Práctica 2.
Inferencia exacta: ventajas y límites.

Docente: Gustavo Landfried

Inferencia Bayesiana Causal 1
1er cuatrimestre 2025
UNSAM

Los temas que se cubren en la semana 2 son: aplicación estricta de las reglas de la probabilidad. Modelos conjugados Beta-Binomial y Gaussianos. El problema histórico de la probabilidad (el costo computacional) y los efectos secundarios de la ruptura de las reglas de la probabilidad (el overfitting). Evaluación de modelos polinomiales de complejidad creciente. Identificabilidad y no Identificabilidad de modelos en datos observables.

Índice

1. ¿En qué negocio conviene comprar? 3
1.1. Interpretación del Galton Board . . . 3
1.2. Graficar el posterior . . . 4
1.3. Intervalos de credibilidad el posterior . . . 5
1.4. Elegir dónde comprar . . . 5

2. La puntería de las arqueras mexicanas. 6
2.1. Predicción sobre la posición de la primera flecha . . . 8
2.2. Posterior sobre la posición del arco . . . 9
2.3. Predicción sobre posición de la última flecha . . . 9

3. Modelos polinomiales de complejidad creciente. 9
3.1. Generar 20 datos alrededor de una período de una sinoidal . . . 11
3.2. Graficar el valor de máxima verosimilitud obtenido por los modelos polinomiales de grado 0 a 9 . . . 11
3.3. Graficar las curvas obtenidas con cada modelo mediante máxima verosimilitud. . . . 12
3.4. Evaluación de la predicción "en línea" que hacen los modelos ajustados por máxima verosimilitud. . . . 13
3.5. Más criterios arbitrarios de selección de hipótesis: los regularizadores . . . 14
3.6. El balance natural de las reglas de la probabilidad. . . . 15
3.7. Cómo se explica el balance natural de las reglas de la probabilidad . . . 16

4. Efecto causal del sexo biológico sobre la altura. 17
4.1. Abrir el archivo `alturas.csv` y visualizar los datos . . . 17
4.2. Definir 3 modelos causales alternativos . . . 17
4.3. Computar la evidencia de los modelos causales alternativos . . . 18
4.4. Computar la media geométrica de los modelos causales alternativos . . . 18
4.5. Computar el posterior de los modelos . . . 18

5. Modelos AcausaB vs Modelo BcausaA **18**
5.1. Generar datos con el modelo AcausaB . . . 19
5.2. Evaluar el desempeño predictivo de los modelos causales alternativos sobre los datos sintéticos generados en el punto anterior . . . 19
5.3. Actualizar la creencia respecto de los modelos causales alternativos luego de ver los datos. . . . 19
5.4. Dar sus conclusiones. . . . 19

6. Anexo 1. Gaussianas en una dimensión **20**
6.1. Producto de gaussianas . . . 20
6.2. Suma de gaussianas . . . 21
6.3. Gaussiana por acumulada de Gaussiana. . . . 22
6.4. Division de gaussianas . . . 23

7. Anexo 2. Gaussiana multivariada **25**
7.1. Distribuciones condicionales y marginales . . . 25
7.1.1. Distribución condicional . . . 25
7.1.2. Distribución marginal . . . 26
7.2. Modelo lineal multivariado . . . 27
7.2.1. Evidencia y posterior . . . 28

1. ¿En qué negocio conviene comprar?

Tenemos que adquirir un producto y debemos elegir un negocio entre tres donde compararlo. Para decidir vamos usar la información provista en una página de internet donde clientes reales y verificados indican si el producto comprado les gustó o no les gustó. En particular, podemos observar que el puntaje de cada uno de los negocio

- **Negocio** A: 100 %. 10 Me gusta de 10 Ratings
- **Negocio** B: 96 %. 48 Me gusta de 50 Ratings
- **Negocio** C: 93 %. 186 Me gusta de 200 Ratings

¿En qué negocio conviene comprar? Para decidir vamos a suponer un modelo simple, en el cual cada negocio j puede ser descrito mediante una variable $p_j \in [0, 1]$, que representa la probabilidad que a un consumidor le guste el producto adquirido, $P(m_i|p)$. Vamos a definir una distribución sobre la característica de los negocios uniforme en el intervalo $[0, 1]$, y una distribución sobre los "me gusta" de los clientes Bernoulli.

$$p(p_j) = U(p_j|0, 1) = 1$$
$$P(m_i|p_j) = \text{Bern}(m_i|p_j) = p_j^{m_i}(1 - p_j)^{1-m_i}$$

En la siguiente figura se muestran 3 representaciones alternativas del modelo propuesto.

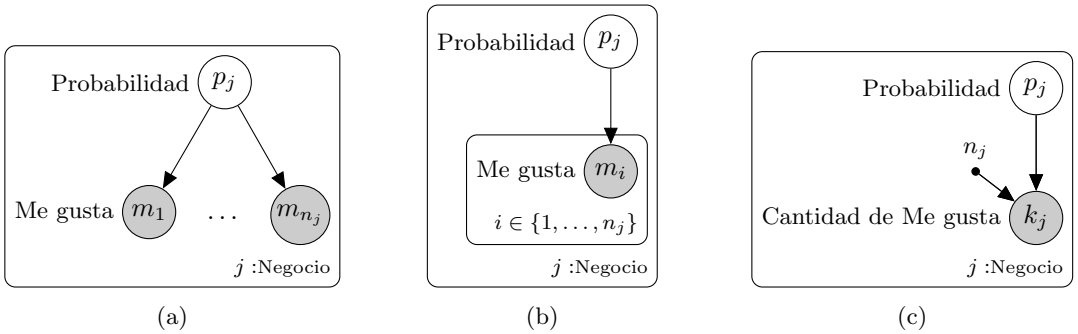


Figura 1: Redes bayesianas. Las variables ocultas aparecen en blanco, las observadas en gris, las flechas representan dependencias causales probabilísticas y las placas representan repeticiones de la misma estructura dada por el subíndice respectivo.

Los tres modelos son equivalentes. En los primeros dos modelamos cada "Me gusta" por separado. En el último, usamos una única variable para representar todos la cantidad de $k_j \leq n_j$ de "Me gusta" que hubo en cada negocio. Dada las definiciones previas, la distribución sobre la cantidad de me gustas k_j es naturalmente una Binomial.

$$P(k_j|n_j, p_j) = \text{Binomial}(k_j|n_j, p_j) = \binom{n_j}{k_j} p_j^{k_j}(1 - p_j)^{n-k_j}$$

1.1. Interpretación del Galton Board

En la figura 2 podemos ver un Galton Board, un dispositivo mecánico diseñado a principios del siglo 20 que permite hacer una aproximación discreta de la distribución normal mediante la caída de bolas a través de una serie de obstáculos regulares. ¿Qué representa cada uno de los obstáculos? ¿Qué representan cada uno de los recipientes en los que se acumulan las bolas al final del recorrido? ¿Qué representa la cantidad de niveles?

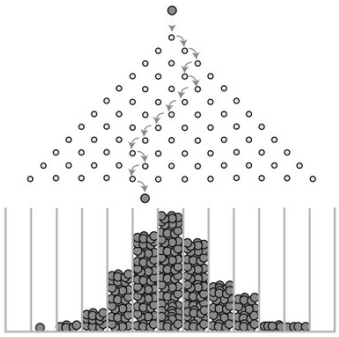


Figura 2: Galton Board.

1.2. Graficar el posterior

Antes de calcular el posterior sobre la variable oculta p_j, vamos a descomponer la distribución conjunta que va en el numerador de la probabilidad condicional.

$$p(p_j|\text{Datos} = \{m_1, \ldots, m_j\}) = \frac{p(p_j, \text{Datos})}{P(\text{Datos})}$$

De forma similar a lo que nos había ocurrido en el modelo NoMontyHall, que diseñamos durante la primera semana, la distribución conjunta se puede descomponer como el producto de todos los likelihood individuales por el prior.

$$p(p_j|\text{Datos} = \{m_1, \ldots, m_j\}) = \frac{\overbrace{p(p_j)}^{\text{Prior}} \overbrace{\prod_{i=1}^{n_j} P(m_i|p_j)}^{\text{Likelihood}}}{P(\text{Datos})}$$

La demostración es directa, ¿la ven? (recordar que la conjunta es el producto de las condicionales).

También podemos calcular el posterior dado la cantidad de me gustas.

$$p(p|k,n) = \frac{P(k|p,n)p(p)}{P(k|n)} = \underbrace{(n+1)}_{1/P(k\,|\,n)} \underbrace{\frac{n!}{k!(n-k)!}\, p^k\,(1-p)^{n-k}}_{P(k\,|\,p,n)} \underbrace{1}_{P(p)}$$

En esta ecuación evitamos el uso de subíndices para no sobrecargar su visualización. En su escrito, Bayes se convence de que el denominador del posterior es $P(k|n) = 1/(n+1)$, pues como tenemos una distribución uniforme sobre la probabilidad p, todos los k son igualmente probables. Una demostración formal se puede realizar notando que, si p es uniforme vale que,

$$\begin{aligned} P(k|n) &= \int P(k|n,p)P(p)dp = \int P(k|n,p)dp \\ &= \binom{n}{k} \underbrace{\int p^k(1-p)^{n-k}\, dp}_{\beta(k+1,n-k+1)} \end{aligned}$$

Donde $\beta()$ se conoce como la función Beta y para números enteros su solución es,

$$\beta(k+1, n-k+1) = \frac{k!(n-k)!}{(n+1)!}$$

Por lo tanto,

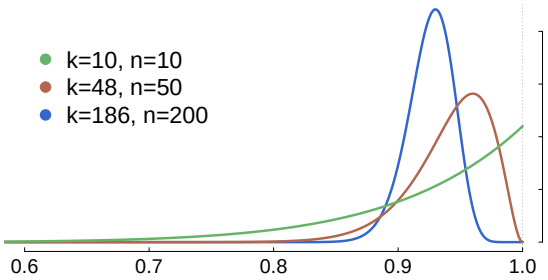
$$P(k|n) = \binom{n}{k} \frac{k!(n-k)!}{(n+1)!} = \frac{1}{(n+1)}$$

Luego, la distribución de probabilidad del posterior es una beta

$$p(p|k,n) = \text{Beta}(k+1, n-k+1) = (n+1)\binom{n}{k} p^k (1-p)^{n-k}$$
$$= \frac{p^k(1-p)^{n-k}}{\beta(k+1, n-k+1)}$$

La distribución Beta está definida en la mayoría de los lenguajes de programación.

Graficar el posterior sobre p_j para los 3 negocios.



