
Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Departamento de Física Juan José Giambiagi

Tesis de Licenciatura

# Inferencia causal en estudios observacionales usando modelos bayesianos

*(aplicada a deserción estudiantil)*

**Leila Sofía Asplanato**

Director: Dr. Rodrigo Diaz

Codirector: Dr. Ezequiel Alvarez

Marzo 2023

TEMA: Inferencia causal en estudios observacionales usando modelos bayesianos (aplicada a deserción estudiantil)
ALUMNA: Leila Sofía Asplanato
LU N°: 521/14
LUGAR DE TRABAJO: ICAS, UNSAM - Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires
DIRECTOR DEL TRABAJO: Dr. Rodrigo Díaz
CODIRECTOR DEL TRABAJO: Dr. Ezequiel Álvarez
FECHA DE INICIACIÓN: Marzo de 2022
FECHA DE FINALIZACIÓN: Marzo de 2023
FECHA DE EXAMEN: 30 de Marzo 2023

# Agradecimientos

A mi familia, por ser mi sostén en en el largo camino para estar acá. *Babi*, sin tus mates y desayunos difícilmente hubiera llegado al final. *Mamele* y *Guadi*, gracias por su apoyo incondicional y su creencia ciega en mi.

A mis amigas de la facultad: *Sofi B*, *Sofi A* y *Flor F*. Nada de esto hubiera sido tan placentero como lo fue sin tenerlas de referencia, compañeras de payasadas y laboratorios, risas, comidas y salidas.

A mis amigos del colegio, quienes me vieron crecer y desaparecer durante períodos estresantes de la carrera, pero aún así siguen rodeándome de cariño. Gracias.

A *Rodrigo Díaz* y *Ezequiel Álvarez* especialmente, por recibirme con los brazos abiertos y darme la oportunidad de elegir mi camino, sabiendo que éíbamos por el desconocido. Gracias por su guía durante esta aventura.

A la UBA que me ofreció educación pública y gratuita de excelencia, y al ICAS en UNSAM, por brindarme el espacio para llevarla a cabo.

A ustedes, por leer esta tesis.

# Resumen

En el siguiente trabajo se emplearon técnicas de inferencia bayesiana y análisis causal buscando explicar la deserción estudiantil de primer año de la Universidad Nacional de San Martín. Con el objetivo de cuantificar la influencia de cada causa de deserción, se aplicó el novedoso método del *deconfounder* que se destaca en los casos con presencia de múltiples causas.

Se logró aplicar el método de este estudio sobre dos bases de datos previamente analizadas por otros autores. Primero se generaron datos semi-sintéticos a partir de una base de datos de fumadores estadounidenses, aplicando el método deconfounder con una de estas variables como confundidora latente a inferir. Aquí pudieron compararse nuestros resultados con los coeficientes reales, y compararlos con la regresión si se hubiese tenido el conjunto completo de datos. También se aplicó el deconfounder sobre la base de datos de cáncer de mamas, de muestras con asignaciones de tumores malignos o benignos a partir de la observación variables morfológicas de muestras celulares. En este caso logramos obtener resultados que coincidían, mayoritariamente, con los reportados por el análisis publicado.

Finalmente, se definió el grupo de causas de la base de datos de estudiantes de la UNSAM entre 2017 y 2021, filtrando los datos de la Escuela de Humanidades en base a estudios preliminares realizados por el Licenciado Pablo Aguila. Definiendo un número de 4 variables confundidoras para 13 covariables causales se pudieron obtener coeficientes para los parámetros de medida causal, además de estimar el impacto de las variables confundidoras latentes.

# Índice general

# Capítulo 1

# Introducción

Los seres humanos nacen preguntándose el por qué de las cosas y continúan con ese comportamiento hacia la adultez. Es natural generar explicaciones de los eventos del entorno razonando *para atrás* en base a lo observado, y luego pensar *para adelante* para predecir ocurrencias futuras [1]. Sin embargo, estas asociaciones pueden devenir en falacias; si se demuestra que los techos de dos casas vecinas aparecen mojados al mismo tiempo, y también que regar un techo genera que esté mojado, podría llegar a decirse que regar un techo ocasiona que ambos se mojen.

El cliché dice que la "*correlación no implica causalidad*". Que dos cosas estén asociadas no significa que la relación entre las mismas sea del tipo causal: el techo del vecino no se moja porque el nuestro lo hace, simplemente sucede que ambos están bajo la acción de una tercera variable, como la lluvia (entre otras). Es decir, que dos variables aparezcan en cierta relación de manera más o menos constante no asegura que una causa la otra. Esta relación puede también ser una coincidencia o estar vinculada por medio de otras variables no consideradas.

La inferencia causal busca establecer la relación entre una variable (a veces llamada *intervención* o *tratamiento*) y el resultado (también llamado efecto o *"outcome"* del inglés). Formalmente se quiere cuantificar cómo el cambio de tratamiento afecta el resultado. Por ejemplo, el efecto de tomar aspirina para que se vaya el dolor de cabeza, comparado con el contrafáctico de haber dejado pasar el tiempo sin tomar dicho remedio. Los experimentos deben ser diseñados para permitir la manipulación de la causa -o variable independiente- y permitir observar si hay cambios en el efecto - la variable dependiente-. Para eso, es necesario controlar por factores que potencialmente pueden estar influyendo en la relación bajo estudio.

El aprendizaje automático (ML de las siglas del inglés *machine learning*) es una de las ramas de la inteligencia artificial que concede a las computadoras la habilidad de aprender sin ser programadas de manera explícita. Se define una tarea específica que deben realizar junto

con un conjunto de reglas que cumplir y permite que encuentre el camino óptimo para lograrlo por su cuenta. Se entrenan a partir de grandes cuerpos de datos, llamados sets o conjuntos de entrenamiento, buscando e identificando patrones para "aprender" y forman predicciones. Estas herramientas vienen desarrollándose desde la década del '50 y están bien establecidas, logrando clasificadores de dígitos, grupos y otros con un alto nivel de precisión al explotar la correlación de los datos. Sin embargo, al buscar los patrones se limita a las asociaciones inherentes de los datos y sirve para formar predicciones, pero no define y prueba las relaciones causales entre variables, identificando causa y efecto. Las redes neuronales, por ejemplo, generan excelentes predicciones según los patrones y relaciones de los datos, pero no permiten la manipulación de las variables para probar causalidad.

La inferencia bayesiana es un método estadístico basado en el teorema de Bayes que permite actualizar la probabilidad de una hipótesis a medida que se descubre nueva evidencia, y cuantifica la incerteza de un modelo estadístico al emplear distribuciones de probabilidades. Mientras que la inferencia causal se centra en entender cómo una intervención afecta un resultado, la inferencia bayesiana define el marco teórico en el que se actualiza la probabilidad de una hipótesis dada nueva evidencia. De esta forma, los métodos bayesianos pueden usarse en el marco de la inferencia causal para determinar la verosimilitud (o *likelihood*) de una relación causal entre dos variables, dados los datos disponibles.

La confluencia de aportes históricos de ciencias económicas, computación y estadística devienen en investigaciones como la de Yixin Wang y David Blei del 2019 donde se define un modelo de *deconfounder* o de-confundidora como método para controlar por posibles variables que estén alterando las relaciones causales reales buscadas, pero que no han sido observadas [2].

Si bien "*todos los modelos están mal*", "*algunos son útiles*" [3]. En la figura 1.1 se muestra de forma esquemática cómo evoluciona el ajuste de un modelo según su complejidad, ya sea por la forma funcional del modelo, o por el número de parámetros empleados. Se observa una curva superior, la bondad del ajuste (*goodness of fit*), que aumenta con la complejidad y una curva inferior, la habilidad de generalizar el modelo (*generalizability*), con máximo en una complejidad media. Los gráficos inferiores muestran el ajuste del modelo a los datos. Puede verse cómo la bondad del ajuste aumenta con la complejidad del modelo, pero se puede llegar a sobre-ajustar y pasar a captar el ruido de los datos. La habilidad de generalizar se da en casos de menor complejidad [4].

El gráfico ilustra el intercambio entre la predicción y la explicabilidad. Para tener mayor poder de interpretar el modelo y comprender por qué y cómo es que funciona se debe limitar la complejidad, perdiendo parte de la exactitud.

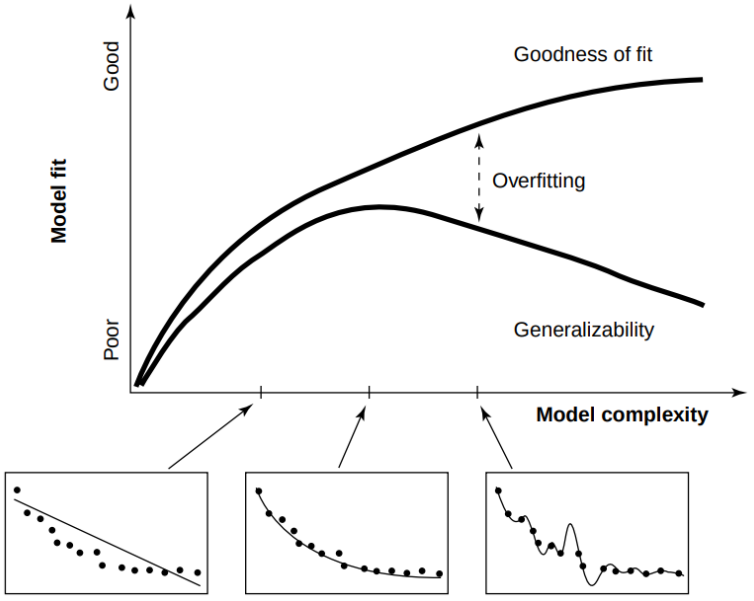



Figura 1.1: Ajuste del modelo en relación a su complejidad. Se observan la bondad del ajuste por arriba y el poder de generalización debajo. La distancia entre las mismas es el sobre-ajuste de los datos. Los cuadros inferiores muestran los datos (*puntos*) y el ajuste obtenido(*línea*).

En este trabajo se estudió la deserción estudiantil universitaria de la Universidad Nacional de San Martín (*UNSAM*). Este tópico es de gran interés en la Argentina y ha sido estudiado por diversos autores con un abordaje histórico y social [5, 6, 7]. Uno de los principales problemas que se encuentra es el de la retención, es decir el porcentaje de personas que continúan sus estudios. Se ha comprobado que sólo un $25,1\,\%$ de estudiantes termina una carrera determinada en el tiempo teórico, y aún mas llamativo, la retención de estudiantes durante el primer año universitario es del $61{,}9\,\%$ [8]. El enfoque causal permite además un estudio de posibles problemas o barreras sociales sistémicas que estén generando deserción dispareja, y puede informar políticas, estrategias y prácticas orientadas al apoyo de grupos vulnerables. Gracias a la colaboración con la Dirección General de Información, Planificación y Evaluación perteneciente a la Secretarıéa de Planificación y Evaluación de la UNSAM se obtuvo una base de datos de estudiantes de la universidad de 2017 a 2020 sobre la cual aplicar nuestro análisis.

