Departamento de Matemáticas, Facultad de Ciencias
Universidad Autónoma de Madrid

Estimación de máxima verosimilitud para modelos gráficos gaussianos

Trabajo de fin de grado

Grado en Matemáticas

Autor: Daniel Brito Sotelo
Tutor: Bojan Mihaljevic

Curso 2021-2022

Resumen

El propósito de este trabajo es presentar los modelos gráficos gaussianos y sus usos para la estimación de parámetros.

Comenzamos introduciendo la teoría de grafos, y el concepto del grafo de independencia de una variable aleatoria, así como diversos resultados que nos permiten analizar las independencias condicionales entre variables en un grafo. A continuación, definimos los modelos gráficos gaussianos y observamos como se relacionan los parámetros de las distribuciones normales multivariantes con la estructura de su grafo.

Posteriormente, nos centraremos en la estimación de máxima verosimilitud de la matriz de covarianzas de una variable aleatoria dada la estructura de su grafo de independencia, tratando dicho problema desde el ámbito de la optimización convexa, para así proponer dos algoritmos iterativos para la maximización de la verosimilitud y obtención de los parámetros del modelo.

Hecho esto, nos centramos en discutir la existencia de la estimación de máxima verosimilitud en función del número de muestras y la estructura del grafo, y daremos algunos resultados que nos permitirán obtener estas estimaciones en escenarios en los que el número de muestras sea menor que el número de variables, siendo de gran interés para aplicaciones en biología o genética. Finalizamos el trabajo con una aplicación de todos los conceptos vistos a un conjunto de datos, empleando el lenguaje de programación estadístico R.

Abstract

The purpose of this paper is to study graphical models and their uses in parameter estimations of random variables, specifically we will be covering gaussian graphical models.

Firstly, we introduce a series of definitions and results regarding graph theory and independence graphs in order to analyze conditional independences between variables given their graph. Next, we will define gaussian graphical models and observe how their gaussanity together with their graph structure can be used for parameter estimation.

Furthermore, we will cover the maximum likelihood estimation of the covariance matrix of a multivariate normal distribution given its grap, presenting this problem as a convex optimization one in order to present two iterative methods to solve this maximization.

Finally, we will discuss the existence of said estimations based on the number of observations and variables, using the arrangement of the graph to give results that guarantee this existence even in cases when the number of observations is lower than the number of variables, a situation that frequently occurs in biological and genetic problems. We end this paper by applying all the aforementioned concepts to a dataset using R as our programming environment.

Índice general

1 Grafos de independencia **1**

2 Modelos gráficos gaussianos **7**

2.1 Covarianza parcial 7

2.2 Normal multivariante 10

2.3 Distribución marginal y condicionada 10

2.4 Información mutua 12

3 Estimación de máxima verosimilitud **15**

3.1 Verosimilitud 15

3.2 Estimaciones de máxima verosimilitud 16

3.3 Optimización convexa 18

3.4 Cálculo del estimador de máxima verosimilitud 20

3.5 Existencia del estimador de máxima verosimilitud 22

4 Aplicación a un conjunto de datos **26**

A Optimización convexa **30**

B Código R **32**

Bibliografía **33**

CAPÍTULO 1
Grafos de independencia

En este capítulo definiremos la relación de independencia e independencia condicional entre variables aleatorias. Además, introduciremos algunas definiciones y proposiciones de la teoría de grafos y los grafos de independencia. Finalmente, presentamos las propiedades de Markov y el concepto de separación en un grafo.

Definición 1.1. Sean X e Y vectores aleatorios con funciones de densidad f_X y f_Y respectivamente. Diremos que X e Y son *independientes*, denotado como $X \perp\!\!\!\perp Y$, si y sólo si la función de densidad conjunta f_{XY} satisface:

$$f_{XY}(x, y) = f_X(x) f_Y(y).$$

La definición de independencia condicional es análoga. Veámosla.

Definición 1.2. Sean X, Y, Z vectores aleatorios. Decimos que X e Y son condicionalmente independientes dado Z, denotado $X \perp\!\!\!\perp Y|Z$, si y sólo si la función de densidad condicional conjunta se puede factorizar en

$$f_{XYZ}(x, y, z) = f_{X|Z}(x; z) f_{Y|Z}(y; z),$$

para todo x, y, y para todo z tal que $f_Z(z) > 0$.

Teorema 1.3. *Criterio de Factorización. Los vectores X e Y son condicionalmente independientes dado Z si y sólo si, existen dos funciones g y h que factoricen la función de densidad conjunta en*

$$f_{XYZ}(x, y, z) = g(x, z)h(y, z),$$

para todo x, y, y para todo z tal que $f_Z(z) > 0$.

Vamos a introducir dos propiedades relacionadas con la independencia condicional que nos serán útiles más adelante.

Proposición 1.4. *Sea un vector aleatorio X particionado en (X_1, X_2, Y, Z) se tiene*

(a) $(X_1, X_2) \perp\!\!\!\perp Y|Z \Rightarrow X_1 \perp\!\!\!\perp Y|Z$ *y también* $X_2 \perp\!\!\!\perp Y|Z$.

(b) $Y \perp\!\!\!\perp (X_1, X_2)|Z$ *sí y sólo si* $Y \perp\!\!\!\perp X_1|(Z, X_2)$ *e* $Y \perp\!\!\!\perp X_2|(Z, X_1)$.

Demostración. (a). Por definición, $(X_1, X_2) \perp\!\!\!\perp Y|Z$ implica que existen funciones g y h que cumplen

$$f_{X_1X_2YZ}(x_1, x_2, y, z) = g(x_1, x_2, z)h(y, z).$$

Como g es función de x_1, x_2 y z, si marginalizamos x_2 la podemos expresar como $g(x_1, z)$

$$f_{X_1YZ}(x_1, y, z) = g(x_1, z)h(y, z),$$

que por definición significa que $X_1 \perp\!\!\!\perp Y|Z$.

(b) $\Rightarrow$. Aplicando 1.3 a $Y \perp\!\!\!\perp (X_1, X_2)|Z$ implica que existen funciones g y h tales que

$$\log f(x_1, x_2, y, z) = g(y, z) + h(x_1, x_2, z).$$

Como en el lado derecho de la igualdad no hay función que contenga como argumentos a y y x_1, de nuevo por 1.3 tenemos $Y \perp\!\!\!\perp X_1|Z$ y repitiendo este argumento para y, x_2 también $Y \perp\!\!\!\perp X_2|Z$.

(b) $\Leftarrow$. $Y \perp\!\!\!\perp X_1|(Z, X_2)$ implica que, por 1.3, existen funciones g y h tales que

$$f(x_1, x_2, y, z) = g(y, z, x_2) * h(x_1, x_2, z).$$

y como $Y \perp\!\!\!\perp X_2|(Z, X_1)$, tenemos que $g(y, x_2, z) = w(y, z)t(x_2, z)$, por tanto,

$$\log f(x_1, x_2, y, z) = w(y, z) + t(x_2, z) + h(x_1, x_2, z)$$

Agrupando t y h como una función de argumentos x_1, x_2, z y aplicando 1.3, obtenemos $Y \perp\!\!\!\perp (X_1, X_2)|Z$. □

Aplicando este lema a la Figura 1.1 podemos ver que $X_1 \perp\!\!\!\perp (Y_1, Y_2)|Z$ y a su vez, $X_1 \perp\!\!\!\perp Y_1|(Z, X_2)$ y $X_1 \perp\!\!\!\perp Y_2|(Z, X_1)$.

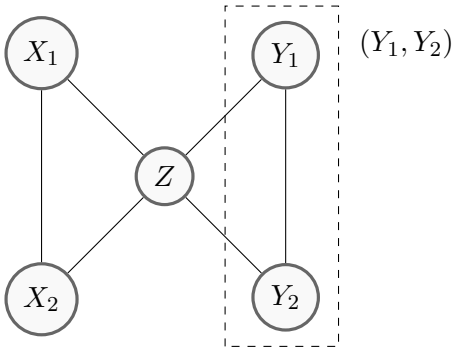


Figura 1.1: Grafo de Independencia de $X = (X_1, X_2, Y_1, Y_2, Z)$.

Ahora vamos a definir los conceptos básicos de la teoría de grafos.

Definición 1.5. Un grafo no dirigido simple es el par $G = (V, E)$ donde

- V el conjunto de los vértices
- E el conjunto de las aristas, siendo cada arista un par (v_1, v_2) de dos vértices v_1 y v_2 diferentes.

Definición 1.6. Camino. Un camino es una secuencia de vértices donde cada par de vértices consecutivos en la secuencia pertenece al conjunto de las aristas E. Es decir que si la secuencia es $\{v_1, v_2, ..., v_n\}$ entonces $(v_i, v_{i+1}) \in E$, para $i \in \{1, \ldots, n-1\}$. Volviendo a la Figura 1.1, un camino entre X_1 y Y_1 puede ser el formado por las aristas (X_1, Z) y (Z, Y_1).

Definición 1.7. Ciclo. Dado un grafo de independencia G un ciclo es un camino cerrado, es decir, un camino en el que el primer y último vértice son el mismo. En la Figura 1.1 un ciclo es el camino $\{X_1, X_2, Z, X_1\}$

Definición 1.8. El grafo G es cordal si cada ciclo de longitud mayor o igual a 4 tiene una arista uniendo dos vértices no consecutivos de dicho ciclo.

Definición 1.9. Separación. Dado un grafo de independencia G, cuyo conjunto de vértices puede particionarse en (A, B, C). Decimos que A y B están separados por C si cualquier camino entre un vértice de A y uno de B contiene al menos un vértice de C.

Definición 1.10. Vecinos. Dado un grafo de independencia G, decimos que los vecinos de un vértice i son los vértices que tienen una arista que los conecta con i.

Definición 1.11. Clique. Dado un grafo G, decimos que un clique es un subgrafo de G completamente conexo, es decir, que entre cada par de vértices del subgrafo, existe una arista conectándolos. En la Figura 1.1 un clique podría ser el subgrafo formado por los vértices X_1, X_2 y Z.

Definición 1.12. Dado un grafo de independencia G, denotamos por $\omega(G)$ como el tamaño del mayor clique de G.

Definición 1.13. Grafo de Independencia. Un grafo no dirigido $G = (V, E)$, donde los vértices $v_1, v_2, ..., v_n$ corresponden a las variables aleatorias $(X_1, X_2, ..., X_n)$ que forman un vector aleatorio X, es un grafo de independencia para X si se cumple que

$$(v_i, v_j) \notin E \Rightarrow X_i \perp\!\!\!\perp X_j | X \backslash \{X_i, X_j\}.$$

Para simplificar la notación podemos reescribir

$$X_i \perp\!\!\!\perp X_j | X \backslash \{X_i, X_j\} \text{ como } i \perp\!\!\!\perp j | V \backslash \{i, j\}.$$

Cuando estemos trabajando con vectores, o particiones de vectores de la forma $X_1, X_2, ..., X_n$. Escribiremos el grafo de independencia, tomando el conjunto de vértices como el conjunto de los subíndices de X, es decir, $\{1, 2, ..., n\}$ en vez de $\{X_1, X_2, ..., X_n\}$.

En la Figura 1.1, si consideramos el grafo como un grafo de independencia para el vector (X_1, X_2, Y_1, Y_2, Z), podemos observar que como no existe la arista (X_1, Y_1) entonces $X_1 \perp\!\!\!\perp Y_1 | \{X_2, Z, Y_2\}$.

Ejemplo 1.1. *Sea* $X = (X_1, X_2, X_3, X_4, X_5, X_6)$ *un vector aleatorio con función de densidad conjunta*

$$f_X(x) = Ce^{x_1+x_1x_4x_5+x_1x_2x_3+x_2x_6},$$

donde C *es una constante de normalización. Factorizando* $f_X = Ce^{x_1}e^{x_1x_4x_5}e^{x_1x_2x_3}e^{x_2x_6}$ *y aplicando 1.3 tenemos las siguientes dependencias condicionales entre dos variables dadas el resto:*

$$\begin{aligned} 4 \perp\!\!\!\perp 2|(1,3,5,6) \qquad & 4 \perp\!\!\!\perp 3|(1,2,5,6) \\ 5 \perp\!\!\!\perp 2|(1,3,4,6) \qquad & 5 \perp\!\!\!\perp 3|(1,2,4,6) \\ 6 \perp\!\!\!\perp i|(1,2,3,4,5)\backslash\{i\} \ \ \forall i \in \{1,3,4,5\}. & \end{aligned}$$

Por tanto, un posible grafo de independencia de X *será la Figura 2.1, que no contiene las aristas* $(2,4)$, $(2,5)$, $(3,4)$, $(3,5)$, $(1,6)$, $(3,6)$, $(4,6)$, $(5,6)$.

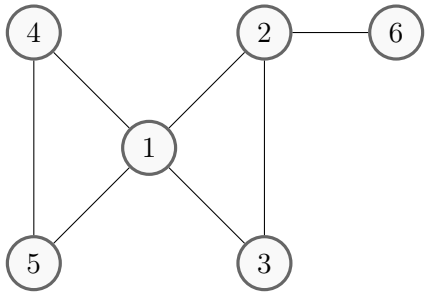


Figura 1.2: Grafo de independencia de X con densidad $f_X = Ce^{x_1}e^{x_1x_4x_5}e^{x_1x_2x_3}e^{x_2x_6}$.

Definición 1.14. Propiedades de Markov. Dado un grafo de independencia $G = (V, E)$.

(G) Markov global. Para subconjuntos de V disjuntos A, B, C donde A y C están separados por B, se cumple que $A \perp\!\!\!\perp C|B$.

(L) Markov local. Para cada vértice i, si llamamos N al conjunto de sus vecinos y R al resto de vértices, se cumple $i \perp\!\!\!\perp R|N$.

(P) Markov por pares. Para cada par de vértices no adjuntos i, j tomando R como el resto de vértices, se cumple $i \perp\!\!\!\perp j|R$.

Ahora introduciremos el concepto de separación, que además utilizaremos en la posterior demostración de la equivalencia de las tres propiedades de Markov. Este concepto es especialmente útil en la predicción de variables dadas otras. Supongamos que tenemos el vector aleatorio X con grafo de independencia G, que contiene un subconjunto de variables que queremos predecir, llamémoslo X_P. A partir de G podemos inferir cierta información sobre dicho subconjunto. Si llamamos,

- X_N: vecinos de X_P en el grafo G

- X_R: al resto de variables de X que no están en X_P o X_N

Entonces, dado X toda la información que necesitamos para predecir X_P está contenida en X_N, y X_R no puede proporcionarnos más.

Lema 1.15. *Sea $G = (V, E)$ un grafo de independencia con $V = \{1, 2, ..., n\}$. Si V se puede separar en dos subconjuntos A y B, tales que no existe un camino entre ellos, entonces:*

$$i \perp\!\!\!\perp j \ \forall i \in A, j \in B.$$

Lema 1.16. *Sea $G = (V, E)$ un grafo de independencia con $V = \{1, 2, ..., n\}$. Si V se puede particionar en dos subconjuntos A y B, tales que no existe un camino entre ellos, entonces para cada $i \in A, j \in B$ tenemos:*

$$i \perp\!\!\!\perp j|S,$$

donde S es cualquier subconjunto de V que no contenga a i o j.

Demostración. Tomamos dos vértices i, j de A y B, respectivamente, y escogemos otro vértice p, que, sin perder en generalidad, suponemos estará en B. Sabemos que, como G es un grafo de independencia para V, por 1.13:

$$i \perp\!\!\!\perp j|V\backslash\{i, j\} \text{ y también } i \perp\!\!\!\perp p|V\backslash\{i, p\}.$$

Aplicando 1.4.b y 1.4.a a $i \perp\!\!\!\perp j|V\backslash\{i, j\}$ obtenemos $i \perp\!\!\!\perp j|V\backslash\{i, j, p\}$. Por tanto, para cada vértice p, queda probado que hemos podido retirar dicho vértice del conjunto de variables que condicionan, ya que pasó de ser $V\backslash\{i, j\}$ a $V\backslash\{i, j, p\}$. Repitiendo para todos los p que forman S hasta llegar a tener $i \perp\!\!\!\perp j|S$. □

Lema 1.17. *Sea $G = (V, E)$ donde i, j son vértices y S un subconjunto de vértices que los separa, entonces*

$$i \perp\!\!\!\perp j|S.$$

Demostración. La demostración de este lema emplea sucesivamente el lema 1.11. Para una demostración detallada, ver el lema 3.3.1 de [1]. □

Teorema 1.18. *Teorema de Separación. Sea el vector aleatorio X, con grafo de independencia G y A, B, C vectores compuestos por conjuntos disjuntos de variables aleatorias que conforman X. Si en G cada vértice de A y C está separado por uno de B entonces*

$$A \perp\!\!\!\perp C|B.$$

Probemos el teorema más relevante de esta sección.

Teorema 1.19. *Las tres propiedades de Markov, global, local y por pares, son equivalentes.*

Demostración. Separando por partes veamos que:

(I) (G) $\Rightarrow$ (L). Podemos tomar como conjunto separador de dos vértices cualesquiera i, j el conjunto de los vecinos de i.

(II) (P) $\Rightarrow$ (G). Particionamos V en (A, B, C) con A y C separados por B. Entonces por 1.18 tenemos que

$$A \perp\!\!\!\perp C|B.$$

(III) (L) $\Rightarrow$ (P). Sea $V = \{1, 2, ..., n\}$. Dado que satisface Markov local, sabemos que para $i \in V$ con conjunto de vecinos N

$$i \perp\!\!\!\perp R|N, \ R \text{ resto de vértices } V\backslash N. \qquad (*)$$

Cogiendo $j \in R$ entonces podemos reescribir (*) como $i \perp\!\!\!\perp (j, R\backslash\{j\})|N$. Usando 1.4.b tenemos

$$i \perp\!\!\!\perp j|(N, R\backslash\{j\}) \quad \text{y} \quad i \perp\!\!\!\perp R|\{N, j\}.$$

Fijándonos en la primera podemos ver que $(N, R\backslash\{j\}) = R\backslash\{j\} = V\backslash\{i, j\}$, por tanto, queda

$$i \perp\!\!\!\perp j|V\backslash\{i, j\},$$

por lo cual se cumple la propiedad de Markov por Pares.

□

En Figura 1.1 aplicando las propiedades de Markov, tendríamos, además de las independencias condicionales por pares dadas en el Ejemplo 1.1, correspondientes a (P), las siguientes:

- Aplicando (G), separando V en los subconjuntos formados por los vértices $(4, 5)$, (1) y $(2, 3, 6)$. Obtenemos $(4, 5) \perp\!\!\!\perp (2, 3, 6)|1$.
- Por (L), y tomando por ejemplo el vértice 2, cuyos vecinos son $(1, 3, 6)$, obtenemos $2 \perp\!\!\!\perp (4, 5)|(1, 3, 6)$.

CAPÍTULO 2
Modelos gráficos gaussianos

Ahora estudiaremos las matrices de covarianzas y su inversa, la matriz de precisión, y qué información podemos extraer de sus elementos.

2.1. Covarianza parcial

Definición 2.1. Varianza. La varianza de un vector aleatorio X, denotada por $\mathrm{Var}(X)$ se define como $\mathrm{E}(X - \mathrm{E}X)^2$.

Definición 2.2. La covarianza entre dos vectores aleatorios X e Y, denotada por $\mathrm{Cov}(X, Y)$ se define como $\mathrm{Cov}(X, Y) = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y)$. Además, podemos expresar la varianza de una variable aleatoria como $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.

Sea (X, Y, Z) un vector aleatorio su matriz de covarianzas, que denotaremos por Σ, será

$$\Sigma = \mathrm{Var}(X, Y, Z) = \begin{bmatrix} \mathrm{Var}(X) & Cov(X, Y) & Cov(X, Z) \\ Cov(Y, X) & \mathrm{Var}(Y) & Cov(Y, Z) \\ Cov(Z, X) & Cov(Z, Y) & \mathrm{Var}(Z) \end{bmatrix}.$$

Definición 2.3. Covarianza parcial. Dadas las variables aleatorias X, Y, Z definimos la covarianza parcial entre Y y Z dada X como

$$Cov(Y, Z|X) = Cov(Y, Z) - Cov(Y, X)\mathrm{Var}(X)^{-1}Cov(X, Z).$$

Definición 2.4. Varianza Parcial. Dadas las variables aleatorias X, Y la varianza parcial de Y dado X es

$$\mathrm{Var}(Y|X) = \mathrm{Var}(Y) - Cov(Y, X)\mathrm{Var}(X)^{-1}Cov(X, Y).$$

Definición 2.5. Correlación parcial. Dadas las variables aleatorias X, Y la correlación parcial de Y y Z dado X es

$$Corr(Y, Z|X) = \frac{Cov(Y, Z|X)}{\sqrt{Var(Y|X)Var(Z|X)}}.$$

Proposición 2.6. *Identidad matricial de Woodbury. Dadas las matrices A, B, C, D de tamaños $n \times n, n \times k, k \times k$ y $k \times n$ respectivamente, se cumple que $(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$.*

Proposición 2.7. *Dada una matriz particionada en bloques de la siguiente forma*

$$\begin{bmatrix} P & Q \\ R & S \end{bmatrix},$$

donde P, S son matrices con tamaños $n \times n$ y $k \times k$, mientras que Q y R, tienen tamaño $k \times n$ y $n \times k$ respectivamente, entonces su matriz inversa es

$$\begin{bmatrix} (P - QS^{-1}R)^{-1} & -(P - QS^{-1}R)^{-1}QS^{-1} \\ -(S - RP^{-1}Q)^{-1}RP^{-1} & (S - RP^{-1}Q)^{-1} \end{bmatrix}.$$

Demostración. Sea la inversa de $\begin{bmatrix} P & Q \\ R & S \end{bmatrix}$ la matriz $\begin{bmatrix} X & Y \\ Z & W \end{bmatrix}$ es decir que

$$\begin{bmatrix} P & Q \\ R & S \end{bmatrix}\begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ 0 & I_k \end{bmatrix} \text{ por tanto } \begin{bmatrix} PX + QZ & PY + QW \\ RX + SZ & RY + SW \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ 0 & I_k \end{bmatrix}.$$

De las igualdades de los bloques no diagonales obtenemos que $Y = -P^{-1}QW$ y $Z = -S^{-1}RX$, sustituyendo en las igualdades de la diagonal, tenemos $PX + QZ = PX - QS^{-1}RX = I_n$, despejando $X = (P - QS^{-1}R)^{-1}$, y de forma análoga $W = (S - RP^{-1}Q)^{-1}$. Volviendo a las expresiones iniciales de Y y Z y sustituyendo los valores de X y W que acabamos de hallar, obtenemos $Z = -S^{-1}R(P - QS^{-1}R)^{-1}$, $Y = -P^{-1}Q(S - RP^{-1}Q)^{-1}$, por tanto hemos obtenido la matriz inversa

$$\begin{bmatrix} (P - QS^{-1}R)^{-1} & -P^{-1}Q(S - RP^{-1}Q)^{-1} \\ -S^{-1}R(P - QS^{-1}R)^{-1} & (S - RP^{-1}Q)^{-1} \end{bmatrix},$$

que aplicando 2.6 es equivalente a $\begin{bmatrix} (P - QS^{-1}R)^{-1} & -(P - QS^{-1}R)^{-1}QS^{-1} \\ (S - RP^{-1}Q)^{-1}RP^{-1} & (S - RP^{-1}Q)^{-1} \end{bmatrix}$.

□

Ahora aplicamos este resultado a la matriz de covarianzas de un vector aleatorio particionado en 2 bloques.

Proposición 2.8. *Sea X un vector aleatorio particionado en $X = (X_a, X_b)$ entonces su matriz de precisión, denotada como D, es*

$$D = \begin{bmatrix} \mathrm{Var}(X_a|X_b)^{-1} & -\mathrm{Var}(X_a|X_b)^{-1}Cov(X_a, X_b)\mathrm{Var}(X_b)^{-1} \\ -\mathrm{Var}(X_b|X_a)^{-1}Cov(X_b, X_a)\mathrm{Var}(X_a)^{-1} & \mathrm{Var}(X_b|X_a)^{-1} \end{bmatrix}.$$

Demostración. Usando 2.7 sobre la matriz $\Sigma = \begin{bmatrix} \mathrm{Var}(X_a) & \mathrm{Cov}(X_a, X_b) \\ \mathrm{Cov}(X_b, X_a) & \mathrm{Var}(X_b) \end{bmatrix}$ obtenemos que su inversa es, usando la notación compacta $\Sigma_{aa} = \mathrm{Var}(X_a)$ y $\Sigma_{ab} = \mathrm{Cov}(X_a, X_b)$,

$$D = \Sigma^{-1} = \begin{bmatrix} (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} & -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab} \\ (\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})^{-1}\Sigma_{ba}\Sigma_{aa}^{-1} & (\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})^{-1} \end{bmatrix}$$

Usando 2.4 podemos simplificar la matriz a

$$D = \Sigma^{-1} = \begin{bmatrix} \mathrm{Var}(X_a|X_b)^{-1} & -\mathrm{Var}(X_a|X_b)^{-1}Cov(X_a,X_b)\mathrm{Var}(X_b)^{-1} \\ -\mathrm{Var}(X_b|X_a)^{-1}Cov(X_b,X_a)\mathrm{Var}(X_a)^{-1} & \mathrm{Var}(X_b|X_a)^{-1} \end{bmatrix}.$$

□

Los siguientes dos resultados son los más importantes del capítulo, y nos relacionarán los elementos de la matriz de precisión de un vector aleatorio con su grafo de independencia.

Corolario 2.8.1. *Los elementos de la diagonal de la varianza inversa son los recíprocos de la varianza parcial, es decir, sea* $X = (X_1, ..., X_n)$

$$d_{ii} = \mathrm{Var}(X_i|X_2, ..., X_{i-1}, X_{i+1}, ..., X_n)^{-1} para i \in \{1, \dots, n\}.$$

Demostración. Partiendo de $X = (X_a, X_b)$, suponiendo que $|X_b| = 1$, tenemos que, por 2.8

$$d_{bb} = \mathrm{Var}(X_b|X_a)^{-1},$$

para cualquier X_i de tamaño 1, permutamos X de forma que quede $X = (X_i, X_R)$ siendo $R = \{1, 2, ..., i-1, i+1, ..., n\}$ y aplicamos el mismo argumento. □

Corolario 2.8.2. *Los elementos de la matriz de precisión cumplen*

$$-\frac{d_{ij}}{\sqrt{d_{ii}d_{jj}}} = Corr(X_i, X_j|X\backslash\{X_i, X_j\}), \quad para\ i \neq j$$

Demostración. Partiendo de $X = (X_a, X_b)$, donde $X_b = (X_i, X_j)$

$$D_{bb} = \begin{bmatrix} d_{ii} & d_{ij} \\ d_{ji} & d_{jj} \end{bmatrix} = \mathrm{Var}(X_b|X_a)^{-1}$$

Ahora por 2.4

$$\mathrm{Var}(X_b|X_a) = \begin{bmatrix} \mathrm{Var}(X_i|X_a) & Cov(X_i, X_j|X_a) \\ Cov(X_j, X_i|X_a) & \mathrm{Var}(X_j|X_a) \end{bmatrix} = \frac{1}{d_{ii}d_{djj} - d_{ij}d_{ji}} \begin{bmatrix} d_{jj} & -d_{ij} \\ -d_{ji} & d_{ii} \end{bmatrix}$$

Dividiendo la fila y columna i-ésima por $\mathrm{Var}(X_i|X_a)^{\frac{1}{2}} = \left(\frac{d_{jj}}{d_{ii}d_{djj} - d_{ij}d_{ji}}\right)^{\frac{1}{2}}$, y de forma análoga las filas y columna j-ésima por $\mathrm{Var}(X_j|X_a)^{\frac{1}{2}} = \left(\frac{d_{ii}}{d_{ii}d_{djj} - d_{ij}d_{ji}}\right)^{\frac{1}{2}}$, obtenemos

$$\begin{bmatrix} 1 & \frac{Cov(X_i,X_j|X_a)}{\mathrm{Var}(X_i|X_a)^{\frac{1}{2}}\mathrm{Var}(X_j|X_a)^{\frac{1}{2}}} \\ \frac{Cov(X_j,X_i|X_a)}{\mathrm{Var}(X_i|X_a)^{\frac{1}{2}}\mathrm{Var}(X_j|X_a)^{\frac{1}{2}}} & 1 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{d_{ij}}{\sqrt{d_{ii}d_{jj}}} \\ -\frac{d_{ji}}{\sqrt{d_{ii}d_{jj}}} & 1 \end{bmatrix},$$

y usando 2.5 se cumple $-\frac{d_{ij}}{\sqrt{d_{ii}d_{jj}}} = Corr(X_i, X_j|X\backslash\{X_i, X_j\})$, para $i \neq j$. □

Corolario 2.8.3. *Podemos aprovechar este corolario y su demostración para afirmar que, los elementos de la matriz de precisión cumplen*

$$-d_{ij} = \mathrm{Cov}(X_i, X_j|X\backslash\{X_i, X_j\})\ para\ i \neq j.$$

2.2. Normal multivariante

En esta sección introduciremos la distribución normal multivariante, algunas de sus propiedades, veremos como se comporta bajo marginalización y la distribución condicionada entre sus variables marginales.

Definición 2.9. Sea X un vector n-dimensional diremos que sigue una distribución normal multivariante $N(\mu, \Sigma)$ si y sólo si su función de densidad es:

$$f_X(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)},$$

donde μ es el vector de medias de X y Σ es la matriz de varianzas (y covarianzas) de X. Usando la matriz de precisión podemos reescribir la función de densidad como:

$$f_X(x) = |D|^{1/2}(2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}(x-\mu)^t D(x-\mu)}.$$

Ahora veamos algunas proposiciones interesantes relacionadas con dicha distribución

Proposición 2.10. *Cualquier normal multivariante $X \sim N(\mu, \Sigma)$ se puede expresar como una transformación lineal de una normal multivariante estándar $Z \sim N(0, I)$, de la forma:*

$$X = AZ + \mu \text{ para cierta matriz } A.$$

Proposición 2.11. *Dado $X \sim N(\mu, \Sigma)$, A una matriz de tamaño $k \times n$ y b un vector (k-dimensional) entonces la transformación lineal $AX + b$ cumple:*

$$AX + b \sim N(A\mu, A\Sigma A^t).$$

Como consecuencia de estas dos proposiciones tenemos que para cualquier $X \sim N(\mu, \Sigma)$

$$E(X) = E(AZ + \mu) = E(AZ) + E(\mu) = AE(Z) + \mu = \mu.$$
$$Var(X) = Var(AZ + \mu) = Var(AZ) = AVar(Z)A^t = AA^t.$$

AA^t por construcción de $X = AZ + \mu$ usando 2.11 es igual a Σ.

2.3. Distribución marginal y condicionada

Observemos ahora las expresiones de media y varianza para un vector X particionado en (X_a, X_b).

Proposición 2.12. *Dado el vector aleatorio particionado como $X = (X_a, X_b)$, como consecuencia de 2.11, su vector de medias y matriz de covarianzas serán*

$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \ \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$$

Donde $X_a \sim N(\mu_a, \Sigma_{aa})$ y $X_b \sim N(\mu_b, \Sigma_{bb})$ y también $\Sigma_{ab} = \Sigma_{ba} = Var(X_a, X_b)$.

Ahora la distribución de X_b dado $X_a = x_a$ denotado como $X_b|X_a = x_a$, es una variable aleatoria de igual dimensión que X_b y sigue una distribución normal multivariante con media:

$$\mu_{b|a} = \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a),$$

y con varianza:

$$\Sigma_{b|a} = \mathrm{Var}(X_b|X_a) = \Sigma_{bb} + \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}.$$

Lo siguiente es ver algunas condiciones de independencia e independencia condicional y como se relacionan con los ceros de la matriz de precisión.

Corolario 2.12.1. *Los vectores normales marginales X_a y X_b son independientes si y sólo si*

- $Cov(X_a, X_b) = \Sigma_{ab} = 0$, *o equivalentemente,*
- $D_{ab} = 0$ *con D la matriz de precisión.*

Corolario 2.12.2. *Los vectores normales marginales X_a, X_b son condicionalmente independientes dado X_c si y solo si, se cumple:*

- $Cov(X_a, X_b|X_c) = \Sigma_{ab|c} = 0$, *o equivalentemente,*
- *El bloque de la matriz de varianza inversa $D_{ab} = 0$.*

Con lo visto hasta ahora podemos empezar a construir modelos gráficos gaussianos, que son grafos de independencia en los que la muestra de la variable aleatoria dada sigue una distribución normal multivariante.

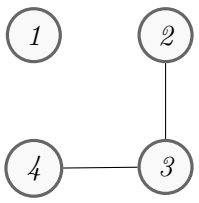
Dado un grafo de independencia G y una variable aleatoria X n-dimensional, su modelo gráfico Gaussiano será una familia de distribuciones normales en las que se cumplan las condiciones de independencia condicionada que nos impone G.

Definición 2.13. Dado un grafo de independencia $G = (V, E)$ con vértices $V = \{1, 2, \ldots, n\}$, decimos que el vector $X \in \mathbb{R}^n$ **satisface** el grafo de independencia si $X \sim N_n(\mu, \Sigma)$ con

$$\Sigma_{i,j}^{-1} = D_{i,j} = 0 \qquad \forall (i,j) \notin E.$$

Veamos un ejemplo.

Ejemplo 2.1. *Dado el siguiente grafo de independencia de un vector aleatorio $X = (X_1, X_2, X_3, X_4)$.*



Un vector aleatorio que satisfaga el grafo deberá tener una matriz de precisión D de la siguiente forma

$$D = \begin{bmatrix} d_{11} & 0 & 0 & 0 \\ 0 & d_{22} & d_{23} & 0 \\ 0 & d_{32} & d_{33} & d_{34} \\ 0 & 0 & d_{43} & d_{44} \end{bmatrix}.$$

Ahora será objeto de nuestro interés el estudio de la estimación de los pesos de las aristas, dada la estructura del grafo.

2.4. Información mutua

Definición 2.14. Entropía. Dada una variable aleatoria X con función de densidad f_X, su entropía, denotada por $H(X)$ se define como $H(X) = -E(\log f_X(x))$.

Definición 2.15. Entropía Conjunta. Dadas X, Y dos variables aleatorias con función de densidad conjunta f_{XY}, definimos su entropía conjunta como $H(X,Y) = -E(\log f_{XY}(x,y))$

Ahora estudiemos la información mutua entre dos variables que particionan un vector normal multivariante.

Proposición 2.16. *La entropía de una variable $X \sim N_n(\mu, \Sigma)$ es*

$$H(X) = \frac{n}{2}(1 + \log 2\pi) + \frac{1}{2}\log|\Sigma|$$

Demostración. Aplicando el logaritmo a la función de densidad de una normal multivariante tenemos que

$$\log f_X(x) = \frac{1}{2}\log|\Sigma^{-1}| - \frac{n}{2}\log 2\pi - \frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu).$$

Tomando la menos esperanza tenemos

$$H(X) = -E\log f_X(x) = -\frac{1}{2}\log|\Sigma^{-1}| + \frac{n}{2}\log 2\pi + \frac{1}{2}E((x-\mu)^t\Sigma^{-1}(x-\mu)),$$

como $(X-\mu)$ tiene media 0, usando que para X con $EX = 0$ y $Var(X) = \Sigma$, se tiene, $E(X^tAX) = \mathrm{tr}(A\Sigma)$. En nuestro caso tendremos que $E((x-\mu)^t\Sigma^{-1}(x-\mu)) = tr(\Sigma^{-1}\Sigma) = tr(I_n) = n$. Por tanto la entopría es

$$\begin{aligned} H(X) &= -\frac{1}{2}\log|\Sigma^{-1}| + \frac{n}{2}\log 2\pi + \frac{n}{2} = -\frac{1}{2}\log|\Sigma^{-1}| + \frac{n}{2}(1 + \log 2\pi) \\ &= \frac{1}{2}\log|\Sigma| + \frac{n}{2}(1 + \log 2\pi). \end{aligned}$$

□

Definición 2.17. Información mutua. Sea un vector aleatorio X particionado en (X_a, X_b) definimos la información mutua entre X_a y X_b como $\text{Inf}(X_a, X_b) = H(X_a) + H(X_b) - H(X_a, X_b)$.

Proposición 2.18. *Sea un vector aleatorio $X \sim N(\mu, \Sigma)$ particionado en (X_a, X_b) la información mutua entre X_a y X_b.*

$$\text{Inf}(X_a, X_b) = -\frac{1}{2} log \frac{|\Sigma|}{|\Sigma aa|\,|\Sigma bb|} = -\frac{1}{2} log \frac{|\Sigma_{bb|a}|}{|\Sigma bb|}.$$

Demostración. Supongamos X_a p-dimensional y X_b q-dimensional, usando 2.14 y 2.15:

$$\text{Inf}(X_a, X_b) = H(X_a) + H(X_b) - H(X_a, X_b) = E(\log f_{ab}(x)) - E(\log f_a(x)) - E(\log f_b(x)).$$

Usando 2.16, tenemos

$$\begin{aligned}\text{Inf}(X_a, X_b) &= \frac{1}{2}\left(-(p+q)(\log 2\pi + 1) - \log|\Sigma| + p(\log 2\pi + 1) + \log|\Sigma_{aa}| + q(\log 2\pi + 1) + \log|\Sigma_{bb}|\right) \\ &= \frac{1}{2}(-\log|\Sigma| + \log|\Sigma_{aa}| + \log|\Sigma_{bb}|) = -\frac{1}{2}\log\frac{|\Sigma|}{|\Sigma_{aa}||\Sigma_{bb}|}.\end{aligned}$$

Dado que el complemento de Schur $\Sigma \backslash \Sigma_{aa}$ de la matriz $\Sigma = \begin{bmatrix}\Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}$ cumple que $|\Sigma| = |\Sigma_{aa}||\Sigma \backslash \Sigma_{aa}|$, despejando $|\Sigma \backslash \Sigma_{aa}|$ podemos ver que se cumple $|\Sigma \backslash \Sigma_{aa}| = |\Sigma_{bb|a}|$, por tanto, aplicando $|\Sigma| = |\Sigma_{aa}||\Sigma_{bb|a}|$ a la expresión de la información anterior obtenemos

$$\text{Inf}(X_a, X_b) = -\frac{1}{2}\log\frac{|\Sigma|}{|\Sigma_{aa}||\Sigma_{bb}|} = -\frac{1}{2}\log\frac{|\Sigma_{aa}||\Sigma_{bb|a}|}{|\Sigma_{aa}||\Sigma_{bb}|} = -\frac{1}{2}\log\frac{|\Sigma_{bb|a}|}{|\Sigma_{bb}|}.$$

□

Definición 2.19. Información Mutua Condicional. Dadas las variables aleatorias X, Y, Z, la información mutua condicional entre X e Y dada Z se define como $\text{Inf}(X, Y|Z) = H(X|Z) + H(Y|Z) - H(X, Y|Z)$.

Proposición 2.20. *Sea un vector aleatorio $X \sim N(\mu, \Sigma)$ particionado en (X_a, X_b, X_c), la información mutua condicional entre X_b y X_c dada X_a es*

$$\text{Inf}(X_b, X_c|X_a) = -\frac{1}{2} log \frac{|\Sigma||\Sigma_{aa}|}{|\Sigma_{ab}||\Sigma_{ac}|} = -\frac{1}{2} log \frac{|\Sigma_{bc|a}|}{|\Sigma_{bb|a}||\Sigma_{cc|a}|} = -\frac{1}{2} log \frac{|\Sigma_{cc|ab}|}{|\Sigma_{cc|a}|}$$

Demostración. Para una demostración detallada ver la proposición 6.4.6 de [1]. □

Veamos un ejemplo de la aplicación de estas medidas de información a la hora de crear un grafo de independencia con pesos en las aristas.

Ejemplo 2.2. *Sea X una variable aleatoria con matriz de varianzas* $\Sigma = \begin{bmatrix} 1 & 0{,}5 & 0{,}4 \\ 0{,}5 & 1 & 0{,}2 \\ 0{,}4 & 0{,}2 & 1 \end{bmatrix}$.

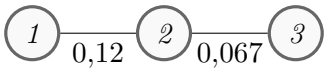
Empleando 2.8.2 obtenemos

$$\begin{bmatrix} 1 & -0{,}47 & -0{,}35 \\ -0{,}47 & 1 & 0{,}00 \\ -0{,}35 & 0{,}00 & 1 \end{bmatrix}.$$

Ahora calculando la información contra la independencia condicional obtenemos

- $\mathrm{Inf}(X_1, X_2|X_3) = \frac{1}{2} log(1 - 0{,}47^2) = 0{,}12.$
- $\mathrm{Inf}(X_1, X_3|X_2) = 0{,}067.$
- $\mathrm{Inf}(X_2, X_3|X_1) = 0.$

Por tanto el grafo de independencia es



CAPÍTULO 3
Estimación de máxima verosimilitud

3.1. Verosimilitud

El objetivo de este capítulo es introducir la estimación de máxima verosimilitud para la matriz de covarianzas Σ de un modelo gráfico gaussiano, partiendo de que conocemos el grafo de independencias de dicha variable. Para ello introduciremos las definiciones de verosimilitud, y su maximización, además de plantear este problema como uno de optimización convexa, para proporcionar métodos iterativos que lo resuelvan.

Definición 3.1. Función de Verosimilitud. Dada una muestra de N variables independientes con distribución normal multivariante con media μ y varianza Σ, es decir

$$x^{(i)} \sim N_n(\mu, \Sigma) \text{ para } i = 1, \ldots, n.$$

La log verosimilitud $l(\mu, \Sigma) = l(\mu, \Sigma; x^{(1)}, ..., x^{(N)}) = \sum_{i=1}^{N} \log f(x^{(i)})$ tenemos

$$2l(\mu, \Sigma) = -N \log |\Sigma| - Nn \log 2\pi - \sum_{i=1}^{n} (x^{(i)} - \mu)^t \Sigma^{-1} (x^{(i)} - \mu).$$

Denotemos la media muestral por $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$ y la varianza muestral por S donde $s_{ij} = \frac{1}{N} \sum_{k=1}^{N} (x_i^{(k)} - \bar{x_i})^t \Sigma^{-1} (x_j^{(k)} - \bar{x_j})$.

Proposición 3.2. *Para una variable $X \sim N_n(\mu, \Sigma)$, su logverosimilitud puede expresarse como*

$$l(\mu, \Sigma) \propto -\frac{N}{2} \text{tr}(\Sigma^{-1} S) - \frac{N}{2} \log |\Sigma| - \frac{N}{2} (\bar{x} - \mu)^t \Sigma^{-1} (\bar{x} - \mu).$$

Demostración. Por 3.1 tenemos

$$l(\mu, \Sigma) = -\frac{N}{2} \log |\Sigma| - \frac{N}{2} n \log (2\pi) - \frac{N}{2} \sum_{i=1}^{n} (x^{(i)} - \mu)^t \Sigma^{-1} (x^{(i)} - \mu),$$

separando el término $\sum_{i=1}^{N}(x^{(i)}-\mu)^t\Sigma^{-1}(x^{(i)}-\mu) = \sum_{i=1}^{N}(x^{(i)}-\bar{x})^t\Sigma^{-1}(x^{(i)}-\bar{x}) + N(\bar{x}-\mu)^t\Sigma^{-1}(\bar{x}-\mu)+2\sum_{i=1}^{N}(x^{(i)}-\bar{x})\Sigma^{-1}(\bar{x}-\mu)$, y usando que este último sumando es igual a 0 queda reducido a $\sum_{i=1}^{N}(x^{(i)}-\bar{x})^t\Sigma^{-1}(x^{(i)}-\bar{x}) + N(\bar{x}-\mu)^t\Sigma^{-1}(\bar{x}-\mu)$. Centrándonos en el término de la izquierda y usando las propiedades de la traza

$$\sum_{i=1}^{N}(x^{(i)}-\bar{x})^t\Sigma^{-1}(x^{(i)}-\bar{x}) = \sum_{i=1}^{N}\mathrm{tr}((x^{(i)}-\bar{x})^t\Sigma^{-1}(x^{(i)}-\bar{x}))$$
$$= tr\underbrace{\left(\sum_{i=1}^{N}(x^{(i)}-\bar{x})^t(x^{(i)}-\bar{x})\Sigma^{-1}\right)}_{NS} = \mathrm{tr}(NS\Sigma^{-1}) = N\mathrm{tr}(S\Sigma^{-1}).$$

Por tanto concluimos que

$$l(\mu,\Sigma) = -\frac{N}{2}n\log(2\pi) - \frac{N}{2}\mathrm{tr}(\Sigma^{-1}S) - \frac{N}{2}\log|\Sigma| - \frac{N}{2}(\bar{x}-\mu)^t\Sigma^{-1}(\bar{x}-\mu),$$

y como el primer término es constante podemos escribir este resultado como

$$l(\mu,\Sigma) \propto -\frac{N}{2}tr(\Sigma^{-1}S) - \frac{N}{2}\log|\Sigma| - \frac{N}{2}(\bar{x}-\mu)^t\Sigma^{-1}(\bar{x}-\mu).$$

□

Proposición 3.3. *Para una variable $X \sim N_n(\mu,\Sigma)$ podemos escribir la logverosimilitud en función de la media μ y matriz de precisión D de la siguiente forma*

$$l(\mu,D) \propto -\frac{N}{2}tr(DS) + \frac{N}{2}\log|D| - \frac{N}{2}(\bar{x}-\mu)^tD(\bar{x}-\mu)$$

3.2. Estimaciones de máxima verosimilitud

Ahora que ya sabemos expresar la función de verosimilitud de una variable normal multivariante el siguiente paso es hallar su máximo valor.

Proposición 3.4. *Cuando la matriz V no tiene restricciones, provenientes de un grafo de independencia, entonces el estimador de máxima verosimilitud de la media μ y varianza V de una muestra de normales multivariantes independientes e idénticamente distribuidas son la media muestral $\bar{x}$ y varianza muestral $\bar{S}$ respectivamente.*

Demostración. Para una demostración detallada de esta proposición ver [9]. □

En lo que resta de nuestro estudio, como los modelos gráficos gaussianos solo imponen restricciones a la matriz de precisión, μ no tendrá restricciones, por lo que podemos asumir que será igual a la media muestral.

Comencemos con un ejemplo en el que obtengamos la estimación de máxima verosimilitud para una muestra dada.

Ejemplo 3.1. *Sea $X = (X_1, X_2, X_3)$ una normal multivariante con matriz de precisión,*

$$D = \begin{bmatrix} d_{11} & d_{22} & d_{33} \\ d_{21} & d_{22} & 0 \\ d_{31} & 0 & d_{32} \end{bmatrix},$$

es decir con la independencia condicional $X_2 \perp\!\!\!\perp X_3 | X_1$. Usando 3.3 y 3.4 el estimador de máxima verosimilitud para la media μ será la media muestral $\bar{x}$, obtenemos

$$l(D) \propto -\underbrace{(s_{11}d_{11} + s_{22}d_{22} + s_{33}d_{33} - 2s_{12}d_{12} - 2s_{13}d_{13})}_{\mathrm{tr}(DS)} + \log \underbrace{(d_{11}d_{22}d_{33} + d_{12}^2 d_{33} + d_{13}^2 d_{22})}_{|D|}$$

Derivando en función de $d_{i,j}$ obtenemos las siguientes ecuaciones

$$\begin{aligned} 0 =& - s_{11} + \frac{d_{22}d_{33}}{|D|} \\ 0 =& - s_{22} + \frac{d_{11}d_{33} - d_{13}^2}{|D|} \\ 0 =& - s_{33} + \frac{d_{11}d_{22} - d_{12}^2}{|D|} \\ 0 =& - s_{12} - 2\frac{d_{12}d_{33}}{|D|} \\ 0 =& - 2s_{13} - 2\frac{d_{12}d_{22}}{|D|}. \end{aligned}$$

Despejando las $s_{i,j}$ nos da la siguiente igualdad entre S y D,

$$\begin{bmatrix} s_{11} & * & * \\ s_{21} & s_{22} & * \\ s_{31} & ? & s_{33} \end{bmatrix} = \begin{bmatrix} d_{22}d_{33} & * & * \\ -d_{12}d_{33} & d_{11}d_{33} - d_{13}^2 & * \\ -d_{13}d_{22} & ? & d_{11}d_{22} - d_{12}^2 \end{bmatrix} \frac{1}{|D|},$$

donde las ? son desconocidas ya que corresponden a los ceros de D, hallemos una expresión para ellas. Usando 2.8.3

$$\begin{aligned} 0 = d_{23} = Cov(X_2, X_3|X_1) &= Cov(X_2, X_3) - Cov(X_2, X_1)\mathrm{Var}(X_1)^{-1}Cov(X_1, X_3) \\ \Leftrightarrow Cov(X_2, X_3) &= Cov(X_2, X_1)\mathrm{Var}(X_1)^{-1}Cov(X_1, X_3). \end{aligned}$$

Dado que esta es una restricción del modelo tenemos que $s_{23} = \frac{s_{12}s_{13}}{s_{11}}$, y en términos de D se puede escribir el término ? como $\frac{d_{12}d_{33}d_{13}d_{33}}{d_{22}d_{33}} = d_{12}d_{13}$. Por tanto la igualdad entre matrices es

$$\begin{bmatrix} s_{11} & * & * \\ s_{21} & s_{22} & * \\ s_{31} & \frac{s_{12}s_{13}}{s_{11}} & s_{33} \end{bmatrix} = \begin{bmatrix} d_{22}d_{33} & * & * \\ -d_{12}d_{33} & d_{11}d_{33} - d_{13}^2 & * \\ -d_{13}d_{22} & d_{12}d_{13} & d_{11}d_{22} - d_{12}^2 \end{bmatrix} \frac{1}{|D|}.$$

La matriz de la derecha es la matriz de precisión invertida, es decir, tenemos que nuestra matriz de varianza estimada es la de la izquierda en la igualdad anterior.

En este ejemplo todos los subvectores correspondientes a cliques en el grafo, (X_1, X_3) y (X_1, X_2) cumplen que su varianza estimada es igual a la varianza muestral.

Teorema 3.5. *El estimador de máxima verosimilitud de un modelo gráfico con grafo de independencia G basado en una distribución normal multivariante, cumple*

$$D_{ij} = 0,$$

cuando i, j no son adyancentes en G, y

$$\Sigma_{AA} = S_{AA},$$

cuando el subconjunto de vértices A de G forme un clique en el grafo.

Demostración. La demostración de este teorema puede encontrarse en 1. □

3.3. Optimización convexa

Otra forma de enfocar el problema de la estimación de máxima verosimilitud es como un problema de optimización convexa en función de la matriz de covarianza, donde el parametro μ se asume igual a la media muestral $\bar{x}$.

$$\begin{aligned} \underset{\Sigma}{\text{maximizar}} \;\; & l(\Sigma) = -\log|\Sigma| - \operatorname{tr}(S\Sigma^{-1}) \\ \text{sujeto a} \;\; & \Sigma \in \Theta, \end{aligned}$$

con Θ las matrices de varianza de los vectores aleatorios que satisfacen el grafo de independencia.

Proposición 3.6. *La función objetivo $l(\Sigma) = -\log|\Sigma| - \operatorname{tr}(S\Sigma^{-1})$ es convexa en la región $\{\Sigma \in \mathbb{S}^n_{\succ 0} \mid 2S - \Sigma \in \mathbb{S}^n_{\succ 0}\}$ del cono $\mathbb{S}^n_{\succ 0}$ (ver A.9).*

Demostración. Para una demostración detallada ver Ejercicio 7.4 de 4. □

Dado que los modelos gráficos gaussianos imponen restricciones en la matriz de precisión tiene sentido expresar el problema de optimización usando dicha matriz, sustituimos Σ por su inversa D y obtenemos:

$$\begin{aligned} \underset{D}{\text{maximizar}} \;\; & l(D) = \log|D| - \operatorname{tr}(SD) \\ \text{sujeto a} \;\; & D \in \mathcal{D} \end{aligned}$$

Donde el espacio muestral $\mathcal{D} = \{D \in \mathbb{S}^n_{\succ 0} | D_{i,j} = 0 \text{ para todo } (i,j) \notin E \text{ con } i \neq j\}$.

Proposición 3.7. *La función objetivo $l(D) = \log|D| - \operatorname{tr}(SD)$ es cóncava en $D \in \mathbb{S}^n_{\succ 0}$*

Demostración. $\operatorname{tr}(SD)$ es lineal en D, por tanto es cóncava.

Para probar que $\log|D|$ es cóncavo, demostramos que es cóncavo sobre cada recta en $\mathbb{S}^n_{\succ 0}$. Consideramos una recta $D = A + tB$ donde $A, B \in \mathbb{S}^n_{\succ 0}$. Tomemos $g(t) = \log|A+tB|$, donde t es tal que $A+tB \in \mathbb{S}^n_{\succ 0}$ y probemos que esta función es cóncava:

$$\begin{aligned} g(t) &= \log|A+tB| = \log|A^{\frac{1}{2}}(I + tA^{-\frac{1}{2}}BA^{-\frac{1}{2}})A^{\frac{1}{2}}| \\ &= \log|A^{\frac{1}{2}}| + \log|I + tA^{-\frac{1}{2}}BA^{-\frac{1}{2}}| + \log|A^{\frac{1}{2}}| \\ &= \log|A| + \log|I + tA^{-\frac{1}{2}}BA^{-\frac{1}{2}}| \\ &= \log|A| + \log\left((t\lambda_1 + 1)\cdot\cdots\cdot(t\lambda_p + 1)\right)) \qquad \lambda_i \text{ autovalor de } A^{-\frac{1}{2}}BA^{-\frac{1}{2}} \\ &= \log|A| + \sum_{i=1}^{p}\log(1+t\lambda_i). \end{aligned}$$

Como el logaritmo $\log(1+t\lambda_i)$ es cóncavo en t, y $\log|A|$ es una constante, hemos expresado $g(t)$ como suma de funciones cóncavas, concluimos que $\log|D|$ es cóncavo para $D \in \mathbb{S}^n_{\succ 0}$. □

Dado que $\mathcal{D}$ es un cono convexo y la función objetivo cóncava, la maximización de la verosimilitud es un problema de optimización convexa. El Lagrangiano de este problema de optimización es

$$\begin{aligned} L(D, v) &= \log|D| - \operatorname{tr}(SD) - 2\sum_{(i,j)\notin E} v_{ij}d_{ij} \\ &= \log|D| - \sum_{i=1}^{n} s_{ii}d_{ii} - 2\sum_{(i,j)\in E} s_{ij}d_{ij} - 2\sum_{(i,j)\notin E, i\neq j} v_{ij}d_{ij}, \end{aligned} \tag{3.*}$$

donde v_{ij} son los multiplicadores de Lagrange.

El problema dual de Lagrange se obtiene hallando el supremo de 3.*, es decir

$$g(v) = \sup_{D\in\mathcal{D}} L(D, v). \tag{3.1}$$

El máximo del lagrangiano en función de D, escrito $\hat{D}$, se obtiene

$$\begin{aligned} 0 = \frac{\partial L(D,v)}{\partial D} &= D^{-1} - \sum_{i=0}^{n} s_{ii} - 2\sum_{(i,j)\in E} s_{ij} - 2\sum_{(i,j)\notin E} v_{ij} \\ &= D^{-1} - \sum_{i=0}^{n} s_{ii} - \sum_{(i,j)\in E} s_{ij} - \sum_{(i,j)\in E} s_{ji} - \sum_{(i,j)\notin E} v_{ij}. \end{aligned}$$

Es decir que

$$(\hat{D}^{-1})_{ij} = \sigma_{ij} = \begin{cases} s_{ij} \text{ cuando } i = j \text{ ó } (i,j) \in E \\ v_{ij} \text{ cuando } (i,j) \notin E \text{ y } i \neq j. \end{cases}$$

Por tanto la función dual de Lagrange es el resultado de sustituir la matriz $\hat{D}$ hallada en el lagrangiano 3.*, obtenemos que el problema de optimización dual correspondiente ess

$$\begin{aligned} &\text{minimizar} \quad -\log|\Sigma| - n \\ &\quad \text{sujeto a} \;\; \sigma_{ij} = s_{ij} \; \forall (i,j) \in E \text{ ó } i = j. \end{aligned} \tag{3.+}$$

Lo cual equivale a

$$\begin{aligned} &\text{maximizar} \quad \log|\Sigma| + n \\ &\quad \text{sujeto a} \;\; \sigma_{ij} = s_{ij} \; \forall (i,j) \in E \text{ ó } i = j. \end{aligned}$$

Observemos que este problema cumple las condiciones de Slater A.6, por tanto es un problema que cumple dualidad fuerte, lo que significa que podemos estudiar el problema dual asociado 3.+ y obtener el mismo valor óptimo que para el problema original 3.*.

3.4. Cálculo del estimador de máxima verosimilitud

Una vez presentado el estimador de máxima verosimilitud como un problema de optimización pasamos a discutir como calcularlo, para ello podemos utilizar métodos de puntos interiores, con complejidad temporal polinómica, para un estudio más detallado de estos métodos ver 4 capítulo 11.

Otra opción eficiente para modelos gráficos de gran tamaño es el descenso de coordenadas. Dado un grafo de independencia $G = (V, E)$ este algoritmo comienza con la matriz de varianza muestral e iterando a través de las aristas $(i, j) \notin E$ minimiza la logverosimilitud en la dirección $\sigma_{i,j}$. El pseudocódigo del algoritmo es el presentado en Algoritmo 1. Podemos dar una solución cerrada para el paso de la maximización

Algorithm 1 Descenso de coordenadas para la estimación de Σ por máxima verosimilitud

Entrada: S, $G = (V, E)$ grafo de independencia, tolerancia
Salida: Estimación de máxima verosimilitud de Σ

1: $\tilde{\Sigma} \leftarrow S$
2: **repeat**
3: **for** $(i, j) \notin E$ **do**

$$\begin{aligned} &\underset{\Sigma \in \mathbb{S}^n_{\succ 0}}{\text{maximizar}} \quad \log|\Sigma| + n \\ &\quad \text{sujeto a} \;\; \Sigma_{uv} = \tilde{\Sigma}_{uv} \; \forall (u,v) \neq (i,j) \end{aligned}$$

4: **end for**
5: $\tilde{\Sigma} \leftarrow \Sigma$
6: **until** $\|\Sigma - \tilde{\Sigma}\|_1 >$ tolerancia

de la logverosimilitud en cada iteración, linea 5 de Algoritmo 1. Para ello usaremos

la expresión del complemento de Schur de tamaño 2×2, $\Sigma\backslash R$ para $R = V\backslash\{i,j\}$, es decir $\Sigma\backslash R = \begin{bmatrix} \sigma_{ii} - \Sigma_{iR}\Sigma_{RR}^{-1}\Sigma_{Ri} & \sigma_{ji} - \Sigma_{jR}\Sigma_{RR}^{-1}\Sigma_{Ri} \\ \sigma_{ij} - \Sigma_{iR}\Sigma_{RR}^{-1}\Sigma_{Rj} & \sigma_{jj} - \Sigma_{jR}\Sigma_{RR}^{-1}\Sigma_{Rj} \end{bmatrix}$.

Usando $|\Sigma| = |\Sigma\backslash R||\Sigma_{RR}|$ y que $|\Sigma_{RR}|$ es constante en la maximización iterativa del algoritmo, línea 3 de Algoritmo 1, tenemos que maximizar $|\Sigma|$ es igual a maximizar $|\Sigma\backslash R|$.

Por lo que el problema de optimización en cada iteración es equivalente a

$$\begin{aligned} &\underset{\Sigma \in \mathbb{S}^n_{\succ 0}}{\text{maximizar}} \quad \log|\Sigma\backslash R| \\ &\quad \text{sujeto a} \quad (\Sigma\backslash R)_{uu} = (\Sigma\tilde{\backslash}R)_{uu} \quad \text{para } u \in \{i,j\}. \end{aligned}$$

Veamos que los elementos de la diagonal no varían, ya que las reestricciones del algoritmo nos dicen, que para cada (i,j) únicamente cambiaremos σ_{ij} y σ_{ji} es decir que los elementos de la diagonal del complemento de Schur no cambian.

Usando que $\Sigma\backslash R$, que es simétrica el máximo de $\log|\Sigma\backslash R|$ se obtendrá cuando $\sigma_{ij} - \Sigma_{iR}\Sigma_{RR}^{-1}\Sigma_{Rj} = 0$ que se consigue cuando $\sigma_{ij} = \Sigma_{iR}\Sigma_{RR}^{-1}\Sigma_{Rj}$.

Dado que en capítulos anteriores hemos tratado siempre con la matriz de precisión, demos un algoritmo análogo ,Algoritmo 2, para hallar la máxima verosimilitud en función de dicha matriz.

Algorithm 2 Descenso de coordenadas para la estimación de D por máxima verosimilitud

Entrada: S, $G = (V,E)$ grafo de independencia, tolerancia
Salida: Estimación de máxima verosimilitud de D

1: $D \leftarrow I_n$
2: **repeat**
3: **for** $(i,j) \notin E$ **do**

$$\begin{aligned} &\underset{D \in \mathbb{S}^n_{\succ 0}}{\textit{maximizar}} \quad \log|D| - \operatorname{tr}(DS) \\ &\quad \text{sujeto a} \quad D_{uv} = \tilde{D}_{uv} \; \forall (u,v) \neq (i,j),(i,i),(j,j),(j,i). \end{aligned}$$

4: **end for**
5: $\tilde{D} \leftarrow D$
6: **until** $\|D - \tilde{D}\|_1 >$ tolerancia

La solución cerrada para el problema de maximización en la dirección (i,j), línea 4 de Algoritmo 2, viene, de nuevo de tomar el complemento de Schur de la matriz D, tomamos $A = \{i,j\}$ y $R = V\backslash A$.

$$\begin{aligned}\log|D| - \operatorname{tr}(DS) &= \log|D\backslash R||D_{RR}| - \operatorname{tr}(DS)\\ &= \log|D\backslash R| + \log|D_{RR}| - \operatorname{tr}(DS)\\ &= \log|D\backslash R| + \log|D_{RR}| - \sum_{a=1}^{n}\sum_{b=1}^{n} d_{ab}s_{ba}\\ &= \log|D\backslash R| + \log|D_{RR}| - \sum_{a,b\in\{u,v\}} D_{ab}S_{ba} - \sum_{\substack{a=1\\a\neq u,v}}^{n}\sum_{\substack{b=1\\b\neq u,v}}^{n} d_{ab}s_{ba}.\end{aligned}$$

Dado que $D\backslash R = D_{AA} - D_{AR}D_{BB}^{-1}D_{RA}$ donde el segundo término es constante en cada iteración, derivamos en función de D_{AA} y obtenemos,

$$0 = \frac{\partial}{\partial D_{AA}}\log|D\backslash R| + \frac{\partial}{\partial D_{AA}}\log|D_{RR}| - \frac{\partial}{\partial D_{AA}}\sum_{a,b\in\{u,v\}} D_{ab}S_{ba} - \frac{\partial}{\partial D_{AA}}\sum_{\substack{a=1\\a\neq u,v}}^{p}\sum_{\substack{b=1\\b\neq u,v}}^{p} D_{ab}S_{ba}.$$

Ahora usamos que los términos segundo y cuarto son constantes en D_{AA} por lo que sus derivadas son 0, obtenemos

$$\begin{aligned}0 &= \frac{\partial}{\partial D_{AA}}\log|D\backslash R| - \frac{\partial}{\partial D_{AA}}\sum_{a,b\in\{u,v\}} D_{ab}S_{ba}\\ &= (D\backslash R)^{-1} - \frac{\partial}{\partial D_{AA}}\operatorname{tr}(D_{AA}S_{AA})\\ &= (D\backslash R)^{-1} - S_{AA}^{T} = (D\backslash R)^{-1} - S_{AA}.\end{aligned}$$

Despejando $D\backslash R$, tenemos $D\backslash R = S_{AA}^{-1}$. Por tanto, $D_{AA} = S_{AA}^{-1} + D_{AR}D_{BB}^{-1}D_{RA}$, dándonos una solución cerrada al problema de optimización en el paso 4 de Algoritmo 2, este método es análogo al *iterative proportional fitting*.

3.5. Existencia del estimador de máxima verosimilitud

El siguiente paso es plantearnos cuando existe el estimador de máxima verosimilitud para la matriz de covarianzas (o de precisión), dicho problema es el de completar una matriz definida positiva. Tomaremos la matriz de varianzas muestrales y retiramos las entradas correspondientes a ejes que no se encuentran en el grafo, para ello introducimos la siguiente proyección.

Definición 3.8. Proyección del grafo. Llamamos proyección del grafo G sobre la matriz S a la aplicación

$$\begin{aligned}&\phi_G : \mathbb{S}^n_{\succ 0} \mapsto M^r(\mathbb{R})\\ &\phi_G(S) = S_G = \{s_{ij} \mid (i,j) \in E \text{ o bien } i = j\},\end{aligned}$$

donde $r = |E| + |V|$.

Ejemplo 3.2. *Dado el gráfo del 2.1 y dada la siguiente matriz S calculemos su proyección por G*

$$S = \begin{bmatrix} 1 & 0 & 3 & 1 \\ 0 & 1 & 2 & 0 \\ 3 & 2 & 1 & 3 \\ 1 & 0 & 3 & 1 \end{bmatrix} \Longrightarrow \phi_G(S) = \begin{bmatrix} 1 & * & * & * \\ * & 1 & 2 & * \\ * & 2 & 1 & 3 \\ * & * & 3 & 1 \end{bmatrix},$$

*donde los * se corresponden con las aristas retiradas por las restricciones del grafo de independencia.*

Ahora veamos cuando una matriz tiene una completación definida positiva.

Proposición 3.9. *Dada una muestra de N variables independientes con distribución normal multivariante de tamaño n, entonces la matriz de covarianzas muestral S tiene rango $min(N, n)$ con probabilidad uno.*

Proposición 3.10. *Dada una matriz parcial $S_G = \phi_G(S)$, si esta tiene una completación definida positiva entonces todas las submatrices de S_G que están completamente definidas, es decir de las que no hemos retirado ninguna entrada, son definidas positivas.*

Definición 3.11. El umbral de observaciones del estimador de máxima verosimilitud de G, denotado como $UMV(G)$, se define como el número de observaciones de N tal que el estimador de máxima verosimilitud del modelo gráfico existe con probabilidad uno. Es decir, el menor N tal que para casi toda $S \in \mathbb{S}^n_{\succ 0}$, se tiene que $\exists\ \Sigma \in \mathbb{S}^p_{\succ 0}$: $\Sigma_G = S_G$, es decir, $\sigma_{ij} = s_{ij}$ para $i = j$ y $(i, j) \in E$.

Veamos cotas útiles para el umbral de observaciones.

Proposición 3.12. *Dado un modelo gráfico gaussiano con grafo de independencia G correspondiente a un vector normal multivariante $\sim N_n(0, \Sigma)$, y dada una muestra de variables $X_1, ..., X_N$ aleatorias idénticamente distribuidas, tenemos que si $N \geq n$ entonces el estimador de máxima verosimilitud de la matriz de covarianzas existe con probabilidad uno.*

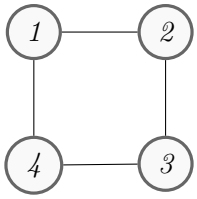
Demostración. Usando 3.9 sabemos que el rango de la matriz de covarianzas muestral S será $min(N, n)$, aplicando que $N \geq n$ tenemos que $\text{rg}(S) \geq n$, por tanto la matriz S tendrá rango máximo n, o lo que es equivalente, su determinante es distinto de 0. Por definición S es semidefinida positiva, entonces usando ambos resultados, concluimos que S es positiva definida. □

Teorema 3.13. *Dado el grafo G la cota inferior del umbral de observaciones es*

$$UMV(G) \geq \omega(G).$$

Veamos un ejemplo de este teorema

Ejemplo 3.3. *Dado el siguiente grafo de independencia con un clique de tamaño 2, veamos que el umbral de observaciones no puede ser 1.*



Es decir por 3.10 queremos ver que para alguna matriz semidefinida positiva de rango 1, no podemos convertirla en una matriz definida positiva ajustando las entradas correspondientes a las aristas que no están en el grafo. Sea dicha matriz la matriz de covarianzas muestral dada una sola observación, por definición, esta matriz tendrá rango 1.

$$S_G = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{12} & x_{22} & x_{23} & x_{24} \\ x_{13} & x_{23} & x_{33} & x_{34} \\ x_{14} & x_{24} & x_{34} & x_{44} \end{bmatrix},$$

donde hemos marcado en rojo las entradas que podemos ajustar. Como dicha matriz tiene rango 1, sabemos que, por ejemplo la submatriz principal $\begin{bmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{bmatrix}$ *no puede ser definida positiva, esto se observa viendo que no tiene rango máximo 2, ya que esto contradiría* $\operatorname{rg}(S_G) = 1$. *Por tanto la matriz* S_G *no podrá ser semidefinida positiva, sin importar los valores que tomen las entradas* x_{13} *y* x_{24}.

Veamos un teorema muy útil para el estudio de una cota superior del $UMV(G)$.

Teorema 3.14. *Dado un grafo de independencia* G *cordal y una matriz parcial* S_G, *dicha matriz tendrá completación positiva definida si y sólo si todas las submatrices completamente definidas son definidas positivas.*

Demostración. La demostración de este teorema puede encontrarse en el Teorema 7 de [2] además de proporcionar un método secuencial para hallar dicha completación. □

Esto nos indica que para grafos cordales el estimador de máxima verosimilitud existe con probabilidad uno, si y sólo si $N \geq \omega(G)$; por tanto en dichos grafos se cumple que $UMV(G) = \omega(G)$. Este resultado nos permite dar una cota superior para el umbral de verosimilitud de un grafo no cordal. Pero antes introduzcamos el concepto de cobertura cordal de un grafo.

Definición 3.15. Cobertura Cordal. Sea $G = (V, E)$ un grafo, Tenemos que $G^+ = (V, E^+)$ es una cobertura cordal de G si

1. G^+ es cordal.
2. $E \subset E^+$.

Definición 3.16. Cobertura Cordal Mínima. Sea G un grafo su cobertura cordal mínima, a la que denotamos por G^*, es una cobertura cordal de G que cumple que

$$\varphi(G^*) = \min_{G^+} \varphi\left(G^+\right), \text{ con } G^+ \text{ cobertura cordal de } G.$$

Por tanto si tenemos un grafo G no cordal, podemos convertirlo en cordal, añadiendo vértices hasta obtener su cobertura cordal mínima G^*. Una vez hallada tenemos la siguiente cota:

Proposición 3.17. *Sea G un grafo y G^* su cobertura cordal mínima. Tenemos,*

$$\varphi(G) \leq UMV(G) \leq \varphi(G^*).$$

Demostración. La cota inferior viene de 3.13. Mientras que para la superior, usamos que para G grafo cualquiera se cumple $UMV(G) \leq UMV(G^+)$, siendo G^+ una cobertura cordal de G, por tanto la cobertura cordal mínima G^* también cumple dicha igualdad. Usando que en grafos cordales el umbral coincide con el tamaño del clique maximal tenemos la cota superior, $UMV(G) \leq UMV(G^*) = \omega(G^*)$. □

CAPÍTULO 4
Aplicación a un conjunto de datos

Ahora vamos a aplicar lo visto en estos últimos capítulos de existencia y calculo del estimador de máxima verosimilitud a unos datos reales, en concreto usaremos el conjunto de datos *carcassall* que contiene 334 observaciones de 17 variables, sobre las mediciones del porcentaje de carne magra en 344 cadáveres de cerdo, además de otra información auxiliar. Para simplificar la visualización de los datos y en especial de los modelos gráficos tomaremos solamente algunas de estas variables. Que serán Fat11, Meat11, Fat12, Meat12, Fat13, Meat13, Fat14, Fat16 y LeanMeat.

Las primeras 6 filas de la muestra son:

	Fat11	Meat11	Fat12	Meat12	Fat13	Meat13	Fat14	Fat16	LeanMeat
1	17	51	12	51	12	61	15	16	56.52475
2	17	49	15	48	15	54	17	19	57.57958
3	14	38	11	34	11	40	13	14	55.88994
4	17	58	12	58	11	58	13	12	61.81719
5	14	51	12	48	13	54	12	10	62.95964
6	20	40	14	40	14	45	14	19	54.57870

El primer paso es calcular la matriz de covarianzas muestrales S, usamos la función *cov.wt* del paquete *stats*. Después calculamos la matriz de correlaciones parciales D a partir de la matriz S, usando el método *cov2pcor* del paquete *gRbase*. La resultante matriz de correlaciones parciales es:

	Fat11	Meat11	Fat12	Meat12	Fat13	Meat13	Fat14	Fat16	LeanMeat
Fat11	1.000	-0.103	0.378	0.318	0.273	-0.189	-0.014	0.160	-0.202
Meat11	-0.103	1.000	0.096	0.402	0.197	0.351	-0.082	0.029	0.113
Fat12	0.378	0.096	1.000	-0.211	0.329	0.124	0.100	0.066	-0.132
Meat12	0.318	0.402	-0.211	1.000	0.049	0.611	-0.027	-0.103	-0.034
Fat13	0.273	0.197	0.329	0.049	1.000	-0.150	0.118	0.117	-0.057
Meat13	-0.189	0.351	0.124	0.611	-0.150	1.000	0.157	0.142	0.192
Fat14	-0.014	-0.082	0.100	-0.027	0.118	0.157	1.000	0.490	-0.299
Fat16	0.160	0.029	0.066	-0.103	0.117	0.142	0.490	1.000	-0.148
LeanMeat	-0.202	0.113	-0.132	-0.034	-0.057	0.192	-0.299	-0.148	1.000

Podemos observar que las correlaciones parciales más pequeñas son: Fat11 – Fat14, Meat11 – Fat16, Meat12 – Fat14, Meat12 – Fat13 y Meat12 – LeanMeat.

Vamos a asumir que un modelo gráfico gaussiano correspondiente a dichos datos es el de la Figura 4.1, donde se han marcado en rojo las aristas a retirar, correspondientes a las cinco correlaciones parciales más pequeñas.

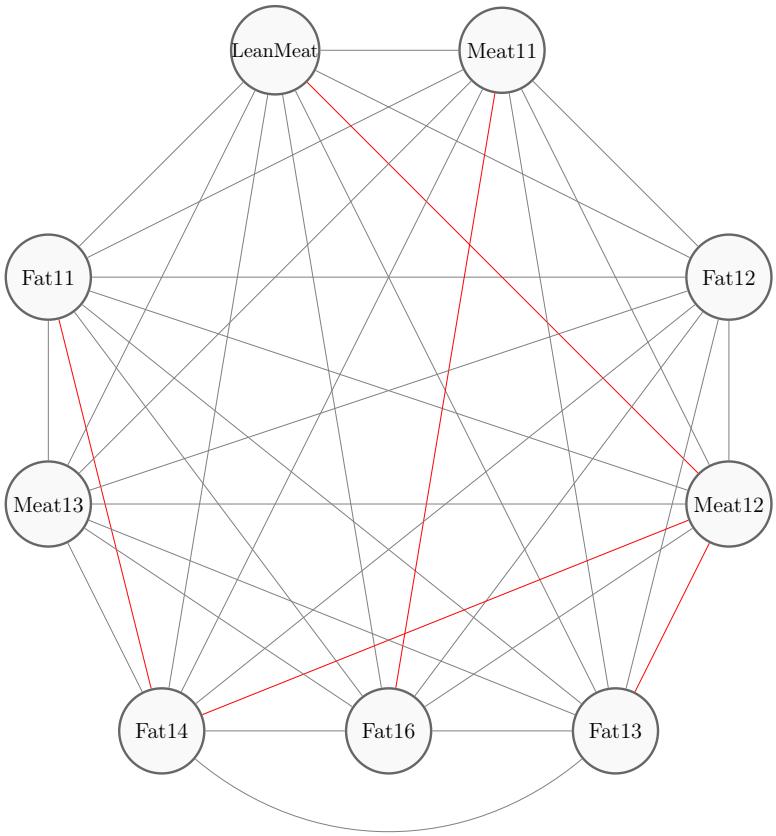


Figura 4.1: Modelo gráfico correspondiente a la matriz de correlaciones parciales de los datos, con 31 aristas, 5 menos que el grafo completo.

El siguiente paso es estimar la matriz de covarianzas usando el estimador de máxima verosimilitud sujeto a las restricciones dadas por el grafo de independencia.

Para ajustar la matriz al modelo gráfico, usamos la función *ggmfit* que implementa el algoritmo *iterative proportional fitting* [10] que, como vimos, es equivalente al descenso de coordenaas de los algoritmos 1 y 2.

Podríamos haber especificado el modelo en vez de como una lista de aristas, especificando los cliques de dicho grafo, lo que haría que el algoritmo convergiera en menos iteraciones. Esto se debe a que estaría iterando sobre bloques, mientras que en el método presentado, análogo al Algoritmo 1, lo hace sobre direcciones individuales, correspondientes a las aristas que generan el modelo gráfico.

La matriz de covarianzas inversa estimada $\hat{D}^{-1}$ usando este método es:

	Fat11	Meat11	Fat12	Meat12	Fat13	Meat13	Fat14	Fat16	LeanMeat
Fat11	0.456	0.029	-0.186	-0.083	-0.147	0.056	0.000	-0.058	0.080
Meat11	0.029	0.161	-0.027	-0.060	-0.067	-0.059	0.023	0.000	-0.022
Fat12	-0.186	-0.027	0.548	0.054	-0.183	-0.034	-0.051	-0.033	0.055
Meat12	-0.083	-0.060	0.054	0.137	0.000	-0.092	0.000	0.020	0.000
Fat13	-0.147	-0.067	-0.183	0.000	0.589	0.039	-0.057	-0.056	0.027
Meat13	0.056	-0.059	-0.034	-0.092	0.039	0.171	-0.038	-0.037	-0.041
Fat14	0.000	0.023	-0.051	0.000	-0.057	-0.038	0.487	-0.201	0.118
Fat16	-0.058	0.000	-0.033	0.020	-0.056	-0.037	-0.201	0.346	0.049
LeanMeat	0.080	-0.022	0.055	0.000	0.027	-0.041	0.118	0.049	0.329

Como era de esperar las entradas correspondientes a las aristas no presentes en el modelo gráfico son ceros. Ha necesitado 917 iteraciones para converger con una tolerancia de e^{-12} y la logverosimilitud obtenida es -7013.124.

Comparemos ahora dicha verosimilitud con la que nos daría si hubiéramos mantenido el modelo completo, es decir sin retirar ninguna arista del modelo gráfico, para ello planteamos un contraste de hipótesis donde

H_0 Se cumple el modelo gráfico completo.

H_1 Se cumple el modelo gráfico reducido.

El estadístico de razón de verosimilitudes tiene distribución aproximada χ^2_d con d la diferencia del número de aristas entre H_0 y H_1, en nuestro caso tenemos que $d = 5$.

Por 3.4 como estamos ante el modelo gráfico completo tenemos que los estimadores de máxima verosimilitud de Σ y D son

$$\hat{\Sigma} = S \text{ y por tanto } \hat{D} = S^{-1}.$$

Por tanto la logverosimilitud en este caso es

$$l_0 = -\frac{N}{2}(\log|\hat{\Sigma}| + tr(S\hat{\Sigma}^{-}1)) = -\frac{N}{2}(\log|\hat{\Sigma}| - n).$$

Ahora el estadístico de razón de verosimilitudes es

$$\chi^2 = 2(l_0 - l_1) = -N(\log|S| - \log|\Sigma_1|) = -N(\log|S| + \log|D_1|) = -N\left(\log|SD_1|\right).$$

Para realizar este calculo, podemos, seguir dos métodos, extraer la matriz ajustada D_1 del resultado de la función $ggmfit()$ y usando que la muestral *carcassall* tiene 344 observaciones realizar los cálculos, o usar el atributo *dev* que forma parte de los valores que devuelve la función y que simplemente es la desviación del modelo propuesto frente al saturado. Obtenemos que la desviación es 1,838735, la correspondiente región de rechazo a nivel de significación α será

$$R = \{\chi^2 > \chi^2_{4;\alpha}\}.$$

Para los valores usuales de α tenemos que $\chi^2_{4;0,05} = 9{,}488, \chi^2_{4;0,025} = 11{,}143$, por lo que concluimos que el modelo reducido que hemos propuesto no se ajusta bien.

También puede sernos útil calcular la información mutua entre variables marginales, para observar como interactuan entre ellas cuando conocemos el valor de otras, usando la funcion *condinformation()* del paquete *infotheo*. Podemos calcular la información mutua condicional, por ejemplo entre las variables LeanMeat, Fat11 dada la variable Meat11, lo cual nos da un resultado de 1,008596 bits. Mientras que la información mutua condicional entre LeanMeat y Fat11 dado Fat13 da un resultado menos, 0,4865057 bits.

APÉNDICE A
Optimización convexa

En este apéndice definimos una serie de conceptos básicos de optimización convexa.

Definición A.1. La función $f : \mathbb{R}^n \mapsto \mathbb{R}$ es convexa si su dominio $\mathrm{Dom}(f)$ es convexo y para cada $x, y \in Dom(f)$ y $\theta \in [0, 1]$ se tiene

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y).$$

Definición A.2. Un problema de optimización convexa en forma estándar es

$$\begin{aligned} &\text{maximizar} \;\; f_0(x) \\ &\quad \text{sujeto a} \;\; a_i^T x = b_i \; i \in \{1, ..., m\}, \end{aligned}$$

donde f_0 es una función convexa.

Cuando estos problemas tengan solución llamaremos valor óptimo y lo denotaremos por p^*, $p^* = \inf\{f_0(x) |\, a_i^T x = b_i i = 1, ..., m\}$

Definición A.3. El lagrangiano $L : \mathbb{R}^n \times \mathbb{R}^m \mapsto R$ asociado al problema de optimización A.2 es

$$L(x, v) = f_0(x) + \sum_{i=1}^{m} v_i f_i(x).$$

donde a $v = (v_1, ..., v_m)$ se les llama multiplicadores de Lagrange.

Definición A.4. La función dual de Lagrange $g : \mathbb{R}^m \mapsto R$ se define como el valor mínimo del lagrangiano en función de x, es decir

$$g(v) = \inf_x L(x, v) = \inf_x \left(f_0(x) + \sum_{i=1}^{m} v_i f_i(x) \right).$$

Proposición A.5. *La función dual de Lagrange es cóncava*

Demostración. La función dual de Lagrange se define como el ínfimo de una familia de funciones afines sobre v es cóncava, sin importar si el problema original es convexo. □

Veamos que la función dual nos da una cota inferior para el valor óptimo p^* de (4.1). Lo que nos lleva a preguntarnos cual es la mejor cota inferior que podemos obtener de la función dual, esto puede expresarse como un problema de optimización

$$\text{maximizar } g(v).$$

El problema de optimización A.2 tiene, a su vez un valor óptimo d^*, se cumple siempre la siguiente desigualdad: $d^* \leq p^*$ conocida como *dualidad débil*. Por otro lado, la *dualidad fuerte* significa que $d^* = p^*$. Veamos una condición que nos garantizara que en un problema de optimización se satisface la dualidad fuerte.

Proposición A.6. *Decimos que el problema 4.1 cumple la condición de Slater si existe un x tal que $a_i^T x = b_i$ para $i \in \{1, ..., m\}$. Si se cumple dicha condición entonces tendremos dualidad fuerte.*

La dualidad fuerte es de gran utilidad ya que cuando se cumple, es equivalente estudiar el problema dual asociado que el problema original, ya que sus valores óptimos coinciden, pudiendo aprovechar si el problema dual tiene una estructura más sencilla.

Definición A.7. Cono Lineal. Sea C un subespacio del espacio vectorial V sobre un cuerpo ordenado K, decimos que es un cono si para cada $x \in C, \alpha \in K$ $\alpha x \in C$.

Definición A.8. Cono Convexo. Sea C un cono lineal, diremos que es además un cono convexo, si $C + C \subset C$, es decir que para $x, y \in C; \alpha, \beta \in K$ $\alpha x + \beta y \in C$.

Definición A.9. Cono $\mathbb{S}^n_{\succ 0}$. El cono $\mathbb{S}^p_{\succ 0}$ se define como el cono de matrices de tamaño $n \times n$ definidas positivas.

APÉNDICE B
Código R

```
library(gRbase)
library(gRim)
library(infotheo)
data(carcassall)
carcassall$sex <- NULL
carcassall$slhouse <- NULL
carcassall$weight <- NULL
carcassall$lengthc <- NULL
carcassall$lengthf <- NULL
carcassall$lengthp <- NULL
carcassall$Fat02 <- NULL
carcassall$Fat03 <- NULL

head(carcassall)

S <- cov.wt(carcassall, method="ML")$cov
D <- cov2pcor(S)
round(D,3)

graphAll <- cmod(~.^., data=carcassall)
graphAll <- update(
    graphAll,
    list(dedge=~Fat11:Fat14
    + Meat11:Fat16
    + Meat12:Fat14
    + Meat12:Fat13
    + Meat12:LeanMeat)
)

fit <- ggmfit(S, n = nrow(carcassall), edgeList(as(graphAll, "graphNEL")))

carcassDisc <- discretize(carcassall)
I <-condinformation(carcassDisc$LeanMeat, carcassDisc$Fat11, carcassDisc$Meat11)
natstobits(I)
I <-condinformation(carcassDisc$LeanMeat, carcassDisc$Fat11, carcassDisc$Fat13)
natstobits(I)
```

Bibliografía

[1] Whittaker, J. *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester, 1990.

[2] Grone, R., Johnson, C., Sá, E., Wolkowicz, H. *Positive definite completions of partial Hermitian matrices. Linear Algebra and Its Applications,*vol. 58, 109-124. 1984.

[3] L Buhl, Søren *On the Existence of Maximum Likelihood Estimators for Graphical Gaussian Models. Scandinavian Journal of Statistics* vol. 20, pp. 263-270, 1993.

[4] Boyd, S., Vandenberghe, L. *Convex optimization*. Cambridge University Press, Cambridge, 2009.

[5] Uhler, C. *Geometry of maximum likelihood estimation in Gaussian graphical models*. The Annals of Statistics, vol. 40, no. 1, 2012.

[6] Højsgaard, S., Edwards, D., Lauritzen, S. *Graphical Models with R* Springer, 2012.

[7] Højsgaard, S. *Package 'gRim'*. 2017

[8] Houduo Qi *Positive Semidefinite Matrix Completions on Chordal Graphs and Constraint Nondegeneracy in Semidefinite Programming*. Linear Algebra and Its Applications, 430(4), 1151-1164. 2009

[9] Josh Meyer *Maximum Likelihood Estimation of Gaussian Parameters*. Aug 18, 2017.

[10] Demspter, A.P. *Covariance Selection*. Biometrics. 1972.