1.3. Intervalos de credibilidad el posterior

Visualizar el posterior que obtenemos de la calidad de los negocios nos permite ganamos bastante intuición respecto de dónde nos conviene y no nos conviene comprar. Por ejemplo, parece ser bastante claro que no sería una buena idea comprar en el negocio con solo 10 puntuaciones porque, a pesar de que tenga el 100 % positivos, tenemos todavía bastante incertidumbre respecto de la calidad del negocio. De hecho, se puede ver que esa distribución de creencias no es cero aún para valores $p = 0{,}7$ de calidad del negocio. Los posterior de los otros negocios, en cambio están concentrados en valores más altos, siendo siempre mayores a $p = 0{,}8$.

Para ayudar interpretar las distribuciones posterior, calcular el intervalo de credibilidad del 95 %

Un intervalo de credibilidad del 95 % tiene que tener 2.5 % de la probabilidad en el extremo izquierdo y otro 2.5 % de la probabilidad en el extremo derecho.

1.4. Elegir dónde comprar

¿Cómo podemos hacer para elegir dónde comparar? Existen muchas formas de resumir los datos, pero cual es la forma correcta? Nuestro objetivo, en definitiva, es maximizar la probabilidad de que el producto nos guste. Es decir, queremos comprar en el negocio que maximice la probabilidad de que nuestra experiencia sea positiva.

$$\arg\max_j P(m = 1|\text{Datos}) = \arg\max_j \int p(p_j, m = 1|\text{Datos})\, dp$$

En términos gráficos, lo que estamos haciendo es agregar una variable oculta adicional al modelo que representa nuestra experiencia de usuario.

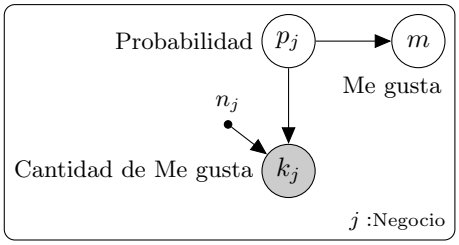


Figura 3: Extensión del modelo, donde se agrega una variable oculta que representa nuestra experiencia en cada uno de los negocios.

Calcular la distribución de probabilidad marginal de nuestra experiencia dado los datos, integrando todos los valores de p.

Notar que,

$$\int p(p_j, m = 1|\text{Datos})\, dp = \int P(m = 1|p_j)\, p(p_j|\text{Datos})\, dp = \int p \cdot p(p_j|\text{Datos})\, dp$$

¿Qué forma tiene la última ecuación?

2. La puntería de las arqueras mexicanas.

En el artículo *Performance assessment in archery: a systematic review* del año 2024 se realizó una búsqueda sistemática de literaturas sobre bases de datos relevantes hasta julio de 2021, identificando 41 estudios que abarcan 35 años (1986-2021). Sería interesante incluir en esta práctica una base de datos real con la que jugar. Sin embargo, el citado artículo científico no es gratuito, por lo que no se pudo revisar las bases de datos. Si alguien consigue una base de datos real, interesante, por favor comunicarse con `glandfried@dc.uba.ar`. Son de particular interés los datos de las olimpiadas 2024 en el que el equipo mexicano femenino de arquería obtuvo una medalla de bronce después por debajo de Corea del Sur (plata) y China (oro).

Figura 4: Olimpiadas 2024. Arquera mexicana

En esta práctica, vamos a estudiar un caso simplificado, que nos va a servir de introducción a los modelos lineales gaussianos, los que son fundamentales tanto para la regresión lineal bayesiana [1, 2], para modelos de estimación de habilidad estado del arte en la industria del video

juego (TrueSkill Through Time) [3], para modelos de recomendación basado en descomposición de matrices (MatchBox Microsoft) [4], y para procesos gaussianos [5], entre otras.

El caso más básico que podemos considerar contiene dos variables, una que representa la posición del arco x y otra que representa la posición de la flecha y_i, ambas descritas en una única dimensión. Vamos a suponer que la posición de la flecha está centrada en la posición del arco con cierto ruido β[1].

$$p(y_i|x) = \mathcal{N}(y_i|x, \beta^2)$$

Las flechas son las variables observables que usaremos para inferir la posición del arco. Como distribución a priori sobre la posición del arco vamos tomar una distribución gaussiana con cierta media μ y suficiente incertidumbre σ^2 de modo tal de no dejar "afuera" el valor real desconocido de la posición del arco.

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

Luego, la especificación matemática del modelo queda definida sin ambigüedades y de forma completa a través del factor graph descrito en la figura 5a. Sin embargo, ese mismo modelo puede ser especificado de forma equivalente mediante el factor graph descrito en la figura 5b.

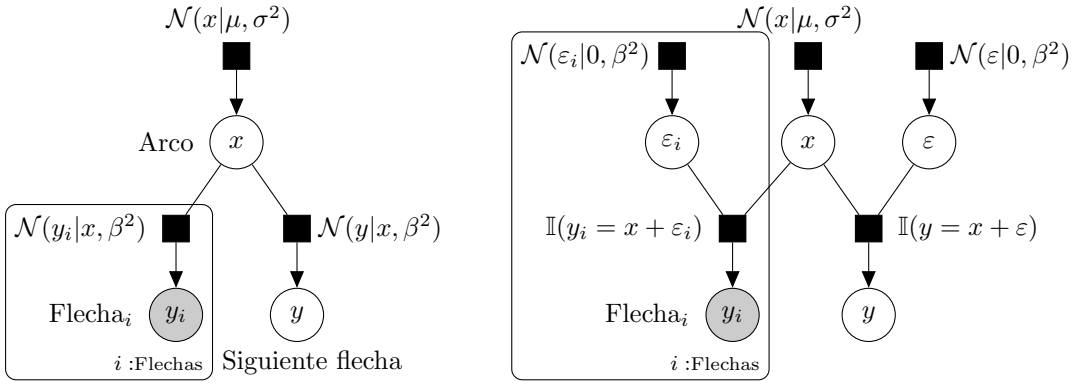


Figura 5: Modelos especificados matemáticamente mediante la notación Factor Graph. Los factor graph tienen dos tipos de nodos, los cuadrados negros que representan distribuciones de probabilidad condicional y los círculos que representan variables. Al igual que las redes bayesianas, las variables ocultas aparecen en blanco, las observadas en gris, las flechas representan dependencias causales probabilísticas y las placas representan repeticiones de la misma estructura dada por el subíndice respectivo.

En el segundo modelo, la posición de la flecha y_i depende de forma determinística de la posición oculta del arco x y de un ruido aleatorio ε_i. Por ese motivo, la distribución de probabilidad que permite describir relaciones deterministicas entre las variables es la función indicadora (o delta de dirac).

$$p(y_i|x, \varepsilon_i) = \mathbb{I}(y_i = x + \varepsilon_i)$$

Para variables discretas la función indicadora vale 1 cuando la igualdad se cumple y 0 en caso contrario. Para variables continuas, como ésta, la función indicadora debe interpretarse como una delta de dirac. En el sección *Anexo 1. Gaussianas en una dimensión* se describen las propiedades de la distribución gaussiana unidimensional. Ambos factor graphs son equivalentes.

[1] Como nota de color, la distribución gaussiana es la que maximiza incertidumbre (o entropía) dada la información de la media y la varianza.

2.1. Predicción sobre la posición de la primera flecha

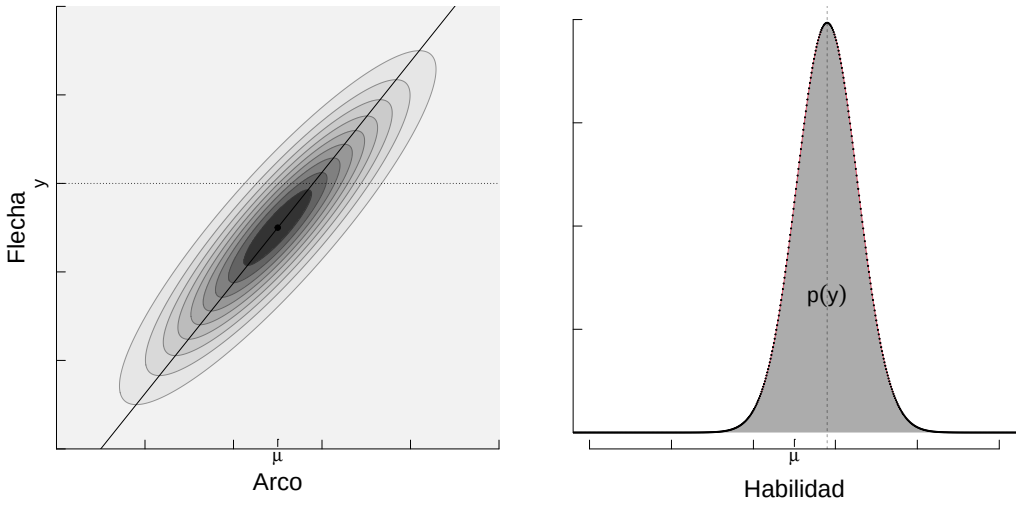
Para predecir la posición de la primera flecha, necesitamos determinar la distribución de probabilidad a priori $p(y)$ sobre la posición de la flecha antes de haber observado ningún dato. Notemos que la distribución de la flecha depende de la posición del arco, que es una variable oculta:

$$\begin{aligned} p(y) = \int p(y|x)p(x)dx &= \int \mathcal{N}(y|x, \beta^2)\mathcal{N}(x|\mu, \sigma^2)\, dx \\ &\overset{6,1}{=} \int \underbrace{\mathcal{N}(y|\mu, \beta^2 + \sigma^2)}_{\text{constante en } x} \underbrace{\mathcal{N}(x|\mu_*, \sigma_*^2)\, dx}_{\text{integra 1}} = \mathcal{N}(y|\mu, \beta^2 + \sigma^2) \end{aligned} \tag{1}$$

Donde la igualdad $\overset{6,1}{=}$ vale por las propiedades descritas en la sección 6.1 (*Producto de gaussianas*). En esa sección también encontrarán la definiciones de μ_* y σ_*. ¿Qué es lo que está ocurriendo? ¿Que significa que la integral del producto de esas dos gaussianas sea a su vez una gaussiana?

Ejercicio. Para ganar intuición del proceso es útil verificar meidante aproximaciones numéricas que los resultados matemáticos del anexo son correctos.

Para ello, inicializar los valores a priori μ y σ y crear una matriz en la que, para todos los valores de x y todos los valores de y se guarde el valor de la multiplicación de las dos gaussianas que representan $p(y|x)$ y $p(x)$. Les debería quedar un gráfico el de la figura 6a. Luego, queremos calcular la integral. Para ello, fijamos un valor determinado de y que se corresponde por ejemplo con la línea horizontal 6a. En la figura 6b graficamos todos los valores que están debajo de la línea que representa un valor de y constante. Se puede ver que su forma se parece a la de una gaussiana no normalizada. ¿Qué gaussiana es? Verificar su respuesta. Calcular la integral significa sumar el área debajo de la curva de esa gaussiana.



(a) (b)

Figura 6

Lo llamativo del resultado matemático es que el valor de esa integral para diferentes puntos de la variable y está parametrizado mediante otra distribución gaussiana! ¿Qué gaussiana es?. Verificar el resultado.

Este mismo resultado se obtiene también cuando calculamos la predicción $p(y)$ en el segundo modelo (figura 5b). Primero debemos definir la integral en de la conjunta x, y, ε, ahora integrando

en dos dimensiones.

$$\begin{aligned} p(y) &= \int\int p(y|\varepsilon, x)p(\varepsilon)p(x)dxd\varepsilon \\ &= \int\int \mathbb{I}(y = x + \varepsilon)\mathcal{N}(\varepsilon|0, \beta^2)\mathcal{N}(x|\mu, \sigma^2)\, dxd\varepsilon \\ &\overset{1}{=} \int \mathcal{N}(y - x|0, \beta^2)\mathcal{N}(x|\mu, \sigma^2)\, dx \\ &\overset{2}{=} \int \mathcal{N}(y|x, \beta^2)\mathcal{N}(x|\mu, \sigma^2)\, dx \end{aligned}$$

Debido a que la distribución de probabilidad sobre y vale 0 en todo el espacio salvo cuando se cumple la igualdad, en $\overset{1}{=}$ reemplazamos variables y reducimos dimensionalidad. Luego, debido a la simetría de la distribución gaussiana, en $\overset{2}{=}$ pasamos el x como media, quedando la misma ecuación que ya hemos evaluado arriba para calcular $p(y)$.

2.2. Posterior sobre la posición del arco

El objetivo de este ejercicio es calcular posición del arco x dado los datos de las flechas. Para ello, vamos a supongamos que el ruido $\beta = 1$. El posterior dada la posición de una única flecha es,

$$\begin{aligned} p(x|y) = \frac{p(x,y)}{p(y)} &= \frac{\mathcal{N}(y|x, \beta^2)\mathcal{N}(x|\mu, \sigma^2)}{\int \mathcal{N}(y|x, \beta^2)\mathcal{N}(x|\mu, \sigma^2)dx} \\ &\overset{(1)}{=} \frac{\mathcal{N}(y|x, \beta^2)\mathcal{N}(x|\mu, \sigma^2)}{\mathcal{N}(y|\mu, \beta^2 + \sigma^2)} \\ &\overset{6,1}{=} \frac{\cancel{\mathcal{N}(y|\mu, \beta^2 + \sigma^2)}\mathcal{N}(x|\mu_*, \sigma_*^2)}{\cancel{\mathcal{N}(y|\mu, \beta^2 + \sigma^2)}} = \mathcal{N}(x|\mu_*, \sigma_*^2) \end{aligned}$$

Donde

$$\sigma_* = \sqrt{\frac{\beta^2\sigma^2}{\beta^2 + \sigma^2}} \qquad \mu_* = \frac{(y\sigma^2 + \mu\beta^2)}{(\beta^2 + \sigma^2)} \tag{2}$$

Determinar cuál es el posterior de x dado el dataset completo

$$p(x|\text{Datos} = \{y_1, \ldots, y_n\}) = ?$$

y calcularlo para datos simulados en base a un arco que tiene posición $x = 2$ y con $\beta = 1$. Graficar como cambia la estimación en el tiempo.

2.3. Predicción sobre posición de la última flecha

Determinar cuál es la predicción de y_{n+1} dado el dataset completo

$$p(y_{n+1}|\text{Datos} = \{y_1, \ldots, y_n\}) = ?$$

3. Modelos polinomiales de complejidad creciente.

Pueden usar cualquier librería que implemente: 1. la regresión lineal por mínimos cuadrados (máxima verosimilitud); y 2. la regresión lineal bayesiana. En el primer caso proponemos el uso del métodos `OLS` de la libraría `statmodels`, en particular los métodos `fit()` que selecciona las hipótesis (atributo `params`) y el logaritmo del likelihood del modelo (atributo `llf`). En el segundo caso proponemos de la librería contenida en el archivo `ModeloLineal.py`, incluido en los materiales de esta práctica, y en particular los métodos `posterior()` que devuelve la media y la covarianza de los hipótesis y `log_evidence()` que devuelve la predicción que hace el modelo del conjunto de datos en escala logarítmica. Ambos librerías tienen un uso similar.

```
from statsmodels.api import OLS
import ModeloLineal as ml          # El archivo ModeloLineal.py debe estar en la misma carpeta
 ...
N = 20
X = np.random.rand(N,1)-0.5
Y = realidad_causal_subyacente(X)
assert Y.shape == (20,1)

# Transformacion de los X
def phi(X, complejidad):
    return(pd.DataFrame({f'X{d}': X[:, 0]**d for d in range(complejidad+1)}))

assert phi(X, 2).shape == (20,3)

modelo = OLS(Y, phi(X,2)).fit()
modelo.params # La hipotesis seleccionadas
modelo.llf     # El likelihood del modelo en escala logaritmica

MU_2, COV_2 = ml.posterior(Y, phi(X, 2))
log_evidence_2 = ml.log_evidence(Y, phi(X, 2))
```

La gran mayoría de los modelos de inteligencia artificial, incluyendo las redes neuronales [6], se construyen a partir de modelos que postulan relaciones "lineales" entre las hipótesis. El modelo más básico es la famosa regresión lineal. Dado un conjunto de datos $\{(x_1, y_1), \ldots, (x_T, y_T)\}$, un modelo causal lineal afirma que el valor de la variable y_i se generar en función de x_i y un conjunto de hipótesis ocultas h, tal que

$$y \leftarrow h_0 + h_1 \cdot x$$

h_0 representa el valor base de y cuando x es neutral ($x = 0$), y h_1 representa el efecto causal que la varibale x tiene sobre y, el cual es proporcional a su propio valor ($h_1 \cdot x$), dando una relación lineal entre el valor de la causa x y el valor de su efecto y. De modo similar, podemos construir relaciones más complejas, como son los polinomios de grado M.

$$y \leftarrow h_0 + h_1 \cdot x + h_2 \cdot x^2 + \cdots + h_M \cdot x^M = \sum_{i=0}^{M} h_i \cdot x^i$$

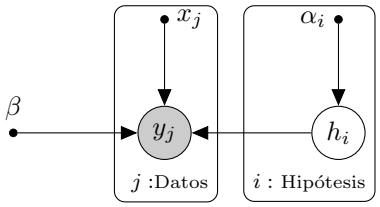
Es interesante notar que un polinomio de grado $M-1$ es un caso especial de un polinomio de grado M en el que el $h_M = 0$. Es decir, cuánto más complejo sea el polinomio, más flexibilidad tiene el modelo para representar la relación entre x e y. Debido a que los datos se miden con un error que se produce con desvío estándar β centrado en cero, la probabilidad condicional de observar un dato es,

$$p(y|x, \boldsymbol{h}, \text{Modelo} = M) = \mathcal{N}\left(y \,\middle|\, \sum_{i=0}^{M} h_i \cdot x^i,\ \beta^2\right)$$

Como tenemos incertidumbre respecto de los valores de las hipótesis h_i, proponemos una distribución que creencias *a priori* alrededor del cero.

$$p(h_i) = \mathcal{N}(h_i \,|\, 0, \alpha_i^2)$$

Luego, la especificación gráfica del modelo lineal es



Quisiéramos actualizar nuestra creencia respecto de las hipótesis h al interior de cada modelo y actualizar nuestra creencia sobre los modelos causales alternativos. Ninguno de estas dos objetivos se pudo realizar de forma exacta hasta las vísperas del siglo 21. En este ejercicio nos interesa comparar los resultados que se obtienen mediante los métodos propuestos durante el siglo 20, basados en estimadores puntuales, respecto de los resultado que se obtienen de evaluar todo el espacio de hipótesis (y modelos) mediante la aplicación estricta de las reglas de la probabilidad. En el primer caso usaremos el método OLS (por *Ordinary Least Squares*) del paquete `statsmodels` y para el segundo caso usaremos nuestra propia implementación disponible en el archivo `ModeloLineal.py`[2].

3.1. Generar 20 datos alrededor de una período de una sinoidal

Supongamos que los datos se generan de una realidad causal subyacente completamente distinta, siguiendo una función sinoidal en el rango $x \in [-\frac{1}{2}, \frac{1}{2}]$.

$$p(x) = \text{Unif}(x| -0{,}5, 0{,}5)$$
$$p(y|x) = \mathcal{N}(y|\texttt{sin}(2\pi x), \beta^2)$$

Al graficar la función objetivo (línea) y los datos (puntos) se soberva lo siguiente.

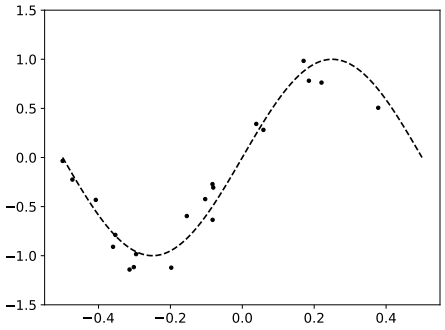


Figura 7: Datos generados de la "función objetivo" (realidad causal subyacente)

3.2. Graficar el valor de máxima verosimilitud obtenido por los modelos polinomiales de grado 0 a 9

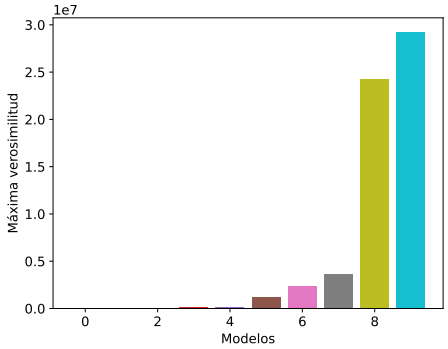
Debido a la complejidad computacional de la aplicación estricta de las reglas de la probabilidad, durante el siglo 20 se propusieron una gran cantidad de criterios arbitrarios para la selección de

[2]La derivación matemática de la regresión lineal la pueden encontrar en el capítulo 2 del libro de Bishop PRML. El presente ejercicio está extraído del capítulo 3 del mismo libro.

una única hipótesis del espacio, como es el principio de máxima verosimilitud.

$$\arg\max_{h} \; p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{h}, M_M) \overset{*}{=} \arg\min_{h} \sum_{j\in\{1,\dots,N\}} (y_j - \sum_{i}^{M} h_i \cdot x_j^i)^2$$

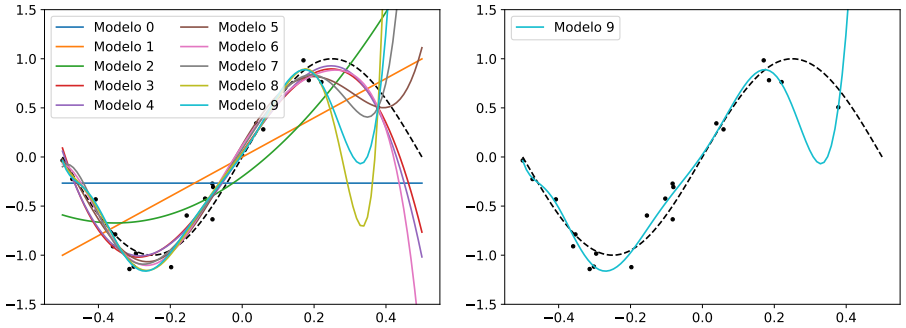
Por propiedades de este tipo de modelos ($\overset{*}{=}$) encontrar las hipótesis que mejor predicen es lo mismo que minimizar la suma de las distancias cuadradas entre el verdadero valor y el valor medio propuesto. Al graficar el valor de máxima verosimilitud obtenido con cada uno de los modelos veremos que a medida que aumentamos la complejidad de los modelos aumenta también el valor de máxima verosimilitud.



Esto ocurre debido a que cuanto más flexible es un modelo, más se puede acercar a todos los puntos. Por lo tanto, si usamos el criterio de máxima verosimilitud para seleccionar modelo, elegiríamos siempre el modelo más complejo.

3.3. Graficar las curvas obtenidas con cada modelo mediante máxima verosimilitud.

Seleccionar el modelo de mayor complejidad no es deseable. Para ver por qué, vamos a graficar las curvas obtenidas por cada uno de los modelos de grado 0 a 9.



(a) Todos los modelos (b) Modelo de máxima verosimilitud

Figura 9: Ajuste de los polinomios por máxima verosimilitud.

En la figura 9a podemos ver que los modelos de grado 0 a 2 no tienen la complejidad suficiente para acercarse a los datos como lo hacen los modelo de complejidad 3 o superior. En la figura 9b destacamos la curva que se obtiene con el modelo más complejo, de grado 9.

A diferencia del modelo 3 (color rojo en figura 9a) que se mantiene cerca de la función objetivo, el modelo de grado 9 se separa significativamente de la función objetivo alrededor de los valores

$x = 0{,}3$ y $x = 0{,}5$. La excesiva flexibilidad hace que los modelos más complejos adopten formas extrañas que suelen alejarse más de la función objetivo oculta, por lo que la predicción sobre nuevos datos del modelo 3 termina siendo superior al del modelo 9 que habíamos elegido por máxima verosimilitud.

Esto se conoce se conocen en el área de inteligencia artificial con el nombre de sobreajuste o *overfitting* y veremos que es efecto secundario de utilizar métodos arbitrarios de selección de hipótesis (cómo máxima verosimilitud) y no una propiedad indeseable del sistema de razonamiento probabilístico.

3.4. Evaluación de la predicción "en línea" que hacen los modelos ajustados por máxima verosimilitud.

Para evaluar correctamente los modelos alternativos deberíamos calcular la evidencia de los modelos (P(Datos|Modelo)), que no es más que la predicción *a priori* que hace el modelo de todos los datos, lo que es lo mismo que el producto de las predicciones individuales del siguiente dato usando los datos ya observados como información previa.

$$P(\text{Datos} = \{d_1, d_2, \dots\}|M_D) = P(d_1|M_d)P(d_2|d_1, M_D)\dots$$

Este producto de predicciones representa lo que nos ocurriría efectivamente si usáramos el modelo en la vida real: ajustamos el modelo con los datos que ya tenemos disponibles (entrenamiento) y evaluamos el desempeño en los datos nuevos (testeo). Para simular este proceso, vamos a implementar un procedimiento en el cual predecimos los datos individuales ajustando los modelos por máxima verosimilitud sobre los datos observados previamente.

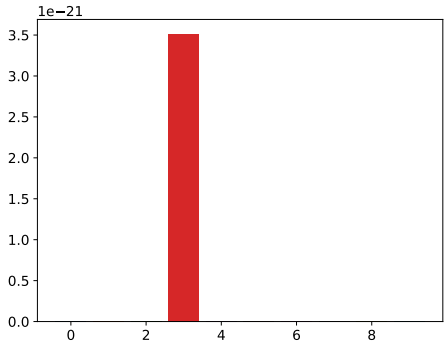


Figura 10: Producto de las predicciones a prior de los modelos ajustados por máxima verosimilitud.

Si evaluáramos los modelos en base al desempeño predictivo, dado por el producto de las predicciones a periori, rechazaríamos todos los modelos salvo el de grado 3. Esto también es un caso de sobreajuste (overfitting), pues en caso de recibir datos por fuera del período, tendremos bajo desempeño predictivo debido a que el modelo de grado 3 no tendrá la flexibilidad suficiente para acercarse a la función objetivo. El problema del sobreajuste (overfitting) ocurre como efecto secundario de seleccionar una única hipótesis del espacio y no poder computar de forma exacta la evidencia de los modelos, P(Datos|Modelo).

Cuando seleccionamos de forma arbitraria una única hipótesis no podemos computar la evidencia de los modelos, P(Datos|Modelo), debido a que las predicciones que hacen los modelos debería realizarse con la contribución de todas las hipótesis, y no con una única hipótesis seleccionada

previamente.

$$P(y_i|\underbrace{x_1, y_1}_{d_1}, \ldots, \underbrace{x_{i-1}, y_{i-1}}_{d_{i-1}}, x_i, M_D) = \int_{\boldsymbol{h}} P(y_i|d_1, \ldots, d_{i-1}, x_i, \boldsymbol{h}, M_D) P(\boldsymbol{h}|d_1, \ldots, d_{i-1}, M_D)$$

$$\approx P(y_i|d_1, \ldots, d_{i-1}, x_i, \underbrace{\arg\max_{\boldsymbol{h}} P(d_1, \ldots, d_{i-1}|\boldsymbol{h}, M_D), M_D)}_{\text{Hipóteis de máxima verosimilitud}}$$

Es decir, aproximamos las predicciones que deberían hacer los modelos con la contribución de todas las hipótesis mediante la predicción que obtenemos con la hipótesis que maximizaba la verosimilitud en el paso anterior.

3.5. Más criterios arbitrarios de selección de hipótesis: los regularizadores

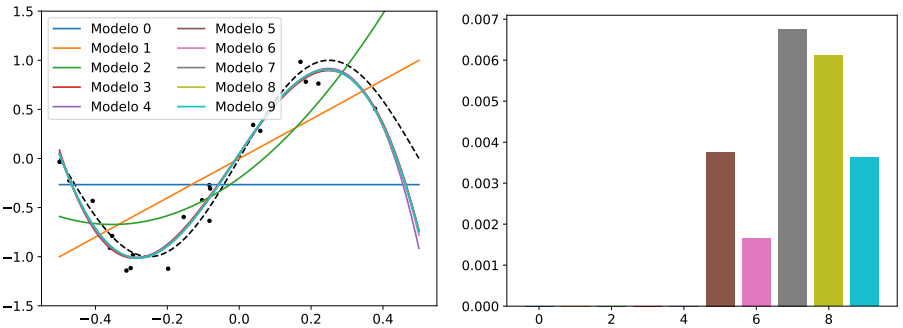
Para evitar algunos de los efectos secundarios de la selección arbitraria de hipótesis mediante máxima verosimilitud, se han propuesto una gran variedad de otros criterios arbitrarios de selección. Una familia alternativa de criterios muy difundida son los llamados regularizadores. En vez de seleccionar la hipótesis que mejor predice los datos (máxima verosimilitud) elegimos la hipótesis que tienen mayor creencia luego de observar los datos (máximo a posteriori).

$$\arg\max_h \underbrace{p(\boldsymbol{h}|\overbrace{\boldsymbol{y}, \boldsymbol{x}}^{\text{Datos}}, M_M)}_{\text{Posterior}} = \arg\max_h \overbrace{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{h}, M_M)}^{\text{Verosimilitud}} \overbrace{p(\boldsymbol{h}|M_M)}^{\text{Prior}}$$

$$\overset{*}{=} \arg\min_h \underbrace{\sum_j (y_j - \sum_i h_i \cdot x_j^i)^2}_{\text{Distancia entre dato y predicción media}} + \underbrace{\frac{\alpha}{2}||h||^2}_{\text{Penalización}}$$

Por propiedades del modelo ($\overset{*}{=}$) maximizar el producto de la verosimilitud con el prior es equivalente a minimizar la suma de distancias cuadradas entre el dato y la predicción, más la suma de la distancia cuadrada de las hipótesis al origen pesada por la precisión a priori sobre las hipótesis α. Esto hace que la hipótesis que se seleccione no sea la que más se acerca a los datos, sino la que mejor balance tenga entre la distancia y la penalización.

Veamos cómo se desempeñan los modelos bajo este nuevo criterio arbitrario de selección de hipótesis. Vamos a proceder del modo similar a lo realizado en el ejercicio anterior. Para calcular el posterior exacto sobre las hipótesis vamos a usar nuestra propia implementación disponible en el archivo `ModeloLineal.py`. Debido a que la penalización depende del prior, vamos a calcular el posterior usando un prior 100 veces más informativo que el prior que definimos al inicio del ejercicio.



(a) Ajuste por Máximo a posteriori (b) Predicciones a priori "en línea"

Figura 11: Selección arbitraria de hipótesis por máximo a posterior con prior informativo.

El prior informativo obliga a que los valores de las hipótesis no puedan alejarse de 0. Esto hace que los modelos más complejos pierdan flexibilidad y se comporten como si fueran modelos más simples. De esta forma, se logra mejorar el desempeño de los modelos complejos incluso cuando evaluamos sus predicciones a priori, entrenando el modelo en línea cada vez que recibe un nuevo dato.

En este caso estamos rechazando todos los modelos simples, quedándonos con los modelos de mayor complejidad. Si bien esto podría parecer una solución al problema del ejercicio anterior, no lo es. Así como el rechazo de todos los modelos salvo el de grado 3 hacía que no pudiéramos predecir potenciales datos futuros que estuvieran fuera del período observado, la selección de los modelos más complejos no ocurre por mérito de su complejidad sino por la reducción de flexibilidad que le impusimos mediante la regularización. En términos prácticos, los modelos complejos están funcionando como una variante apenas más flexible que el modelo 3 regularizado. La regularización de hipótesis entonces no resuelve el problema, pues cuando veamos datos por fuera del período observado, estos modelos tampoco van a ser capaces de predecir los nuevos datos.

$$P(\text{Datos}) = \sum_{m \in \{5,\dots,9\}} P(\text{Datos}|\text{Modelo} = m) P(\text{Modelo} = m) \approx P(\text{Datos}|\text{Modelo} = 3)$$

3.6. El balance natural de las reglas de la probabilidad.

Para evaluar correctamente las hipótesis y modelos causales alternativos es suficiente con actualizar las creencias aplicando estrictamente las reglas de la probabilidad. Para eso necesitamos calcular la probabilidad a posteriori de los los modelos dado los datos.

$$P(\text{Modelo} = m|\text{Datos}) = \frac{\overbrace{P(\text{Datos}|\text{Modelo} = m)}^{\text{Evidencia}} P(\text{Modelo} = m)}{\sum_i P(\text{Datos}|\text{Modelo} = i) P(\text{Modelo} = i)}$$

Para calcular de forma exacta la evidencia del modelo (en órdenes de magnitud) podemos usar el método `lm.log_evidence()` de nuestro paquete `ModeloLineal.py`. Suponiendo que no tenemos preferencia a priori por ningún modelo, vamos a usar la evidencia para calcular el posterior exacto de los modelos.

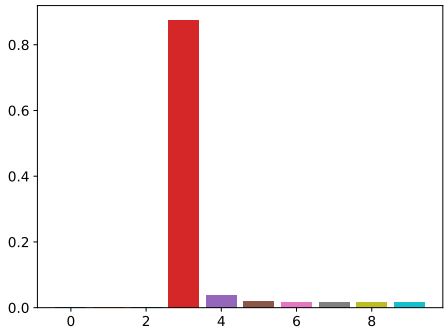


Figura 12: Posterior exacto de los modelos.

Cuando evaluamos correctamente las hipótesis, encontramos que: el modelo que tiene la menor complejidad necesaria (grado 3) es el que obtiene mayor creencia a posteriori; los modelos que tienen menor complejidad de la necesaria (grado 0 a 2) son rechazados; y los modelos que tienen mayor complejidad de la necesaria (grado 4 a 9) tienen baja probabilidad, pero no son rechazados. A diferencia de lo que ocurría cuando seleccionábamos hipótesis de forma arbitraria, evaluando correctamente las hipótesis garantizamos nuestra capacidad para predecir potenciales datos que aparecieran en el futuro por fuera del del período observado gracias a que tenemos disponibles

todavía modelos más complejos capaces de explicarlos.

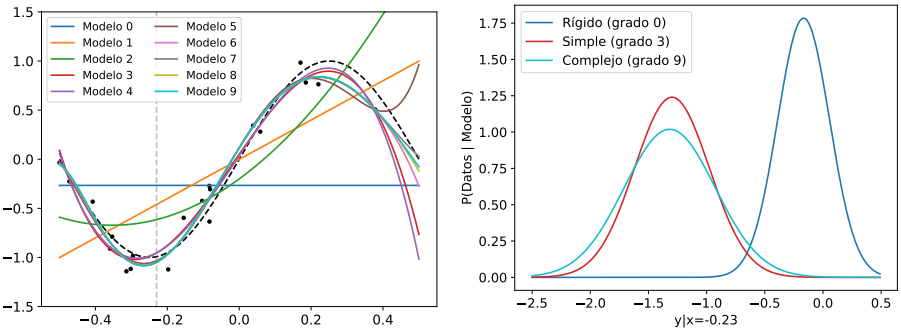
$$P(\text{Datos}) = \sum_m P(\text{Datos}|\text{Modelo} = m)P(\text{Modelo} = m)$$

En definitva, los problemas que emergen cuando se seleccionan las hipótesis mediante criterios arbitrarios simplemente no forman parte del sistema de razonamiento en contextos de incertidumbre. Por eso, la aplicación estricta de las reglas de la probabilidad, siguiendo el principio de "no mentir", es el razonamiento óptimo en contextos de incertidumbre.

3.7. Cómo se explica el balance natural de las reglas de la probabilidad

En probabilidad las predicciones son distribuciones de probabilidad que deben integrar 1. Esto produce una balance natural entre los modelos debido a que ningún modelo es superior a otro en términos absolutos. Todos los modelos tienen una región en donde ganan y todos tienen una región en donde pierden. Veamos esto de forma concreta.

En la figura 13a graficamos la media del posterior de los modelos basado en un prior no informativo. No estamos graficando la incertidumbre de los modelos, que representaría la predicción que hacen los modelos de los datos. Para graficar esa incertidumbre vamos a hacer un corte en la figura 13a cuando x vale $-0{,}23$ (linea vertical).



(a) Media del posterior con priori no informativo

(b) Predicción en $x = -0{,}23$

Figura 13: El balance natural entre modelos

En la figura 13b hacemos un corte en línea punteada ($x = -0{,}23$) y graficamos la predicción que hace el modelo más rígido (grado 0), el modelo simple (grado 3) y el modelo más complejo (grado 9) luego de ver el cuarto dato, vamos a observar tres distribuciones gaussianas con diferente media y diferente desvío estándar. Podemos ver que el modelo más rígidos (grado 0) concentra su creencia en una región lejana al valor verdadero. Los modelos simples (grado 3) y complejo (grado 9) distribuyen su creencia alrededor del verdadero valor pero de forma distinta: cuanto mayor flexibilidad tienen los modelos, mayor tendencia a distribuir la creencia en una región más amplia de valores posibles. En términos un poco más esquemáticos (con $x = 0{,}1$) lo que ocurre es.

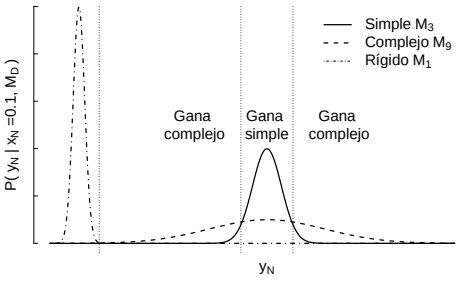


Figura 14: Esquema de la predicción de los modelos

De esta forma, el modelo rígido se rechaza por su incapacidad de llevar su creencia a la región correcta. Pero entre el modelo simple y el modelo complejo hay una balance en el cual existen ciertas regiones donde el modelo simple gana y ciertas regiones donde el modelo más complejo gana.

4. Efecto causal del sexo biológico sobre la altura.

Vamos a utilizar un conjunto de datos sobre alturas de un grupo de personas y 3 variables adicionales: sexo biológico, contextura de la madre, y altura de la madre. En este ejemplo vamos a proponer modelos causales alternativos entre estas variables.

4.1. Abrir el archivo `alturas.csv` y visualizar los datos

El archivo de datos tiene la siguiente estructura,.

	id	altura	sexo	contextura_madre	altura_madre
0	1	172.7	M	mediana	159.8
1	2	171.5	M	mediana	160.3
2	3	162.6	F	mediana	160.5
3	4	174.1	M	mediana	159.8
4	5	168.3	M	mediana	158.3

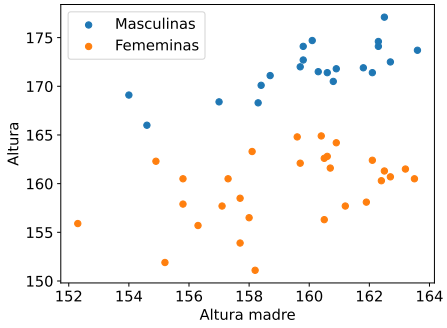


Figura 15: Datos de alturas

4.2. Definir 3 modelos causales alternativos

Si bien los datos son de dudosa procedencia, vamos a jugar un juego de inferencia causal. Vamos a proponer y evaluar tres modelos causales alternativos. En el modelo base vamos a suponer que la altura de la madre tienen un efecto causal lineal sobre la altura de su descendencia.

$$\text{Modelo Base: } \texttt{altura} = h_0 + h_1 \cdot \texttt{altura\_madre}$$

En el modelo biológico vamos a suponer que el sexo tiene un efecto causal adicional sobre la altura.

$$\text{Modelo Biológico: } \texttt{altura} = h_0 + h_1 \cdot \texttt{altura\_madre} + h_2 \cdot \mathbb{I}(\texttt{sexo} = F)$$

Cuando el sexo es masculino, la función identidad $\mathbb{I}(\texttt{sexo} = M)$ vale 1 y en caso contrario vale 0, lo que prende y apaga la hipótesis h_2, que reprersentar el efecto causal adicional del sexo biológico.

Por último vamos a plantear el modelo identitario, en el que vamos a suponer que la identidad de la persona (y no el sexo) tiene efecto causal sobre la altura.

$$\text{Modelo grupos al azar: } \texttt{altura} = h_0 \cdot \texttt{altura\_madre} + h_{1+(\texttt{ID mod (max(ID)/2)})}$$

Notar que los grupos al azar son todos de tamaño 2 y que cada grupo tiene su propia ordenada al origen pero comparten la misma pendiente. Es decir, cuando queramos usar `OLS(Y,phi)` o `log_evidence(Y,phi)`, el `phi` va a ser una matriz con 26 columnas: la primera con la altura de la madre y las siguiente 25 columnas con 0 y 1, indicando quién pertenece a un determinado grupo.

4.3. Computar la evidencia de los modelos causales alternativos

En la siguiente figura graficamos los órdenes de magnitud de evidencia ($P(\text{Datos}|\text{Modelo})$) de los 3 modelos causales alternativos.

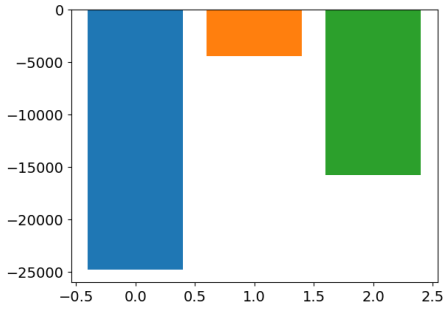


Figura 16: Evidencia en escala logarítmica de los modelos causales alternativos

4.4. Computar la media geométrica de los modelos causales alternativos

La media geométrica

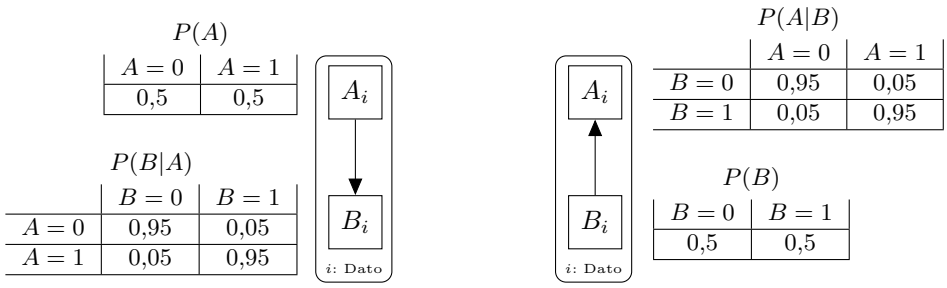
$$P(\text{Datos} = \{d_1, d_2, \ldots, d_N\}|M) = P(d_1|M)P(d_2|d_1, M)\cdots = \prod_{i\in\{1,\ldots,N\}} \underbrace{(P(d_1|M)P(d_2|d_1, M)\ldots)^{1/N}}_{\text{Media geométrica}}$$

4.5. Computar el posterior de los modelos

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Datos}|\text{Modelo})P(\text{Modelo})}{P(\text{Datos})}$$

5. Modelos AcausaB vs Modelo BcausaA

Hasta ahora pudimos usar las reglas de las probabilidad para evaluar modelos causales alternativos. En este ejercicios intentaremos descubrir la dirección de una relación causal entre dos variables A y B.



(a) Modelo AcausaB (b) Modelo BcausaA

Figura 17: Incertidumbre sobre la dirección de la relación causal

5.1. Generar datos con el modelo AcausaB

Vamos a suponer que el modelo AcausaB representa perfectamente la realidad causal subyacente. Generar un conjunto de datos con $T = 16$ episodios.

$$\text{Datos} = \{(a_0, b_0), \ldots, (a_{15}, b_{15})\}$$

5.2. Evaluar el desempeño predictivo de los modelos causales alternativos sobre los datos sintéticos generados en el punto anterior

$$P(\text{Datos}|\text{Modelo}) = \prod_t P(\text{Episodio}_t = (a_t, b_t)|\text{Modelo})$$

5.3. Actualizar la creencia respecto de los modelos causales alternativos luego de ver los datos.

Calcular el posterior de los modelos dado los datos.

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Datos}|\text{Modelo})P(\text{Modelo})}{P(\text{Datos})}$$

5.4. Dar sus conclusiones.

Explicar el posterior de los modelos obtenido.

Referencias

[1] Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012.

[2] Murphy KP. Probabilistic Machine Learning: An introduction. MIT Press; 2022.

[3] Dangauthier P, Herbrich R, Minka T, Graepel T. Trueskill through time: Revisiting the history of chess. In: Advances in Neural Information Processing Systems; 2008. p. 337–344.

[4] Stern DH, Herbrich R, Graepel T. Matchbox: large scale online bayesian recommendations. In: Proceedings of the 18th international conference on World wide web; 2009. p. 111–120.

[5] Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. Adaptive computation and machine learning. MIT Press; 2006.

[6] Bishop CM, Bishop H. Deep learning: Foundations and concepts. Springer Nature; 2023.

6. Anexo 1. Gaussianas en una dimensión

6.1. Producto de gaussianas

El problema que tenemos que resolver es

$$\int \mathcal{N}(x|\mu_1, \sigma_1^2)\mathcal{N}(x|\mu_2, \sigma_2^2)dx \tag{3}$$

Por defnición,

$$\begin{aligned}\mathcal{N}(x|y, \beta^2)\mathcal{N}(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \\ &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\underbrace{\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right)}_{\theta}\right)\end{aligned} \tag{4}$$

Luego,

$$\theta = \frac{\sigma_2^2(x^2 + \mu_1^2 - 2x\mu_1) + \sigma_1^2(x^2 + \mu_2^2 - 2x\mu_2)}{2\sigma_1^2\sigma_2^2} \tag{5}$$

Expando y reordeno los factores por potencias de x

$$\frac{(\sigma_1^2 + \sigma_2^2)x^2 - (2\mu_1\sigma_2^2 + 2\mu_2\sigma_1^2)x + (\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2)}{2\sigma_1^2\sigma_2^2} \tag{6}$$

Divido al numerador y el denominador por el factor de x^2

$$\frac{x^2 - 2\frac{(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)}{(\sigma_1^2+\sigma_2^2)}x + \frac{(\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2)}{(\sigma_1^2+\sigma_2^2)}}{2\frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2+\sigma_2^2)}} \tag{7}$$

Esta ecuación es cuadrática en x, y por lo tanto es proporcional a una función de densidad gausiana con desvío

$$\sigma_\times = \sqrt{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}} \tag{8}$$

y media

$$\mu_\times = \frac{(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)}{(\sigma_1^2 + \sigma_2^2)} \tag{9}$$

Dado que un término $\varepsilon = 0$ puede ser agregado para completar el cuadrado en θ, esta prueba es suficiente cuando no se necesita una normalización.

$$\varepsilon = \frac{\mu_\times^2 - \mu_\times^2}{2\sigma_\times^2} = 0 \tag{10}$$

Al agregar este término a θ tenemos

$$\theta = \frac{x^2 - 2\mu_\times x + \mu_\times^2}{2\sigma_\times^2} + \underbrace{\frac{\frac{(\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2)}{(\sigma_1^2+\sigma_2^2)} - \mu_\times^2}{2\sigma_\times^2}}_{\varphi} \tag{11}$$

Reorganizando φ

$$\begin{aligned}
\varphi &= \frac{\frac{(\mu_1^2\sigma_2^2+\mu_2^2\sigma_1^2)}{(\sigma_1^2+\sigma_2^2)} - \left(\frac{(\mu_1\sigma_2^2+\mu_2\sigma_1^2)}{(\sigma_1^2+\sigma_2^2)}\right)^2}{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}} \\
&= \frac{(\sigma_1^2+\sigma_2^2)(\mu_1^2\sigma_2^2+\mu_2^2\sigma_1^2) - (\mu_1\sigma_2^2+\mu_2\sigma_1^2)^2}{\sigma_1^2+\sigma_2^2}\frac{1}{2\sigma_1^2\sigma_2^2} \\
&= \frac{(\mu_1^2\sigma_1^2\sigma_2^2 + \cancel{\mu_2^2\sigma_1^4} + \cancel{\mu_1^2\sigma_2^4} + \mu_2^2\sigma_1^2\sigma_2^2) - (\cancel{\mu_1^2\sigma_2^4} + 2\mu_1\mu_2\sigma_1^2\sigma_2^2 + \cancel{\mu_2^2\sigma_1^4})}{\sigma_1^2+\sigma_2^2}\frac{1}{2\sigma_1^2\sigma_2^2} \\
&= \frac{(\sigma_1^2\sigma_2^2)(\mu_1^2+\mu_2^2-2\mu_1\mu_2)}{\sigma_1^2+\sigma_2^2}\frac{1}{2\sigma_1^2\sigma_2^2} = \frac{\mu_1^2+\mu_2^2-2\mu_1\mu_2}{2(\sigma_1^2+\sigma_2^2)} = \frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}
\end{aligned} \tag{12}$$

Luego,

$$\theta = \frac{(x-\mu_\times)^2}{2\sigma_\times^2} + \frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)} \tag{13}$$

Colocando θ en su lugar

$$\begin{aligned}
\mathcal{N}(x|y,\beta^2)\mathcal{N}(x|\mu,\sigma^2) &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\underbrace{\left(\frac{(x-\mu_\times)^2}{2\sigma_\times^2} + \frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right)}_{\theta}\right) \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{(x-\mu_\times)^2}{2\sigma_\times^2}\right)\exp\left(-\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right)
\end{aligned} \tag{14}$$

Multiplicando por $\sigma_\times\sigma_\times^{-1}$

$$\frac{\overbrace{\cancel{\sigma_1\sigma_2}}^{\sigma_\times}}{\sqrt{\sigma_1^2+\sigma_2^2}}\frac{1}{\sigma_\times}\frac{1}{2\pi\cancel{\sigma_1\sigma_2}}\exp\left(-\frac{(x-\mu_\times)^2}{2\sigma_\times^2}\right)\exp\left(-\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right) \tag{15}$$

Luego,

$$\frac{1}{\sqrt{2\pi}\sigma_\times}\exp\left(-\frac{(x-\mu_\times)^2}{2\sigma_\times^2}\right)\frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left(-\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right) \tag{16}$$

Retonando a la integral

$$\begin{aligned}
I &= \int \mathcal{N}(x|\mu_\times,\sigma_\times^2)\ \overbrace{\mathcal{N}(\mu_1|\mu_2,\sigma_1^2+\sigma_2^2)}^{\text{Escalar independiente de x}}\ dx \\
&= \mathcal{N}(\mu_1|\mu_2,\sigma_1^2+\sigma_2^2)\underbrace{\int \mathcal{N}(x|\mu_\times,\sigma_\times^2)dx}_{\text{Integra 1}} \\
&= \mathcal{N}(\mu_1|\mu_2,\sigma_1^2+\sigma_2^2)
\end{aligned} \tag{17}$$

6.2. Suma de gaussianas

Demostración por inducción,

Casos base

$$P(1) := \int \delta(t_1 = x_1)\mathcal{N}(x_1|\mu_1,\sigma_1^2)dx_1 = \mathcal{N}(t_1|\mu_1,\sigma_1^2) \tag{18}$$

La proposición $P(1)$ es verdadera dada las propiedades de la función delta de dirac.

$$\begin{aligned} P(2) &:= \iint \delta(t_2 = x_1 + x_2)\mathcal{N}(x_1|\mu_1, \sigma_1^2)\mathcal{N}(x_2|\mu_2, \sigma_2^2)dx_1dx_2 \\ &\overset{19,1}{=} \int \mathcal{N}(x_1|\mu_1, \sigma_1^2)\mathcal{N}(t_2 - x_1|\mu_2, \sigma_2^2)dx_1 \\ &\overset{19,2}{=} \int \mathcal{N}(x_1|\mu_1, \sigma_1^2)\mathcal{N}(x_1|t_2 - \mu_2, \sigma_2^2)dx_1 \\ &\overset{*}{=} \int \underbrace{N(t_2|\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)}_{\text{const.}}\underbrace{\mathcal{N}(x_1|\mu_*, \sigma_*^2)dx_1}_{1} \\ &= \mathcal{N}(t_2|\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \end{aligned} \tag{19}$$

Donde $\overset{19,1}{=}$ vale por las propiedades de la función delta de dirac, $\overset{19,2}{=}$ vale por la simetría de las gaussianas, y $\overset{*}{=}$ vale por la demostración de miltiplicación de normales en la sección 6.1. Luego, vale $P(2)$.

Paso inductivo $P(n) \Rightarrow P(n+1)$

Dado,

$$P(n) := \int \cdots \int \delta(t_n = \sum_{i=1}^{n} x_i)\left(\prod_{i=1}^{n}\mathcal{N}(x_i|\mu_i, \sigma_i^2)\right)dx_1 \ldots dx_n = \mathcal{N}(t|\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2) \tag{20}$$

Queremos ver que $P(n+1)$ es válida.

$$P(n+1) := \int \cdots \int \delta(t_{n+1} = x_{n+1} + \sum_{i=1}^{n} x_i)\left(\prod_{i=1}^{n}\mathcal{N}(x_i|\mu_i, \sigma_i^2)\right)\mathcal{N}(x_{n+1}|\mu_{n+1}, \sigma_{n+1}^2)dx_1 \ldots dx_n dx_{n+1} \tag{21}$$

Por independencia

$$\int \mathcal{N}(x_{n+1}|\mu_{n+1}, \sigma_{n+1}^2)\left(\int \cdots \int \delta(t_{n+1} = x_{n+1} + \sum_{i=1}^{n} x_i)\left(\prod_{i=1}^{n}\mathcal{N}(x_i|\mu_i, \sigma_i^2)\right)dx_1 \ldots dx_n\right)dx_{n+1} \tag{22}$$

Por hipótesis inductiva

$$\int \mathcal{N}(x_{n+1}|\mu_{n+1}, \sigma_{n+1}^2)\mathcal{N}(t_{n+1} - x_{n+1}|\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2)dx_{n+1} \tag{23}$$

Por la demostración de la sección 6.1

$$\mathcal{N}(t_{n+1}|\mu_{n+1} + \sum_{i=1}^{n}\mu_i, \sigma_{n+1}^2\sum_{i=1}^{n}\sigma_i^2)dx_{n+1} \tag{24}$$

Luego, $P(n+1)$ es válida.

6.3. Gaussiana por acumulada de Gaussiana.

Queremos resolver la integral

$$f(x) = \int \mathcal{N}(y; \mu_1, \sigma_1^2)\Phi(y + x; \mu_2, \sigma_2^2)dy \tag{25}$$

Para ello trabajamos con la drivada $\frac{\partial}{\partial x}f(x) = \theta(x)$,

$$\theta(x) = \frac{\partial}{\partial x}\int \mathcal{N}(y; \mu_1, \sigma_1^2)\Phi(y + x|\mu_2, \sigma_2^2)dy \tag{26}$$

$$\theta(x) = \int \mathcal{N}(y|\mu_1, \sigma_1^2) \frac{\partial}{\partial x} \Phi(y + x|\mu_2, \sigma_2^2) dy \tag{27}$$

La derivada de Φ es justamente una Gaussiana,

$$\begin{aligned} \theta(x) &= \int \mathcal{N}(y|\mu_1, \sigma_1^2) \mathcal{N}(y + x|\mu_2, \sigma_2^2) dy \\ &= \int \mathcal{N}(y|\mu_1, \sigma_1^2) \mathcal{N}(y|\mu_2 - x, \sigma_2^2) dy \end{aligned} \tag{28}$$

Por la demostración de la sección 6.1 sabemos

$$\theta(x) = \mathcal{N}(\mu_1|\mu_2 - x, \sigma_1^2 + \sigma_2^2) \tag{29}$$

Por simetría

$$\theta(x) = \mathcal{N}(x|\mu_2 - \mu_1, \sigma_1^2 + \sigma_2^2) \tag{30}$$

Retornando a $f(x)$

$$f(x) = \Phi(x|\mu_2 - \mu_1, \sigma_1^2 + \sigma_2^2) \tag{31}$$

6.4. Division de gaussianas

$$\kappa = \frac{\mathcal{N}(x|\mu_f, \sigma_f^2)}{\mathcal{N}(x|\mu_g, \sigma_g^2)} = \mathcal{N}(x|\mu_f, \sigma_f^2) \mathcal{N}(x|\mu_g, \sigma_g^2)^{-1} \tag{32}$$

Por definición

$$\begin{aligned} \kappa &= \frac{1}{\sqrt{2\pi}\sigma_f} e^{-\left(\frac{(x-\mu_f)^2}{2\sigma_f^2}\right)} \left(\frac{1}{\sqrt{2\pi}\sigma_g} e^{-\left(\frac{(x-\mu_g)^2}{2\sigma_g^2}\right)} \right)^{-1} \\ &= \frac{1}{\cancel{\sqrt{2\pi}}\sigma_f} e^{-\left(\frac{(x-\mu_f)^2}{2\sigma_f^2}\right)} \frac{\cancel{\sqrt{2\pi}}\sigma_g}{1} e^{\left(\frac{(x-\mu_g)^2}{2\sigma_g^2}\right)} \\ &= \frac{\sigma_g}{\sigma_f} \exp\Bigg(- \underbrace{\left(\frac{(x-\mu_f)^2}{2\sigma_f^2} - \frac{(x-\mu_g)^2}{2\sigma_g^2} \right)}_{\theta} \Bigg) \end{aligned} \tag{33}$$

Reorganizando θ

$$\begin{aligned} \theta &= \frac{(x-\mu_f)^2}{2\sigma_f^2} - \frac{(x-\mu_g)^2}{2\sigma_g^2} = \frac{\sigma_g^2(x-\mu_f)^2 - \sigma_f^2(x-\mu_g)^2}{2\sigma_f^2\sigma_g^2} \\ &= \frac{\sigma_g^2(x^2 + \mu_f^2 - 2\mu_f x) - \sigma_f^2(x^2 + \mu_g^2 - 2\mu_g x)}{2\sigma_f^2\sigma_g^2} \end{aligned} \tag{34}$$

Expandimos y ordenamos en base x,

$$\begin{aligned} \theta &= \left((\sigma_g^2 - \sigma_f^2)x^2 - 2(\sigma_g^2\mu_f - \sigma_f^2\mu_g)x + (\sigma_g^2\mu_f^2 - \sigma_f^2\mu_g^2) \right) \frac{1}{2\sigma_f^2\sigma_g^2} \\ &= \left(x^2 - \frac{2(\sigma_g^2\mu_f - \sigma_f^2\mu_g)}{(\sigma_g^2 - \sigma_f^2)} x + \frac{(\sigma_g^2\mu_f^2 - \sigma_f^2\mu_g^2)}{(\sigma_g^2 - \sigma_f^2)} \right) \frac{(\sigma_g^2 - \sigma_f^2)}{2\sigma_f^2\sigma_g^2} \end{aligned} \tag{35}$$

Esto es cuadrático en x. Dado que un término $\varepsilon = 0$, independiente de x puede ser agregado para completar el cuadrado en θ, esta prueba es suficiente para dterminar la media y la varianza cuando no es necesario normalizar.

$$\sigma_\div = \sqrt{\frac{\sigma_f^2\sigma_g^2}{(\sigma_g^2 - \sigma_f^2)}} \tag{36}$$

$$\mu_\div = \frac{(\sigma_g^2\mu_f - \sigma_f^2\mu_g)}{(\sigma_g^2 - \sigma_f^2)} \tag{37}$$

agregado $\varepsilon = \frac{\mu_\div^2 - \mu_\div^2}{2\sigma_\div^2} = 0$

$$\theta = \frac{x^2 - 2\mu_\div x + \mu_\div^2}{2\sigma_\div^2} + \underbrace{\frac{\frac{(\sigma_g^2\mu_f^2 - \sigma_f^2\mu_g^2)}{(\sigma_g^2 - \sigma_f^2)} - \mu_\div^2}{2\sigma_\div^2}}_{\varphi} \tag{38}$$

Reorganizando φ

$$\begin{aligned}
\varphi &= \left(\frac{(\sigma_g^2\mu_f^2 - \sigma_f^2\mu_g^2)}{(\sigma_g^2 - \sigma_f^2)} - \left(\frac{(\sigma_g^2\mu_f - \sigma_f^2\mu_g)}{(\sigma_g^2 - \sigma_f^2)}\right)^2\right)\frac{(\sigma_g^2 - \sigma_f^2)}{2\sigma_f^2\sigma_g^2} \\
&= \left((\sigma_g^2\mu_f^2 - \sigma_f^2\mu_g^2)(\sigma_g^2 - \sigma_f^2) - \left((\sigma_g^2\mu_f - \sigma_f^2\mu_g)\right)^2\right)\frac{1}{2\sigma_f^2\sigma_g^2(\sigma_g^2 - \sigma_f^2)} \\
&= \left(\cancel{\sigma_g^4\mu_f^2} - \sigma_f^2\sigma_g^2\mu_f^2 - \sigma_f^2\sigma_g^2\mu_g^2 + \cancel{\sigma_f^4\mu_g^2} - \left(\cancel{\sigma_g^4\mu_f^2} + \cancel{\sigma_f^4\mu_g^2} - 2\sigma_f^2\sigma_g^2\mu_f\mu_g\right)\right)\frac{1}{2\sigma_f^2\sigma_g^2(\sigma_g^2 - \sigma_f^2)}
\end{aligned} \tag{39}$$

Cancelando $\sigma_f^2\sigma_g^2$

$$\varphi = \frac{-\mu_g^2 - \mu_f^2 + 2\mu_f\mu_g}{2(\sigma_g^2 - \sigma_f^2)} = \frac{-(\mu_g - \mu_f)^2}{2(\sigma_g^2 - \sigma_f^2)} \tag{40}$$

Luego θ

$$\theta = \frac{(x - \mu_\div)^2}{2\sigma_\div^2} - \frac{(\mu_g - \mu_f)^2)}{2(\sigma_g^2 - \sigma_f^2)} \tag{41}$$

Volviendo a la expresión original

$$\begin{aligned}
\kappa &= \frac{\sigma_g}{\sigma_f}\exp\left(-\frac{(x - \mu_\div)^2}{2\sigma_\div^2} + \frac{(\mu_g - \mu_f)^2)}{2(\sigma_g^2 - \sigma_f^2)}\right) \\
&= \frac{\sigma_g}{\sigma_f}\exp\left(-\frac{(x - \mu_\div)^2}{2\sigma_\div^2}\right)\exp\left(\frac{(\mu_g - \mu_f)^2)}{2(\sigma_g^2 - \sigma_f^2)}\right)
\end{aligned} \tag{42}$$

Multiplicando por $\frac{\sqrt{2\pi}}{\sqrt{2\pi}}\frac{\sigma_\div}{\sigma_\div}\frac{\sqrt{\sigma_g^2 - \sigma_f^2}}{\sqrt{\sigma_g^2 - \sigma_f^2}} = 1$,

$$\begin{aligned}
\kappa &= \frac{1}{\sqrt{2\pi}\sigma_\div}e^{-\frac{(x-\mu_\div)^2}{2\sigma_\div^2}}\left(\frac{1}{\sqrt{2\pi(\sigma_g^2 - \sigma_f^2)}}e^{-\frac{(\mu_g - \mu_f)^2)}{2(\sigma_g^2 - \sigma_f^2)}}\right)^{-1}\frac{\sigma_\div}{\sqrt{\sigma_g^2 - \sigma_f^2}}\frac{\sigma_g}{\sigma_f} \\
&= \frac{\mathcal{N}\left(x|\mu_\div, \sigma_\div\right)}{\mathcal{N}\left(\mu_g|\mu_f, \sigma_g^2 - \sigma_f^2\right)}\frac{\sigma_g^2}{\sigma_g^2 - \sigma_f^2}
\end{aligned} \tag{43}$$

7. Anexo 2. Gaussiana multivariada

La distribución gaussiana multivariada del vector D-dimensional $\mathbf{x}$ es

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

donde $\boldsymbol{\mu}$ es un vector D-dimensional que representa la media, $\boldsymbol{\Sigma}$ es una matriz $D \times D$ que representa la covarianza y $|\Sigma|$ su determinante.

7.1. Distribuciones condicionales y marginales

Dado la distribución conjunta

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$$

tal que

$$\mathbf{x} = \begin{pmatrix}\mathbf{x}_a\\ \mathbf{x}_b\end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix} \quad \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Lambda} = \begin{pmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab}\\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{pmatrix}$$

Queremos encontrar,

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b},\boldsymbol{\Sigma}_{a|b})$$
$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_a)$$

Notar que $\boldsymbol{\Lambda}_{aa} \neq \boldsymbol{\Sigma}_{aa}^{-1}$

7.1.1. Distribución condicional

En primer lugar notamos que la distribución condicional es proporcional a la conjunta, y que ella es proporcional al exponente de la distribución gaussiana multivariada.

$$p(\mathbf{x}_a|\mathbf{x}_b) \propto p(\mathbf{x}_a,\mathbf{x}_b) \propto \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)$$

Primero reordenamos los términos en su interior y los agrupamos por el grado en $\mathbf{x}$.

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\underbrace{\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}_{\text{Cuadrático}} + \underbrace{\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}_{\text{Lineal}} + \text{const}$$

Esto mismo lo podemos hacer con un nivel de detalle superior, usando la definición de la precisión $\boldsymbol{\Sigma}^{-1}$ (inversa de la covarianza) para descomponer los términos, agrupando por el grado en $\mathbf{x}_a$.

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\\
&-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)\\
&=\\
&-\frac{1}{2}\mathbf{x}_a^T\,\boldsymbol{\Lambda}_{aa}\,\mathbf{x}_a + \mathbf{x}_a^T\big(\,\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\,\big) + \text{const}\\
&-\frac{1}{2}\mathbf{x}_a^T\underbrace{\boldsymbol{\Lambda}_{aa}}_{\boldsymbol{\Sigma}_{a|b}^{-1}}\mathbf{x}_a + \mathbf{x}_a^T\big(\underbrace{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)}_{\boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}}\big) + \text{const}
\end{aligned}$$

Donde $\mathbf{\Lambda}_{aa}$ va a representar la precisión de la distribución condicional $\mathbf{\Sigma}_{a|b}^{-1}$ y el factor que acompaña a $\mathbf{x}$ va a representar el producto entre la precisión y la media de la distribución condicional $\mathbf{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}$.

$$\begin{aligned}\mathbf{\Sigma}_{a|b} = \mathbf{\Lambda}_{aa}^{-1} \qquad \boldsymbol{\mu}_{a|b} &= \mathbf{\Sigma}_{a|b}\Big(\mathbf{\Lambda}_{aa}\boldsymbol{\mu}_a - \mathbf{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\Big) \\ &= \boldsymbol{\mu}_a - \mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\end{aligned}$$

Por lo tanto

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \mathbf{\Lambda}_{aa}^{-1})$$

Distribución condicional en función de la covarianza Éste resultado está expresado usando uno de los elementos que componen la matriz de precisión de la distribución conjunta. También podríamos expresar el resultado usando la matriz de covarianzas de la distribución conjunta. Para ello es necesario revisar cómo la inversa de la matriz de covarianzas está relacionada con la matriz de precisión.

$$\begin{pmatrix}\mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb}\end{pmatrix}$$

Es importante notar primero que la inversa de cualquier matriz cumple con la siguiente propiedad.

$$\begin{pmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}\end{pmatrix}$$

Por lo tanto,

$$\begin{pmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{M} & -\mathbf{MBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1}\mathbf{CMBD}^{-1}\end{pmatrix} = \begin{pmatrix}\mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}\end{pmatrix}$$

$$\text{con } \mathbf{M} = (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}$$

Lo que define la siguiente relación.

$$\begin{aligned}\mathbf{\Lambda}_{aa} &= (\mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba})^{-1} \\ \mathbf{\Lambda}_{ab} &= -(\mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba})^{-1}\mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\end{aligned}$$

7.1.2. Distribución marginal

Vimos que si la distribución conjunta $p(\mathbf{x}_a, \mathbf{x}_b)$ es gaussiana, la condicional $p(\mathbf{x}_a|\mathbf{x}_b)$ también lo es. En esta sección veremos que la distribución marginal $P(\mathbf{x}_a)$ también será gaussiana.

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$$

Para integrar sobre $\mathbf{x}_b$ vamos a considerar los términos que involucren a $\mathbf{x}_b$ para facilitar la integración.

$$-\frac{1}{2}\mathbf{x}_b^T\mathbf{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b^T \underbrace{(\mathbf{\Lambda}_{bb}\boldsymbol{\mu}_b - \mathbf{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a))}_{\mathbf{m}}$$

Si completamos los cuadrados.

$$-\frac{1}{2}\mathbf{x}_b^T\mathbf{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b^T\mathbf{m} = -\frac{1}{2}\mathbf{x}_b^T\mathbf{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b^T\mathbf{m}\overbrace{-\frac{1}{2}\mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1}\mathbf{m} + \frac{1}{2}\mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1}\mathbf{m}}^{\mathbf{0}}$$
$$= \underbrace{-\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^T\mathbf{\Lambda}_{bb}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})}_{\text{Exponente de la Gaussiana}} + \underbrace{\frac{1}{2}\mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1}\mathbf{m}}_{\text{Libre de } \mathbf{x}_b}$$

Luego, la integración se realizará sobre una gaussiana no normalizada, y por lo tanto el resultado va a ser el recíproco del coeficiente de normalización, que depende de $\mathbf{x}_a$.

$$p(\mathbf{x}_a) = f(\mathbf{x}_a)\int \exp\left(-\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^T\mathbf{\Lambda}_{bb}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})\right) d\mathbf{x}_b$$

Esta integración se realiza fácilmente cuando notamos que corresponde a la integral de una Gaussiana no normalizada, por lo que el resultado será el recíproco del coeficiente de normalización. Sabemos, por la definición de la Gaussiana normalizada, que este coeficiente es independiente de la media y depende solo del determinante de la matriz de covarianza. Así, al completar el cuadrado con respecto a $\mathbf{x}_b$, podemos integrar $\mathbf{x}_b$, quedando los siguientes términos como recíproco.

$$f(\mathbf{x}_a)\frac{1}{2}\mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1}\mathbf{m} - \frac{1}{2}\mathbf{x}_a^T\mathbf{\Lambda}_{aa}\mathbf{x} + \mathbf{x}_a^T(\mathbf{\Lambda}_{aa}\boldsymbol{\mu}_a + \mathbf{\Lambda}_{ab}\boldsymbol{\mu}_b) + \text{const}$$
$$\overset{*}{=} -\frac{1}{2}\mathbf{x}_a^T\underbrace{(\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})}_{\mathbf{\Sigma}_a^{-1}}\mathbf{x}_a + \mathbf{x}_a^T\underbrace{(\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})^{-1}\boldsymbol{\mu}_a}_{\mathbf{\Sigma}_a^{-1}\boldsymbol{\mu}_a} + \text{const}$$

donde const refiere a elementos independientes de $\mathbf{x}_a$. Luego,

$$\mathbf{\Sigma}_a = \underbrace{(\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})^{-1}}_{\mathbf{\Sigma}_{aa}}$$

y por lo tanto la distribución gaussiana marginal es,

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \mathbf{\Sigma}_{aa})$$

7.2. Modelo lineal multivariado

Dadas un modelo lineal

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{w} + \mathbf{b}, \mathbf{L}^{-1})$$

Queremos probar,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\cdot, \cdot)$$
$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{y}|\cdot, \cdot)$$

Vamos a usar la probabilidad conjunta

$$\mathbf{z} = \begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}$$

Tomando el logaritmo

$$\log p(\mathbf{z}) = \log p(w) + \log p(y|w)$$
$$= -\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^T\mathbf{\Lambda}(\mathbf{w} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{w} - \mathbf{b})^T\mathbf{L}(\mathbf{w} - \mathbf{A}\mathbf{w} - \mathbf{b}) + \text{const}$$

Términos cuadráticos Analizamos primero los términos cuadráticos

$$
\begin{aligned}
&-\frac{1}{2}\mathbf{w}^T(\mathbf{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A})\mathbf{w}-\frac{1}{2}\mathbf{y}^T(\mathbf{L})\mathbf{y}+\frac{1}{2}\mathbf{y}^T(\mathbf{L}\mathbf{A})\mathbf{w}+\frac{1}{2}\mathbf{w}^T(\mathbf{A}^T\mathbf{L})\mathbf{y}\\
&=-\frac{1}{2}\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}^T\underbrace{\begin{pmatrix}\mathbf{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A} & -\mathbf{A}^T\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}}_{\mathbf{R}}\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}
\end{aligned}
$$

Donde la covarianza

$$
\text{cov}[\mathbf{z}]=\mathbf{R}^{-1}=\begin{pmatrix}\mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^T\\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{L}^{-1}+\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T\end{pmatrix}
$$

Términos lineales Analizamos ahora los términos lineales.

$$
\mathbf{w}^T\mathbf{\Lambda}\boldsymbol{\mu}-\mathbf{w}^T\mathbf{A}^T\mathbf{L}\mathbf{b}+\mathbf{y}^T\mathbf{L}\mathbf{b}=\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}^T\begin{pmatrix}\mathbf{\Lambda}\boldsymbol{\mu}-\mathbf{A}^T\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix}
$$

Donde la media es

$$
\text{E}[\mathbf{z}]=\mathbf{R}^{-1}\begin{pmatrix}\mathbf{\Lambda}\boldsymbol{\mu}-\mathbf{A}^T\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix}=\begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{A}\boldsymbol{\mu}+\mathbf{b}\end{pmatrix}
$$

7.2.1. Evidencia y posterior

Hemos visto que dado,

$$
p(\mathbf{w})=\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\mathbf{\Lambda}^{-1})\qquad p(\mathbf{y}|\mathbf{w})=\mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{w}+\mathbf{b},\mathbf{L}^{-1})
$$

la distribución conjunta es,

$$
\mathbf{z}=\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}\quad \text{E}_{\mathbf{z}}=\begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{A}\boldsymbol{\mu}+\mathbf{b}\end{pmatrix}\quad \text{cov}_{\mathbf{z}}=\begin{pmatrix}\mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^T\\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{L}^{-1}+\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T\end{pmatrix}\quad \text{cov}_{\mathbf{z}}^{-1}=\begin{pmatrix}\mathbf{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A} & -\mathbf{A}^T\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}
$$

Luego, usando las propiedades de la marginal y condicional sabemos que,

$$
p(\mathbf{y})=\mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\ \ \mathbf{L}^{-1}+\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T)
$$

$$
p(\mathbf{w}|\mathbf{y})=\mathcal{N}(\mathbf{w}|\mathbf{\Sigma}\left(\mathbf{A}^T\mathbf{L}(\mathbf{y}-\mathbf{b})+\mathbf{\Lambda}\boldsymbol{\mu}\right),\ \ \mathbf{\Sigma})
$$

$$
\text{con } \mathbf{\Sigma}=(\mathbf{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}
$$