**La tesis está organizada de la siguiente manera.**

Los primeros tres capítulos detallan el marco teórico de este trabajo. El capítulo 2 introduce conceptos de inferencia bayesiana, la definición de densidades de distribución a priori para informar los conocimientos de la situación, la verosimilitud para establecer la relación entre los parámetros de interés con las variables aleatorias observadas y la manera en que se

invierte la relación para definir distribuciones posteriores. También introduce los modelos gráficos probabilísticos. El capítulo 3 complementa al anterior, formalizando nociones de inferencia causal para considerar los resultados potenciales de cada intervención. Estos modelos pueden representarse con los modelos gráficos del capítulo previo, estableciendo enlaces causales entre las variables, cuyas distribuciones pueden inferirse con análisis bayesiano. Aquí se explaya el método del deconfounder, propuesto por los autores Wang y Blei, para cuantificar causas aún sabiendo que faltan observaciones del problema. Finalmente, el capítulo 4 muestra las herramientas computacionales que se usarán con el fin aplicar el método del deconfounder, y una introducción al lenguaje de programación empleado para llevarlo a cabo.

Las primeras dos aplicaciones del método deconfounder fueron sobre casos concretos previamente estudiados por otras publicaciones. Las bases de datos usadas y los resultados de estos dos estudios mencionados se describen en el capítulo 5.

Por último, el capítulo 6 presenta la base de datos de estudiantes universitarios de la Universidad Nacional del General San Martín, el filtrado de los mismos, la selección y manipulación de causas, así como la definición del número de variables confundidoras a analizar. Se describen los modelos seleccionados, los procedimientos y los resultados de los parámetros de interés.

Las conclusiones del trabajo y consideraciones para avances futuros se comentan en el capítulo 7.

# Capítulo 2

# Inferencia Bayesiana

Hay dos grandes escuelas empleadas para realizar inferencia estadística: la inferencia frecuentista y la inferencia bayesiana [9].

El frecuentismo se basa en usar la frecuencia de aparición de un evento "A" para dar su probabilidad de ocurrencia $P_F(A)$. Para un número N grande de repeticiones idénticas de un experimento, si se observan M casos favorables de un valor de la variable aleatoria $A = a$, entonces

$$P_F(A = a) = \frac{M}{N}.$$

Las herramientas frecuentistas pueden servir para realizar predicciones en función de datos ya existentes, pero empiezan a ser problemáticas para obtener una respuesta a nuevos eventos, en instancias que sucedan una única vez y también cuando los experimentos no pueden repetirse o por motivos éticos no pueden manipularse con las medidas requeridas.

La rama bayesiana, por su cuenta, toma distribuciones de probabilidad como una medida de la incerteza de parámetros de modelos o teorías contrarias. En este marco la *probabilidad* es una representación de nuestro conocimiento de la realidad y la *frecuencia* el valor de una propiedad medible de dicha realidad[9, Capítulo 1]. La probabilidad de "A",

$$P_B(A|B),$$

será una medida de la plausibilidad de la proposición, hipótesis o instancia de variable aleatoria "A" condicionada a la información presente dada la proposición "B". Ante la falta de conocimiento de expertos en un área, sin evidencia previa, la probabilidad bayesiana puede

inicializarse con una perspectiva frecuentista

$$P_B(A|B) \sim P_F(A).$$

Para responder, por ejemplo, la probabilidad de un accidente aéreo ambos métodos de inferencia pueden definir una probabilidad de accidente como el cociente entre el número de vuelos accidentados y el número total de vuelos hasta la fecha. Sin embargo, el análisis bayesiano permite incluir los conocimientos pre-existentes que se tengan del dominio a estudiar como punto de partida, pudiendo especificar grados de confianza. Puede ser computacionalmente más costoso, lo que ha dificultado su uso en algunas disciplinas, pero es una herramienta muy útil y flexible, y es la empleada en este trabajo.

## 2.1. Propiedades probabilísticas

Dos variables aleatorias $X$ e $Y$ pueden tomar valores $x_i$ e $y_j$, donde $i = 1, ..., M$ y $j = 1, ..., L$. Ante la repetición de un experimento $N$ veces (con $N \to \infty$), $c_i$ y $r_j$ representan la cantidad de veces que suceden $x_i$ e $y_j$ respectivamente. La cantidad $n_{ij}$ denota el número de instancias donde ocurren $x_i$ e $y_j$ en simultáneo.

La probabilidad *conjunta* de que la variable $X$ tome el valor $x_i$ cuando la variable $Y$ toma el valor $y_j$ es

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}, \tag{2.1}$$

y tiene propiedad de simetría, de manera que

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = P(Y = y_j, X = x_i),$$

es decir, la probabilidad de $X$ e $Y$ es la misma que la probabilidad de $Y$ y $X$.

La probabilidad *independiente* de la variable $X$ se expresa

$$P(X = x_i) = \frac{c_i}{N}, \tag{2.2}$$

y se extiende la definición de manera análoga para $Y$.

Las ecuaciones 2.1 y 2.2, junto con la relación entre las frecuencias

$$c_i = \sum_j n_{ij} \quad y \quad r_j = \sum_i n_{ij}$$

permiten establecer la "*regla de la suma*"

$$P(X = x_i) = \sum_{j=1}^{J} P(X = x_i, Y = y_j) \tag{2.3}$$

de donde se nota que la probabilidad independiente de una variable $X$ también se denomina *probabilidad marginal*, ya que se obtiene al marginalizar la variable $Y$ sumando sobre su espacio de posibilidades $Y = y_j$ [10, Capítulo 1].

La probabilidad *condicional* establece una relación entre dos variables

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}, \tag{2.4}$$

e implica observar sólo las instancias donde sucede $y_j$ habiendo restringido el espacio de posibilidades a donde haya ocurrido necesariamente $x_i$. Tomando las ecuaciones 2.1, 2.2 y 2.4 se llega a la "*regla del producto*"

$$\begin{aligned} P(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N} \\ &= P(Y = y_j | X = x_i) P(X = x_i). \end{aligned} \tag{2.5}$$

### 2.1.1. Densidades probabilísticas

Al considerar variables con eventos continuos se pasa de distribuciones de probabilidad a *densidades* probabilísticas. En estos casos, la probabilidad $P(x)$ pasa a ser la probabilidad que la variable $X$ se encuentre dentro del intervalo $(x, x + \delta x)$, en lugar de idéntica a un único valor. De esta forma, la probabilidad se expresa

$$P(X \in (a, b)) = \int_a^b p(x) dx,$$

donde $p(x)$ es la función de densidad de probabilidad, que debe ser siempre positiva e integrar a 1 en todo su dominio, midiendo la probabilidad relativa que una variable aleatoria tome un dado valor. Las reglas de la suma y del producto de probabilidades se transforman pasando de las sumas sobre contadores de los valores discretos a integrales sobre el espacio de eventos para el caso de variables continuas,

$$P(x) = \int p(x, y) dy, \tag{2.6}$$

$$P(x, y) = P(y|x) P(x). \tag{2.7}$$

### 2.1.2. Valores de expectación y covarianzas

El valor medio de una función $f(x)$ de una distribución o densidad de probabilidades $P(x)$ se conoce como *valor de expectación*. La ecuación para el caso discreto es

$$E[f] = \sum_x P(x)f(x), \tag{2.8}$$

mientras que en el caso de variables continuas se tiene

$$E[f] = \int p(x)f(x)dx. \tag{2.9}$$

El valor de expectación de la variable se encuentra usando $f(x) = x$.

En caso de funciones con más de una variable como $f(x, y)$ se aclara respecto a qué variable se realiza el promedio en el subíndice, de manera que $E_x[f(x, y)]$ es el valor de expectación sobre la variable $x$ y depende de la variable $y$.

Los *valores de expectación condicionales* emplean distribuciones condicionales de las variables

$$E_x[f|y] = \sum_x P(x|y)f(x) \quad o \quad E_x[f|y] = \int p(x|y)f(x)dx. \tag{2.10}$$

Otro valor de interés es la medida de dispersión que la función $f$ tiene alrededor de su valor de expectación. La *varianza* de $f(x)$ está definida como

$$var[f] = E[(f(x) - E[f(x)])^2] = E[f(x)^2] - E[f(x)]^2. \tag{2.11}$$

De la misma manera que para la expectación, si $f(x) = x$, se tiene la varianza de la variable estadística. Si se tienen dos variables aleatorias su *covarianza* se calcula mediante

$$\begin{aligned} cov[x, y] &= E_{x,y}[\{x - E[x]\}\{y - E[y]\}] \\ &= E_{x,y}[xy] - E[x]E[y] \end{aligned} \tag{2.12}$$

y es nula si $x$ e $y$ son variables independientes y, **en ese caso**, $P(x, y) = P(x)P(y)$. Para vectores de variables, la expresión de la covarianza queda

$$\begin{aligned} cov[\mathbf{x}, \mathbf{y}] &= E_{\mathbf{x},\mathbf{y}}[\{\mathbf{x} - E[\mathbf{x}]\}\{\mathbf{y}^T - E[\mathbf{y}^T]\}] \\ &= E_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T] - E[\mathbf{x}]E[\mathbf{y}^T]. \end{aligned} \tag{2.13}$$

## 2.2. Teorema de Bayes

Tomando la expresión de 2.5 e introduciendo $r_j$ en lugar de $c_i$ en el cálculo intermedio se obtiene una expresión análoga de la regla del producto

$$P(X = x_i, Y = y_j) = P(X = x_i|Y = y_j)P(Y = y_j),$$

que uniéndola con 2.5, se llega a la ecuación del "*Teorema de Bayes*"

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}, \tag{2.14}$$

así nombrado en honor a Thomas Bayes, matemático, reverendo y estadístico británico del siglo 18 quien fue el primero en establecer dicha propiedad [11]. Es un pilar fundamental de la estadística que permite la inversión de dependencias entre variables aleatorias, según las observaciones con las que se cuente en cada caso.

Se tiene una versión simplificada de la misma,

$$P(Y|X) \propto P(X|Y)P(Y) \tag{2.15}$$

ya que el denominador de la ecuación 2.14 es una constante normalizadora **independiente de Y**, que se calcula a partir de la ecuación 2.3.

En 2.15 se observan tres estructuras:

- $P(Y|X)$ se considera la probabilidad *posterior* de $Y$ y cuantifica la incerteza que se tiene de la variable $Y$ una vez consideradas las observaciones de $X$,
- $P(X|Y)$ es la llamada *verosimilitud* y expresa la probabilidad de aparición de esas observaciones $X$ dada la variable $Y$, y
- $P(Y)$ es la probabilidad *a priori* o *prior* de $Y$, captura las creencias e hipótesis sobre esta variable, antes de conocer las observaciones de $X$.

Esta dependencia de la posterior, verosimilitud y prior ilustran la idea iterativa de la inferencia bayesiana, donde se forma un cuerpo de evidencia que actualiza la distribución posterior continuamente. En cada experimento puede tomarse como prior la posterior obtenida previamente.

### 2.2.1. Construcción de densidades

## Priors

Es en los *priors* donde se codifica el nivel de conocimiento que se tiene sobre el área de interés. Dentro del espectro continuo entre distribuciones a priori no informativas e informativas se habla de tres clasificaciones: difusos, débilmente informativos e informativos. Un prior informativo es el que mayor nivel de confianza tiene sobre su comprensión del parámetro y su relación con las variables de interés, de manera que suele tener menor desviación estándar y representan información muy específica de los parámetros. Un prior débilmente informativo permite una mayor varianza en la densidad de probabilidad seleccionada, por lo que hay un balance entre la información usada y la incerteza sobre la misma. Para reflejar la máxima incerteza posible sobre los parámetros se usa un prior difuso que suele ser una densidad uniforme, dando igual probabilidad a todos los valores posibles del dominio del parámetro [12].

A continuación se introducen algunas densidades de probabilidad de interés al ser las empleadas en este trabajo.

#### Densidad uniforme

La densidad uniforme puede ser débilmente informativa o difusa según el rango de valores permitidos al definir los parámetros mínimo ($min$) y máximo ($max$). Su expresión se detalla en 2.16, tiene como valor medio el promedio de los parámetros, y la varianza es $Var(x) = \frac{(max-min)^2}{12}$.

$$U(x|min, max) = \frac{1}{max - min} \tag{2.16}$$

#### Densidad gaussiana o normal

Es una densidad determinada de la siguiente manera

$$N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\ exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{2.17}$$

Tiene un parámetro de *centralidad*, la media o valor medio $\mu$, y otro de *escala*, la dispersión $\sigma$, a partir de la cual se tiene la varianza $Var(x) = \sigma^2$.

### Densidad media normal

La media normal es una normal centrada en cero truncada, definida sólo sobre los valores $x \in \mathbb{R}_{\geq 0}$.

$$HN(x|\sigma) = \sqrt{\frac{2}{\pi\sigma^2}}\ exp\left(\frac{-x^2}{2\sigma^2}\right). \tag{2.18}$$

El valor de expectación de una variable con densidad de distribución media normal es $\mathbb{E}(x) = \sqrt{\frac{2\sigma^2}{\pi}}$ y su varianza es $Var(x) = \sigma^2 - \mathbb{E}^2$.

### Distribución de Bernoulli

Esta función de probabilidad discreta mide la tasa de éxitos ($x = 1$) y fracasos ($x = 0$) de un suceso.

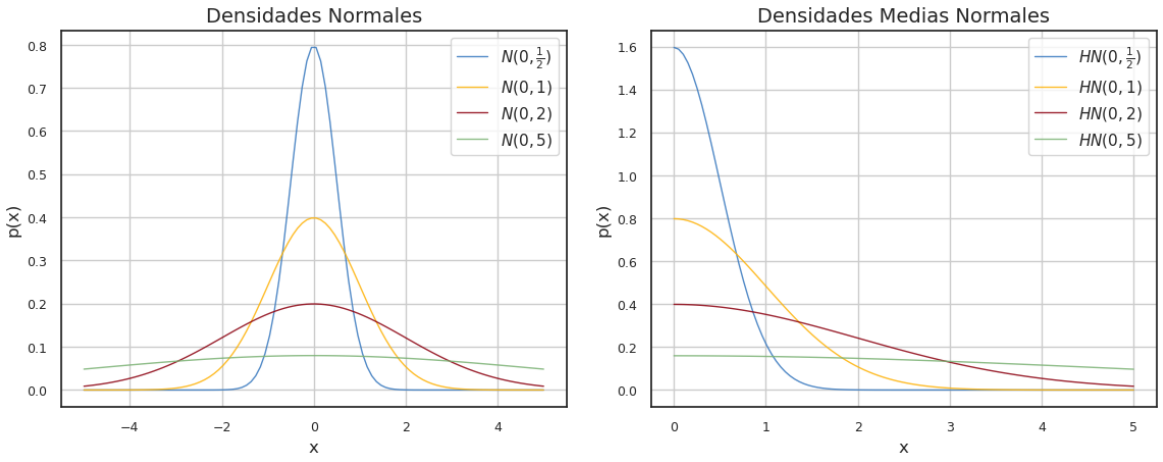
$$Bern(x|t) = t^x(1-t)^{1-x} \tag{2.19}$$

Tiene como valor medio $t$, que es la probabilidad de éxito y pertenece entre 0 y 1. Su varianza es $t(1-t)$.

Las densidades normales y media normales pueden hacerse más o menos informativas de acuerdo a sus parámetros de dispersión, como se aprecia en la figura 2.1. A mayor valor de $\sigma$ (curvas roja y verde) la densidad de probabilidades se hace menor, pero abarcando un mayor número de parámetros posibles. Una manera de verificar la elección de prior es con el chequeo del prior predictivo [12, 13, 14]. Definida la densidad de probabilidades, se toman muestras aleatorias de la misma y se contrastan los estadísticos la distribución generada con los estadísticos de los datos observados, esperando que haya cierto nivel de concordancia en los mismos indicando una definición acorde de los priors. En cualquier selección de prior que se realice, la dimensión de los datos observados tendrá un gran peso en la transformación a la distribución posterior a través de la verosimilitud, puesto que a mayor tamaño de muestra observada, más informada estará la verosimilitud.

## Verosimilitud y posterior

La elección de la verosimilitud se realiza a partir de la relación propuesta entre los parámetros a inferir y las observaciones de la variable, de acuerdo a algún marco teórico y sus hipótesis sobre la dependencia de la generación de los datos con los parámetros de interés. Para ejemplificar, dada una variable que se considera provenir de forma $x_i = \alpha y_i + \mu + \epsilon_i$, es decir, tener una




(a) Densidades normales.

(b) Densidades medias normales.

Figura 2.1: Variación en las densidades de probabilidad normales y medias normales en función del parámetro de desviación estándar. Cuanto menor es el parámetro $\sigma$, más informativo es el prior.

dependencia lineal de error gaussiano $\epsilon_i \sim N(\mu_\epsilon = 0, \sigma_\epsilon = 1)$, la verosimilitud será

$$p(x|y, \mu, \sigma) = N\left(x|\mu_x = \alpha y + \mu, sigma = \sigma_\epsilon\right).$$

Dado el prior y la verosimilitud, el método de Bayes nos permite obtener la densidad de probabilidad posterior de los parámetros a inferir a partir de la definición dada por la ecuación 2.14. En muchos casos, estas densidades no tienen forma cerrada e incluso quedan expresadas salvo alguna constante de proporcionalidad, como para algunos perfiles de expresión molecular o modelos inversos de Potts [15, 16], y se sostiene la ecuación 2.15. Estas distribuciones posteriores suelen ser resumidas según las medias, medianas, desviaciones u otros estadísticos que se consideren relevantes para el caso. Bajo estas condiciones, y especialmente para altas dimensionalidades, los algoritmos y métodos presentados en el capítulo 4 permiten aproximaciones de estas densidades a partir de un gran número de simulaciones.

De manera similar al chequeo predictivo con el prior, se puede realizar un chequeo posterior tomando muestras de generadas por el modelo, dadas las distribuciones posteriores de parámetros inferidos, y comparándolas con las observaciones reales.

## 2.3. Modelos gráficos probabilísticos

Esta herramienta permite representar las relaciones de distribuciones de probabilidad, visualizar la estructura del modelo y determinar propiedades del mismo, a partir de diagramas simples. Constan de nodos, o vértices, conectados por aristas o enlaces. Cada nodo representa una variable aleatoria y los enlaces expresan que hay una relación probabilística entre ellas. En el caso en que las aristas tengan flechas se habla de enlaces dirigidos y definen modelos gráficos dirigidos, que son de especial interés para expresar relaciones causales entre variables aleatorias [10, Capítulo 8].

### 2.3.1. Redes Bayesianas

Dadas tres variables a,b y c una manera de descomponer su distribución conjunta $P(a, b, c)$ es aplicar iterativamente la regla del producto de 2.7

$$\begin{aligned} P(a, b, c) &= P(c|a, b)P(a, b) \\ &= P(c|a, b)P(b|a)P(a). \end{aligned} \tag{2.20}$$

La representación de esta relación probabilística es la provista en la figura 2.2a, donde los nodos se ven conectados por enlaces dirigidos. Los nodos de donde parten las flechas se denominan nodos *progenitores*, mientras que la punta llega a los nodos *hijos*, marcando una jerarquía de dependencias que refleja la relación de probabilidades de la ecuación 2.20. De esta forma, se ve que el nodo de la variable $c$ es hijo de $a$ y $b$, el nodo $b$ es hijo sólo de $a$ y $a$ no tiene ascendencia. Si en la regla del producto se usa otro orden de las variables, el modelo gráfico cambia sus vértices para dar otra relación probabilística.

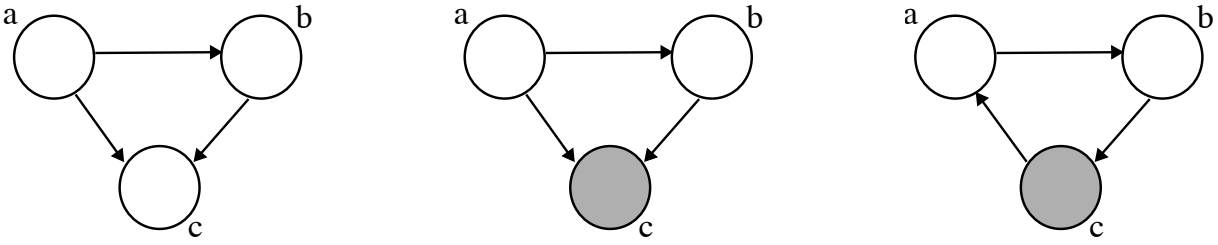
La generalización de 2.20 para una distribución de probabilidad conjunta $p(\mathbf{z})$ de $M$ nodos es

$$p(\mathbf{z}) = \prod_{i=1}^{M} p(\mathbf{z}_i|pa_i), \tag{2.21}$$

donde $\mathbf{z}_i$ son las variables asociadas al nodo $i$ y $pa_i$ denota los progenitores de ese nodo.

Los modelos gráficos permiten ilustrar qué variables han sido observadas, como lo muestra 2.2b, donde el sombreado del nodo $c$ denota que se tienen mediciones sobre dicha variable aleatoria. En el caso que la variable no pueda ser observada se denominan variables *latentes*.

La figura 2.2a ilustra un ejemplo de gráfico aciclico dirigido -o DAG por sus siglas del inglés "*directed acyclical graph*"- ya que las aristas son dirigidas, y no se puede realizar un




(a) *DAG* (b) *DAG*, $c$ observada (c) *DCG*, $c$ observada

Figura 2.2: Red bayesiana para las variables $a$, $b$ y $c$. Los nodos representan las variables, y los enlaces dirigidos la relación de las probabilidades condicionales con probabilidad conjunta establecida por la ecuación 2.20.

camino circular que empiece y termine en la misma variable; no admite ciclos. En 2.2c se puede realizar un camino cíclico $a \to b \to c$ con cualquier permutación de orden, pero dicho gráfico no representa la ecuación de probabilidad conjunta de 2.20, ya que acá $a$ es hija de la variable $c$ y progenitora de $b$. Éste sería un gráfico cíclico dirigido, - o DCG por sus siglas en inglés "*directed cyclical graph*".

En la figura 2.3a se tiene un gráfico acíclico dirigido que ilustra la estructura de una cadena de variables aleatorias $X_i$, donde cada una depende únicamente de la instancia anterior.


(a) *DAG*: cadena (b) *DAG*: PPCA

Figura 2.3: Gráficos acíclicos dirigidos. Cadena de variables aleatorias y modelo PPCA, con nodos grandes representando las variables aleatorias, sombreados en caso de ser observadas, y nodos pequeños para indicar hiper-parámetros de dichas variables. Los enlaces dirigidos parten de la variable progenitora hacia el hijo, indicando la dependencia probabilística. El recuadro es una placa y denota la repetición (N veces) de la estructura que encierra.

El gráfico en 2.3b ilustra la relación de un modelo de análisis de componentes principales probabilístico (PPCA), donde las $N$ variables observadas $Y_n \in \mathbb{R}^D$ dependen del valor de las variables $X_n \in \mathbb{R}^M$ latentes a través de una matriz de mezcla $W \in \mathbb{R}^{DxM}$, valor medio $\mu \in \mathbb{R}^D$

y ruido normal $\epsilon$,

$$Y = WX + \mu + \epsilon,$$

de manera que la distribución condicional queda definida como

$$p(Y|X) = N\left(Y|WX + \mu, \sigma^2 I^{DxD}\right).$$

Notar el uso de la placa (la caja con índice N), que ilustra que te tiene la estructura dentro de la misma repetida $N$ veces.

# Capítulo 3

# Inferencia Causal

Como se mencionó en la introducción, este trabajo busca calificar y cuantificar las relaciones entre variables de interés, informando implicaciones causales. En el capítulo 2 se mostró que gráficamente dichas relaciones se pueden diferenciar de las meramente asociativas al representarse mediante enlaces dirigidos entre la variable causa y la variable efecto.

Un problema central en nuestro estudio es la falta de eventos *contrafácticos*, que devienen en el 'problema fundamental', que se explicará en detalle más adelante. Para ejemplificarlo, ante un dolor de cabeza una persona puede tomar un remedio o no. Si toma la medicación, esta realidad es el hecho, y no tomarla sería el evento contrafáctico. Para poder asegurar que el dolor de cabeza se fué debido a que se tomó la medicina, se debería tener acceso a las dos realidades, donde la única diferencia sea si se tomó o no el remedio. Sólo puede determinarse que dicho remedio es la causa de la sanación si la misma persona se cura ante la exposición al remedio y persiste en dolor sin la exposición a tratamiento. Esto debiera determinarse para una población heterogénea para reportar resultados estadísticos sobre la medicación. Como puede esperarse, no es posible la observación simultánea de situaciones opuestas.

En este capítulo se establecerán las bases necesarias para un análisis causal y listarán las definiciones necesarias para la comprensión del trabajo. Al final, se explicará el método del 'deconfounder' como la herramienta utilizada en esta tesis para resolver el 'problema fundamental' al que nos enfrentamos en estos tipos de estudios.

## 3.1. Definiciones básicas de la inferencia causal

El objetivo principal es cuantificar el efecto sobre una variable de salida o resultado (*target*) $Y$ en base al valor de una variable de entrada (*feature*) $T$ de tratamiento, que se quiere iden-

tificar como causa del resultado observado. En un estudio se tiene un número determinado de individuos o *unidades* $U$ donde la mitad son expuestas a un tratamiento $T = t$ y el resto a un tratamiento de control $T = c$. Cuando la selección de tratamiento es aleatoria cada individuo tiene igual probabilidad de recibir $t$ o $c$, y entonces se tiene un estudio experimental. Bajo otro régimen de asignación de tratamientos se tiene un cuasi-experimento o estudio observacional.

Si bien hay una predilección natural por los métodos experimentales ya que los mismos permiten generar una base de datos con mayor control sobre la distribución de los features, de manera que estén balanceados y la población de cada tratamiento tenga distribuciones similares del resto de las covariables a observar, son muchos los casos donde se cuenta únicamente con información observacional. La manera en que se recopila la información depende de factores como el tiempo, los recursos económicos disponibles y la ética del accionar, entre otros. El estudio de la deserción universitaria en la UNSAM es un cuasi-experimento.

### 3.1.1. El modelo de Rubin y el problema fundamental

El modelo de Rubin propone ensayos sobre cada unidad $U = u$ de manera que a un tiempo $t_1$ se aplica el tratamiento $T$ asignado, y a un tiempo $t_2 > t_1$ se mide el target $Y$. Por simplicidad se asume que el tratamiento admite dos opciones, tratamiento o control. De esta manera se puede observar $Y(U = u_t, T = t)$ para las unidades $u_t$ que recibieron el tratamiento, o $Y(U = u_c, T = c)$ en las que recibieron el control $u_c$. Una manera de medir el efecto causal del tratamiento $T = t$ sobre el control $T = c$ es tomar la diferencia del target $Y$ ante los tratamientos posibles en una *misma* unidad $u$, también llamados resultados potenciales:

$$Y(u, t) - Y(u, c). \tag{3.1}$$

Para simplificar la notación, se usa $Y(U = u, T = c) \equiv Y(u, c)$ y lo mismo para $T = t$. Esto introduce lo que se conoce como 'el problema fundamental de la inferencia causal' dado que $u_c \cap u_t = \emptyset$. Como ya se adelantó, no es posible observar para una misma unidad el resultado potencial de dos tratamientos distintos; en cada ensayo se admite uno de ellos [17].

Una solución posible es tomar el efecto causal promedio $\tau$ a partir del valor de expectación de esta diferencia sobre las unidades:

$$\begin{aligned} \tau &= \mathbb{E}(Y(u, t) - Y(u, c)) \\ &= \mathbb{E}(Y(u, t)) - E(Y(u, c)). \end{aligned} \tag{3.2}$$

El mismo suele llamarse $ATE$ por sus siglas del inglés average treatment effect.

Para calcular los valores de expectación arriba detallados, debe darse que

$$\mathbb{E}(Y(u,t)) = \mathbb{E}(Y(u,t)|T=t),$$

mientras que la asignación del tratamiento en cada unidad sea estadísticamente independiente de las demás variables. Un estudio experimental controlado con asignación aleatoria se aproxima a esta condición, pero en general no necesariamente se cumple, ya que puede existir una variable oculta que influye en las causas medidas y en el resultado de la exposición al tratamiento. Estas variables ocultas se llaman *confundidoras*. Por ejemplo, si se considera un tratamiento médico para controlar la presión, puede ser que factores como el sexo, peso o la edad de una persona afecten la probabilidad de recibir un remedio, y además pueden también cambiar la manera en que esa persona responde al tratamiento. En el caso de estudios observacionales, no se puede controlar la asignación independiente y aleatoria de las unidades, por lo que establecer la relación entre causas y efectos a partir de 3.2 se torna aún más complejo y se deben explotar otras técnicas para realizar inferencia causal.

## 3.2. Modelo Causal

Todo modelo causal puede describirse a partir de dos estructuras básicas: el modelo de resultados y el modelo de asignación o factores, también llamados *outcome model* y *factor* o *assignment models* del inglés [18].

Tómese el problema con asignación de tratamiento binario, $T \in \{c,t\}$ que deviene en dos resultados potenciales $(Y(c), Y(t))$ para cada unidad $u$. Además del tratamiento, cada individuo tiene sus covariables $x_u$.

Llamamos $\psi$ a los parámetros de asignación del tratamiento que siendo desconocidos rigen, junto con las covariables, la asignación para cada individuo. El assignment model es, entonces:

$$\begin{aligned} \psi &\sim p(\psi), \\ x_u &\sim p(x), \\ T_u &\sim p(T|x_u,\psi). \end{aligned} \tag{3.3}$$

Cuando los resultados potenciales son independientes de la asignación del tratamiento, condicionados en las covariables, se cumple la *hipótesis de inconfundibilidad*, a veces llamada ignorabilidad fuerte, y se expresa

$$(Y(c), Y(t)) \perp\!\!\!\perp T|X.$$

De manera central, esta hipótesis teóricamente no puede ser comprobada [19].

Una vez asignado un tratamiento, los resultados potenciales están gobernados por el outcome model. Se tienen parámetros del resultado $\theta$ no observables. Así, el outcome model queda ilustrado por:

$$\begin{aligned} \theta &\sim p(\theta), \\ x_u &\sim p(x), \\ (Y(c), Y(t)) &\sim p(Y(c), Y(t)|\theta, x_u). \end{aligned} \tag{3.4}$$

Dados los parámetros $\theta$ y covariables $x_u$, los resultados potenciales son intercambiables entre individuos al ser condicionalmente independientes bajo esas observaciones.

Observando las covariables $\mathbf{x}$, el resultado $Y(T)$ sabiendo qué tratamiento $T$ es asignado, el modelo causal completo es:

$$p\left(\psi, \theta, \mathbf{x}, Y(c), Y(t), T\right) = p(\psi)p(\theta)p(\mathbf{x})p\left(Y(c), Y(t)|\mathbf{x}, \theta\right)p\left(T|\psi, \mathbf{x}\right) \tag{3.5}$$

Llamando $\bar{T}$ al tratamiento contrafáctico no observado, se pueden marginalizar los mismos para expresar la posterior conjunta de los parámetros dada la observación de $\{\mathbf{x}, \mathbf{T}, \mathbf{Y}(\mathbf{T})\}$,

$$p(\psi, \theta|\mathbf{Y}(\mathbf{T}), \mathbf{T}, \mathbf{x}) \propto p(\psi)p(\theta)p(\mathbf{T}|\psi, \mathbf{x}) \int p(\mathbf{Y}(\mathbf{T}), \mathbf{Y}(\bar{\mathbf{T}})|\psi, \theta)\, d\mathbf{Y}(\bar{\mathbf{T}}). \tag{3.6}$$

Bajo la hipótesis de inconfundibilidad los parámetros del resultado y los de la asignación son independientes entre sí y se factoriza la ecuación 3.6:

$$\begin{aligned} p(\psi|\mathbf{Y}(\mathbf{T}), \mathbf{T}, \mathbf{x}) &\propto p(\psi)p(\mathbf{T}|\psi, \mathbf{x}), \\ p(\theta|\mathbf{Y}(\mathbf{T}), \mathbf{T}, \mathbf{x}) &\propto p(\theta) \int p(\mathbf{Y}(\mathbf{T}), \mathbf{Y}(\bar{\mathbf{T}})|\psi, \theta)\, d\mathbf{Y}(\bar{\mathbf{T}}), \end{aligned} \tag{3.7}$$

por lo que el modelo causal consta de dos estructuras con mecanismos de inferencia independientes entre sí y se pueden estudiar de manera separada en ausencia de variables confundidoras.

Lo postulado hasta ahora implica tener un estudio donde un único tratamiento binario trata de explicarse como causante de un efecto o resultado sobre individuos de los que se tiene un conjunto de covariables observadas. Es directamente extensible a tratamientos no binarios. Muchos estudios observacionales son del tipo de causalidad múltiple, donde existe mas de una variable que puede ser la explicación del resultado observado, un ejemplo típico es

el estudio de asociación de genomas [20]. Bajo estas condiciones se debe pasar al estudio de inferencia de múltiples causas, los cuales traen desventajas como lo son la presencia de variables confundidoras latentes y un aumento en la dimensionalidad. También suelen requerir de otras hipótesis bajo las cuales operar, descriptas más adelante.

## 3.3. Deconfounder

La herramienta del 'deconfounder' propuesta por Wang y Blei en su publicación "The Blessings of Multiple Causes" [2] aprovecha la estructura de múltiples causas y se detalla al ser el algoritmo empleado en este trabajo. Este método permite pasar de un estudio bajo hipótesis incomprobables como la de ignorabilidad - tener todas las variables confundidoras observadas- a una estructura con pasos comprobables para construir un modelo de causas asignadas que concuerde con los datos observados. La figura 3.1 ilustra la relación entre las variables confundidoras latentes $\boldsymbol{Z}_n$, las causas $\boldsymbol{X}_n$ y el *target* $Y_n$ para una dada unidad $n$.

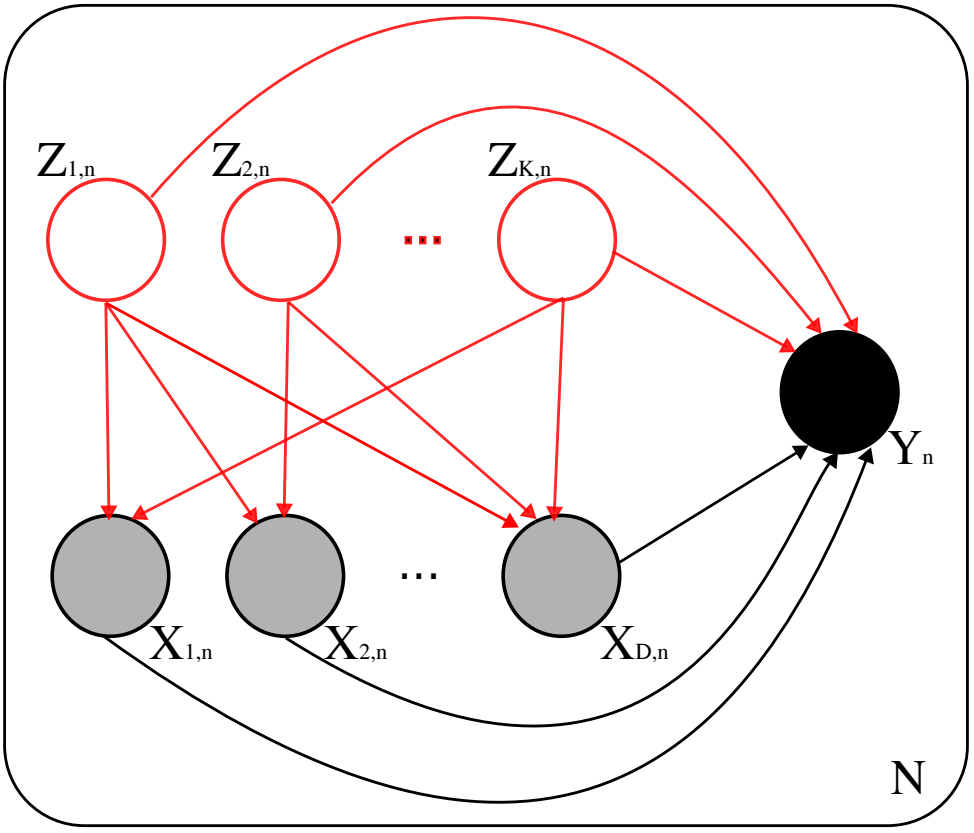



Figura 3.1: Modelo gráfico para $K$ variables confundidoras latentes $\boldsymbol{Z}_k$, y $D$ causas $\boldsymbol{X}_d$ y un resultado $Y$ observados. La placa denota que esto se repite para las $N$ unidades de observación.

Cada causa puede tener influencia por un número de variables confundidoras en distinta medida y todas impactan en el *target*. Al tener más de una causa posible, tenemos un caso de múltiples causas, y con la idea de correctamente cuantificar el efecto de cada una de ellas, independientemente del impacto de las variables confundidoras, el '*deconfounder*' busca proponer un modelo de factores para las variables latentes confundidoras, verificar que sea el adecuado a los datos, obtener estimadores de las mismas para cada unidad, y usarlos como confundidoras sustitutas observadas para controlar la inferencia en el modelo de resultados. De esta manera extrae la información provista por las variables de tratamiento, o causas, y las separa del efecto de las variables confundidoras, que son compartidas por más de una variable y el resultado [2].

### 3.3.1. Estructura del *deconfounder*

**Modelo de factores**

De todas las estructuras posibles, la más sencilla es un modelo de factores con la estructura de PCA probabilístico para las causas. Con $D$ causas y $K$ variables confundidoras para un caso de $N$ unidades, se tiene un vector $x_n \in \mathbb{R}^D$ observado y un vector $z_n \in \mathbb{R}^K$ latente para cada unidad, donde $n = 1, .., N$. Se asume que los mismos son muestras independientes e idénticamente distribuidas a partir de las definiciones detalladas en 3.8

$$\begin{aligned} p(\mathbf{z}_n) &= N(\mu_z, \sigma_z = \sigma^2 \mathbf{I}_K), \\ p(\mathbf{x}_n|\mathbf{z}_n) &= N(\mu_x = \mu^\star, \sigma_x = \sigma^2 \mathbf{I}_D), \end{aligned} \tag{3.8}$$

donde $\mathbf{I}_L$ es la matriz identidad de dimensión $L$ y $\mu^\star$ está definido como

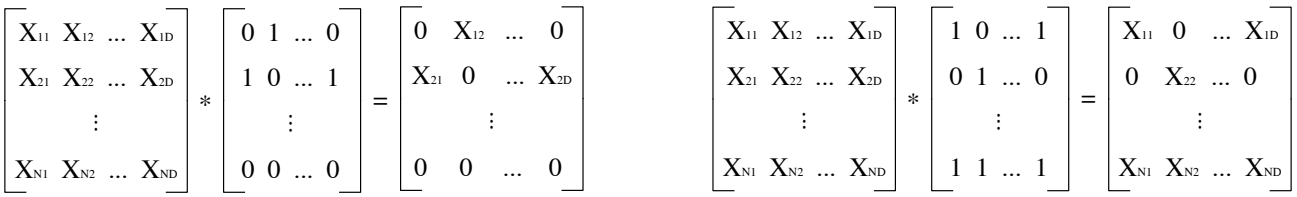
$$\begin{aligned} \mu_x^{lin} &= \mathbf{z}_n \mathbf{W}, \\ \mu_x^{cuad} &= \mathbf{z}_n \mathbf{W} + \mathbf{z}_n^2 \mathbf{W}_2, \end{aligned} \tag{3.9}$$

con $W, W_2 \in \mathbb{R}_{KxD}$, lo que define un modelo de factores lineal o cuadrático respectivamente. La estructura gráfica del mismo se visualiza en 3.3a.

Al realizar inferencia sobre los datos observados, se obtienen distribuciones $p(\mathbf{z}_n)$ y definen los 'confounders sustitos' $\hat{\boldsymbol{z}}_n$ para cada unidad observada como el valor de expectación.

**Chequeo predictivo**

Si bien el modelo lineal de factores tiene la ventaja de ser más simple, y por lo tanto permite una comprensión directa de la conformación de relaciones entre confounders y causas, puede presentar relaciones limitadas para la complejidad de algunas bases de datos. A modo de veri-




(a) $\boldsymbol{X} * \boldsymbol{H} = \boldsymbol{X}_{held}$ (b) $\boldsymbol{X} * \tilde{\boldsymbol{H}} = \boldsymbol{X}_{obs}$

Figura 3.2: Generación matrices de datos retenidos ($\boldsymbol{X}_{held}$) y observados ($\boldsymbol{X}_{obs}$) a partir de las máscaras para el predictive check ante una posible máscara $\boldsymbol{H}$ aleatoria.

ficación se plantea un chequeo predictivo, o *predictive check*, que genera un valor $p$ a contrastar con un rango fijado de manera empírica; si se tiene $p > 0{,}1$ los autores determinan la prueba del chequeo predictivo como superada, y se considera que dicho modelo se ajusta suficientemente bien a los datos como para que los confounders inferidos puedan dar lugar a las causas observadas.

El proceso comienza con la creación de una matriz de retención o máscara (matriz *holdout* por el inglés) $\mathbf{H}_{NxD}$, que tiene las mismas dimensiones que la matriz de covariables $\mathbf{X}_{NxD}$, con los vectores de causas observadas en cada fila. Se define una fracción de datos a enmascarar, $f$, y seleccionan $f.N.D$ coordenadas al azar donde ingresar un 1 mientras que el resto de la matriz tiene un 0. De esta manera, cada individuo tiene una cantidad aleatoria de causas enmascaradas, entre 0 y $D$. $\tilde{\mathbf{H}}_{NxD}$ es el complemento de la matriz *holdout*, $\tilde{\mathbf{H}} = \mathbb{I} - \mathbf{H}$. Al multiplicar la matriz de covariables por $H_{NxD}$ se tiene el set de datos retenidos $\boldsymbol{X}_{held}$, y al enmascarar con su complemento $\tilde{\mathbf{H}}_{NxD} = \mathbf{I}_{NxD} - H_{NxD}$ se obtiene el conjunto de datos observados $\boldsymbol{X}_{obs}$, para el testeo del predictive check. La figura 3.2 ilustra este proceso para un ejemplo aleatorio de la matriz $\boldsymbol{H}$.

Se ajusta el modelo de factores a los datos observados y obtiene $p(\boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{X})$, con $\boldsymbol{\theta} = \boldsymbol{W}$ o $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{W_2})$, según corresponda. Para cada unidad se calcula la distribución posterior de $p(\boldsymbol{Z}_n|\boldsymbol{X}_{n,obs})$. El testeo entonces consta de generar muestras replicadas $X^{rep}_{n,held}$ de las causas retenidas a partir de sus distribuciones predictivas, que fueron informadas sólo por los datos observados, como muestra la ecuación 3.10.

$$p(\boldsymbol{X}^{rep}_{n,held}|\boldsymbol{X}_{n,obs}) = \int p(\boldsymbol{X}_{n,held}|\boldsymbol{Z}_n)p(\boldsymbol{Z}_n|\boldsymbol{X}_{n,obs})d\boldsymbol{Z}_n. \tag{3.10}$$

A continuación se contrastan las muestras replicadas a los datos retenidos del modelo de factores

con la función de discrepancia, $t$, a partir de la ecuación 3.11, donde $logp$ es la log-probabilidad.

$$\begin{aligned} t(\boldsymbol{X}_{n,held}) &= \mathbb{E}_{\boldsymbol{Z}}[logp(\boldsymbol{X}_{n,held}|\boldsymbol{Z})|\boldsymbol{X}_{n,obs}], y \\ t(\boldsymbol{X}^{rep}_{n,held}) &= \mathbb{E}_{\boldsymbol{Z}}[logp(\boldsymbol{X}^{rep}_{n,held}|\boldsymbol{Z})|\boldsymbol{X}_{n,obs}] \end{aligned} \tag{3.11}$$

El puntaje predictivo $p - score$ se construye al integrar las instancias donde la discrepancia de las causas replicadas sea menor que la discrepancia de los datos retenidos, como se plantea en 3.12. Un resultado es óptimo en el caso de tener $p - score \sim 0{,}5$, pero es suficiente un $p - score > 0{,}1$, puesto que no se tendría suficiente evidencia para definir un desajuste entre los datos y el modelo.

$$p - score = p(t(\boldsymbol{X}^{rep}_{n,held}) < t(\boldsymbol{X}_{n,held})) \tag{3.12}$$

En cada estudio se separa la base de datos en conjuntos de causas observadas y retenidas de manera aleatoria, se realiza inferencia con el conjunto de observados y generan replicas para construir el puntaje predictivo con el cual seleccionar el modelo de factores de la librería de modelos disponibles para cada situación. El modelo de factores seleccionado capta la distribución de causas asignadas en la población y su dependencia estructural, mas no necesariamente es el 'verdadero' modelo.

**Modelo de resultados**

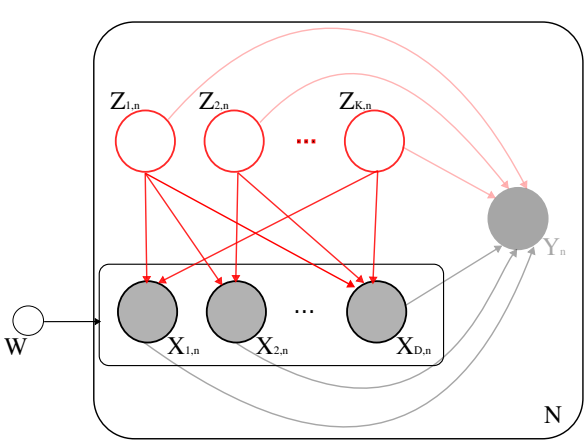
Con los estimadores obtenidos en la primera etapa, a partir del modelo seleccionado que supera el chequeo predictivo, se plantea el modelo de resultados más simple posible para la distribución de los targets observados $Y_n$:

$$p(Y_n|\mathbf{x}_n, \hat{\boldsymbol{z}}_n, \boldsymbol{\beta}, \boldsymbol{\gamma}) = N(\boldsymbol{\beta}\mathbf{x}_n + \boldsymbol{\gamma}\hat{\boldsymbol{z}}_n, \sigma^2). \tag{3.13}$$

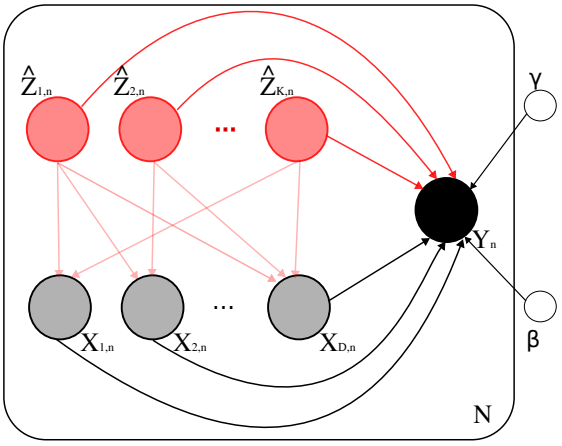
Con este modelo, la relación deseada entre las causas y el target queda codificada en la matriz $\boldsymbol{\beta}$, mientras que el peso de los confounders está medido con $\boldsymbol{\gamma}$. Esta estructura puede observarse en la figura 3.3b.

Existe una alternativa que emplea las distribuciones posteriores del modelo de factores de la matriz $W$ (y $W_2$ si corresponde) para definir estimaciones a partir de los valores de expectación de cada coordenada, y lo mismo para los confounders. Con estos estimadores, se generan las causas

$$\begin{aligned} \hat{\boldsymbol{x}}^{lin}_n &= \hat{\boldsymbol{z}}_n\hat{\boldsymbol{W}} \\ \hat{\boldsymbol{x}}^{cuad}_n &= \hat{\boldsymbol{z}}_n\hat{\boldsymbol{W}} + \hat{\boldsymbol{z}}^2_n\hat{\boldsymbol{W}}_2, \end{aligned}$$




(a) Modelo de factores según 3.8 y 3.9. Hiperparámetro $W$ (y $W_2$ para modelo cuadrático).




(b) Modelo de resultados según 3.13. Hiperparámetros $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$.

Figura 3.3: Modelos de factores y de resultados según las etapas del algoritmo del deconfounder. Las variables y los enlaces ignorados en cada instancia están denotados con menor opacidad. Los círculos menores por fuera de la placa son los hiper- parámetros a definir. En el modelo de resultados, las observaciones de las variables confundidoras son sobre los valores sustitutos, inferidos a partir del modelo de factores, $\hat{\boldsymbol{z}}_n$.

según corresponda. Con estos valores, se tiene una verosimilitud

$$p(Y_n|\mathbf{x}_n, \hat{\boldsymbol{z}}_n, \boldsymbol{\beta}, \boldsymbol{\gamma}) = N(\boldsymbol{\beta}\mathbf{x}_n + \boldsymbol{\gamma}\boldsymbol{x}(\hat{\boldsymbol{z}}_n), \sigma^2),$$

pero la misma no es analizada en este trabajo dado que limita la posibilidad de explicar los resultados obtenidos al tener valoraciones del efecto causal más entrelazadas, con parámetros relacionados con las causas y otros con las causas reconstruidas.

### 3.3.2. Hipótesis del *deconfounder*

De manera general, las estrategias para identificaciones causales requieren de hipótesis para identificar los resultados potenciales. Este algoritmo debe cumplir con ciertos los supuestos detallados más abajo. En primer lugar se debe satisfacer la suposición SUTVA, de sus siglas del inglés *single unit treatment value assumption*, también conocida como suposición de estabilidad [21]. La misma implica que los resultados potenciales de una unidad sean independientes de las causas asignadas a cualquier otra unidad, es decir, que son estables ante los tratamientos de otras unidades. Asume que no hay interferencia entre individuos y que hay una única versión de cada causa asignada. Es decir, para cada unidad $n$ y todos los posibles pares de tratamientos

$T$ y $T'$, se tiene que

$$Y_n(T) = Y_n(T') \; si \; T_n = T'_n.$$

En segundo lugar, la condición 'relajada' de ignorabilidad, a veces llamada ignorabilidad única, permite que se observen únicamente las variables confundidoras de causa única, mientras que las de múltiples causas pueden permanecer como variables latentes. Las causas individuales $X_{i,j}$ deben ser marginalmente independientes de la respuesta $Y(\boldsymbol{x})$ condicionando sobre covariables (no necesariamente confundidoras) observadas $\boldsymbol{C}_i$.

$$X_{i,j} \perp\!\!\!\perp Y_i(\boldsymbol{x})|\boldsymbol{C}_i, \;\; j = 1, ..., m$$

Esta condición impone menos restricciones que la *inconfundibilidad* general, puesto que allí la restricción es para la independencia conjunta de las causas. De esta manera, el deconfounder permite obtener estimaciones no sesgadas debido a las variables confundidoras, pero aumenta la incerteza en sus estimaciones al tener mayor varianza.

También se pide que haya solapamiento (*overlap*) de los confounders sustitutos $\hat{\boldsymbol{Z}}_i$ y covariables observadas $\boldsymbol{C}_i$. Es decir que la probabilidad de las causas condicionadas sobre los confounders sustitutos y sobre las covariables observadas sea mayor que cero para todo conjunto de causas definido positivo.

$$p(X_{i,j} \in \mathbb{X}|\hat{Z}_i) > 0 \quad \wedge \quad p(X_{i,j} \in \mathbb{X}|\boldsymbol{C}_i) > 0 \;\; \forall \;\; \mathbb{X}/p(\mathbb{X}) > 0 \tag{3.14}$$

Este requerimiento favorece a la identificación del resultado potencial $Y_i(\boldsymbol{x})$ dado que la varianza de sus estimadores disminuye con el aumento de la superposición. Esta condición se vuelve más compleja de satisfacer con la dimensionalidad de $Z_i$.

Al cumplirse estas tres hipótesis, el deconfounder obtiene un estimador no sesgado del efecto causal promedio, es decir, la versión independiente de 3.2.

# Capítulo 4

# Métodos Computacionales

Como se mencionó en el capítulo 2, realizar el tipo de inferencia buscado en este trabajo requiere de gran capacidad de cómputo. Se busca replicar el método del *deconfounder* adaptado a nuestra base de datos de deserción en la UNSAM. Para ello, se debe realizar un gran número de iteraciones sobre las distribuciones de probabilidades, ya que buscamos las distribuciones posteriores de los parámetros de un modelo lineal. Si bien la resolución de los mismos puede abordarse mediante métodos analíticos, esto no puede hacerse de manera general. Ya mencionamos previamente trabajos con distribuciones no analíticas. En particular, para considerar las dispersiones relacionadas con las variables latentes, el problema no puede resolverse exactamente. Es por eso que recurrimos a algoritmos de muestreo para obtener representantes de la distribución posterior conjunta de los parámetros de interés, a partir de la cual se calculan, entre otros, valores medios y dispersiones.

Con estas consideraciones, en este capítulo se discutirán las bases de Monte Carlo Hamiltoniano - HMC por sus siglas del inglés Hamiltonian Monte Carlo- un 'sabor' del método de muestreo MCMC (Markov Chain Monte Carlo) y su variante NUTS. Además se introducirá el lenguaje empleado para realizar inferencias bayesianas causales con aproximaciones a partir de sampleos con esta variante. Finalmente se detallará un primer modelo sencillo con variables confundidoras para motivar el uso del algoritmo 'deconfounder'.

## 4.1. Métodos Monte Carlo

Las técnicas de Monte Carlo son algoritmos computacionales que obtienen inferencias aproximadas mediante muestreos numéricos aleatorios de las variables. Son especialmente útiles para integraciones sobre múltiples dimensiones que, ya sea porque son insolubles o por una

gran complejidad de cómputo, requieren buscar alternativas a la solución exacta.

El valor de expectación de una función $f(\mathbf{q})$ respecto una distribución de probabilidades $p(\mathbf{q})$

$$\mathbb{E}[f] = \int f(\mathbf{q})p(\mathbf{q})d\mathbf{q},$$

puede aproximarse mediante muestras $\mathbf{q}^s$, con $s = 1, ..., S$ independientes de la distribución $p(\mathbf{q})$. De esta forma, con el estimador $\hat{f}$ definido como

$$\hat{f} = \frac{1}{S}\sum_{s=1}^{S} f(\mathbf{q}^s),$$

vale que $\mathbb{E}[f] = \mathbb{E}[\hat{f}]$ y la varianza del estimador queda determinada como

$$var[\hat{f}] = \frac{1}{S}\mathbb{E}[(f - \mathbb{E}(f))^2],$$

lo que es independiente de la dimensión de la variable $\mathbf{q}$ y disminuye con el número de muestras. El muestreo puede ser problemático en casos donde la función $f(\mathbf{q})$ tenga valores muy grandes en regiones de poca probabilidad, porque pueden generar un valor de expectación dominado por muestras de esa región, requiriendo de un mayor número de muestras.

En los DAGs se tiene especificada la distribución conjunta descrita en 2.21, por lo que obtener una muestra de la distribución conjunta implica hacer un barrido por las variables en orden $\mathbf{q}_1, ..., \mathbf{q}_M$, tomando muestras de las distribuciones condicionales $p(\mathbf{q}_i|pa_i)$, instanciando valores desde el primero nodo progenitor hasta el último hijo. En el caso donde algunas de las variables están observadas, se tiene el paso agregado de contrastar la muestra con el valor real y se continúa instanciando las variables nodo a nodo sólo si la muestra del nodo coincide con la distribución proveniente del valor observado. Se tiene la muestra de la distribución conjunta al llegar a la instancia del último nodo del gráfico [10, Capítulo 11].

### 4.1.1. Markov Chain Monte Carlo

Para que una transición entre estados de un sistema sea clasificado como proceso de *Markov*, la probabilidad de un estado debe depender únicamente del estado anterior. Así, se forma una cadena de Markov de variables aleatorias $\mathbf{q}^1, ..., \mathbf{q}^M$ cuando se cumple la independencia

condicional

$$\begin{aligned} P(\mathbf{q}^{m+1}|\mathbf{q}^1,...,\mathbf{q}^m) &= P(\mathbf{q}^{m+1}|\mathbf{q}^m) \;\; \forall m = 1,...,M-1 \\ &\equiv T_m(\mathbf{q}^m,\mathbf{q}^{m+1}), \end{aligned} \tag{4.1}$$

y se tiene la probabilidad de transición $T$.

La cadena de Markov es *homogénea* si las probabilidades de transición son iguales para todas las $m$.

Una distribución $p^\star(\mathbf{q})$ es *invariante* respecto de una cadena Markov homogénea si

$$p^\star(\mathbf{q}) = \sum_{\mathbf{q}'} T(\mathbf{q}',\mathbf{q})p^\star(\mathbf{q}').$$

Una condición suficiente pero no necesaria para tal invarianza es que se cumpla el balance detallado:

$$p^\star(\mathbf{q})T(\mathbf{q},\mathbf{q}') = p^\star(\mathbf{q}')T(\mathbf{q}',\mathbf{q}).$$

Para obtener muestras de una dada distribución usando cadenas de Markov la misma debe ser invariante y además la distribución debe converger a la deseada con suficientes pasos, sin importar la elección de distribución inicial $p(\mathbf{q}^0)$. Es decir,

$$p(\mathbf{q}^m) \xrightarrow{m\to\infty} p^\star(\mathbf{q}),$$

en cuyo caso la cadena es *ergódica*, y la distribución invariante es la distribución en *equilibrio.*

Teniendo esto en cuenta, se van a armar cadenas de Markov para muestrear la distribución de probabilidades. Al tener el estado $\mathbf{q}^\tau$ para una cierta instancia $\tau$, se propone la distribución $q(\mathbf{q}|\mathbf{q}^\tau)$ para el siguiente paso de manera iterativa. En cada ciclo se genera un candidato a muestra, $\mathbf{q}^\star$, desde la distribución $q$ y se acepta sólo si cumple con un criterio de aprobación. En los casos más sencillos, se va a aceptar la nueva muestra con probabilidad de una cierta tasa de aceptación $A$ que depende de la muestra en un instante y la propuesta para el siguiente. Una definición sencilla para esta tasa puede ser

$$A(\mathbf{q}^\star,\mathbf{q}^\tau) = min\left(1, \frac{\tilde{p}(\mathbf{q}^\star)}{\tilde{p}(\mathbf{q}^\tau)}\right),$$

donde $\tilde{p}(\mathbf{q}) = p(\mathbf{q})Z_p$, con $Z_p$ una constante de normalización desconocida [22]. En caso de aceptar la muestra, $\mathbf{q}^{\tau+1} = \mathbf{q}^\star$, y si se rechaza $\mathbf{q}^{\tau+1} = \mathbf{q}^\tau$, pudiendo tener una cadena con valores repetidos.

### 4.1.2. Hamiltonian Monte Carlo

Se tiene que encontrar la manera de transicionar de un estado a otro de manera óptima, recorriendo el espacio de parámetros sin tener que descartar muchas muestras y generando un gran número de muestras independientes. Hay múltiples algoritmos de muestreo disponibles para armar las cadenas de Markov, como el Metropolis-Hastings [23] o Gibbs [24]. El de interés en este trabajo es el Monte Carlo híbrido o Hamiltoniano, así nombrado al basarse en esta dinámica, que presenta mejoras respecto de los anteriores, incluso si los mismos se realizan de manera adaptativa.

En la dinámica Hamiltoniana, un sistema tiene coordenadas de posición $\boldsymbol{q} \in R^d$ y momento $\boldsymbol{p} \in R^d$ para un objeto en el espacio con $d$ dimensiones y una matriz de masa $M \in R^{dxd}$ diagonal y definida positivas. Los momentos vienen dados como la tasa de cambio de la variable $q$ en el tiempo, $p_d = \frac{dq_d}{dt}$. La energía total del sistema es el Hamiltoniano

$$H(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}) + K(\boldsymbol{p}),$$

con ecuaciones Hamiltonianas

$$\begin{aligned} \frac{d\boldsymbol{q}}{dt} &= \frac{\partial H}{\partial \boldsymbol{p}} \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{q}}. \end{aligned} \tag{4.2}$$

Por el teorema de existencia y unicidad de ecuaciones diferenciales ordinarias se tiene que debe existir una única función $T_t$ que evolucione el sistema de manera

$$T_t(\boldsymbol{q}_0, \boldsymbol{p}_0) \longrightarrow (\boldsymbol{q}_t, \boldsymbol{p}_t).$$

La dinámica Hamiltoniana es reversible en $\boldsymbol{p}$, preserva el volumen y conserva la energía ($\frac{dH}{dt} = 0$) (desarrollo en el apéndice A.1). Bajo estas condiciones, la dinámica Hamiltoniana deja invariante la distribución

$$p(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{Z_H} exp(-H(\boldsymbol{q}, \boldsymbol{p})), \tag{4.3}$$

donde $Z_H$, llamada función de partición, es una constante de normalización.

Para generar muestras de la distribución posterior de la variable de interés ($\boldsymbol{q}$), el método

Monte Carlo Hamiltoniano o HMC aumenta el espacio de parámetros agregando la variable auxiliar Gaussiana $\boldsymbol{p}$. Se construye la energía global como el logaritmo negativo de la distribución de probabilidad conjunta de la variable de interés y dicha variable auxiliar.

$$\begin{aligned} H(\boldsymbol{q},\boldsymbol{p}) &= -log\left(\pi(\boldsymbol{q}) \times \phi_M(\boldsymbol{p})\right) \\ &\propto -log(\pi(\boldsymbol{q})) + \frac{1}{2}\boldsymbol{p}^T M^{-1}\boldsymbol{p}, \end{aligned} \tag{4.4}$$

La ecuación 4.4 detalla cómo, bajo la hipótesis de independencia, la energía global tiene un término potencial que va con la distribución de probabilidades buscada ($\pi\left(\boldsymbol{q}\right) = p(\boldsymbol{q})$), y un término cinético que es cuadrático en $\boldsymbol{p}$. Por lo tanto, la dinámica Hamiltoniana conserva la distribución conjunta de la variable.

Dada un estado inicial de nuestro parámetro de interés $\boldsymbol{q}_0$, se propone un nuevo estado para la cadena al evolucionar en el tiempo las ecuaciones Hamiltonianas a un tiempo $t$.

Se toma un $\boldsymbol{p}_0 \sim N(0, M)$, se obtiene una nueva muestra $\boldsymbol{q}_t$ evolucionando en el tiempo $(\boldsymbol{q}_t, \boldsymbol{p}_t) = T_t(\boldsymbol{q}_0, \boldsymbol{p}_0)$, que será aceptada según la tasa A:

$$A(\boldsymbol{q}_t, \boldsymbol{p}_t, \boldsymbol{q}_0, \boldsymbol{p}_0) = min\left(1, \frac{\pi(\boldsymbol{q}_t)\phi_M(\boldsymbol{p}_t)}{\pi(\boldsymbol{q}_0)\phi_M(\boldsymbol{p}_0)}|\nabla T|\right), \tag{4.5}$$

donde recordamos que $|\nabla T| = 1$ por la conservación del volumen. De manera exacta, se tendría $A = 1$ (ver apéndice A.2). En la práctica, la evolución continua se aproxima usando el método de discretización del salto de rana, o *leapfrog*. Este método actualiza de manera alternada la posición y el momento con $L$ pasos, o *leaps*, de tamaño $\epsilon$, como se detalla en la ecuación 4.6. Se sigue teniendo reversibilidad en $\boldsymbol{p}$ y se preserva el volumen (ver apéndice A.3) pero se logran variaciones de la energía.

$$\begin{aligned} \boldsymbol{p}\left(t + \frac{\epsilon}{2}\right) &= \boldsymbol{p}(t) + \frac{\epsilon}{2}\nabla log\ \pi\left(\boldsymbol{q}(t)\right), \\ \boldsymbol{q}\left(t + \epsilon\right) &= \boldsymbol{q}(t) + \epsilon M^{-1}\boldsymbol{p}\left(t + \frac{\epsilon}{2}\right), \\ \boldsymbol{p}\left(t + \epsilon\right) &= \boldsymbol{q}\left(t + \epsilon\right) + \frac{\epsilon}{2}\nabla log\ \pi\left(\boldsymbol{q}\left(t + \epsilon\right)\right) \end{aligned} \tag{4.6}$$

Este algoritmo usará entonces información de la distribución de probabilidades, pero también de su gradiente. Con este método se busca recorrer el espacio de fases y llegar a una muestra que sea independiente del estado inicial, manteniendo una tasa de aceptación alta. La elección de los parámetros debe ser tal que el tamaño del paso $\epsilon$ no sea muy pequeño por costos computacionales, pero tampoco muy grande porque deja de ser correcta la discretización.

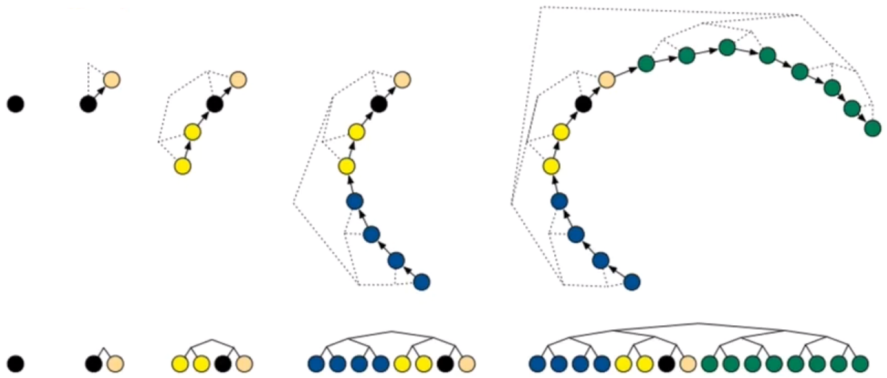


Figura 4.1: Cadena de muestreo según método NUTS. Sección superior muestra la distribución de los puntos en el espacio (2D) de fases, mientras que la sección inferior denota la cadena formada. Imagen tomada de Hoffman y Gelman [25].

Además, $L$ no puede ser muy grande, para evitar la posibilidad de regresar al punto de partida, pero tampoco puede ser muy chico, para evitar comportamientos de caminata azarosa.

Para eliminar la necesidad de definir el número de pasos $L$, se emplea el No-U-Turn-Sampler HMC, o NUTS [25]. Comenzando en un punto en cada iteración, se realizan pasos en profundidad de árbol, dando tantos pasos -hacia adelante o hacia atrás aleatoriamente- como el número de puntos se tenga en este momento. La figura 4.1 ilustra este proceso, partiendo del punto negro inicial, tomando una muestra (naranja) hacia adelante, luego dos muestras (amarillas) y cuatro (azules) hacia atrás y finalmente ocho (verdes) hacia adelante. Se frena el muestreo cuando ocurre alguna de las siguientes condiciones: la dirección del último paso respecto del primero de la cadena es mayor o igual a 90°, o hay una divergencia en la trayectoria, que implica energía potencial infinita. Finalmente, se toma como propuesta el paso más alejado que cumpla con estos criterios, evaluado según la condición de Metropolis definida en la ecuación 4.5.

## 4.2. Lenguaje de programación probabilística - PyMC

Un lenguaje de programación probabilística -*PPL* por sus siglas del inglés Probabilistic Programming Languages- funciona como el nexo entre la persona que investiga y la computadora. Permite plantear modelos probabilísticos de variables aleatorias y realizar inferencia sobre ellas. El objetivo es tener un marco sobre el cual establecer el medio de comunicación para poder transcribir los modelos a código ejecutable. Desde una perspectiva de ingeniería de sistemas, se puede considerar que cualquier PPL consiste de dos componentes: una interfaz para que el usuario defina el modelo, y algoritmos computacionales empleados para realizar la inferencia,

donde entra la elección de método para calcular la distribución posterior según la velocidad de inferencia buscada, el hardware necesario, la complejidad deseada, entre otras cosas [26].

En el marco de este trabajo se utilizó el paquete de *PyMC* como PPL en *Python*, que permitió realizar inferencia sobre modelos bayesianos mediante estimaciones de distribuciones con *HMC* [27, 28, 29, 30].

### 4.2.1. Desarmando la caja negra/Preliminares

#### Definición de modelo simple

A modo de introducción, utilizaremos el caso ilustrado por el modelo en 4.2 donde se tienen $N$ datos observados de una variable unidimensional 'x' que se suponen generados idénticamente a partir de alguna distribución de media $\mu$ y desviación $\sigma$, ambas variables latentes con valores a inferir. Eligiendo un prior normal y uniforme para $\mu$ y $\sigma$ respectivamente, podemos escribir este modelo como

$$\begin{aligned} \mu &\sim N(\mu|media\_prior, desvio\_prior), \\ \sigma &\sim U(\sigma|minimo, maximo), \\ x &\sim N(x|\mu, \sigma), \end{aligned} \tag{4.7}$$

En el ejemplo de código 4.2.1 se ve la estructura básica necesaria para definir un modelo generativo en *PyMC* como el de la ecuación 4.7, así como también cómo tomar muestras de la distribución posterior para realizar la inferencia.

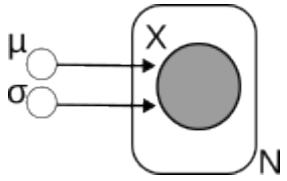



Figura 4.2: Modelo gráfico simple de la variable '**x**', observada $N$ veces, dependiente de los parámetros $\mu$ y $\sigma$.

Se comienza llamando a la clase de modelos para asignarle un nombre, en este caso "*modelo_simple*". Al modelo como objeto se le definen los *priors* para las los parámetros; a la media $\mu$ se le asigna una distribución normal con algún valor de media_prior y desvío_prior y a la desviación $\sigma$ una densidad uniforme entre valores mínimo y máximo. En la definición de la verosimilitud se especifica cómo se supone que la variable 'x' está siendo generada desde una

```
import pymc as pm
#x = vector de valores medidos de la variable aleatoria X
#Modelo
with pm.Model() as modelo_simple:
    with pm.Model() as model:
    #Priors
    media = pm.Normal(''media'', mu = media_prior, sigma =
        desvio_prior)
    sigma = pm.Uniform(''sigma'', lower = minimo, upper = maximo)
    #Verosimilitud
    likelihood = pm.Normal(''x'', mu = media, sigma=sigma, observed=
        data_observada)
# Llamado a inferencia para distribuciones posteriores de las
    variables
with modelo_simple:
    idata = pm.sample(draws, chains, tune)
```

Ejemplo de código 4.1: Modelo simple para una variable. Definición del modelo y posterior llamado al sampleo para generar muestras de las distribuciones posteriores.

distribución normal cuya media y desviación son los parámetros a inferir que ya fueron definidos, y también se fijan sus valores a los datos observados, con el argumento de "*observed*". Para realizar la inferencia propiamente dicha, y establecer las distribuciones posteriores a partir de sampleos iterativos de muestras para las variables $\mu$ y $\sigma$, se debe llamar a la función *pm.sample* que toma como argumento el número de muestras (draws), de cadenas (chains) y de puntos de ajuste (tune).

Dado un vector de 8000 observaciones de $x$ tomadas de una distribución $N(x|\mu = 5, \sigma = 4)$ se modela con priors poco informativos, asignando a la media

$$\mu \sim N(0, 100)$$

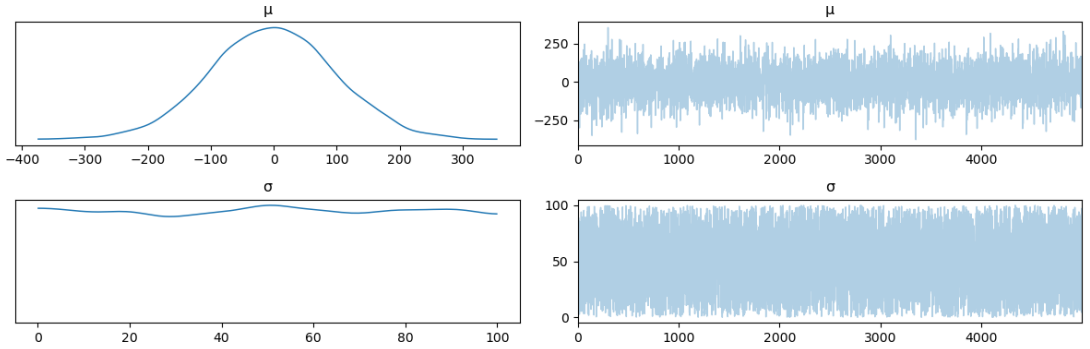
y al desvío

$$\sigma \sim U(0, 100),$$

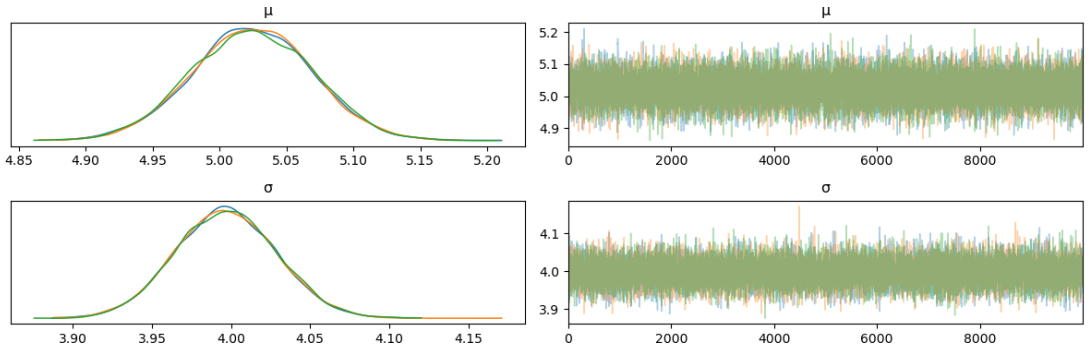
mientras que la verosimilitud de la variable queda definida como

$$x \sim N(x|\mu, \sigma),$$

y se maximiza con respecto a $\mu$ y $\sigma$ según los $\mathbf{X}$ observados. Bajo estas definiciones, se realiza




(a) Priors sampleados desde una cadena informada por la definición de funciones del modelo.




(b) