UNIVERSIDAD POLITÉCNICA DE MADRID

E.T.S. DE INGENIEROS INFORMÁTICOS

TESIS DE FIN DE MÁSTER

Fusión de redes Bayesianas Gaussianas

Autora:
Irene Córdoba-Sánchez

Tutores:
Concha Bielza
Pedro Larrañaga

Diciembre 2015

Agradecimientos

Este trabajo ha sido financiado por la Comunidad de Madrid mediante el proyecto S2013/ICE-2845-CASI-CAM-CM.

Resumen

Las redes Bayesianas constituyen un modelo ampliamente utilizado para la representación de relaciones de dependencia condicional en datos multivariantes. Su aprendizaje a partir de un conjunto de datos o expertos ha sido estudiado profundamente desde su concepción. Sin embargo, en determinados escenarios se demanda la obtención de un modelo común asociado a particiones de datos o conjuntos de expertos. En este caso, se trata el problema de fusión o agregación de modelos. Los trabajos y resultados en agregación de redes Bayesianas son de naturaleza variada, aunque escasos en comparación con aquellos de aprendizaje. En este documento, se proponen dos métodos para la agregación de redes Gaussianas, definidas como aquellas redes Bayesianas que modelan una distribución Gaussiana multivariante. Los métodos presentados son efectivos, precisos y producen redes con menor cantidad de parámetros en comparación con los modelos obtenidos individualmente. Además, constituyen un enfoque novedoso al incorporar nociones exploradas tradicionalmente por separado en el estado del arte. Futuras aplicaciones en entornos escalables hacen dichos métodos especialmente atractivos, dada su simplicidad y la ganancia en compacidad de la representación obtenida.

Abstract

Bayesian networks are a widely used model for the representation of conditional dependence relationships among variables in multivariate data. The task of learning them from a data set or experts has been deeply studied since their conception. However, situations emerge where there is a need of obtaining a consensuated model from several data partitions or a set of experts. This situation is referred to as model fusion or aggregation. Results about Bayesian network aggregation, although rich in variety, have been scarce when compared to the learning task. In this context, two methods are proposed for the aggregation of Gaussian Bayesian networks, that is, Bayesian networks whose underlying modelled distribution is a multivariate Gaussian. Both methods are effective, precise and produce networks with fewer parameters in comparison with the models obtained by individual learning. They constitute a novel approach given that they incorporate notions traditionally explored separately in the state of the art. Future applications in scalable computer environments make such models specially attractive, given their simplicity and the gaining in sparsity of the produced model.

Índice general

Índice de figuras

Índice de tablas

1. Introducción

Un *modelo gráfico probabilístico* (PGM) es un par $(\mathcal{G}, \mathcal{F})$ en el que $\mathcal{G}$ es un grafo y $\mathcal{F}$ una familia de distribuciones de probabilidad que satisfacen ciertas independencias (marginales o condicionales) correspondientes a ciertas características del grafo. Por su capacidad de síntesis y representación gráfica de relaciones entre variables, los PGMs se utilizan con frecuencia en el modelado de datos multivariantes. Cuando los datos son continuos y se asume que provienen de una distribución normal, el PGM resultante se denomina *modelo gráfico Gaussiano* (GGM)[1]. En este trabajo nos centraremos en los GGMs que representan independencias condicionales mediante grafos dirigidos acíclicos (DAGs): las *redes Bayesianas Gaussianas* (GBN).

Se considera a Judea Pearl el primero en usar el término *red Bayesiana* [Pearl, 1988]; sin embargo, el desarrollo de estos modelos tiene fuentes en distintos campos. Sus orígenes como representación dirigida de relaciones entre múltiples variables se remontan al trabajo del genetista Sewall Wright en 1918 con el método de *coeficientes de caminos* [Wright, 1934], más adelante llamado *análisis de caminos*, consistente en representar variables relacionadas de forma lineal con un DAG y seguir las flechas a modo de caminos según determinadas reglas para hacer computaciones. Su método fue aplicado más tarde de forma extensiva en áreas como economía [Wold, 1954], ciencias sociales [Blalock, 1971], e incluso hoy en día es un método popular, ver por ejemplo Van Rheenen et al. [2014] en psicología. Las redes Bayesianas han sido dotadas de una interpretación causal al ser fácil asociar su naturaleza dirigida a una relación causa-efecto entre las variables. Este hecho ha dado lugar a algunas controversias [McKim y Turner, 1977, Williamson, 2005] que se remontan a la concepción del análisis de caminos. Recientemente, Pearl ha publicado un libro [Pearl, 2009] en el que desarrolla formalmente la *teoría de la causalidad* y las *redes causales*; sin embargo, en este documento se prescindirá de dicha interpretación.

En el área de toma de decisiones encontramos otra corriente importante en el desarrollo de las redes Bayesianas: los diagramas de influencia. Se puede considerar a un diagrama de influencia como una red Bayesiana aumentada para modelar procesos de toma de decisiones. El conjunto de nodos se encuentra en este caso dividido en tres tipos: representantes de variables aleatorias, de decisiones y de utilidad. Este tipo de diagramas fue introducido en 1981 por

[1] En la actualidad algunos autores utilizan indiscriminadamente este término para referirse a GGMs de diferentes tipos, por ejemplo [Zhang y Wang, 2010, Zhou et al., 2011].

Howard y Matheson (artículo reimpreso en Howard y Matheson [2005]). Los diagramas de influencia son especialmente relevantes en el caso Gaussiano: Shachter y Kenley [1989] son considerados los padres de las GBNs, y fueron de los primeros en estudiar sus propiedades.

El aprendizaje de redes Bayesianas a partir de un conjunto de datos se ha ido consolidando desde la década de 1990 [Geiger y Heckerman, 1994, Friedman y Goldszmidt, 1998, Heckerman et al., 1995], incluyendo una cantidad importante de libros de revisión [Lauritzen, 1996, Neapolitan, 2003, Koller y Friedman, 2009]. En el presente documento se propone un método para su agregación. Esta tarea ha sido menos explorada y los resultados relacionados son más escasos, especialmente en el caso continuo como es el de las GBNs que nos ocupa. La exposición aquí presentada es una versión ampliada de la investigación preliminar realizada en Córdoba-Sánchez et al. [2015].

1.1. Motivación

La necesidad para la agregación de modelos surge en varios contextos. Supóngase que disponemos de datos que provienen de distintas fuentes (ver Figura 1.1); por ejemplo en el caso de datos particionados en varias máquinas pero sobre el mismo fenómeno, datos que provienen de distintos expertos realizando aserciones sobre el mismo modelo conceptual, etc. En todos estos casos asumiremos que la partición es horizontal; es decir, con respecto a las instancias. El caso de agregación en particiones sobre el espacio de variables queda fuera del alcance del presente documento.

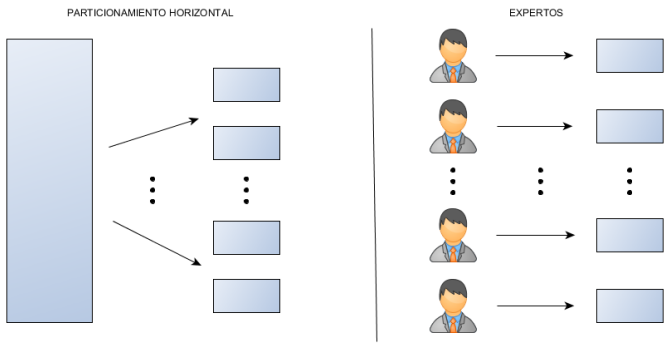

Figura 1.1. Contextos donde surge la agregación de modelos.

Otro contexto donde las particiones horizontales surgen de forma natural, muy popular hoy en día tanto en el mundo de la empresa como de investigación, es con *Big Data* o grandes cantidades de datos. Éstos se caracterizan principalmente por su complejidad a la hora de

realizar procedimientos analíticos: contienen ruido, se actualizan de forma extremadamente rápida o continua, gran volumen, necesidad de análisis continuo, alta precisión de los modelos obtenidos, etc. Una caracterización más precisa y comúnmente aceptada[2] consiste en las llamadas *tres Vs*: alto volumen, variedad y velocidad. A medida que el tiempo pasa y nuevos conjuntos de datos complejos son encontrados, esta definición se está intentando expandir añadiendo nuevas Vs: variabilidad, veracidad, etc. Los conjuntos considerados Big Data han supuesto un reto para la mayoría de algoritmos tradicionales de minería de datos hasta entonces comúnmente aceptados, haciendo que una gran parte del esfuerzo investigador de hoy en día se centre en el desarrollo de nuevos métodos y paradigmas de aprendizajes capaces de tratarlos.

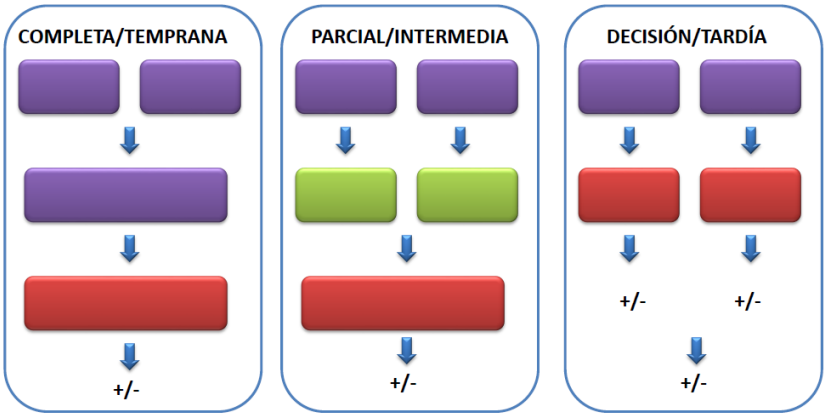

Figura 1.2. Integración completa, parcial y de decisión para datos particionados.

Cuando se trabaja con conjuntos de datos particionados, existen tres principales enfoques que se pueden seguir, ilustrados en la Figura 1.2. El primero, llamado integración *completa* o *prematura*, consiste en concatenar todas las fuentes disponibles y aprender un modelo del conjunto de datos unificado. Cuando no es posible realizar dicha unión de los datos, ya sea por su tamaño, disponibilidad u otros tipos de restricciones, una alternativa es lo que se conoce como integración *parcial* o *intermedia*. En este caso, un modelo se construye a partir de cada una de las particiones de datos y posteriormente un modelo final, que será el utilizado para las inferencias, es obtenido en base al conjunto de modelos intermedios. Finalmente, nos encontramos con la integración *tardía* o de *decisión*, compartiendo con la anterior el hecho de obtener un modelo parcial de cada fuente. En este último caso sin embargo los modelos se mantienen y son los que se utilizarán para realizar las inferencias. Posteriormente la agregación se realiza a nivel de inferencia, obteniendo una conclusión final. Este tipo de método es también conocido como aprendizaje *ensemble*, y, aunque se consigue evitar la selección de modelos, puede suponer un problema en función del número y el tamaño de cada modelo intermedio

[2]No existe consenso sobre una definición universal de Big Data en el momento de escritura de este documento.

que es necesario mantener, además de la sobrecarga de los múltiples procesos de inferencia.

En este documento el método propuesto corresponde a una agregación intermedia.

1.2. Estructura

Este documento se organiza de la siguiente manera. Los grafos dirigidos acíclicos se introducen en el Capítulo 2, mientras que las nociones básicas de probabilidad e independencia aparecen en el Capítulo 3. Estos dos conceptos se juntan en el Capítulo 4 para dar lugar a las redes Bayesianas Gaussianas. En el Capítulo 5 se revisa el estado del arte en fusión de redes Bayesianas Gaussianas, para pasar a presentar nuestro modelo propuesto en el Capítulo 6. Los resultados obtenidos se presentan y discuten en el Capítulo 7, para finalizar el documento con las conclusiones que se puedan extraer y las líneas abiertas de trabajo futuro, en el Capítulo 8. En el Apéndice A se encuentran las representaciones gráficas de las redes de referencia utilizadas para los experimentos.

2. Grafos dirigidos acíclicos

En este capítulo se introducirán algunos conceptos de teoría de grafos necesarios para la comprensión del resto del documento. La teoría de grafos es un campo altamente desarrollado, por lo que aquí daremos una breve introducción a algunos conceptos concernientes a los grafos dirigidos acíclicos (DAGs). Más información se puede encontrar en Chartrand y Lesniak [1986], Bang-Jensen y Gutin [2008] y Thulasiraman y Swamy [2011].

2.1. Introducción

Existen múltiples tipos de grafos. En general un grafo es un par ordenado $\mathcal{G} = (V, E)$ en el que V es el conjunto de *vértices* y E consiste en *pares* de V, el conjunto de *aristas*. La definición más estándar de grafo establece E como pares *no ordenados*; es decir $E \subseteq \{\{u, v\} : u, v \in V\}$.

En ocasiones la naturaleza simétrica del tipo de grafos que hemos definido no será apropiada, y es por ello que se dota a las aristas de un sentido de *dirección*. Esto se consigue considerando ahora E como un conjunto de pares *ordenados* de V, es decir, $E \subseteq V \times V$. Este nuevo tipo de grafo se denomina *dirigido*; y el anterior, *no dirigido*. Las aristas de un grafo dirigido suelen denominarse *arcos* o aristas *dirigidas*.

Para evitar casos patológicos, consideraremos V como un conjunto finito no vacío, aunque es importante mencionar que las nociones de grafo nulo [Harary y Read, 1974] e infinito [Mohar y Woess, 1989] han sido consideradas. En el caso del conjunto de aristas, existen así mismo formulaciones que permiten aristas repetidas (E deja de ser un conjunto) y aristas sobre un mismo elemento (llamadas bucles o *loops*), que también omitiremos de la presente exposición [Chartrand y Lesniak, 1986].

Los grafos dirigidos se representan como un conjunto de puntos o círculos unidos por flechas consistentes con la dirección de cada arista; es decir, para $u, v \in V$, si $(u, v) \in E$ entonces existe una flecha de u a v. Por el contrario, los grafos no dirigidos se representan mediante líneas, dado que la arista $\{u, v\} \in E$ es la misma que $\{v, u\}$. En la Figura 2.1 se muestran algunos ejemplos de grafos dirigidos y no dirigidos.

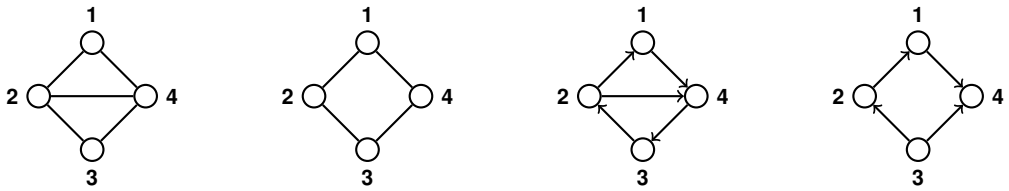

Figura 2.1. Representaciones gráficas de grafos dirigidos y no dirigidos.

2.2. Grafos dirigidos

A continuación expondremos propiedades relevantes de los grafos dirigidos. Los grafos no dirigidos comparten algunas de ellas en analogía con los dirigidos, mientras que otras son específicas de los dirigidos.

En un grafo dirigido $\mathcal{G} = (V, E)$, un arco $a = (u, v) \in E$ se dice que *une* los vértices u y v. En ese caso, el vértice u se dice *adyacente hacia* v; de forma análoga, el vértice v se dice *adyacente desde* u.

Si consideramos dos grafos dirigidos $\mathcal{H} = (V_\mathcal{H}, E_\mathcal{H})$ y $\mathcal{G} = (V_\mathcal{G}, E_\mathcal{G})$, es natural preguntarse si existe algún tipo de relación entre ellos. Una de dichas relaciones es la de *subgrafo*.

$\mathcal{H}$ se dice *sub-grafo* de $\mathcal{G}$ (escrito como $\mathcal{H} \subseteq \mathcal{G}$) si $V_\mathcal{H} \subseteq V_\mathcal{G}$ y $E_\mathcal{H} \subseteq E_\mathcal{G}$; es decir, si tanto los vértices como los arcos de uno se encuentran incluidos en el conjunto de vértices y arcos del otro, respectivamente.

En el caso en el que se encuentren todas los arcos originariamente presentes en el grafo original entre los vértices del subgrafo, dicho subgrafo se denomina subgrafo inducido por vértices, y se denota como $\mathcal{G}_{V_\mathcal{H}}$. Formalmente, la condición se expresa $E_\mathcal{H} = E_\mathcal{G} \cap V_\mathcal{H} \times V_\mathcal{H}$.

Dos vértices $u, v \in V$ en un grafo dirigido $\mathcal{G} = (V, E)$ se pueden considerar *unidos* de diferentes formas.

Un *paseo* entre u y v es una secuencia de arcos y vértices entre u y v de forma que las direcciones de cada arco se respetan. Formalmente, $u = u_0, e_1, u_1, e_2, \ldots, u_{k-1}, e_k, u_k = v$, donde $u, v, u_i \in V$, $e_i \in E$ y $e_i = (u_{i-1}, u_i)$ para $i \in \{1, \ldots, k\}$. El número k es la *longitud* del paseo.

Subtipos de paseos surgen en función de las características del mismo. En el caso de que $u = v$, el paseo se dice *cerrado*, mientras que si $u \neq v$ el paseo se dice *abierto*. Cuando ninguno de los arcos se repite en el paseo, se denomina *sendero*; cuando ninguno de los vértices se repite, se denomina *camino*.

Los subtipos anteriores se pueden combinar y obtenemos las formas de paseo más conocidas. Un *circuito* es un sendero cerrado con al menos un arco. Un circuito de tamaño $n \geq 3$ que también es un camino se denomina *ciclo* o *n-ciclo*.

Un grafo dirigido $\mathcal{G} = (V, E)$ se dice *simétrico* si cada arco se encuentra en ambas

direcciones; es decir, si cada vez que $(u, v) \in E$ se tiene que también $(v, u) \in E$. Por el contrario, si cada vez que $(u, v) \in E$ se tiene que $(v, u) \notin E$, el grafo $\mathcal{G}$ se dice *asimétrico* o *grafo orientado*.

Existe por tanto una correspondencia uno-a-uno entre el conjunto de grafos dirigidos simétricos y el conjunto de grafos no dirigidos, ya que en un grafo simétrico *se pierde* la noción de direccionalidad. Además, un grafo dirigido asimétrico se puede obtener fácilmente de un grafo no dirigido mediante la *orientación* de cada una de sus aristas; es decir, transformando cada arista en un arco. Análogamente, si $\mathcal{G} = (V, E)$ es un grafo dirigido, existe un grafo no dirigido $(V, \{\{u, v\} : (u, v) \in E\})$ que se denomina el grafo no dirigido *subyacente* de $\mathcal{G}$.

2.3. Ausencia de ciclos

Un grafo dirigido acíclico (DAG) es un grafo que no contiene ningún ciclo. Esta característica especial da lugar a la definición de varios conjuntos y propiedades específicas de forma natural, que veremos a continuación, y que se utilizan de forma extensiva en la teoría de las redes Bayesianas.

En un DAG $\mathcal{G} = (V, E)$ se definen conjuntos especiales en torno a un vértice $v \in V$.

Los más inmediatos son el conjunto de *padres* de v, $\text{pa}(v)$, y el conjunto de *hijos* de v, $\text{ch}(v)$, consistentes en los vértices adyacentes hacia y desde v, respectivamente; es decir, $\text{pa}(v) = \{u \in V : (u, v) \in E\}$ y $\text{ch}(v) = \{u \in V : (v, u) \in E\}$.

En un nivel superior encontramos a los *ancestros* de v, $\text{an}(v)$, que son aquellos vértices que pueden alcanzar v mediante un camino; análogamente los *descendientes* de v, $\text{de}(v)$, son aquellos que v alcanza siguiendo un camino. Es común utilizar la notación $\text{nd}(v) = V \setminus (\{v\} \cup \text{de}(v))$ para denotar el conjunto de vértices *no descendientes* de v, dado que juega un papel importante y es utilizado con frecuencia en la teoría de redes Bayesianas.

Por otro lado, un conjunto ancestral se define como un subconjunto A de V que contiene a sus propios ancestros; es decir, que cumple $\text{an}(v) \subseteq A$ para todo $v \in A$. Utilizaremos la notación $\text{An}(A)$ para referirnos al subconjunto ancestral mínimo de V que cumpla $A \subseteq \text{An}(A)$.

Finalmente, un tipo de grafo no dirigido especial fue introducido por Lauritzen y Spiegelhalter [1988]: el grafo *moral*. Se obtiene a partir de un DAG, uniendo el grafo no dirigido subyacente con un nuevo conjunto de aristas. Estas aristas deben unir los vértices que en el grafo dirigido original tenían un hijo en común. Formalmente, si $\mathcal{G}$ es un DAG, el grafo moral es el grafo no dirigido $\mathcal{G}^m = (V, E^m)$ con $E^m = E_U \cup \{uv : \exists w \in V \text{ s.t. } (u, w), (v, w) \in E\}$, donde $E_U = \{uv : (u, v) \in E\}$.

Un DAG cumple propiedades de ordenación de sus vértices que son útiles en algoritmos de inferencia y aprendizaje de redes Bayesianas. El resultado principal se enuncia en la Proposición

1.

Proposición 1. *El conjunto de vértices V de todo DAG $\mathcal{G} = (V, E)$ puede dotarse de un orden total $<$ de tal forma que para cada $(u, v) \in E$, $u < v$.*

El orden total de la Proposición 1 se denomina *orden ancestral*. En un grafo dirigido general su existencia no está garantizada.

En un DAG $\mathcal{G} = (V, E)$ se puede definir así mismo de forma natural un orden parcial en V como $u \leq v$ si $u \in \text{an}(v)$, es decir, situando los ancestros de un vértice por delante del propio vértice. El orden ancestral de la Proposición 1 es entonces una extensión lineal de dicho orden parcial; es decir, un orden total compatible con el mismo.

Un orden ancestral en un DAG $\mathcal{G} = (V, E)$ da lugar a definiciones de nuevos conjuntos especiales. El conjunto de elementos posteriores a un vértice $v \in V$ en el orden se denomina conjunto de sucesores de v y se denota por $\text{su}(v)$. Análogamente, el conjunto de elementos anteriores se denomina conjunto de predecesores, y se denota por $\text{pr}(v)$.

3. Independencia y probabilidad

En este capítulo se introducirán los conceptos de probabilidad condicionada e independencia que son esenciales en la teoría de modelos gráficos probabilísticos en general, y en particular en el caso de las redes Bayesianas.

3.1. Variables aleatorias

El concepto de variable aleatoria es central en el análisis de datos. Previo a su definición es necesario conocer el concepto de espacio de probabilidad.

3.1.1. Espacio de probabilidad

Todas las nociones relacionadas con una variable aleatoria parten de un conjunto Ω. Sobre él se define una σ-*algebra* $\mathcal{F}$, que es un conjunto no vacío de subconjuntos de Ω cerrado bajo uniones contables y toma de complementos. Es decir, si $\{E_i\}_{i\in\mathbb{N}} \subset \mathcal{F}$ entonces $\bigcup_{i\in\mathbb{N}} E_i \in \mathcal{F}$ y si $E \in \mathcal{F}$ entonces $\Omega \setminus E \in \mathcal{F}$, respectivamente.

Al par $(\Omega, \mathcal{F})$ se le denomina *espacio medible*. Una función $f : \Omega \mapsto \Psi$ entre los espacios medibles $(\Omega, \mathcal{F})$ y $(\Psi, \mathcal{H})$ se denomina *función medible* si cumple $f^{-1}(F) \in \mathcal{F}$ para todo $F \in \mathcal{H}$.

El espacio $(\Omega, \mathcal{F})$ se dice medible porque sobre él se pueden definir *medidas*. Una función $m : \mathcal{F} \mapsto [0, +\infty]$ es una *medida* sobre $(\Omega, \mathcal{F})$ si $m(\emptyset) = 0$ (el conjunto vacío mide 0) y se cumple la propiedad de σ-*aditividad*, es decir, para cada colección disjunta por pares $\{E_i\}_{i\in\mathbb{N}} \subset \mathcal{F}$, $m\left(\bigcup_{i\in\mathbb{N}} E_i\right) = \sum_{i\in\mathbb{N}} m\left(E_i\right)$. Si se cumple $m(\Omega) = 1$, es decir, la medida está normalizada, entonces m se denomina una *probabilidad*.

Un espacio medible sobre el que se ha definido una medida, $(\Omega, \mathcal{F}, m)$, se denomina *espacio de medida*. Por tanto, un *espacio de probabilidad* no es más que un espacio de medida donde la medida que se ha definido es una probabilidad. En dicho caso, se utiliza una terminología especial, motivada por los experimentos aleatorios: Ω es el *espacio muestral*, cada elemento de $\mathcal{F}$ es un *evento*, y un elemento de Ω es un *resultado*.

3.1.2. Distribuciones de probabilidad

Una vez definido un espacio de probabilidad podemos introducir el concepto central de variable aleatoria. Partimos de un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, de donde procede la aleatoriedad, y llegamos a un espacio que podemos medir $(\Psi, \mathcal{H})$. Una función medible $X : \Omega \mapsto \Psi$ es una *variable aleatoria*. De nuevo surge terminología: $(\Psi, \mathcal{H})$ se denomina el *espacio de estados*, y un elemento de $X(\Omega)$, una *realización* de la variable aleatoria.

Podemos pensar en X como lo que nos permite manejar cuantitativamente el espacio muestral Ω. Es por ello que normalmente se trabaja en el conjunto de números reales, es decir, $\Psi = \mathbb{R}$, y así es como lo asumiremos en adelante por simplicidad.

La *función de distribución* de una variable aleatoria es una función $F : \mathbb{R} \mapsto [0, 1]$ que asigna a cada número real $x \in \mathbb{R}$ la probabilidad $P(X^{-1}(I))$, donde $I = (-\infty, x]$ y P es la probabilidad que recordemos está definida sobre el espacio muestral. Es decir, con la función de distribución tenemos una correspondencia explícita entre el análisis real de la variable y el espacio muestral.

La caracterización de una variable aleatoria continua viene de conceptos más elaborados de cálculo y teoría de la medida como son la *continuidad absoluta* y funciones *integrables respecto a la medida de Lebesgue*. Nosotros daremos una versión simplificada. Una variable aleatoria X se dice *continua* si existe una función integrable no negativa f tal que para cada $x \in \mathbb{R}$, $F(x) = \int_{-\infty}^{x} f(t)dt$. La función f se denomina *función de densidad* de X, y es única en *casi todo punto*[1].

De las múltiples cantidades que se pueden calcular asociadas a una variable aleatoria continua, son de especial interés la media y la varianza, puesto que juegan un papel muy importante en la distribución Gaussiana y en general en el análisis de datos.

La *media* de una variable aleatoria continua X se define como $\mathrm{E}[X] = \int_{-\infty}^{\infty} t f(t)dt$ en caso de que la integral exista. Se cumple que si $g : \mathbb{R} \mapsto \mathbb{R}$ y la media de X existe, entonces $\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(t)f(t)dt$. Este resultado nos permite definir la *varianza* de X como $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$, de nuevo en caso de que la integral exista.

Dos variables aleatorias sobre el mismo espacio muestral se pueden relacionar mediante la *covarianza* $\mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$. Nótese que $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$.

[1] Si existe otra función g que satisface lo mismo que f, el conjunto $D = \{x \in \mathbb{R} : f(x) \neq g(x)\}$ tiene medida de Lebesgue cero.

3.2. Vectores aleatorios

El caso multivariante, frecuente en el análisis de conjuntos de datos, se da cuando el espacio muestral de la variable aleatoria es un producto cartesiano. La variable aleatoria en este caso tiene varias componentes y se denomina comúnmente *vector aleatorio*. Cada una de las componentes del vector es, en sí misma, una variable aleatoria, es por ello que un vector aleatorio se suele escribir como $\boldsymbol{X} = (X_1, ..., X_p)^t$, aunque omitiremos el indicador de transposición a no ser que sea necesario por el contexto.

3.2.1. Marginales y conjuntas

Los resultados y las definiciones presentados en el capítulo anterior se aplican de forma análoga en el caso de los vectores aleatorios. Al igual que entonces, consideraremos aquellos vectores cuyo espacio de estados sea $\mathbb{R}^p$, es decir, asumimos que existe una función de densidad. Las funciones y cantidades asociadas a un vector aleatorio se suelen apellidar *conjuntas* para distinguirlas del caso en el que estemos considerando las de las variables por separado, a veces apodadas a su vez *marginales*.

La Proposición 2 proporciona un método por el cual obtener las distribuciones marginales a partir de la distribución conjunta del vector aleatorio. Este proceso se conoce como *marginalización*.

Proposición 2. *Sea $\boldsymbol{X} = (X_1, ..., X_p)$ un vector aleatorio. La función de densidad de cada variable aleatoria X_i, f_{X_i} se obtiene de la función de densidad conjunta mediante la integral múltiple*

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(t_1, ..., t_{i-1}, x_i, t_{i+1}, ..., t_n) dt_1 ... dt_{i-1} dt_{i+1} ... dt_p.$$

La media de un vector aleatorio se define en función de las medias marginales, mientras que la varianza es análoga al caso univariante. En concreto, la *media* de $\boldsymbol{X} = (X_1, ..., X_p)^t$ se define como el vector $\mathbf{E}[\boldsymbol{X}] = (\mathrm{E}[X_1], ..., \mathrm{E}[X_n])$; la *matriz de varianza-covarianza* como $\mathbf{Var}[\boldsymbol{X}] = \mathbf{E}[(\boldsymbol{X} - \mathbf{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbf{E}[\boldsymbol{X}])^t]$. Nótese que la entrada (i, j) de la matriz $\mathbf{Var}[\boldsymbol{X}]$ es justamente $\mathrm{Cov}[X_i, X_j]$.

3.2.2. Independencia

Las nociones de independencia e independencia condicional son centrales en los PGMs, pues, como se explicó en la Introducción, un PGM es precisamente una representación gráfica de dichas relaciones. Sin embargo, la definición formal requiere de una teoría más elaborada,

pues parte del espacio muestral y la σ-álgebra definida sobre este. Por tanto, en el presente documento ilustraremos únicamente resultados que caracterizan de forma unívoca la independencia (condicional) entre variables aleatorias.

Teorema 1. *Sea $\boldsymbol{X}$ un vector aleatorio continuo. Las variables aleatorias $\{X_i\}_{i=1}^n$ son independientes si y sólo si su función de distribución conjunta $F_{\boldsymbol{X}}$ satisface*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} F_{X_i}(x_i),$$

donde F_{X_i} denota la función de distribución de X_i para $i \in \{1, ..., p\}$; o, de forma equivalente, si su función de densidad conjunta satisface

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

De ahora en adelante no utilizaremos subíndices para distinguir las funciones de densidad y distribución cuando quede claro a quién se aplican por los argumentos de las mismas.

El Teorema 1 nos proporciona una condición de factorización para el caso de independencia entre variables aleatorias.

3.2.3. Condicionamiento

Para caracterizar la independencia condicional entre dos variables aleatorias es necesario conocer la noción de espacio de probabilidad condicionada.

Podemos pensar en un espacio de probabilidad condicionada como aquel en el que se ha fijado un evento. Intuitivamente, esto equivale a suponer que una parte del espacio muestral ha sido *observada* y por tanto ha *perdido aleatoriedad*. Formalmente, consideramos un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ y un conjunto $F \in \mathcal{F}$ tal que $P(F) \neq 0$.

Proposición 3. *El conjunto $\mathcal{F} \cap F = \{E \cap F : E \in \mathcal{F}\}$ es una σ-álgebra sobre F. La función $P(\cdot \mid F) : \mathcal{F} \cap F \mapsto [0, 1]$ definida como $P(E \mid F) = P(E \cap F)/P(F)$, para $E \in \mathcal{F} \cap F$, es una probabilidad sobre $(F, \mathcal{F} \cap F)$.*

La probabilidad definida en la Proposición 3 se denomina *probabilidad condicional*. El espacio de probabilidad que surge es $(F, \Omega \cap F, P(\cdot \mid F))$ y se llama *espacio de probabilidad condicionada*.

Los resultados sobre variables aleatorias y funciones de distribución hasta ahora vistos se pueden extender de forma análoga al espacio de probabilidad condicionada. Por tanto, el Teorema 1 nos da también una condición de factorización en el caso de las distribuciones condicionadas para caracterizar la independencia condicional entre dos variables aleatorias.

4. Redes Bayesianas Gaussianas

En este apartado veremos propiedades de la distribución Gaussiana multivariante respecto de las independencias (condicionales) que han dado lugar al desarrollo de los modelos gráficos Gaussianos, y que será la distribución que asumiremos sobre los modelos a agregar.

4.1. Densidad Gaussiana multivariante

Un vector aleatorio $\boldsymbol{X} = (X_1, ..., X_p)$ se dice que tiene una *distribución Gaussiana multivariante* si su función de densidad es, para cada $\boldsymbol{x} \in \mathbb{R}^p$,

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{4.1}$$

donde $\boldsymbol{\Sigma}$ es una matriz simétrica definida positiva y $\boldsymbol{\mu}$ es un vector de dimensión p, ambos denominados *parámetros* de la distribución. Un resultado conocido es que $\boldsymbol{\Sigma}$ es precisamente la matriz de covarianza $\mathbf{Var}[\boldsymbol{X}]$ y $\boldsymbol{\mu}$ el vector de medias $\mathbf{E}[\boldsymbol{X}]$. Cuando un vector aleatorio tiene la distribución Gaussiana multivariante se suele denotar como $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. El elemento i de $\boldsymbol{\mu}$ se escribe μ_i; el elemento (i, j) de $\boldsymbol{\Sigma}$ se escribe σ_{ij}. Es decir, para $i, j \in \{1, ..., p\}$, $\mu_i = \mathrm{E}[X_i]$, $\sigma_{ij} = \mathrm{Cov}[X_i, X_j]$ y $\sigma_{ii} = \mathrm{Var}[X_i]$.

4.1.1. Condicionamiento y marginalización

En este apartado asumiremos un vector aleatorio $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Un poco de notación adicional es necesaria para los próximos resultados. Para $I \subseteq \{1, ..., p\}$, la notación $\boldsymbol{X}_I$ se refiere al sub-vector $(X_i)_{i \in I}$. Para una matriz $\boldsymbol{\Sigma}$ de dimensión p, $I, J \subseteq \{1, ..., p\}$, la notación $\boldsymbol{\Sigma}_{IJ}$ se corresponde con la sub-matriz $(\sigma_{ij})_{i \in I, j \in J}$; así, si $J = \{1, ..., p\} \setminus I$, la matriz $\boldsymbol{\Sigma}$ se consideraría particionada como

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{II} & \boldsymbol{\Sigma}_{IJ} \\ \boldsymbol{\Sigma}_{JI} & \boldsymbol{\Sigma}_{JJ} \end{pmatrix}.$$

En los resultados siguientes veremos como las distribuciones marginales y condicionales de un vector aleatorio que sigue una distribución Gaussiana multivariante son a su vez Gaussianas multivariantes en las que los parámetros han cambiado.

Teorema 2 (Distribución marginal)**.** *Si* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, se cumple que* $\boldsymbol{X}_I \sim \mathcal{N}(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_{II})$ *para todo* $I \subseteq \{1, ..., p\}$.

Teorema 3 (Distribución condicionada)**.** *Sea* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *y los conjuntos* $I \subseteq \{1, ..., p\}$, $J = \{1, ..., p\} \setminus I$*. Sea* $\boldsymbol{x}_J$ *una realización de* $\boldsymbol{X}_J$*. Se tiene que* $\boldsymbol{X}_I \mid \boldsymbol{x}_J \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$*, donde*

$$\boldsymbol{\mu}' = \boldsymbol{\mu}_I + \boldsymbol{\Sigma}_{IJ}\boldsymbol{\Sigma}_{JJ}^{-1}(\boldsymbol{x}_J - \boldsymbol{\mu}_J), \tag{4.2}$$

$$\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}_{II} - \boldsymbol{\Sigma}_{IJ}\boldsymbol{\Sigma}_{JJ}^{-1}\boldsymbol{\Sigma}_{JI}. \tag{4.3}$$

El Teorema 3 da lugar a una serie de definiciones claves en las redes Gaussianas. Un elemento (i, j) de la matriz $\boldsymbol{\Sigma}_{IJ}\boldsymbol{\Sigma}_{JJ}^{-1}$, que aparece tanto en la Ecuación 4.2 como en 4.3, se denomina *coeficiente de regresión* de X_i en X_j fijado $\boldsymbol{X}_{J\setminus\{j\}}$, y se denota como $\beta_{ij\cdot J\setminus\{j\}}$ (notación introducida por Yule [1907] para los modelos de regresión multivariantes, que permite hacer explícitas las variables involucradas en el modelo y la relación entre ellas de forma compacta). El elemento (i, k) de $\boldsymbol{\Sigma}'$, la matriz de varianza-covarianza de la nueva distribución condicional (Ecuación 4.3) se denomina covarianza parcial, de nuevo de X_i y X_k fijado $\boldsymbol{X}_J$; y se denota como $\sigma_{ik\cdot J}$. Si $i = k$, $\sigma_{ii\cdot J}$ es la varianza parcial de X_i fijado $\boldsymbol{X}_J$. Finalmente, la correlación parcial entre X_i y X_k, $i, k \in I$, fijado $\boldsymbol{X}_J$, se define como

$$\rho_{ik\cdot J} = \frac{\sigma_{ik\cdot J}}{\sqrt{\sigma_{ii\cdot J}\sigma_{kk\cdot J}}}.$$

Los resultados del Teorema 3, aunque son usados por todos los modelos gráficos Gaussianos, en el caso de las redes Gaussianas se suelen referenciar en la forma no matricial: para cada $i \in I$,

$$X_i \mid \boldsymbol{x}_J \sim \mathcal{N}\left(\mu_i + \sum_{j\in J} \beta_{ij\cdot J\setminus\{j\}}(x_j - \mu_j), \sigma_{ii} - \sum_{j\in J} \beta_{ij\cdot J\setminus\{j\}}\sigma_{ji}\right).$$

4.1.2. Matriz de concentración e independencia

Los resultados de independencia e independencia condicional asociados a la distribución Gaussiana multivariante están fuertemente relacionados con la matriz de covarianza, $\boldsymbol{\Sigma}$, y su inversa, $\boldsymbol{\Sigma}^{-1}$, como veremos en este apartado. Denotaremos a $\boldsymbol{\Sigma}^{-1}$ como $\boldsymbol{\Lambda}$ en lo sucesivo, siendo λ_{ik} su elemento (i, k). A $\boldsymbol{\Lambda}$ también se la denomina *matriz de concentración* o de *precisión*.

Los Teoremas 4 y 5 establecen criterios para la independencia y la independencia condicional, respectivamente, en un vector aleatorio con distribución Gaussiana. Denotaremos, para $I, J, K \subseteq \{1, ..., p\}$, el hecho $\boldsymbol{X}_I$ *y* $\boldsymbol{X}_J$ *son independientes* como $\boldsymbol{X}_I \perp \boldsymbol{X}_J$; análogamente, $\boldsymbol{X}_I$ *y* $\boldsymbol{X}_J$ *son independientes dado* $\boldsymbol{X}_K$ se escribirá $\boldsymbol{X}_I \perp \boldsymbol{X}_J \mid \boldsymbol{X}_K$.

Teorema 4 (Independencia)**.** *Sea* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *y los conjuntos* $I \subseteq \{1, ..., p\}$ *y* $J = \{1, ..., p\} \setminus I$*. Entonces* $\boldsymbol{X}_I \perp \boldsymbol{X}_J$ *si y sólo si* $\sigma_{ij} = 0$ *para todo* $i \in I$*,* $j \in J$*.*

Teorema 5 (Independencia condicional)**.** *Sea* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *y los conjuntos* $I \subseteq \{1, ..., p\}$ *y* $J = \{1, ..., p\} \setminus I$*. Entonces, para* $i, k \in I$*,* $X_i \perp X_k \mid \boldsymbol{X}_J$ *si y sólo si* $\sigma_{ik \cdot J} = 0$*.*

Por tanto las independencias e independencias condicionales en una distribución Gaussiana multivariante están determinadas por las varianzas y las varianzas parciales nulas, respectivamente. Obsérvese que el Teorema 5 es una consecuencia directa de los Teoremas 3 y 4.

Finalmente, las relaciones del Teorema 6 nos proporcionan relaciones que permiten caracterizas las independencias (condicionales) mediante otras cantidades asociadas a la varianza (parcial).

Teorema 6. *Sea* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*,* $i, k \in \{1, ..., p\}$ *y* $J = \{1..., p\} \setminus \{i, k\}$*. Se cumplen las siguientes igualdades*

$$\rho_{ik \cdot J} = -\frac{\lambda_{ik}}{\sqrt{\lambda_{ii}\lambda_{kk}}}, \qquad \rho^2_{ik \cdot J} = \beta_{ik \cdot J}\beta_{ki \cdot J}, \qquad \beta_{ik \cdot J} = \frac{\sigma_{ik \cdot J}}{\sigma_{kk \cdot J}} = \rho_{ik \cdot J}\frac{\sqrt{\sigma_{ii \cdot J}}}{\sqrt{\sigma_{kk \cdot J}}}.$$

Gracias al Teorema 6, las siguientes relaciones se tienen para el caso de la independencia condicional (el de interés en las redes Bayesianas Gaussianas):

$$X_i \perp X_k \mid \boldsymbol{X}_J \iff \sigma_{ik \cdot J} = 0 \iff \beta_{ik \cdot J} = 0 \iff \rho_{ik \cdot J} = 0.$$

4.2. Representación gráfica

A continuación se combinarán las nociones presentadas en las Secciones 2 y 3, así como las de esta sección, para dar lugar a un modelo basado en grafos dirigidos acíclicos (DAGs) que representará un distribución conjunta Gaussiana de un vector aleatorio: las redes Bayesianas Gaussianas.

4.2.1. Propiedades de Markov

Antes de definir lo que es una red Bayesiana es necesario conocer lo que se denominan *propiedades de Markov*, que recordemos del Capítulo 1, establecen la correspondencia entre las independencias en la distribución y las características estructurales del grafo.

Consideramos un DAG $\mathcal{G}$ y una familia de distribuciones de probabilidad $\mathcal{F}$ (genérica, aunque luego particularizaremos en la Gaussiana multivariante). Se dice que la familia $\mathcal{F}$ cumple la propiedad

(a) *de buen orden de Markov* con respecto a $\mathcal{G}$ si $X_v \perp \boldsymbol{X}_{\mathrm{pr}(v)\setminus\mathrm{pa}(v)} \mid \boldsymbol{X}_{\mathrm{pa}(v)}$ para todo $v \in V$;

(b) *local de Markov* con respecto a $\mathcal{G}$, si $X_v \perp \boldsymbol{X}_{\mathrm{nd}(v)\setminus\mathrm{pa}(v)} \mid \boldsymbol{X}_{\mathrm{pa}(v)}$ para cada $v \in V$;

(c) *global de Markov* con respecto a $\mathcal{G}$ si $\boldsymbol{X}_A \perp \boldsymbol{X}_B \mid \boldsymbol{X}_S$ para todos subconjuntos disjuntos A, B, S de V tal que A y B están separados por S en $(\mathcal{G}_{\mathrm{An}(A\cup B\cup S)})^m$, el grafo moral asociado al grafo inducido por el conjunto ancestral $\mathrm{An}(A \cup B \cup S)$.

El criterio de separación utilizado para la propiedad global de Markov basado en el grafo moral es equivalente al conocido criterio de d-separación de Geiger et al. [1990], como lo demuestra Lauritzen et al. [1990].

Una vez introducidas las propiedades de Markov para DAGs, podemos definir lo que es una *red Bayesiana*: un par ordenado $(\mathcal{G}, \mathcal{F})$ donde $\mathcal{G}$ es un DAG y $\mathcal{F}$ tiene la propiedad local de Markov con respecto a $\mathcal{G}$.

El Teorema 7 establece la equivalencia entre las tres propiedades de Markov. Es por ello que a veces se denomina *modelos de Markov dirigidos* a las redes Bayesianas.

Teorema 7. [Lauritzen et al., 1990] *Las propiedades de Markov local, global y de buen orden son equivalentes.*

Cuando la familia $\mathcal{F}$ en una red Bayesiana consiste en distribuciones Gaussianas multivariantes, hablamos de una *red Bayesiana Gaussiana* o *red Gaussiana* simplemente.

4.2.2. Factorización

Las propiedades de Markov en una red Gaussiana junto con las características de la independencia y la independencia condicional en la distribución Gaussiana multivariante permiten caracterizar de forma sencilla la ausencia o presencia de arcos en la red.

Consideramos una red Gaussiana $(\mathcal{G}, \mathcal{N})$. Si asumimos que los vértices de $\mathcal{G}$ se encuentran ordenados ancestralmente; es decir, $1 < \ldots < p$, la distribución conjunta de un vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ se puede escribir como

$$f(\boldsymbol{x}) = \prod_{i=1}^{p} f(x_i \mid x_1, \ldots, x_{i-1}) = \prod_{i=1}^{p} f(x_i \mid \boldsymbol{x}_{\mathrm{pr}(i)}) = \prod_{i=1}^{p} f(x_i \mid \boldsymbol{x}_{\mathrm{pa}(i)}), \tag{4.4}$$

gracias a la propiedad de buen orden de Markov.

Recordando los Teoremas 2 y 3, la densidad en cada factor del término final en la Ecuación 4.4 es una distribución Gaussiana univariante:

$$X_i \mid \boldsymbol{x}_{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mu_i + \sum_{j\in\mathrm{pa}(i)} \beta_{ij}(x_j - \mu_j), \sigma_{ii} - \sum_{j\in\mathrm{pa}(i)} \beta_{ij}\sigma_{ji}\right).$$

Nótese que hemos omitido la notación de Yule para los coeficientes de regresión, y así haremos de ahora en adelante, ya que en una red Gaussiana siempre vamos a considerar la variable X_i condicionada en sus padres $\boldsymbol{X}_{\mathrm{pa}(i)}$, de forma que $\beta_{ij\cdot\mathrm{pa}(i)\setminus j} = \beta_{ij}$ es el coeficiente de regresión correspondiente al padre j, y por tanto no es necesario hacerlo explícito en cada una de sus ocurrencias.

4.2.3. Correspondencia con regresión

Como se vio al final del Capítulo 4, en una red Gaussiana la distribución de probabilidad se factoriza como

$$f(\boldsymbol{x}) = \prod_{i=1}^{p} f(x_i \mid \boldsymbol{x}_{\mathrm{pa}(i)}),$$

donde, gracias a las propiedades de la distribución Gaussiana multivariante,

$$X_i \mid \boldsymbol{x}_{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mu_i + \sum_{j\in\mathrm{pa}(i)} \beta_{ij}(x_j - \mu_j), \sigma_{ii} - \sum_{j\in\mathrm{pa}(i)} \beta_{ij}\sigma_{ji}\right). \tag{4.5}$$

Además, recuérdese también que las independencias condicionales en una distribución Gaussiana multivariante quedan caracterizadas de formas alternativas; en efecto, para $i, k \in \{1, \ldots, p\}$ y el conjunto $J = \{1, \ldots, p\} \setminus \{i, k\}$, se tiene, entre otras condiciones (como también vimos en el Capítulo 4),

$$X_i \perp X_k \mid \boldsymbol{X}_J \iff \beta_{ik\cdot J} = 0.$$

Por tanto, para cada $i, j \in \{1, \ldots, p\}$, supuesto el orden ancestral asociado a la factorización de la Ecuación 4.4, y $j \in \mathrm{pr}(i)\setminus\mathrm{pa}(i)$, entonces $\beta_{ij\cdot\mathrm{pa}(i)} = 0$, y por tanto no existe una arista entre i y j. Además, si $j \in \mathrm{nd}(i)$, también se tiene $\beta_{ij\cdot\mathrm{pa}(i)} = 0$, por la propiedad local de Markov. Es decir, un arco (u, v) en el grafo $\mathcal{G}$ de la red Gaussiana está asociado con el coeficiente de regresión β_{uv} de la Ecuación 4.5 correspondiente a $X_v \mid \boldsymbol{x}_{\mathrm{pa}(v)}$.

5. Trabajo relacionado

En este apartado se dará una visión de los modelos gráficos Gaussianos más relevantes en la actualidad, se revisarán los métodos más comunes de aprendizaje de redes Gaussianas y exploraremos el estado del arte en la fusión de las mismas.

5.1. Redes de Markov Gaussianas

Se puede decir que tanto las propiedades de Markov como los modelos gráficos probabilísticos se originaron y desarrollaron en la teoría de los campos de Markov. Un campo de Markov es una generalización del modelo de Ising sobre ferromagnetismo para variables binarias (y, a su vez, el modelo de Ising es una generalización de una cadena de Markov). En este área existe un resultado que establece la equivalencia entre los campos de Markov y un modelo de vecinos de Gibbs, que Darroch et al. [1980] utilizaron para introducir lo que denominaron *modelos gráficos*. Al definirlos, ellos mismos admiten que su artículo en realidad no representa nada nuevo, sino una reformulación y asociación de resultados ya existentes.

Darroch et al. [1980] definieron sus modelos gráficos, que posteriormente se llamaron redes de Markov, para el caso de variables discretas (*tablas de contingencia*). Quienes lo trasladaron al caso Gaussiano fueron Speed y Kiiveri [1986], apoyándose en resultados previos que Dempster [1972] había obtenido para la estimación de matrices de covarianza en una distribución Gaussiana multivariante (y que dio origen a que, durante un tiempo, el nombre más común utilizado para ellos fuera *selección de covarianza*).

Una red de Markov es similar a una red Bayesiana, pero en vez de un grafo dirigido se utiliza uno no dirigido para su representación. Al igual que la red Bayesiana, se asocian con el grafo independencias condicionales de la distribución subyacente. En el caso Gaussiano, las redes de Markov son especialmente relevantes, alcanzando un nivel similar de popularidad respecto a las redes Gaussianas. Sin embargo, no existe una equivalencia entre los conjuntos de modelos de independencia que se pueden representar con cada uno de los modelos gráficos probabilísticos; es decir, existen modelos de independencia que se pueden representar con una red Bayesiana y no con una red de Markov, y viceversa.

Las propiedades de Markov en el caso de las redes de Markov son análogas a las de las redes Bayesianas, si bien como ya se ha comentado son previas a estas últimas. Se explicará

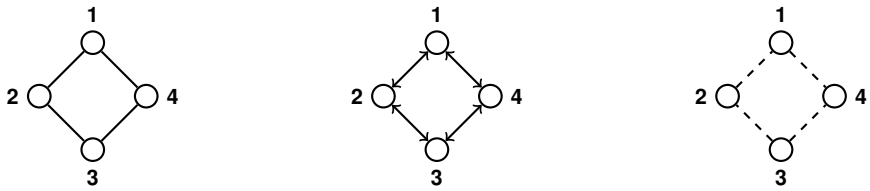

Figura 5.1. Distintas representaciones gráficas de un grafo de covarianza.

solamente la *propiedad entre pares de Markov*, ya que el resto requieren conceptos de teoría de grafos no dirigidos que se ha preferido omitir del presente documento.

Considerando un grafo no dirigido $\mathcal{G}$ y una familia de distribuciones de probabilidad $\mathcal{F}$, se dice que la familia cumple la *propiedad entre pares de Markov* con respecto a $\mathcal{G}$, si $X_u \perp X_v \mid \boldsymbol{X}_{V\setminus\{u,v\}}$ para cada $u, v \in V$ no conectados en $\mathcal{G}$, es decir, tal que no exista una arista uniéndolos.

En el caso de la distribución Gaussiana multivariante, la propiedad enunciada es equivalente a la local y a la global, tal y como demuestran Speed y Kiiveri [1986], si bien este es un resultado consecuencia de un teorema más complejo, de Hammersley y Clifford [1971], originado en la teoría de los campos de Markov. Las propiedades no son equivalentes en el caso de distribuciones genéricas distintas de la Gaussiana, al contrario de lo que ocurre en el caso de las redes Bayesianas.

Finalmente, lo que hace atractivo a las redes de Markov Gaussianas en el campo de minería de datos es la facilidad de interpretación en la distribución Gaussiana multivariante de la propiedad entre pares de Markov. Efectivamente, si recordamos el Teorema 6, dicha propiedad es equivalente a la presencia de un cero en la posición (u, v) de la matriz de concentración.

5.2. Grafos de covarianza Gaussianos

Otro modelo gráfico probabílistico basado en grafos no dirigidos, también popular en el contexto de aprendizaje a partir de datos, aunque no tanto como las redes Bayesianas y de Markov, es el grafo de covarianza.

Este tipo de grafos son los que se han desarrollado de forma más tardía. Fueron mencionados por primera vez en el trabajo de Cox y Wermuth [1993], en el contexto de una unificación de los tipos de relaciones lineales entre varias variables, en la que el nivel superior lo ocupaban los grafos de cadena, y los grafos de covarianza eran un subtipo de los mismos que representaban dependencias marginales.

Para diferenciarlos gráficamente de las redes de Markov, dado que ambos se representan como grafos no dirigidos, existen diversas formas de representar las aristas en un grafo de

covarianza. En sus orígenes, en el artículo de Cox y Wermuth [1993] se utilizaban aristas punteadas (Figura 5.1, derecha). En caso de que el contexto no de origen a confusión, porque solamente se esté tratando con grafos de covarianza, se utilizan las aristas habituales en grafos no dirigidos (Figura 5.1, izquierda). Las aristas con una punta de flecha a cada lado, también llamadas bi-dirigidas (Figura 5.1, centro), fueron introducidas por Richardson y Spirtes [2002] en el contexto de la definición de un nuevo modelo gráfico dirigido más general, los grafos ancestrales.

Como en el caso de las redes de Markov, se enunciará solamente la *propiedad entre pares* para evitar la definición de conceptos adicionales asociados a grafos no dirigidos. Una distribución $\mathcal{F}$ se dice que satisface la *propiedad entre pares de Markov* respecto al grafo no dirigido $\mathcal{G}$ si $X_u \perp X_v$ para cada $u, v \in V$ no conectados en $\mathcal{G}$. Nótese como efectivamente, en el caso de un grafo de covarianza, las relaciones entre la distribución de probabilidad y el grafo se establecen mediante la independencia marginal entre las variables.

Al igual que en el caso de las redes de Markov, las propiedades de Markov para grafos de covarianza son equivalentes en el caso Gaussiano, lo demuestra Kauermann [1996], que fue el que las definió en primer lugar, y Banerjee y Richardson [2003] mediante una corrección posterior.

5.3. Aprendizaje estructural de redes Bayesianas

Considérese un vector aleatorio $\boldsymbol{X}$ de dimensión p siguiendo una distribución Gaussiana multivariante cuya estructura de independencia condicional se representa mediante una red Bayesiana.

Los algoritmos de aprendizaje suelen asumir media cero, es decir, $\mu_i = 0$ para cada $i \in \{1, ..., p\}$. De esta forma, a partir de la distribución condicional

$$X_i \mid \boldsymbol{x}_{\mathrm{pa}(i)} \sim \mathcal{N}\left(\sum_{j \in \mathrm{pa}(i)} \beta_{ij} x_j, \sigma_{ii} - \sum_{j \in \mathrm{pa}(i)} \beta_{ij} \sigma_{ji} \right),$$

se obtiene la conocida ecuación de la regresión múltiple

$$\mathrm{E}[X_i \mid \boldsymbol{x}_{\mathrm{pa}(i)}] = \sum_{j \in \mathrm{pa}(i)} \beta_{ij} x_j,$$

sobre la que se pueden aplicar métodos de aprendizaje ya conocidos para regresión lineal múltiple. Dado que un arco (u, v) se corresponde con un coeficiente de regresión β_{uv} distinto de cero (ver Capítulo 4), el aprendizaje estructural, es decir, la obtención de $\mathcal{G}$, se centra en la mayoría de los casos en averiguar cuáles son dichos coeficientes.

Se conocen dos grandes grupos de algoritmos de aprendizaje estructural: aquellos que buscan en un espacio de redes Gaussianas, puntuando cada red y moviéndose acorde a las puntuaciones obtenidas mediante distintas operaciones (en inglés este grupo se suele denominar *score and search*), y métodos basados en tests estadísticos que restringen la estructura aprendida.

El ejemplo más representativo y famoso de los algoritmos de aprendizaje basado en tests es el algoritmo PC [Spirtes et al., 2000]. Este algoritmo parte de una red completa y va eliminando aristas en función de la evidencia estadística obtenida. La salida de este algoritmo sin embargo no es una red Bayesiana, sino algo más general, una *clase de equivalencia*. El concepto intuitivo de clase de equivalencia de redes Bayesianas es un conjunto de redes que, teniendo estructura diferente, representan el mismo conjunto de independencias condicionales. Así, la salida del algoritmo PC es un grafo con aristas dirigidas y no dirigidas que unifica todas las redes en dicha clase de equivalencia, por tanto, es necesario utilizar posteriormente un algoritmo de orientación de aristas para obtener la red final.

Por otro lado, en los métodos *score and search* es necesario definir la métrica a utilizar para ponderar las estructuras, el espacio de búsqueda sobre el que se va a explorar y el procedimiento de exploración a seguir. El espacio de búsqueda que se escoge en la mayoría de trabajos es el de las redes Bayesianas; sin embargo, otros como Nielsen et al. [2003] utilizan el espacio de clases de equivalencia, que suele suponer una representación más robusta y eficiente [Vidaurre et al., 2010], aunque todavía existe controversia al respecto. Múltiples métricas y procedimientos de búsqueda han aparecido desde los inicios del aprendizaje estructural en redes Bayesianas (por ejemplo Schwarz [1978], Geiger y Heckerman [1994], Heckerman et al. [1995], Tsamardinos et al. [2006]) y la combinación de los mismos dan lugar a un amplio espectro de métodos en este campo.

5.4. Agregación de redes Bayesianas

La agregación de redes Bayesianas, se ha tratado desde que dichos modelos fueron estudiados para la representación del conocimiento de expertos. Aunque los trabajos en este área son escasos, y la mayoría centrados en el caso discreto, se han enfocado desde perspectivas diversas, como se verá a continuación.

Matzkevich y Abramson [1992] considera el problema de fusionar redes que han sido obtenidas de diferentes expertos y que comparten un subconjunto de variables. A partir de ellas, obtienen un grafo que contiene todos los nodos de las redes individuales, además de todos sus arcos, o sus inversos. La necesidad de incluir los inversos de arcos surge por la posibilidad de creación de ciclos; circunstancia no permitida en las redes Bayesianas.

Recientemente, Peña [2011] ha demostrado que el problema de obtención de una red consenso tal y como se define en Matzkevich y Abramson [1992] tiene varias soluciones no equivalentes (en el sentido de las clases de equivalencia), y encontrar una de ellas es NP-completo. Además, muestra que los métodos originalmente presentados por Matzkevich y Abramson [1992] no son correctos y proporciona su corrección.

Por otro lado, del Sagrado y Moral [2003] se centran en la estructura de independencia condicional que representa la red Bayesiana. Así, trabajan con las afirmaciones del tipo X_i *es condicionalmente independiente de* X_j *dado* X_K que se pueden deducir de cada red individual gracias a las propiedades de Markov. Con este enfoque, estudian las características de la intersección y unión de dichas propiedades de independencia condicional, analizando en qué casos se mantienen las características de *grafoide* y *semi-grafoide* (ver Pearl [1988] para una explicación detallada de estas estructuras algebraicas), que se asumen satisfechas por las redes individuales.

Un enfoque Bayesiano se encuentra en el trabajo de Richardson y Domingos [2003]. En este caso, el conocimiento de un grupo de expertos es utilizado para la obtención de una distribución a priori sobre las estructuras Bayesianas candidatas. Motivan su propuesta observando que el proceso de elicitación de conocimiento de un experto se facilita si se le permite ser algo *ruidoso* en sus afirmaciones sobre la estructura de la red Bayesiana; para contrarrestar dicho ruido, se emplean varios expertos. Este argumento es interesante puesto que se puede trasladar al caso de conjuntos de datos grandes y ruidosos, donde en vez de elegir un submuestreo y aprender la red de la muestra reducida, se elija particionar el conjunto y aplicar fusión.

Un caso práctico reciente en el que se hace necesaria la agregación de redes Bayesianas es el encontrado por López-Cruz et al. [2014]. En este caso un conjunto de expertos en neurociencia debían clasificar un conjunto de neuronas, dando lugar a un conjunto para aprendizaje supervisado por cada experto, del que se obtuvo una red Bayesiana. Un proceso de *clustering* se aplicó posteriormente sobre el conjunto de redes obtenidas, eligiendo una red Bayesiana representativa de cada clúster. Finalmente, la agregación se realiza en el nivel de estas redes representantes, combinándolas en una multi-red Bayesiana.

Por último, aunque queda fuera del alcance de este documento, existen algunos ejemplos de agregación de parámetros, también escasos. Estrategias populares de agregación de parámetros son los sondeos lineales (LinOP, Maynard-Reid y Chajewska [2001]) y los sondeos logarítmicos (LogOP, Pennock y Wellman [1999]). Recientemente, Etminani et al. [2013] proponen efectuar *clustering* sobre parámetros obtenidos de expertos y agregar mediante sondeo solamente aquellos que corresponden al clúster con el mayor número de miembros (DemocraticOP).

6. Métodos

Como hemos visto en el capítulo anterior, no existen muchas propuestas para la agregación de redes Bayesianas, y las que existen trabajan con redes discretas. En este trabajo se proponen dos nuevos métodos de fusión o agregación de redes Bayesianas Gaussianas que a continuación se detallarán.

Se asumirán datos estandarizados, práctica habitual en los procedimientos de aprendizaje (véase Capítulo 5).

6.1. Visión general

Antes de proceder a detallar los dos métodos propuestos en esta tesis, se dará una visión general del esquema de agregación de las redes.

En la Figura 6.1 aparece representada la agregación de estructuras. Se parte de conjuntos de datos (o expertos, véase Capítulo 1) de los que se asumen obtenidas sendas redes Gaussianas. Nótese que como se asume una distribución teórica de probabilidad única subyacente a las distintas redes estimadas a partir de los datos, se espera que las redes difieran en un número no muy amplio de arcos. A continuación, se aplica el método de agregación que se haya elegido. La red resultante será la entrada del posterior método de agregación de parámetros, representado en la Figura 6.2.

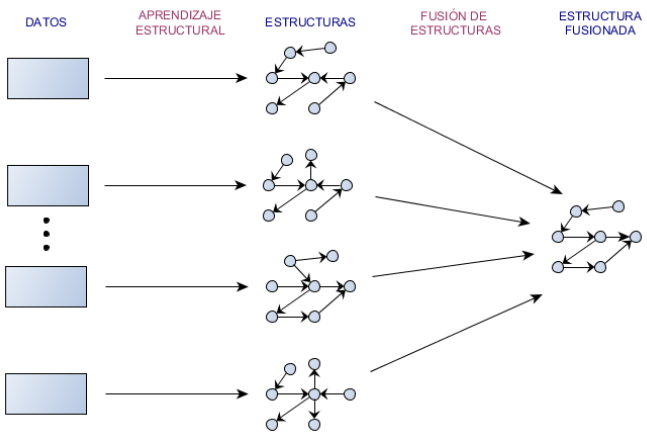

Figura 6.1. Esquema de fusión de las estructuras.

Para agregar los parámetros el procedimiento es análogo. Existen dos opciones: la estructura ya fusionada es empleada para aprender los parámetros a partir de cada conjunto de datos o experto; o se utiliza cada una de las estructuras individuales fusionadas en el paso anterior. Este último caso es el único viable cuando no se puede repetir el aprendizaje (común si se trabaja con expertos); si se dispone de conjuntos de datos o acceso repetido a los expertos, cualquiera de las dos alternativas es posible.

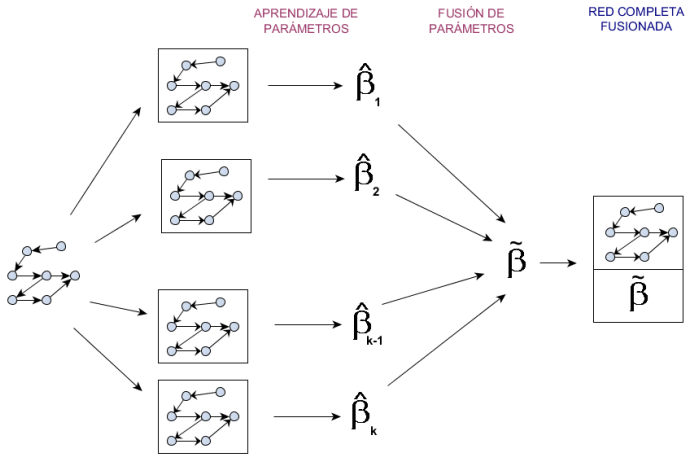

Figura 6.2. Esquema de fusión de los parámetros.

6.2. *GBNFuseSVote*: método basado en votación

El método basado en votaciones (*GBNFuseSVote*, Algoritmo 1) realiza la fusión de las estructuras de redes Bayesianas siguiendo un esquema similar al caso de los meta-clasificadores (*ensembles*), en los que cada modelo individual aporta su voto y el voto mayoritario es elegido.

En el caso de las redes Gaussianas, la votación se efectúa sobre los arcos. La red agregada quedará determinada por el número de votos sobre cada posible arco.

Las votaciones se representan utilizando una matriz que consiste en la suma de las matrices de adyacencia de cada grafo dirigido. Los elementos de una matriz de adyacencia $\boldsymbol{M}^{\mathcal{G}}$ de dimensión $p \times p$ asociada a un grafo $\mathcal{G} = (V = \{1, ..., p\}, E)$ se definen, para cada $i, j \in \{1, ..., p\}$, como

$$m_{ij}^{\mathcal{G}} = \begin{cases} 1 \text{ si } (i,j) \in E \\ 0 \text{ en otro caso.} \end{cases}$$

De esta forma, si denotamos por $\{\mathcal{G}_1, ..., \mathcal{G}_k\}$ el conjunto de grafos dirigidos acíclicos obtenidos a partir de cada partición de datos o experto, la matriz de votaciones $\boldsymbol{V}$ queda

Algoritmo 1 GBNFuseSVote

Entrada: *datasets*. Datos de los que se aprenderán las redes individuales.

Entrada: *threshold*. Umbral para el voto mayoritario.

Salida: Red Bayesiana agregada con el umbral dado.

```
1: n_bn ←size(datasets);
2: bn_list ←list();
3: for i ∈ {1, n_bn} do
4:     bn_list[i] ←learn_struc(datasets[i])
5: end for
6: v_matrix ← get_votes(bn_list);
7: result ←bn_aggr(threshold, v_matrix);
8: return result;
```

definida como

$$\boldsymbol{V} = \sum_{j=1}^{k} \boldsymbol{M}^{\mathcal{G}_j},$$

donde la suma está bien definida al estar todas las redes Bayesianas definidas sobre el mismo conjunto de variables aleatorias $\{X_1, ..., X_p\}$. Este procedimiento de obtención de votos se ilustra en el Algoritmo 2.

Algoritmo 2 get_votes

Entrada: *bn_list*. Lista de redes Bayesianas individuales.

Salida: Matriz de votos.

```
1: n_nodes ←nodes(bn);
2: v_matrix ←matrix(n_nodes, n_nodes);
3: for bn ∈ bn_list do
4:     bn_arcs ←arcs(bn);
5:     for arc ∈ bn_arcs do
6:         from ←from(arc);
7:         to ←to(arc);
8:         v_matrix[from][to] ← v_matrix[from][to] + 1;
9:     end for
10: end for
11: return v_matrix;
```

Obtenida la matriz de votaciones $\boldsymbol{V}$, cada una de los elementos es comparado con el umbral especificado, parámetro del método, y aquellos arcos que lo superen o igualen son incluidos.

Es decir, la matriz de adyacencia final, $\boldsymbol{M}^{\mathcal{G}_F}$, para el grafo fusionado $\mathcal{G}_F$, queda definida por sus elementos $m_{ij}^{\mathcal{G}_F}$, $i, j \in \{1, ..., p\}$, suponiendo que no se ocasionan ciclos, como

$$m_{ij}^{\mathcal{G}_F} = \begin{cases} 1 \text{ si } v_{ij} \geq t \\ 0 \text{ en otro caso,} \end{cases}$$

donde v_{ij} es el elemento (i, j) de la matriz de votos $\boldsymbol{V}$ y t es el umbral especificado. El método descrito para la construcción de la matriz de adyacencia final se muestra en pseudocódigo en el Algoritmo 3. Nótese que el grafo final queda completamente determinado por la matriz de adyacencia obtenida. En caso de que la adición de un arco cree un ciclo, éste es descartado. Este último apunte se discute en el Capítulo 8.

Algoritmo 3 bn_aggr

Entrada: *threshold*. Umbral para la votación.
Entrada: *v_matrix*. Matriz de votos.
Salida: *bn*. Estructura agregada de la red Gaussiana.

```
1:  bn ← empty_dag();
2:  for i ∈ cols(v_matrix) do
3:     for j ∈ rows(v_matrix) do
4:        if v_matrix[i][j] ≥ threshold then
5:           if not arc_causes_cycle(bn, i, j) then
6:              add_arc(bn, i, j);
7:           end if
8:        end if
9:     end for
10: end for
11: return bn;
```

6.3. *GBNFuseSInter*: método basado en intersección

Este método se basa en la representación de la distribución normal multivariante en la forma factorizada y en un teorema de del Sagrado y Moral [2003], que se enunciará a continuación. En el enunciado original del teorema se utiliza la noción de *I-map* y los axiomas de los *grafoides* o *modelos de independencia* [Geiger et al., 1990]. Un I-map es una forma alternativa de definir una red Bayesiana, basada en ciertos axiomas sobre enunciados lógicos. De este modo, los axiomas de probabilidad son un caso particular de los axiomas de grafoides, siendo un I-map

un grafo que representa un modelo de independencia de probabilidad condicionada, aunque no necesariamente de forma completa (como sucede con la definición de red Bayesiana vía las propiedades de Markov).

Teorema 8. [del Sagrado y Moral, 2003] *Sean $\mathcal{G} = (V, E_{\mathcal{G}})$ y $\mathcal{H} = (V, E_{\mathcal{H}})$ dos grafos dirigidos acíclicos correspondientes a sendas redes Bayesianas con modelos de independencia $I_{\mathcal{G}}$ e $I_{\mathcal{H}}$. Si es posible encontrar un orden ancestral compatible con $\mathcal{G}$ y $\mathcal{H}$, entonces $\mathcal{G} \cap \mathcal{H} = (V, E_{\mathcal{G}} \cap E_{\mathcal{H}})$ es una red Bayesiana del cierre del modelo $I_{\mathcal{G}} \cup I_{\mathcal{H}}$ bajo los axiomas de la probabilidad condicional.*

El Teorema 8 da una intuición sobre agregación de modelos de independencia condicional, en la que se basa este segundo método, *GBNFuseSInter*. En concreto, este teorema implica que, bajo ciertas condiciones, la intersección del conjunto de arcos de las redes individuales aporta una representación de la unión de los modelos de independencia condicional. Por tanto, si consideramos el conjunto $\{\mathcal{G}_1, ..., \mathcal{G}_k\}$ de grafos iniciales, con $\mathcal{G}_j = (V, E_j)$ para cada $j \in \{1, ..., k\}$, se define de forma inicial el conjunto de arcos $E_{\mathcal{G}_F}$ de la red agregada como

$$\tilde{E}_{\mathcal{G}_F} = \bigcap_{j=1}^{k} E_j.$$

Sin embargo, el criterio de intersección, aunque teóricamente justificado, es restrictivo en la práctica, puesto es común que existan arcos presentes en la mayoría, pero no todas, de las redes individuales, y que pertenecen al modelo teórico. Para mejorar el criterio de intersección, recuérdese (Capítulo 4), en primer lugar, la forma factorizada de la distribución Gaussiana multivariante en una red Bayesiana:

$$f(\boldsymbol{x}) = \prod_{i=1}^{p} f(x_i \mid \boldsymbol{x}_{\mathrm{pa}(i)}),$$

donde

$$X_i \mid \boldsymbol{x}_{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mu_i + \sum_{j \in \mathrm{pa}(i)} \beta_{ij}(x_j - \mu_j), \sigma_{ii} - \sum_{j \in \mathrm{pa}(i)} \beta_{ij}\sigma_{ji}\right). \tag{6.1}$$

Como se mostró en el Capítulo 4, un coeficiente $\beta_{ij} = 0$ implica que el nodo j no es padre del nodo i en el grafo dirigido acíclico de la red Gaussiana. Por tanto, una medida de la relevancia de un arco excluido se obtiene a partir de los coeficientes de regresión. Concretamente, si $(u, v) \in E_j$ para algún $j \in \{1, ..., p\}$ y tal que $(u, v) \notin \tilde{E}_{\mathcal{G}_F}$, $\hat{\beta}_{uv}$ es un factor relevante en la decisión sobre la inclusión de (u, v) en el modelo final. De esta forma, el conjunto de arcos de la red fusionada queda definido como

$$E_{\mathcal{G}_F} = \tilde{E}_{\mathcal{G}_F} \cup \left\{(u, v) : (u, v) \in \bigcup_{j=1}^{k} E_j \setminus \tilde{E}_{\mathcal{G}_F}, \hat{\beta}_{uv} \geq t\right\},$$

donde $t \in [0, 1]$ es un parámetro de umbral y $\hat{\beta}_{uv}$ se asume normalizado.

Algoritmo 4 GBNFuseSInter

Entrada: *datasets*. Datos de los que se aprenderán las redes individuales.
Entrada: *threshold*. Umbral para la adición del resto de arcos.
Salida: Red Bayesiana agregada con el umbral dado.

```
n_bn ←size(datasets);
bn_list ←list();
for i ∈ {1, n_bn} do
    bn_list[i] ←learn_struc(datasets[i])
end for
bn_base ← dag_intersection(bn_list);
result ←bn_aggr(threshold, bn_base, bn_list);
return result;
```

En el Algoritmo 4 se describe en pseudocódigo el método expuesto. La intersección de los grafos se efectúa de nuevo en base a la matriz de votos $\boldsymbol{V}$; sin embargo, en este caso, los elementos $\tilde{m}_{uv}$ de la matriz del grafo intersección, $u, v \in \{1, ..., p\}$, se definen como

$$\tilde{m}_{uv} = \begin{cases} 1 \text{ si } v_{ij} = k \\ 0 \text{ en otro caso,} \end{cases}$$

siendo k el número de grafos individuales $\mathcal{G}_1, ..., \mathcal{G}_k$. Al ser este procedimiento y el de agregación en base al umbral para el coeficiente de regresión análogos a los Algoritmos 2 y 3, respectivamente, se ha omitido de este apartado su pseudocódigo.

6.4. Agregación de parámetros

Aunque el objetivo principal de este documento es la agregación de estructuras de redes Gaussianas, se propone a continuación un método de agregación de parámetros, quedando de esta forma completado el proceso de fusión descrito al comienzo de este capítulo. En el método propuesto se considera la estructura fusionada y un nuevo conjunto de coeficientes de regresión obtenido en base a la misma, en cada partición de datos.

Supóngase un modelo de regresión lineal múltiple en el que los predictores son $\{X_1, ..., X_q\}$, y denótese por $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{1j}, ..., \hat{\beta}_{qj})$ el vector de estimadores para los coeficientes de regresión en el conjunto de datos j, $j \in \{1, ..., k\}$. Sean muestras de igual tamaño $\boldsymbol{X}_j = \{\boldsymbol{X}_j^{(1)}, ..., \boldsymbol{X}_j^{(N)}\}$ en cada partición $j \in \{1, ..., k\}$, donde $\boldsymbol{X}_j^{(n)} = (X_{1j}^{(n)}, ..., X_{qj}^{(n)})$. En este contexto, se usará la

notación λ_{ij} para el elemento (i,i) en la matriz $(\boldsymbol{X}_j \boldsymbol{X}_j^t)^{-1}$. De este modo, para cada predictor X_i, $i \in \{1, ..., q\}$, se define un estimador agregado de los coeficientes de regresión como

$$\tilde{\beta}_i = \sum_{j=1}^{k} w_{ij} \hat{\beta}_{ij},$$

donde

$$w_{ij} = \frac{r_{ij}^{-1}}{\sum_{j=1}^{k} r_{ij}^{-1}}, \; r_{ij} = \lambda_{ij} \sigma_{jj},$$

siendo σ_{jj} la varianza del error en la partición j. Se cumple que $\tilde{\beta}_i$ tiene varianza mínima entre los estimadores lineales de la forma

$$\sum_{j=1}^{k} w_{ij} \hat{\beta}_{ij}, \text{ tal que } \sum_{j=1}^{k} w_{ij} = 1,$$

y tiene una distribución asintóticamente Gaussiana [Fan et al., 2007].

El estimador introducido es un candidato para el caso de las redes Gaussianas, en las que asumiendo datos estandarizados se tiene para cada variable un modelo de regresión lineal múltiple (véase Capítulo 4). El esquema de agregación de parámetros esbozado se enmarca dentro de lo que del Sagrado y Moral [2003] denominan *fusión topológica*: obtener una estructura consenso y posteriormente estimar los parámetros. La alternativa, agregar o fusionar las distribuciones de probabilidad de cada red y obtener posteriormente una estructura que la represente, se suele denominar *representación gráfica del consenso*.

7. Resultados

En este capítulo se presentarán los resultados obtenidos tras implementar los métodos descritos en el Capítulo 6 y utilizarlos sobre distintos conjuntos de datos. Para dicha tarea se han empleado utilidades del paquete de R `bnlearn` [Scutari, 2010].

7.1. *GBNFuseSVote*: ejemplo ilustrativo

A continuación se introducirá un ejemplo de la actuación del método de votación del Capítulo 6 (*GBNFuseSVote*) sobre una red objetivo de tamaño pequeño, mostrada en la Figura 7.1.

A partir de la red original, asumiendo una distribución Gaussiana multivariante, cuya estructura de independencia viene dada por el grafo de la Figura 7.1, se han generado 8 conjuntos de datos sintéticos con ruido, de 50 muestras cada uno (las particiones de datos).

Sobre cada conjunto de datos se ha simulado el proceso de aprendizaje de una red individual (análogamente, el proceso de educción de información de un experto), utilizando el algoritmo de Tsamardinos et al. [2006], por su buen rendimiento general. En la Figura 7.2 se observan las diferentes estructuras que se han obtenido sobre cada conjuntos de datos, resaltando en verde los arcos correctamente aprendidos, en rojo los sobrantes y en azul los faltantes.

Como se puede observar, en la mayoría de los casos, una porción sustancial de la red es correctamente aprendida, siendo los falsos positivos (arcos rojos) el error más común. Se

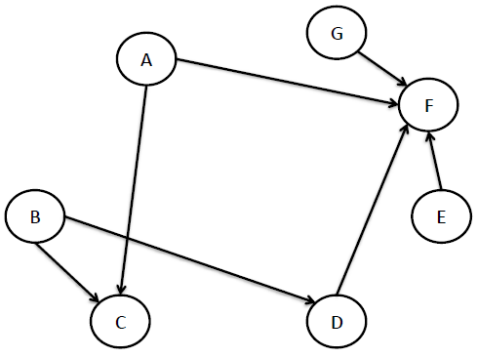

Figura 7.1. Estructura original usada para los experimentos.

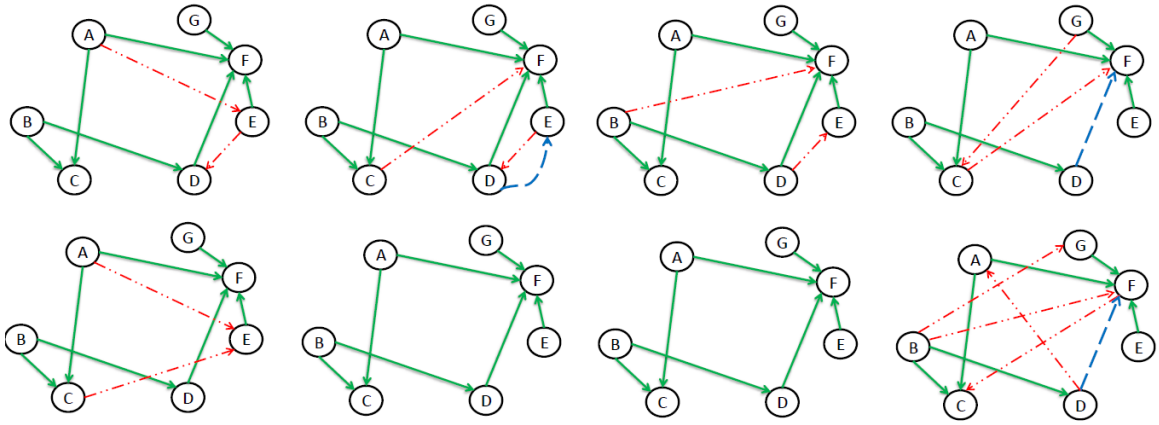

Figura 7.2. Estructura aprendida sobre cada conjunto de datos. Los arcos verdes son aquellos correctamente aprendidos, los rojos son los falsos positivos y los azules los falsos negativos.

denotará este conjunto de redes por $\{\mathcal{G}_1, ..., \mathcal{G}_8\}$, con $\mathcal{G}_k = (V, E_k)$ para $k \in \{1, ..., 8\}$.

Obtenidas las ocho redes individuales, se aplica el método de fusión basado en votaciones. Dado que el umbral de votos $t \in \{1, ..., 8\}$ es un parámetro del algoritmo, se ejecuta para cada valor, obteniendo las distintas redes fusionadas mostradas en la Figura 7.3.

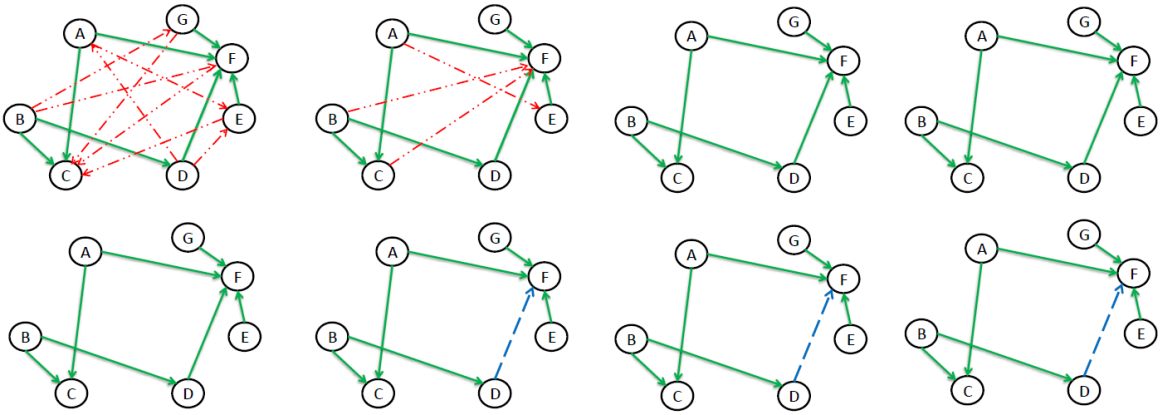

Figura 7.3. Estructuras agregadas para los distintos umbrales $\{1, ..., 8\}$, creciente de izquierda a derecha comenzando por la primera fila. Los arcos verdes son aquellos en los que se da una coincidencia con la estructura original; en rojo aparecen los falsos positivos y en azul los falsos negativos.

Como es de esperar, a medida que se aumenta el umbral para los votos, la red agregada es más restrictiva y contiene menos arcos. Esto es así porque se verifica que la matriz de adyacencia del grafo fusionado, $M^{\mathcal{G}_F}$, con umbral $t = 1$ (véase Capítulo 6) da lugar a un grafo $\mathcal{G}_F = (V, E_F)$ con $E_F = \cup_{k=1}^{8} E_k$ (salvo un conjunto E_C de arcos que hayan originado ciclos, dándose en este caso $E_C = \emptyset$).

Red	SHD	TP	FP	FN
1	3	7	2	0
2	3	6	2	1
3	2	7	2	0
4	3	6	2	1
5	2	7	2	0
6	0	7	0	0
7	0	7	0	0
8	5	6	4	1

Umbral (t)	SHD	TP	FP	FN
1	8	7	8	0
2	3	7	3	0
3	0	7	0	0
4	0	7	0	0
5	0	7	0	0
6	1	6	0	1
7	1	6	0	1
8	1	6	0	1

Tabla 7.1. Resultados de las redes aprendidas en cada partición (izquierda) y las agregadas bajo distintos umbrales (derecha) comparadas con la estructura original. TP, FP y FN denotan el número de aciertos, falsos positivos y falsos negativos, respectivamente. SHD denota la distancia estructural de Hamming. Las redes están numeradas de acuerdo a su aparición en las respectivas figuras, de izquierda a derecha comenzando por la fila superior.

Para evaluar los resultados obtenidos en este ejemplo (Figura 7.3) y en los siguientes apartados, se utilizarán diferentes métricas. En primer lugar, recordando la relación de equivalencia que se da entre las redes Bayesianas respecto al modelo de independencia condicional que representan (véase Capítulo 5), puede que se obtengan redes que a simple vista difieren de la original, pero que son equivalentes. La métrica definida como *distancia estructural de Hamming* (SHD, Tsamardinos et al. [2006]), solventa este problema. Contabiliza el número de operaciones sobre arcos necesarias para igualar los grafos dirigidos parciales, representantes de las clases de equivalencia de las estructuras comparadas. Las operaciones consideradas son: añadir, quitar e invertir un arco. Gracias a esta métrica, se obtiene una comparación entre las redes que no penaliza a aquellas estadísticamente indistinguibles.

En la Tabla 7.1 se muestran los resultados para la métrica SHD así como las tasas de acierto, falsos positivos y falsos negativos, respecto de la red original de la Figura 7.1, tal y como se ha ido mostrando en las Figuras 7.2 y 7.3.

Se observa que, exceptuando el caso $t = 1$, los resultados de la agregación son en general superiores con respecto a los resultados del aprendizaje individual de cada red, obteniendo para $t \geq 2$ una tasa nula de falsos positivos y $SHD \in \{0, 1\}$, es decir, la red agregada representa, salvo una operación de arcos en el peor caso, el mismo modelo de independencia condicional. Los mejores resultados se dan para valores intermedios de t: en $t \in \{3, 4, 5\}$ la estructura original es recuperada de forma exacta. También se observa que para valores altos de t, los falsos negativos aumentan, siendo de esperar por ser más restrictiva la votación.

7.2. *GBNFuseSInter*: redes *Alarm*, *Insurance* y *Hailfinder*

El ejemplo del apartado anterior constituye una red pequeña sobre la que ya se visualiza el desarrollo y las ventajas del método de votaciones. Para experimentar también con redes de mayor tamaño, se han elegido tres de referencia, incluidas en el repositorio de `bnlearn`[1] y comúnmente utilizadas para validar métodos sobre redes Gaussianas (véase Vidaurre et al. [2010], Huang et al. [2013]). Sus características se muestran en la Tabla 7.2 y su representación gráfica aparece en el Apéndice A.

Red (t)	Nodos	Arcos
Alarm	37	46
Insurance	27	52
Hailfinder	56	66

Tabla 7.2. Características de las redes Bayesianas de referencia a utilizar

Análogamente al apartado anterior, se ha simulado en cada bloque de la partición 5000 instancias y se han considerado dos casos, una partición de 10 y otra de 100 bloques. Nótese que las redes utilizadas de referencia son tradicionalmente discretas, por lo que los coeficientes de regresión de referencia han sido simulados de forma uniforme. Se han utilizado las siguientes métricas sobre los resultados: el número de arcos de la red obtenida, la distancia estructural de Hamming (SHD) y los falsos positivos.

En la Figura 7.4 aparecen distintas gráficas con diagramas de cajas para cada una de las redes representando la distribución del número de arcos. En cada gráfica aparecen dos diagramas, el de la izquierda correspondiente a la distribución sobre las redes individuales aprendidas en cada bloque, y, el de la derecha, a la distribución sobre las redes agregadas en función de todos los umbrales posibles para la comparación con los coeficientes de regresión.

La primera conclusión a extraer de este conjunto de gráficas es que, independientemente del valor del umbral, la red fusionada contiene un número de arcos menor que las redes aprendidas de forma individual, llegando a ser la diferencia en varios casos igual o superior a 20. También se observa cómo los resultados obtenidos presentan, excepto en un caso, poca variabilidad en función del umbral (el número de arcos en la red fusionada se encuentra cercano a su media); mientras que la variabilidad en el caso de aprendizaje individual es superior (cajas más anchas).

Si bien las redes obtenidas tienen menos arcos, es necesario estudiar la distancia estructural de Hamming para comprobar si en ese proceso de reducción de arcos, que ha provocado el

[1] http://www.bnlearn.com/bnrepository/

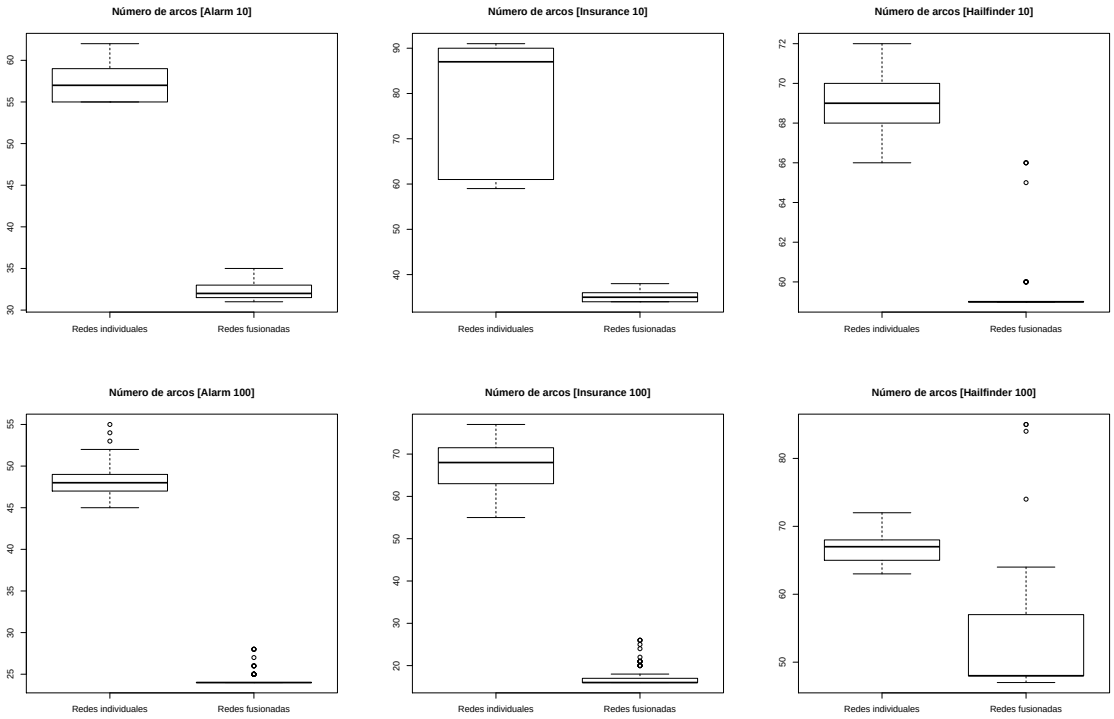

Figura 7.4. Arcos para todos los casos de prueba del método de intersección. En la fila superior aparece el caso de una partición de 10 elementos, de izquierda a derecha se muestran las redes *Alarm*, *Insurance* y *Hailfinder*. En la fila inferior, en el mismo orden, aparece el caso de una partición de 100 elementos.

método de fusión basado en intersección, se ha producido un acercamiento al modelo original con respecto a lo obtenido mediante el aprendizaje individual. En la Figura 7.5 se muestran los diagramas de cajas análogos a la Figura 7.4, anteriormente explicada, pero en este caso para la distancia estructural de Hamming.

Observamos que, salvo en un caso, la distancia estructural de Hamming obtenida es inferior (excluyendo elementos aislados) a la presente en el caso de aprendizaje individual de redes. Con esto se confirma la hipótesis de Richardson y Domingos [2003], que mencionamos en el Capítulo 5, acerca de cómo permitiendo a los modelos obtenidos de forma individual ser algo ruidosos, se compensa dicho ruido obteniendo una estructura consenso. En este caso, la estructura fusionada tiene, en general, un menor número de arcos respecto a las redes individuales, pero se aproxima de forma igual o más precisa a la red original. Esto implica que los arcos eliminados constituían en su mayoría ruido.

Finalmente, los falsos positivos, representados de forma análoga en la Figura 7.6, confirman la discusión de resultados efectuada hasta el momento. En este caso, no sólo de forma general los falsos positivos en las redes agregadas se encuentran por debajo de aquellos en las redes

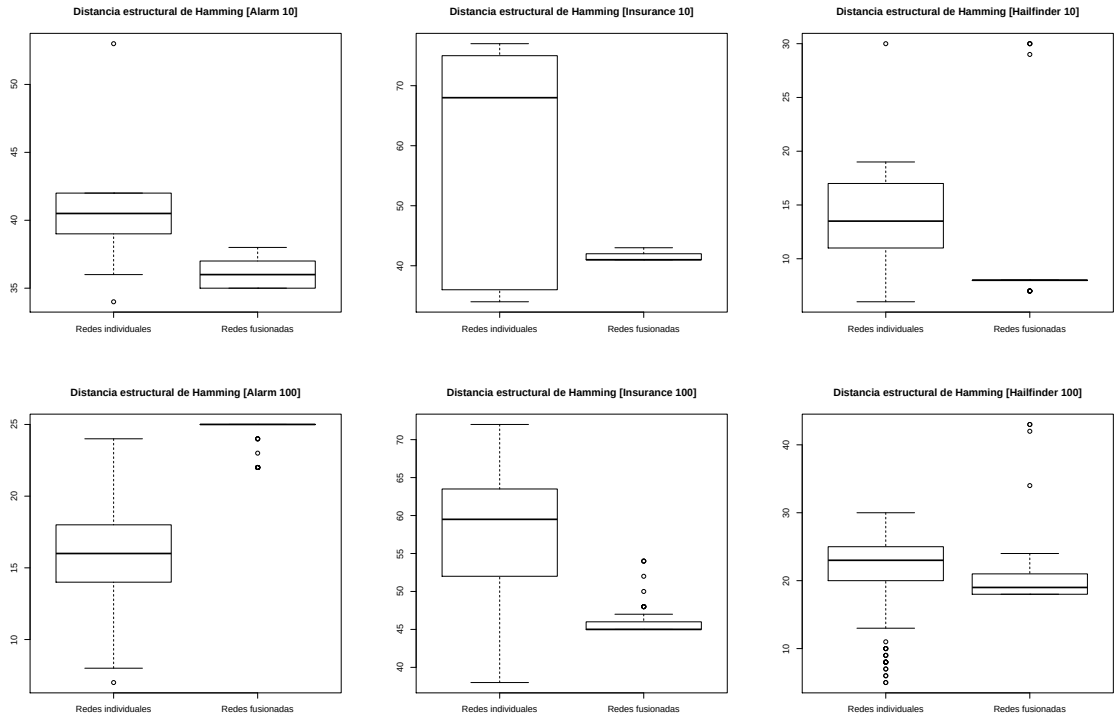

Figura 7.5. Distancia estructural de Hamming (SHD) para todos los casos de prueba del método de intersección. En la fila superior aparece el caso de una partición de 10 elementos, de izquierda a derecha se muestran las redes *Alarm*, *Insurance* y *Hailfinder*. En la fila inferior, en el mismo orden, aparece el caso de una partición de 100 elementos.

aprendidas individualmente, sino que además en muchos casos la media es cercana o igual a cero. Se obtienen por tanto redes de gran precisión, cercanas a la red original (véase Figura 7.5) y con un menor número de parámetros a aprender por tener una menor cantidad de arcos (véase Figura 7.4).

7.3. Comparación: *GBNFuseSVote* y *GBNFuseSInter*

En los apartados anteriores se han realizado experimentos con los dos métodos propuestos en el Capítulo 6, analizándolos de forma separada. En esta sección se realizará una comparación de ambos métodos sobre las mismas redes Bayesianas individuales.

Las redes de referencia utilizadas para la comparativa son las de la Tabla 7.2. Se utilizarán para este apartado las redes individuales obtenidas en el apartado anterior (a las que se aplicó *GBNFuseSInter*), sobre las que en este caso también se aplicará el método de agregación basado en votación (*GBNFuseSVote*).

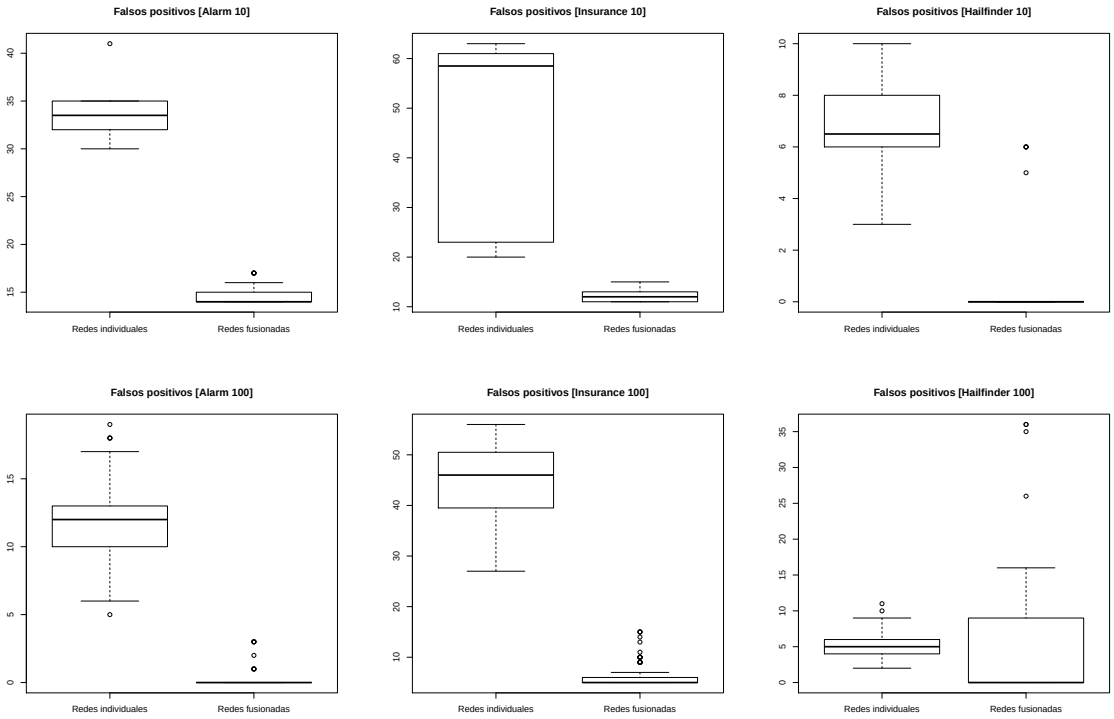

Figura 7.6. Falsos positivos (FP) para todos los casos de prueba del método de intersección. En la fila superior aparece el caso de una partición de 10 elementos, de izquierda a derecha se muestran las redes *Alarm*, *Insurance* y *Hailfinder*. En la fila inferior, en el mismo orden, aparece el caso de una partición de 100 elementos.

En la Figura 7.7 se muestra la evolución para cada red de referencia del número de arcos con respecto al parámetro de cada método: umbral de votación en el caso de *GBNFuseSVote* y umbral sobre el coeficiente de regresión en el caso de *GBNFuseSInter*. Se observa que siempre son funciones decrecientes, siendo más abruptas con *GBNFuseSInter*. Este hecho implica que la intersección inicial ($\tilde{E}_{\mathcal{G}_F}$, véase Capítulo 6) contiene la mayoría de los arcos *relevantes*, es decir, aquellos (u, v) en los que $\beta_{uv} \geq t$ para $t \in [0, 1]$. La independencia del parámetro en cuanto al número de arcos es por tanto mayor en *GBNFuseSInter* que en *GBNFuseSVote*.

La evolución de la distancia estructural de Hamming, mostrada de forma análoga en la Figura 7.8, sigue un patrón similar en el que las curvas son más suaves para *GBNFuseSVote*. Sin embargo, se observa en este caso que, para ambos métodos, las funciones no son decrecientes, sino que decrecen hasta un punto óptimo para después volver a crecer ligeramente o mantenerse. Este resultado es esperado, dado que, en ambos casos, a medida que el parámetro t (umbral) crece, se hacen más estrictos los requisitos sobre las aristas a incluir en el modelo agregación. En el caso de *GBNFuseSVote*, se requieren mayor cantidad de votos para la inclusión del

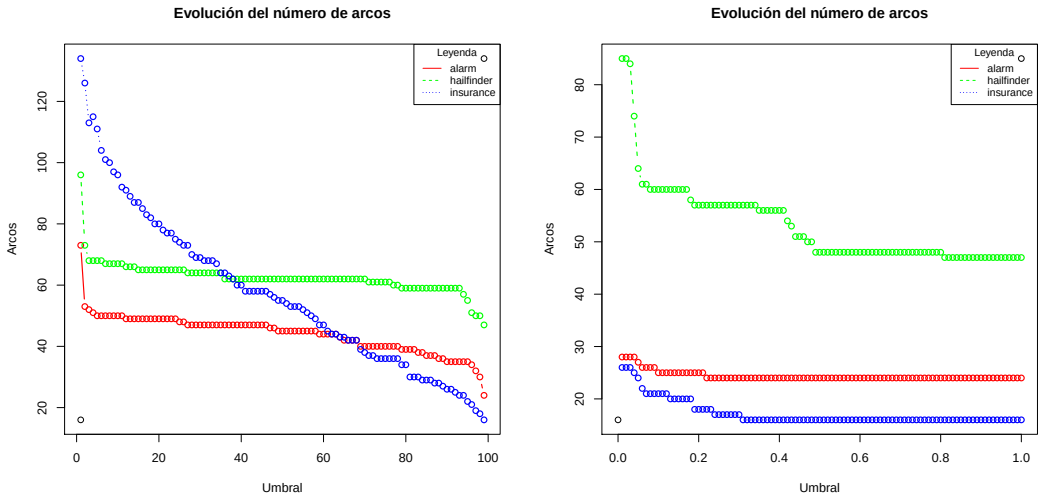

Figura 7.7. Evolución, para cada red, del número de nodos con *GBNFuseSVote* (izquierda) y *GBNFuseSInter* (derecha).

arco; en el caso de *GBNFuseSInter*, se requiere que el valor del correspondiente coeficiente de regresión sea más alto, y por tanto la dependencia entre las variables sea más fuerte.

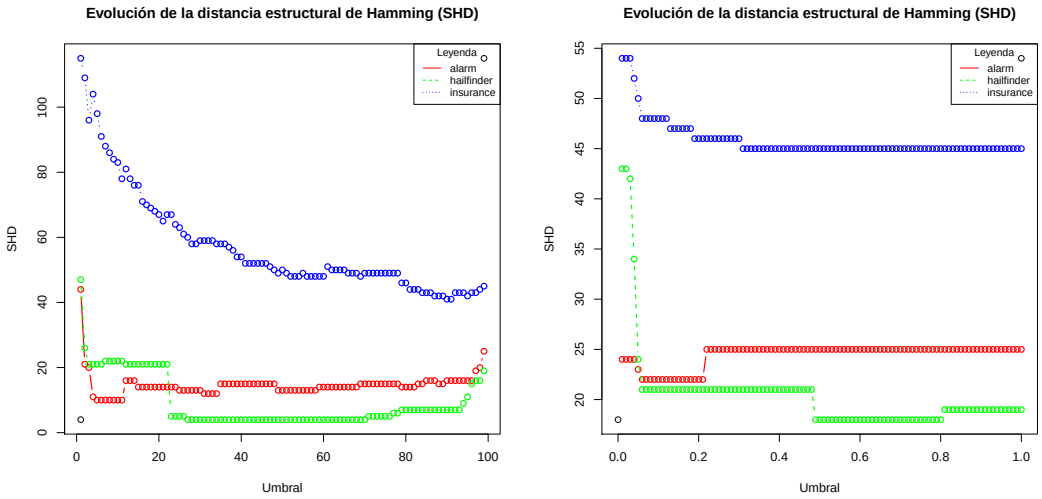

Figura 7.8. Evolución, para cada red, de la distancia estructural de Hamming (SHD) con *GBNFuseSVote* (izquierda) y *GBNFuseSInter* (derecha).

En cualquier caso, para la distancia estructural de Hamming, los valores óptimos son los más frecuentes, y, si omitimos aquellos valores de umbral demasiado bajos (es decir, que relajan demasiado las restricciones sobre los arcos a incluir), se observa en *GBNFuseSInter* un comportamiento casi constante en función del umbral. Para *GBNFuseSVote*, en cambio,

existe mayor variabilidad del resultado con respecto al parámetro, siendo la matriz de votos más susceptible a los arcos ruidosos.

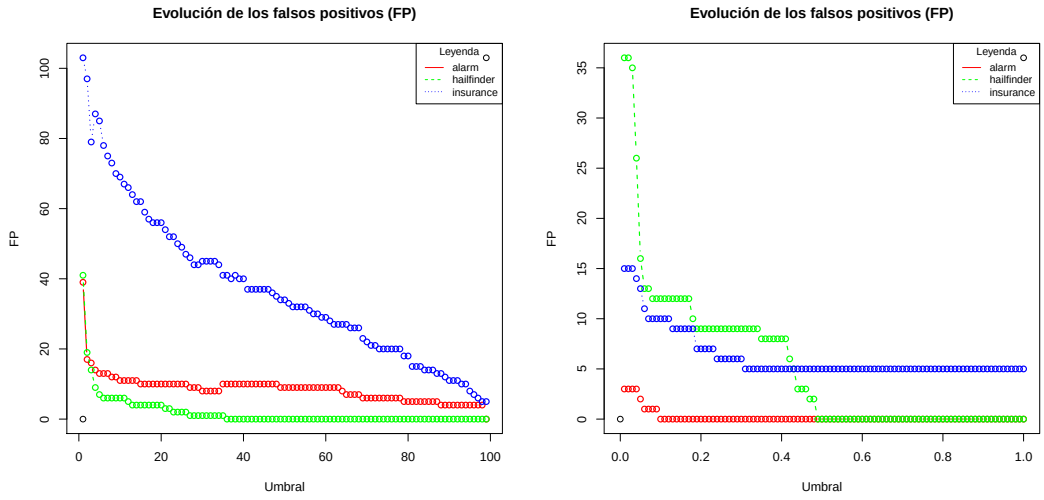

Figura 7.9. Evolución, para cada red, de los falsos positivos (FP) con *GBNFuseSVote* (izquierda) y *GBNFuseSInter* (derecha).

Finalmente, en la Figura 7.9 se muestra la evolución de los falsos positivos con respecto del umbral. La tendencia es similar a la comentada para las Figuras 7.7 y 7.8: funciones decrecientes y más abruptas para *GBNFuseSInter*. Una característica a destacar en este caso particular es que, en ambos métodos, se alcanza la tasa cero de falsos positivos, siendo especialmente precisos los modelos obtenidos por *GBNFuseSInter* y desde valores tempranos del umbral.

Como se ha observado en la exposición realizada, *GBNFuseSVote* es un método más simple que *GBNFuseSInter*. No obstante, generalmente obtiene resultados mejores que las redes individuales (compárense las figuras de este apartado con las del apartado anterior). Sin embargo, las métricas de número de arcos, distancia estructural y falsos positivos decrecen más rápidamente con respecto del umbral, aunque de forma abrupta, en el caso de *GBNFuseSInter*. Por tanto, en caso de disponer de datos sobre los coeficientes de regresión de las redes individuales, *GBNFuseSInter* es la elección a efectuar. Si, por el contrario, solamente se dispone de datos acerca de la estructura, *GBNFuseSVote* obtiene mejores resultados que la selección arbitraria de una de las estructuras individuales obtenidas.

8. Conclusiones y trabajo futuro

Se ha explorado el trabajo existente en dicho campo, encontrándolo escaso y principalmente centrado en los modelos de independencia condicional y redes discretas. Debido a ello, se han propuesto y evaluado dos métodos para la agregación de redes Bayesianas, obteniendo resultados prometedores con respecto a la selección arbitraria de redes individuales. A continuación se concluirá exponiendo las aportaciones de este trabajo a la literatura existente presentada en el Capítulo 5.

8.1. Conclusiones

En la literatura sobre fusión de redes Bayesianas, del Sagrado y Moral [2003] estudian propiedades teóricas de la unión e intersección de modelos de independencia condicional sin tener en cuenta la parte cuantitativa o las aplicaciones en el aprendizaje. Matzkevich y Abramson [1992], pioneros en la fusión de redes Bayesianas con su algoritmo *FuseDAG*, y Peña [2011], ignoran de igual forma este hecho al pretender encontrar la red que englobe todos los arcos originales o sus inversos. Esto es inviable en presencia de datos ruidosos, ya sean en forma de conjunto particionado o de expertos, puesto que la red resultante contendría un gran número de arcos (más incluso que la unión de las redes originales si se tienen en cuenta las inversiones), y, además, se estarían incluyendo con alta probabilidad arcos no relevantes, como muestran los experimentos en el Capítulo 7 (umbral 1 en método por votación). Los métodos propuestos en esta tesis tienen en cambio mayor flexibilidad al no perseguir la inclusión total de los arcos y tener en cuenta la estructura de regresión lineal múltiple inducida por la distribución Gaussiana.

El enfoque Bayesiano de Richardson y Domingos [2003], aunque con el objetivo de obtener una estructura *consenso*, es esencialmente distinto al resto, puesto que lo que se fusiona no son las estructuras de las redes, sino el conocimiento de los expertos, obteniendo una distribución *a priori* como resultado. Es con esta distribución con la que luego se efectúa una búsqueda avariciosa estándar en un conjunto de datos separado. En el caso de disponer de conjuntos de datos, esto sería análogo a realizar un sub-muestreo de los datos a partir de las estructuras individuales y de ahí aprender una nueva red. En definitiva, la fusión no se hace con los modelos de las redes Bayesianas, como es el objetivo de esta tesis.

Por último, el modelo de López-Cruz et al. [2014] está desarrollado sobre el problema específico de neurociencia que se está tratando, sin estudiar su generalidad. Los modelos presentados en esta tesis son en cambio aplicables a cualquier contexto en el que se disponga de datos Gaussianos y han sido probados en redes de referencia.

Dentro del escaso trabajo efectuado en este área se ha determinado que los métodos propuestos en este documento constituyen los primeros en el ámbito de las redes Gaussianas. Además, en el caso más general de redes Bayesianas, *GBNFuseSInter* es el primer método orientado tanto al aprendizaje y presencia de ruido, como a la independencia condicional y estructura cuantitativa de las redes subyacentes; aspectos notoriamente distanciados en los trabajos anteriores, como se ha destacado.

8.2. Trabajo futuro

Dado el volumen de trabajo y resultados en el área, existen muchas líneas interesantes de investigación que se podrían perseguir en un futuro. Un ejemplo es el tratamiento de los ciclos. En nuestro caso, al asumir una distribución y modelo subyacente común a las redes individuales, las cuales difieren solamente por la presencia de ruido, se ha elegido descartar los arcos. Una posibilidad sería la inversión de arcos utilizada por Matzkevich y Abramson [1992], Peña [2011]; sin embargo, se debe valorar si la penalización en la inferencia y el aprendizaje de parámetros por el mayor número de arcos compensa con respecto a los resultados de proximidad al modelo original (véase Capítulo 7 sobre discusión entre número de arcos, estructura condicional y precisión del modelo). Una posibilidad alternativa sería efectuar un *ranking* parcial de los arcos candidatos a causar ciclos (según el coeficiente de regresión asociado) y añadirlos de acuerdo a dicho *ranking* hasta cierto nivel de relevancia. El *ranking*, por otro lado, podría ser completo y los arcos añadidos de acuerdo al mismo, evitando así descartar arcos relevantes por haberlos considerado más tarde. Existen múltiples heurísticas, como las mencionadas, que se podrían estudiar para este problema, si bien es poco frecuente dadas las asunciones de los modelos.

Si las redes individuales se han obtenido a partir de una partición de los datos, se podría almacenar información asociada tanto a las particiones como a los procesos de aprendizaje utilizados en cada una (algo análogo es posible en el caso de conocimiento de expertos). Estos datos serían utilizados posteriormente para mejorar los métodos de fusión propuestos; por ejemplo, en el caso de las votaciones, con el objetivo de ponderar las mismas en función de la calidad de la partición, su adecuación para el procedimiento de aprendizaje utilizado, etc.

Finalmente, los métodos de agregación propuestos, dada su simpleza y buenos resultados, son especialmente atractivos para su utilización en un entorno escalable, pues proporcionan

una forma sencilla de obtener un modelo único, consensuado, más próximo al original que si de un solo aprendizaje se tratase, y con una menor complejidad al tener un menor número de arcos. Con el auge del *Big Data*, cabe esperar mayor experimentación en este ámbito, dando lugar a más propuestas como las contenidas en este documento.

A. Redes de referencia

Las redes *Alarm* (Figura A.1), *Insurance* (Figura A.2) y *Hailfinder* se han obtenido del repositorio http://www.bnlearn.com/bnrepository/ [Scutari, 2010]. La representación gráfica de *Hailfinder* se ha omitido dada su dimensión.

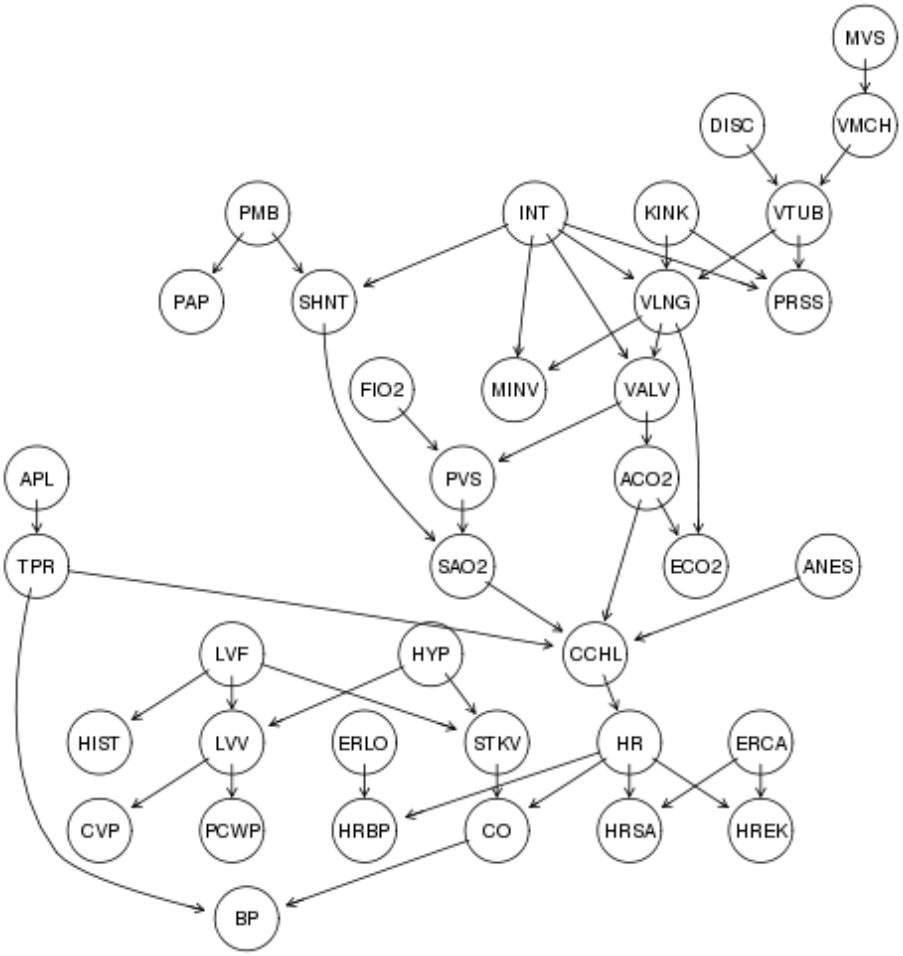

Figura A.1. Red *Alarm*.

Figura A.2. Red *Insurance*.

Bibliografía

M. Banerjee and T. Richardson. On a dualization of graphical Gaussian models: A correction note. *Scandinavian Journal of Statistics*, 30(4):817–820, 2003.

J. Bang-Jensen and G. Z. Gutin. *Digraphs: Theory, algorithms and applications*. Springer, 2008.

H. Blalock. *Causal Models in the Social Sciences*. Macmillan, 1971.

I. Córdoba-Sánchez, C. Bielza, and P. Larrañaga. Towards Gaussian Bayesian network fusion. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, LNAI 9161, pages 519–528. Springer, 2015.

G. Chartrand and L. Lesniak. *Graphs & Digraphs*. Wadsworth, 1986.

D. R. Cox and N. Wermuth. Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218, 1993.

J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, 8(3):522–539, 1980.

J. del Sagrado and S. Moral. Qualitative combination of Bayesian networks. *International Journal of Intelligent Systems*, 18(2):237–249, 2003.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

K. Etminani, M. Naghibzadeh, and J. M. Peña. DemocraticOP: A democratic way of aggregating Bayesian network parameters. *International Journal of Approximate Reasoning*, 54 (5):602 – 614, 2013.

T.-H. Fan, D. K. Lin, and K.-F. Cheng. Regression analysis for massive datasets. *Data & Knowledge Engineering*, 61(3):554–562, 2007.

N. Friedman and M. Goldszmidt. Learning Bayesian networks with local structure. In *Learning in Graphical Models*, volume 89 of *NATO ASI Series*, pages 421–459. Springer, 1998.

D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, UAI'94, pages 235–243. Morgan Kaufmann, 1994.

D. Geiger, T. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.

J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished manuscript, 1971.

F. Harary and R. C. Read. Is the null-graph a pointless concept? In *Graphs and Combinatorics*, volume 406 of *Lecture Notes in Mathematics*, pages 37–44. Springer, 1974.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

R. A. Howard and J. E. Matheson. Influence diagrams. *Decision Analysis*, 2(3):127–143, 2005.

S. Huang, J. Li, J. Ye, A. Fleisher, K. Chen, T. Wu, E. Reiman, A. D. N. Initiative, et al. A sparse structure learning algorithm for Gaussian Bayesian network identification from high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1328–1342, 2013.

G. Kauermann. On a dualization of graphical Gaussian models. *Scandinavian Journal of Statistics*, 23(1):105–116, 1996.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

P. L. López-Cruz, P. Larrañaga, J. DeFelipe, and C. Bielza. Bayesian network modeling of the consensus between experts: An application to neuron classification. *International Journal of Approximate Reasoning*, 55(1, Part 1):3 – 22, 2014.

S. L. Lauritzen. *Graphical Models*. Oxford, 1996.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):157–224, 1988.

S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

I. Matzkevich and B. Abramson. The topological fusion of Bayes nets. In *Proceedings of the Eighth International Conference on Uncertainty in Artificial Intelligence*, UAI'92, pages 191–198. Morgan Kaufmann, 1992.

P. Maynard-Reid and U. Chajewska. Aggregating learned probabilistic beliefs. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 354–361. Morgan Kaufmann, 2001.

V. McKim and S. Turner, editors. *Causality in Crisis?*, Proceedings of the Notre Dame Conference on Causality, 1977.

B. Mohar and W. Woess. A survey on spectra of infinite graphs. *Bulletin of the London Mathematical Society*, 21(3):209–234, 1989.

R. E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Inc., 2003.

J. Nielsen, T. Kocka, and J. Peña. On local optima in learning Bayesian networks. In *Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence*, UAI'03, pages 435–442. Morgan Kaufmann, 2003.

J. M. Peña. Finding consensus Bayesian network structures. *Journal of Artificial Intelligence Research*, 42(1):661–687, 2011.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Palo Alto, CA, 1988.

J. Pearl. *Causality*. Cambridge University Press, 2009.

D. M. Pennock and M. P. Wellman. Graphical representations of consensus belief. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 531–540. Morgan Kaufmann, 1999.

M. Richardson and P. Domingos. Learning with knowledge from multiple experts. In *Proceedings of the 20th International Conference on Machine Learning*, ICML'03, pages 624–631. AAAI Press, 2003.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4): 962–1030, 2002.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

R. D. Shachter and C. R. Kenley. Gaussian influence diagrams. *Management Science*, 35(5): 527–550, 1989.

T. P. Speed and H. T. Kiiveri. Gaussian Markov distributions over finite graphs. *Annals of Statistics*, 14(1):138–150, 1986.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

K. Thulasiraman and M. N. S. Swamy. *Graphs: Theory and Algorithms*. John Wiley & Sons, 2011.

I. Tsamardinos, L. Brown, and C. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

T. E. Van Rheenen, D. Meyer, and S. L. Rossell. Pathways between neurocognition, social cognition and emotion regulation in bipolar disorder. *Acta Psychiatrica Scandinavica*, 130 (5):397–405, 2014.

D. Vidaurre, C. Bielza, and P. Larrañaga. Learning an L1-regularized Gaussian Bayesian network in the equivalence class space. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(5):1231–1242, 2010.

J. Williamson. *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford University Press, 2005.

H. Wold. Causality and econometrics. *Econometrica*, 22:162–177, 1954.

S. Wright. The method of path coefficients. *Annals of Mathematical Statistics*, 5(3):161–215, 1934.

G. U. Yule. On the theory of correlation for any number of variables, treated by a new system of notation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 79(529):182–193, 1907.

B. Zhang and Y. Wang. Learning structural changes of Gaussian graphical models in controlled experiments. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, UAI'10, pages 701–708. AUAI Press, 2010.

S. Zhou, P. Rütimann, M. Xu, and P. Bühlmann. High-dimensional covariance estimation based on Gaussian graphical models. *Journal of Machine Learning Research*, 12:2975–3026, 2011.