproximación a la RBG correspondiente a las mujeres, se muestra en la Figura 6.8.

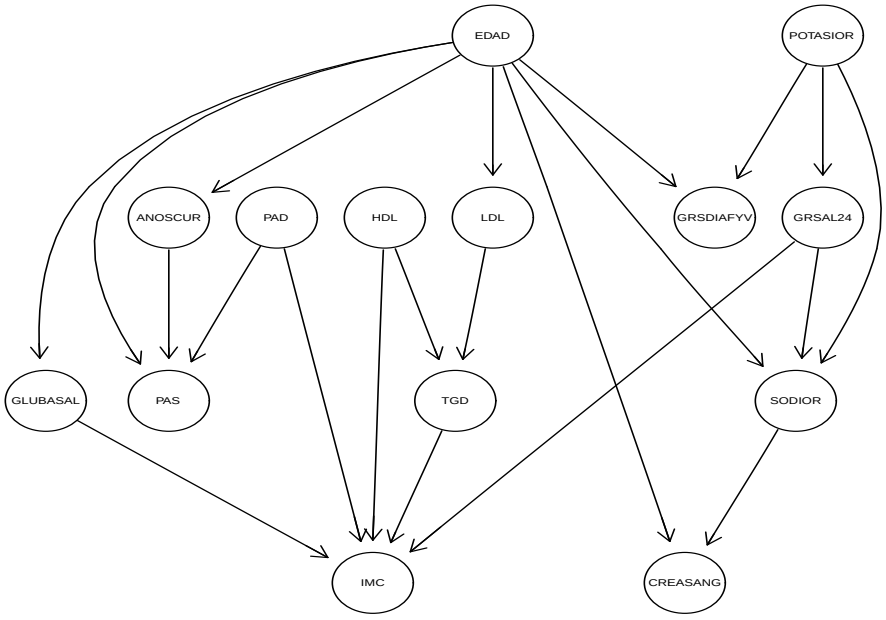

Figura 6.8: RBG inicial ENS mujeres

A continuación, al igual que se hizo para los hombres, en la Tabla 6.10 se muestran los coeficientes de importancia de cada enlace del DAG presentado en la Figura 6.8.

De la Tabla 6.10 se puede observar que sólo el enlace 21 tiene un coeficiente muy débil (menor a -1). En la Figura 6.9, obtenida directamente con el programa *bnlearn*, se observa la fuerza de cada enlace y con linea de puntos el enlace más débil.

Debido al bajo coste de eliminar el enlace 21, se ha decidido eliminarlo de la red. En la Figura 6.10 se muestra el DAG definitivo para las mujeres.

Tabla 6.10: Coeficientes de fuerza RBG mujeres

	from	to	strength
1	$EDAD$	PAS	−55.0
2	PAD	PAS	−73.8
3	$POTASIOR$	$SODIOR$	−62.5
4	$GRSAL24$	$SODIOR$	−43.8
5	$EDAD$	$ANOSCUR$	−27.9
6	$EDAD$	$GLUBASAL$	−27.2
7	$EDAD$	$CREASANG$	−16.7
8	TGD	IMC	−8.9
9	$EDAD$	$SODIOR$	−15.2
10	HDL	TGD	−19.0
11	LDL	TGD	−14.7
12	$GLUBASAL$	IMC	−7.2
13	$POTASIOR$	$GRSAL24$	−11.3
14	$GRSAL24$	IMC	−4.9
15	$EDAD$	$GRSDIAFYV$	−7.0
16	$EDAD$	LDL	−2.9
17	$POTASIOR$	$GRSDIAFYV$	−2.5
18	$SODIOR$	$CREASANG$	−1.6
19	PAD	IMC	−1.7
20	HDL	IMC	−1.1
21	$ANOSCUR$	PAS	−0.7

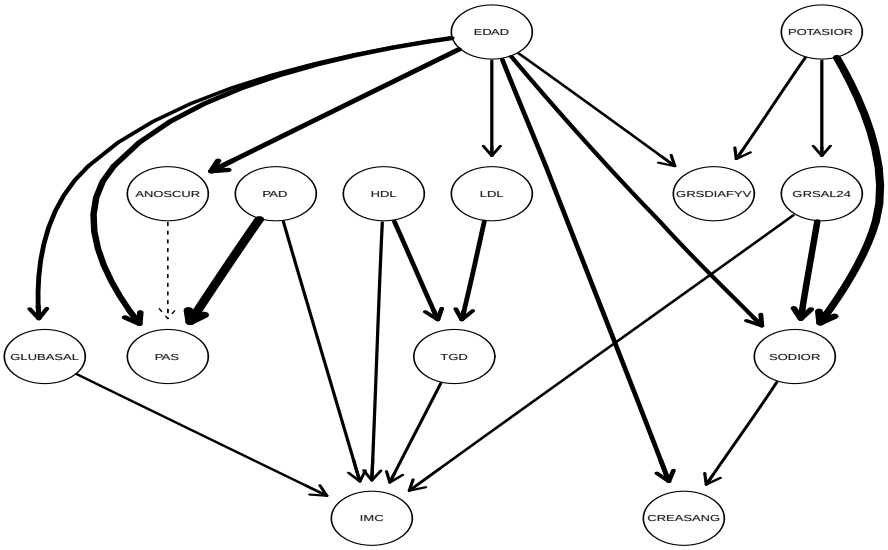

Figura 6.9: Fuerza de enlaces en RBG mujeres

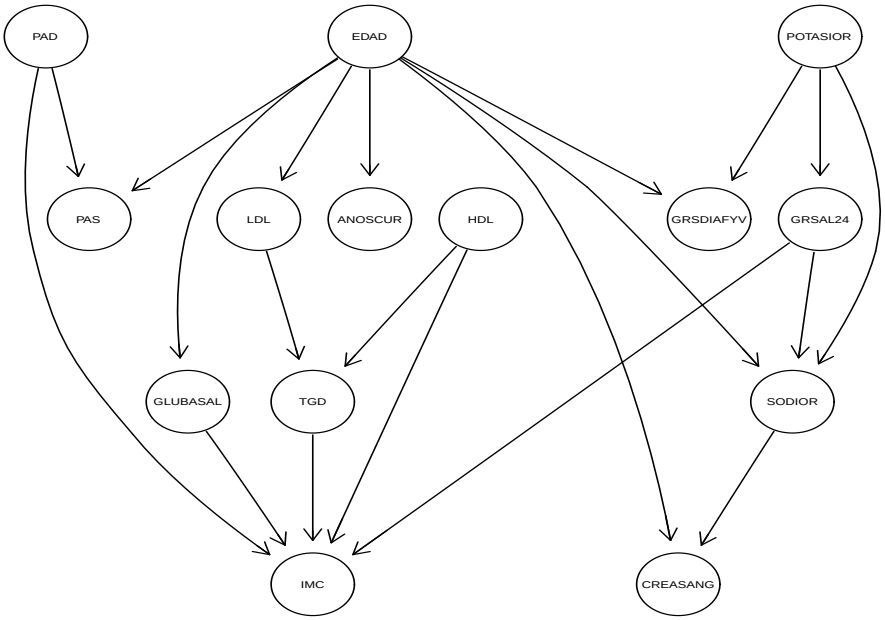

Figura 6.10: RBG final ENS mujeres

En las siguientes Figuras 6.11 y 6.12 se muestra para cada variable, el gráfico de probabilidad Normal y el gráfico de residuos, respectivamente.

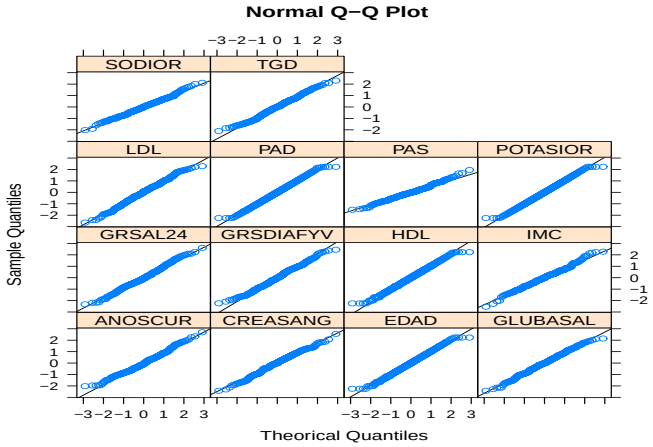

Figura 6.11: Normal Q-Q Plot red mujeres

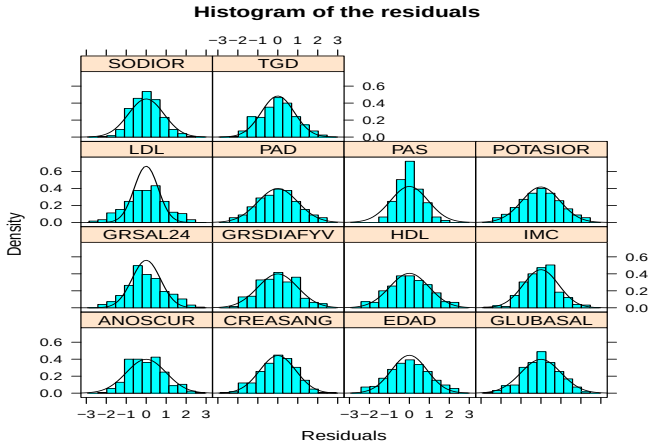

Figura 6.12: Histogram of residuals red mujeres

Finalmente, antes de continuar con la siguiente Sección, es necesario asignar nombre a cada variable de acuerdo a la notación utilizada durante este trabajo. De acuerdo con esto, la notación que será utilizada de aquí en adelante, es la siguiente:

- X_1 **: Edad** Edad de la persona al momento de ser aplicada la encuesta.
- X_2 **: GluBasal** Glucosa basal en ayuno.
- X_3 **: CREASANG** Creatinina en sangre.
- X_4 **: POTASIOR** Electrolito K (orina)
- X_5 **: SODIOR** Electrolito Na (orina)
- X_6 **: HDL** Colesterol HDL
- X_7 **: LDL** Colesterol LDL
- X_8 **: TGD** Triglicéridos
- X_9 **: PAS** Presión arterial sistólica
- X_10 **: PAD** Presión arterial diastólica
- X_{11} **: GRSAL24** Consumo diario de sal (gramos)
- X_{12} **: GRSDIAFYV** Consumo diario de frutas y verduras (gramos)
- X_{13} **: ANOSCUR** Número de años cursados
- X_{14} **: imc** Indice de masa corporal ($Peso/Talla^2$). Variable de interés del modelo.

6.3.2. Parámetros estimados

De acuerdo a lo presentado en la Sección 3.1.2, y a las redes Bayesianas Gaussianas recién ajustadas para hombres y mujeres, cuyos DAG se muestran en las Figuras 6.5 y 6.10 respectivamente; a continuación se introducen los valores estimados de cada red considerando que la densidad conjunta se puede escribir como el producto de las siguientes densidades condicionadas:

$$f(x_i \mid pa(x_i)) \sim N\left(\mu_i + \sum_{j=1}^{i-1} \beta_{ji}(x_j - \mu_j), v_i\right), \tag{6.1}$$

donde β_{ji} es el coeficiente de regresión de X_j en la regresión de X_i condicionada en sus padres y $v_i = \Sigma_i - \Sigma_{ipa(X_i)}\Sigma^{-1}_{pa(X_i)}\Sigma^T_{ipa(X_i)}$ es la varianza condicionada de X_i dado sus padres, con Σ_i la varianza de X_i, $\Sigma_{ipa(X_i)}$ el vector de covarianzas entre X_i y las variables del conjunto $pa(X_i)$, y $\Sigma_{pa(X_i)}$ la matriz de covarianzas de los $pa(X_i)$.

La función *bn.fit* implementada en el programa *bnlearn*, permite obtener los valores estimados de cada red para las densidades condicionadas escritas como:

$$f(x_i \mid pa(x_i)) \sim N\left(Intercepto + \sum_{j=1}^{i-1} \beta_{ji} x_j, v_i\right), \tag{6.2}$$

Luego, es trivial demostrar que se cumple que, de acuerdo a la notación formulada en 6.3:

$$Intercepto = \mu_i - \sum_{j=1}^{i-1} \beta_{ji}\mu_j$$

De acuerdo a esta última consideración, a continuación se determinan los parámetros estimados para la red correspondiente a los hombres y luego para la de mujeres.

Parámetros estimados RBG para hombres

Se obtienen los siguientes valores estimados para el DAG presentado en la Figura 6.5; esto corresponde a la parte cuantitativa de la red Bayesiana Gaussiana ajustada. En este caso, de acuerdo al correspondiente DAG, se considera la ordenación
$\mathbf{X_H} = \{X_1, X_2, X_3, X_4, X_5, X_{12}, X_7, X_6, X_9, X_{10}, X_{11}, X_{13}, X_8, X_{14}\}$

$$\boldsymbol{\mu} = \begin{pmatrix} 0.00878 \\ 0.01093 \\ 0.01087 \\ 0.01061 \\ 0.01061 \\ 0.01117 \\ 0.01061 \\ 0.01098 \\ 0.01061 \\ 0.01056 \\ 0.01353 \\ 0.00780 \\ 0.01061 \\ 0.01059 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 0 & 0.368 & 0 & 0 & 0 & 0 & 0.306 & 0 & 0.507 & 0 & 0 & -0.467 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.285 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.349 & 0 & 0 & 0 & 0 & 0 & -0.582 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.555 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.175 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 269 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.457 & -0.192 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.769 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.301 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.170 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.264 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.859 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.992 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.994 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.876 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0.990 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.904 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.992 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.741 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.407 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.547 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.690 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.707 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.653 \end{pmatrix},$$

con $\boldsymbol{\mu}$ vector de medias, $\mathbf{D}$ matriz diagonal de varianzas v_i y $\mathbf{B}$ matriz triangular superior con los coeficientes de regresión β_{ji} para X_j padre de X_i.

Considerando que la matriz $\boldsymbol{\Sigma}$ se puede calcular como:

$$\boldsymbol{\Sigma} = [(\mathbf{I} - \mathbf{B})^{-1}]^T \mathbf{D} (\mathbf{I} - \mathbf{B})^{-1}$$

Se obtiene que el vector aleatorio
$\mathbf{X_H} = \{X_1, X_2, X_3, X_4, X_5, X_{12}, X_7, X_6, X_9, X_{10}, X_{11}, X_{13}, X_8, X_{14}\}$ es una variable aleatoria $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con parámetros,

$$\boldsymbol{\mu} = \begin{pmatrix} 0.00878 \\ 0.01093 \\ 0.01087 \\ 0.01061 \\ 0.01061 \\ 0.01117 \\ 0.01061 \\ 0.01098 \\ 0.01061 \\ 0.01056 \\ 0.01353 \\ 0.00780 \\ 0.01061 \\ 0.01059 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.368 & 0 & 0 & 0 & 0 & 0.306 & 0 & 0.507 & 0.389 & 0 & -0.467 & 0.082 & 0.139 \\ 0.368 & 0.995 & 0 & 0 & 0 & 0 & 0.113 & 0 & 0.187 & 0.144 & 0 & -0.172 & 0.030 & 0.051 \\ 0 & 0 & 0.992 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.283 & 0 & 0 \\ 0 & 0 & 0 & 0.994 & 0.346 & 0 & 0 & 0 & 0 & 0 & -0.387 & 0 & 0 & -0.066 \\ 0 & 0 & 0 & 0.346 & 0.997 & 0 & 0 & 0 & 0 & 0 & 0.351 & 0 & 0 & 0.059 \\ 0 & 0 & 0 & 0 & 0 & 0.990 & 0 & 0 & 0 & 0 & 0 & 0.174 & 0 & 0 \\ 0.306 & 0.113 & 0 & 0 & 0 & 0 & 0.998 & 0 & 0.155 & 0.119 & 0 & -0.143 & 0.269 & 0.107 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.992 & 0 & 0 & 0 & 0 & -0.454 & -0.310 \\ 0.507 & 0.187 & 0 & 0 & 0 & 0 & 0.155 & 0 & 0.998 & 0.768 & 0 & -0.236 & 0.042 & 0.242 \\ 0.389 & 0.143 & 0 & 0 & 0 & 0 & 0.119 & 0 & 0.768 & 0.998 & 0 & -0.182 & 0.032 & 0.309 \\ 0 & 0 & 0 & -0.387 & 0.351 & 0 & 0 & 0 & 0 & 0 & 0.968 & 0 & 0 & 0.164 \\ -0.467 & -0.172 & 0.283 & 0 & 0 & 0.174 & -0.143 & 0 & -0.236 & -0.182 & 0 & 1.019 & -0.038 & -0.065 \\ 0.082 & 0.030 & 0 & 0 & 0 & 0 & 0.269 & -0.454 & 0.042 & 0.032 & 0 & -0.038 & 0.987 & 0.358 \\ 0.139 & 0.051 & 0 & -0.066 & 0.059 & 0 & 0.107 & -0.310 & 0.242 & 0.309 & 0.164 & -0.065 & 0.358 & 0.928 \end{pmatrix}$$

Parámetros estimados RBG para mujeres

En este caso, se obtienen los valores estimados para el DAG presentado en la Figura 6.9. De acuerdo a la figura, esta vez se considera la ordenación
$\mathbf{X_M} = \{X_1, X_2, X_4, X_{11}, X_5, X_6, X_7, X_8, X_{10}, X_9, X_3, X_{12}, X_{13}, X_{14}\}$

$$\boldsymbol{\mu} = \begin{pmatrix} 0.00819 \\ 0.00831 \\ 0.00798 \\ 0.00799 \\ 0.00801 \\ 0.00823 \\ 0.00801 \\ 0.00803 \\ 0.00798 \\ 0.00801 \\ 0.00816 \\ 0.00797 \\ 0.00229 \\ 0.00797 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 0 & 0.439 & 0 & 0 & -0.264 & 0 & 0.201 & 0 & 0 & 0.522 & 0.367 & 0.258 & -0.439 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.235 \\ 0 & 0 & 0 & -0.305 & 0.578 & 0 & 0 & 0 & 0 & 0 & 0 & 0.188 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.474 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.199 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.169 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.364 & 0 & 0 & 0 & 0 & 0 & -0.146 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.323 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.265 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.519 & 0 & 0 & 0 & 0.154 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.002 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.885 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.508 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.999 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.964 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.789 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.001 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.363 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.799 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.912 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.789 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.681 \end{pmatrix},$$

con $\boldsymbol{\mu}$ vector de medias, $\mathbf{D}$ matriz diagonal de varianzas v_i, $\mathbf{B}$ matriz triangular superior con los coeficientes de regresión β_{ji} para X_j padre de X_i.

Considerando que la matriz $\boldsymbol{\Sigma}$ se puede calcular como:

$$\boldsymbol{\Sigma} = [(\mathbf{I} - \mathbf{B})^{-1}]^T \mathbf{D} (\mathbf{I} - \mathbf{B})^{-1}$$

Se obtiene que el vector aleatorio
$\mathbf{X_M} = \{X_1, X_2, X_4, X_{11}, X_5, X_6, X_7, X_8, X_{10}, X_9, X_3, X_{12}, X_{13}, X_{14}\}$
es una variable aleatoria $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con parámetros,

$$\boldsymbol{\mu} = \begin{pmatrix} 0.00819 \\ 0.00831 \\ 0.00798 \\ 0.00799 \\ 0.00801 \\ 0.00823 \\ 0.00801 \\ 0.00803 \\ 0.00798 \\ 0.00801 \\ 0.00816 \\ 0.00797 \\ 0.00229 \\ 0.00797 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.439 & 0 & 0 & -0.264 & 0 & 0.201 & 0.065 & 0 & 0.522 & 0.412 & 0.258 & -0.439 & 0.120 \\ 0.439 & 1.002 & 0 & 0 & -0.116 & 0 & 0.088 & 0.028 & 0 & 0.229 & 0.181 & 0.113 & -0.193 & 0.243 \\ 0 & 0 & 1.002 & -0.305 & 0.434 & 0 & 0 & 0 & 0 & 0 & -0.073 & 0.188 & 0 & -0.061 \\ 0 & 0 & -0.305 & 0.979 & 0.287 & 0 & 0 & 0 & 0 & 0 & -0.048 & -0.057 & 0 & 0.195 \\ -0.264 & -0.116 & 0.434 & 0.287 & 0.965 & 0 & -0.053 & -0.017 & 0 & -0.138 & -0.259 & 0.013 & 0.116 & 0.025 \\ 0 & 0 & 0 & 0 & 0 & 0.999 & 0 & -0.364 & 0 & 0 & 0 & 0 & 0 & -0.242 \\ 0.201 & 0.088 & 0 & 0 & -0.053 & 0 & 1.005 & 0.324 & 0 & 0.105 & 0.083 & 0.052 & -0.088 & 0.107 \\ 0.065 & 0.028 & 0 & 0 & -0.017 & -0.364 & 0.324 & 1.026 & 0 & 0.034 & 0.027 & 0.017 & -0.028 & 0.332 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.001 & 0.519 & 0 & 0 & 0 & 0.154 \\ 0.522 & 0.229 & 0 & 0 & -0.137 & 0 & 0.105 & 0.034 & 0.519 & 0.905 & 0.215 & 0.135 & -0.229 & 0.143 \\ 0.412 & 0.181 & -0.073 & -0.048 & -0.259 & 0 & 0.083 & 0.027 & 0 & 0.215 & 0.994 & 0.092 & -0.181 & 0.039 \\ 0.258 & 0.113 & 0.188 & -0.057 & 0.013 & 0 & 0.052 & 0.017 & 0 & 0.135 & 0.092 & 1.014 & -0.113 & 0.019 \\ -0.439 & -0.193 & 0 & 014 & 0.116 & 0 & -0.088 & -0.028 & 0 & -0.229 & -0.181 & -0.113 & 0.981 & -0.053 \\ 0.120 & 0.243 & -0.061 & 0.195 & 0.025 & -0.242 & 0.107 & 0.332 & 0.154 & 0.143 & 0.039 & 0.019 & -0.053 & 0.923 \end{pmatrix}$$

Con estos parámetros ya estimados, se puede aplicar los procedimientos propuestos en este trabajo.

Es importante notar que al aplicar la transformación a la normalidad, las varianzas y covarianzas de la matriz Σ son pequeñas porque corresponde a las variables transformadas. Sin embargo, para efecto comparativo de importancia relativa, esto no afecta al análisis y escaladas de sensibilidad.

6.4. Análisis de sensibilidad

Como ya se dijo, la variable de interés de la red es $X_{14} : IMC$. A continuación se aplicarán los procedimientos propuestos en 5.3.1 y 5.4.1.

Análisis de sensibilidad para RBG de los hombres

La entropía diferencial inicial calculada es $h(X_{14}) = 1.3817$; los valores de entropía, información mutua e información mutua normalizada de las variables no evidenciales con la variable respuesta, se muestran en la Tabla 6.11. Se mantiene el orden original de las variables para facilitar la comparación entre hombres y mujeres, aunque no se corresponde en algunos casos con el orden ancestral de la red correspondiente.

Tabla 6.11: Aplicación. Paso 2. RBG hombres.

Y_{-i}	$h(Y_{-i})$	$h(X_{14}, Y_{-i})$	$I(X_{14}; Y_{-i})$	$IN(X_{14}; Y_{-i})$
$X_1 : Edad$	1.4189	2.7901	0.0105	0.0075
$X_2 : GluBasal$	1.4165	2.7967	0.0014	0.0010
$X_3 : Creasang$	1.4149	2.7966	0	0
$X_4 : PotasioR$	1.4159	2.7953	0.0023	0.0017
$X_5 : SodioR$	1.4174	2.7971	0.0019	0.0014
$X_6 : HDL$	1.4149	2.7413	0.0553	0.0396
$X_7 : LDL$	1.4181	2.7936	0.0062	0.0044
$X_8 : TGD$	1.4124	2.7189	0.0752	0.0538
$X_9 : PAS$	1.4179	2.7668	0.0328	0.0234
$X_{10} : PAD$	1.4179	2.7451	0.0545	0.0389
$X_{11} : GrSal24$	1.4025	2.7689	0.0153	0.0109
$X_{12} : GrsDiaFyV$	1.4139	2.7956	0	0
$X_{13} : AnosCur$	1.4287	2.8081	0.0022	0.0016

Análisis de sensibilidad para RBG de las mujeres

La entropía diferencial inicial calculada es $h(X_{14}) = 1.379$; los valores de entropía, información mutua e información mutua normalizada de las variables no evidenciales con la variable respuesta, se muestran en la Tabla 6.12.

A partir de las Tablas 6.11 y 6.12, se pueden sacar algunas conclusiones respecto a la importancia relativa de cada variable en las RBG ajustadas para hombres y mujeres. Sin embargo, con el objetivo de facilitar esta comparación, en la siguiente Tabla 6.13 se muestra la razón entre la IN de hombres y la IN de mujeres.

A continuación, se presentan las principales conclusiones que se pueden obtener a partir de la Tabla 6.13:

Tabla 6.12: Aplicación. Paso 2. RBG mujeres.

Y_{-i}	$h(Y_{-i})$	$h(X_{14}, Y_{-i})$	$I(X_{14}; Y_{-i})$	$IN(X_{14}; Y_{-i})$
$X_1 : Edad$	1.4189	2.7904	0.0079	0.0056
$X_2 : GluBasal$	1.4202	2.7666	0.0329	0.0236
$X_3 : Creasang$	1.4161	2.7946	0.0008	0.0006
$X_4 : PotasioR$	1.4198	2.7972	0.0020	0.0014
$X_5 : SodioR$	1.4012	2.7802	0.0003	0.0003
$X_6 : HDL$	1.4186	2.7651	0.0328	0.0235
$X_7 : LDL$	1.4213	2.7945	0.0062	0.0044
$X_8 : TGD$	1.4320	2.7497	0.0617	0.0439
$X_9 : PAS$	1.3689	2.7359	0.0123	0.0089
$X_{10} : PAD$	1.4195	2.7859	0.0130	0.0093
$X_{11} : GrSal24$	1.4082	2.7661	0.0214	0.0154
$X_{12} : GrsDiaFyV$	1.4259	2.8051	0.0002	0.0001
$X_{13} : AnosCur$	1.4096	2.7874	0.0015	0.0011

Tabla 6.13: Razón de Información mutua normalizada hombres vs. mujeres.

Variable	$IN_H(X_{14}; Y_{-i})$	$IN_M(X_{14}; Y_{-i})$	$IN_H(X_{14}; Y_{-i})/IN_M(X_{14}; Y_{-i})$
$X_1 : Edad$	0.0075	0.0056	1.339
$X_2 : GluBasal$	0.0010	0.0236	0.042
$X_3 : Creasang$	0	0.0006	≈ 1
$X_4 : PotasioR$	0.0017	0.0014	1.214
$X_5 : SodioR$	0.0014	0.0003	4.667
$X_6 : HDL$	0.0396	0.0235	1.685
$X_7 : LDL$	0.0044	0.0044	1
$X_8 : TGD$	0.0538	0.0439	1.225
$X_9 : PAS$	0.0234	0.0089	2.629
$X_{10} : PAD$	0.0389	0.0093	4.183
$X_{11} : GrSal24$	0.0109	0.0154	0.708
$X_{12} : GrsDiaFyV$	0	0.0001	≈ 1
$X_{13} : AnosCur$	0.0016	0.0011	1.454

- En el caso de los hombres, las variables que proporcionan un mayor aporte a la reducción de entropía del objetivo IMC son los Triglicéridos (TGD), la Presión arterial diastólica (PAD) y el Colesterol HDL (HDL).
- En las mujeres, las principales variables, en el mismo sentido, son los Triglicéridos (TGD), la Glucosa basal (GluBasal) y el Colesterol HDL (HDL).
- Las variables antes mencionadas, contribuyen mejor a la disminución de entropía de IMC en la RBG ajustada para los hombres que en la red ajustada para las mujeres.
- Tanto en los hombres como en las mujeres, las variables Creatinina en sangre (Creasang) y grs. diarios de consumo de frutas y verduras (GrsDiaFyV), no intervienen en la disminución de entropía de IMC. De hecho, si se observan las Figuras 6.5 y 6.10 es posible notar que ambas variables, condicionadas por sus padres, son independientes de IMC.
- La importancia relativa de Glucosa basal (GluBasal) es claramente mayor en las mujeres que en los hombres, por lo que sería recomendable medir esta variable, para estos efectos, sólo en las mujeres.
- Por otro lado, las presiones sistólica (PAS) y diastólica (PAD) son claramente más importantes en los hombres, de hecho en las mujeres no contribuyen a la disminución de entropía de IMC.
- Se debe tener cuidado al observar un índice muy alto, no se debe dejar de mirar en forma paralela la información mutua normalizada que aporta la variable en cada caso; así, por ejemplo, es claro que la variable Sodio (SodioR) es más importante para los hombres que para las mujeres, sin embargo, en ambos casos su aporte es muy pequeño para ser considerado.

Capítulo 7

Comentarios y futuras líneas de investigación

En este trabajo se ha propuesto una metodología para análizar la evidencia en redes Bayesianas Gaussianas, basada en la Teoría de la Información.

La principal motivación de esta propuesta nació de la idea de que no todas las variables de una RBG influyen de igual manera en la variable de interés. Es importante recalcar en este contexto, que el objetivo del procedimiento consiste en identificar entre las variables disponibles, cuál o cuáles, intervienen más para disminuir la entropía del objetivo, entendiendo ésto como una disminución de la incertidumbre.

Con respecto a la Aplicación presentada en el Capítulo 6, esto se puede describir de la siguiente manera:

Si interesa modelizar el Indice de Masa Corporal, de acuerdo a variables provenientes de tres fuentes distintas: encuesta a los participantes, mediciones biofisiológicas y exámenes de laboratorio. Es posible en primer lugar ajustar la red y conocer la estructura de interacción entre las variables. Sin embargo, ante la posibilidad de observar algunas de esas variables en un paciente de interés, ¿es necsario observarlas todas para entender su IMC? o ¿será suficiente con observar algunas? En ese caso, ¿cuáles? Claramente, obtener esta información para un paciente, tiene un coste asociado, razón por la cuál tiene sentido responder a estas preguntas.

Este trabajo responde a estas cuestiones a partir del procedimiento propuesto, basa-

do en Teoría de la Información, que permite priorizar la evidencia para disminuir la entropía de la variable de interés, en este caso, el IMC. De esta manera, y entendiendo la disminución de la entropía como un aumento en la certeza, este procedimiento permite identificar exactamente no solo qué variables aportan mayor información al objetivo, sino además, cuáles no aportan significativamente, lo que implica que no hay razón alguna que justifique su medición.

Una segunda aportación del trabajo, consiste en la utilización de la información mutua normalizada para detectar el valor informativo de cada variable respecto del objetivo. Esta extensión al procedimiento, permite que además de priorizar variables en el sentido ya expuesto, sea posible comparar el aporte de diferentes variables alternativas en una misma red, o variables similares en redes distintas. Así, se presenta un ejemplo en que interesa saber si el reemplazamiento de un nodo en la red tiene un aporte informativo mayor o menor que el nodo original. En relación a la aplicación hecha a datos reales, su utilización permitió realizar el siguiente análisis:

Bajo el supuesto de que las variables consideradas no interactúan de igual manera en hombres y mujeres al explicar el IMC, se ajustaron dos RBG, una para cada sexo. Efectivamente, en los correspondientes DAGs fue posible confirmar esta hipótesis. Luego, para comparar la aportación informativa de las diferentes variables sobre la entropía de IMC, se calculó la información mutua normalizada. De esta manera, para cada variable, se pudo comparar su aporte en cada una de las redes. De manera adicional, se calculó un índice que corresponde a la razón entre ambas informaciones mutuas normalizadas, de manera que se ha obtenido una medida de importancia relativa de cada variable en hombres versus mujeres.

Hasta ahora, en general, los análisis en RBG desarrollados, habían sído de tipo teórico. Luego, una aportación importante de este trabajo fue realizar una aplicación a datos reales. Sin embargo, este procedimiento llevó a enfrentar algunos hechos que no estaban considerados de manera teórica. En primer lugar, la revisión del supuesto de Normalidad. Al trabajar con redes ajustadas bajo el enfoque tradicional, se asume la Normalidad y se desarrollan los análisis sin problema. En cambio, al considerar una base de datos real, se hace necesario revisar el cumplimiento del supuesto de Normalidad. Así, en este trabajo se hizo una búsqueda de diferentes metodologías, y se aplicó un test de Normalidad Multivariante que permite responder a este supuesto.

Por otro lado, frente al rechazo de la hipótesis de Normalidad, queda planteado un nuevo problema: cómo resolver la falta de Normalidad para poder ajustar una Red Bayesiana Gaussiana. Para responder, se propone considerar para la matriz de datos la distribución Nonparanormal, que a través de una transformación semiparamétrica, permite obtener una matriz de datos que cumple el supuesto de Normalidad. Este hecho toma gran importancia ante la posibilidad de utilizar las redes Bayesianas como

una verdadera herramienta de modelado de la realidad, ya que es muy frecuente que datos reales no sigan una distribución Normal Multivariante.

Respecto a futuras líneas de investigación, esta es un área en que queda mucho por desarrollar, sobre todo frente a un escenario de tratamiento de datos reales.

Una línea de investigación interesante, es extender estos resultados a redes Bayesianas con distribución MEP (Multivariate Power Exponential). Como ya se mencionó, la falta de Normalidad es un problema habitual en datos reales, por lo que considerar una distribución MEP podría relajar algunos supuestos en la distribución y hacerlo más viable. Sin embargo, esto requiere no sólo rehacer los cálculos para el análsis de sensibilidad a la evidencia, sino sobre todo, abrir una amplia línea de investigación respecto a la revisión del supuesto de distribución MEP y su reparación en caso de salir rechazado.

La distribución de Nonparanormal propuesta en este trabajo para reparar la falta de Normalidad, fue introducida por los autores en un escenario de datos de alta dimensión (*high dimensional data*), es decir, para un número de variables muy grande en relación a la cantidad de datos. La Aplicación de este trabajo, no se corresponde con ese escenario, por lo que es una línea de investigación interesante conocer las propiedas y comportamiento de la metodología propuesta en estas situaciones.

Por último, la principal línea de interés a desarrollar en el futuro se refiere al tipo de Redes ajustadas. Cuando interesa modelar situaciones reales, lo más cercano a dicha realidad es considerar que se presentarán escenarios donde interactúan tanto variables continuas como discretas. En este sentido, es muy interesante no sólo tener un buen entrenamiento en ajuste de redes mixtas, sino también, es importante ver si todos los procedimientos propuestos en este trabajo son aplicables a este tipo de redes.

Bibliografía

[1] Aho, K.; many thanks to V. Winston and D. Roberts (2013). asbio: A collection of statistical tools for biologists. R package version 1.1. http://cran.r-project.org/web/packages/asbio/index.html

[2] Ben-Gal, I., Ruggeri, F., Faltin, F. and Kenett, R. (Eds.),(2007) *Bayesian Networks*. Encyclopedia of Statistics in Quality and Reliability, New York: John Wiley and Sons.

[3] Besson, P., Richiardi, J., Bourdin, C., Bringoux, L., Mestre, D.R. and Vercher, J.-L.(2010) *Bayesian networks and information theory for audio-visual perception modeling*. Biological Cybernetics, vol 103, 213-226.

[4] Bowman, K.O. and Shenton, L.R. (1975) *Omnibus test contours for departures from normality based on* $\sqrt{b_1}$ *and* b_2. Biometrika, vol 2, 103-117.

[5] Castillo, E., Gutiérrez, J.M. and Hadi, A.S. (1997) *Expert Systems and Probabilistic Network Models*. New York: Springer Verlag.

[6] Castillo, E. and Kjaerulff, U.(2003) *Sensitivity analysis in Gaussian Bayesian networks using a symbolic-numerical technique*. Reliab Eng Syst Safety 79(2),139-48.

[7] Chan, H. and Darwiche, A. (2004) *Sensitivity analysis in Bayesian networks: From single to multiple parameters*. Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (UAI), AUAI Press, Arlington, Virginia, pp. 67-75.

[8] Cobb, B.R., Rumí, R. and Salmerón, A. (2007) *Bayesian networks models with discrete and continuous variables*. Advances in Probabilistic Graphical Models, Studies in Fuzziness and Soft Computing. Springer 81-102.

[9] Cover, T. and Thomas, J. (1991) *Elements of Information Theory*. New York: John Wiley and Sons.

[10] Dawid, A.P. (1979). *Conditional Independence in Statistical Theory*. Journal of the Royal Statistical Society, Series B 41, 1-31.

[11] Diez, F.J.(2010) *Introducción a los Modelos Gráficos Probabilistas.* Departamento de Ingeniería Artificial. UNED.

[12] Doornik, J.A. and Hansen, H. (2008). *An Omnibus Test for Univariate and Multivariate Normality.* Oxford Bulletin of Economics and Statistics, 70, 927-939

[13] Driver, E. and Morrell, D. (1995). *Implementation of continuous bayesian networks using sums of weighted Gaussians.* In Proceedings of the Eleventh UA1 Conference, 134- 140. Morgan Kaufmann.

[14] ENS 2010

`http://epi.minsal.cl/epi/html/presenta/Taller2011/Dia1/02_ENS_2000-2010.pdf`

[15] Genz, A., Bretz, F. and Hotborn, T. (2007) mvtnorm: Multivariate Normal and t distribution. R package version 0.9-9994, (available at http://cran.r-project.org/doc/packages/mvtnorm.pdf)

[16] Gómez, E., Gómez-Villegas, M.A. and Marín, J.M. (1998). *A multivariate generalization of the power exponential family of distributions.* Communications in Statistics-Theory and Methods, 27:3, 589-600.

[17] Gómez-Villegas, M.A., Main, P. and Susi, R. (2007). *Sensitivity analysis in Gaussian Bayesian networks using a divergence measure.* Communications in Statistics-Theory and Methods, vol 36, 523-539.

[18] Gómez-Villegas, M.A., Main, P., Navarro, H. and Susi, R. (2011). *Evaluating the difference between graph structures in Gaussian Bayesian networks.* Expert Systems with Applications, vol 38, 12409-12414.

[19] Gómez-Villegas, M.A., Maín, P. and Viviani, P. (2014) *Sensitivity to evidence in Gaussian Bayesian networks using mutual information.* Information Sciences (article in press).

[20] BMI http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

[21] Ihara, S. (1993) *Information Theory for Continuous Systems.* Singapore: World Scientific.

[22] Jensen, F. and Nielsen, L. B. (1999). *Hugin Lite.* Aalborg, Denmark, Hugin Expert A/S.

[23] Jensen, F.V. (2001) *Bayesian Networks and Decision Graphs.* Barcelona: Springer.

[24] Kjaerulff, U. and Madsen, A. (2007) *Probabilistics Networks for Practitioners, A Guide to Construction and Analysis of Bayesian Networks and Influence Diagrams.* New York: Springer.

[25] Korb, K., Nicholson, A. (2004). *Bayesian Artificial Intelligence*. Boca Ratón, FL: Chapman and Hall.

[26] Lafferty, J., Liu, H. and Wasserman, L. (2012) *Sparse nonparametric graphical models.* Statistical Science, 27(4), 519-537.

[27] Laskey, K.B. (1995). *Sensitivity analysis for probability assessments in Bayesian networks*. IEEE Trans. Syst., Man, Cybern, vol 25, 901-909.

[28] Lauritzen, S. (1992) *Propagation of probabilities, means and variances in mixed graphical association models*. Journal of the American Statistical Association, 87, 1098-1108.

[29] Liu, H., Lafferty, J. and Wasserman, L. (2009). *The Nonparanormal: Semiparametric estimation of high dimensional undirected graphs.* J. Mach. Learn. Res. 10 2295-2328. MR2563983

[30] Main, P. and Navarro, H. (2009). *Analyzing the effect of introducing a kurtosis parameter in Gaussian Bayesian networks*. Reliability Engineering and System Safety, 94, 922-926.

[31] Malhas, R. and Al Aghbari, Z. (2008) *Using sensitivity of a Bayesian network to discover interesting patterns.* International Conference on Computer Systems and Applications, vols 1-3, 196-205.

[32] Moral, S., Rumí, R., Salmerón, A. (2001) *Mixtures of truncated exponentials in hybrid Bayesian networks*. Lecture Notes in Artificial Intelligence, Vol. 2143, pp. 135-143.

[33] Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann.

[34] R Development Core Team (2013). R: A Language and Environment for Statistical Computing. R Core Team Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org

[35] Richiardi, J. (2007) *Probabilistics models for multi-classifier biometric authentication using quality measures.*These no. 3954, Ecole Polytechnique Fédérale de Laussane, Lausanne, Switzerland.

[36] Russell, SJ., Norvig, P. (2009). *Artificial Intelligence: A Modern Approach.* Boca ratón, FL: Prentice Hall, 3rd edition.

[37] Scutari, M. (2010). *Learning Bayesian Networks with the bnlearn R Package.* Journal of Statistical Software, 35(3):1-22.

[38] Shachter, R. and Kenley, C. (1989). *Gaussian influence diagrams.* Management Science, 35(5),527-550.

[39] Shannon, C.E. (1948) *A mathematical theory of communication* The Bell System Technical Journal, 27, 379-423 and 623-656.

[40] Shenton, L.R. and Bowman, K.O. (1977) *A bivariate model for the distribution of* $\sqrt{b_1}$ *and* b_2. Journal of the American Statistical Association, vol 72, 206-211.

[41] SKLAR, M. (1959). *Fonctions de répartition a n dimensions et leurs marges.* Publ. Inst. Statist. Univ. Paris 8 229-231. MR0125600

[42] Strehl, A. and Ghosh, J. (2002) *Cluster ensembles- a knowledge reuse framework for combining multiple partitions.* J Mach Learn Res, vol 3, 583-617.

[43] Tsamardinos, I., Aliferis, C. F. and Statnikov, A. (2003). *Time and sample efficient discovery of Markov blankets and direct causal relations.* In The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 673-678.

[44] Tsamardinos, I., Brown, LE., Aliferis, CF. (2006). *The max-min hill-climbing Bayesian network structure learning algorithm.* Machine Learning, 65:31-78.

[45] Zhao, T., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2012). *The huge package for high-dimensional undirected graph estimation in r.* The Journal of Machine Learning Research, 98888:1059-1062. R package http://cran.r-project.org/web/packages/huge/index.html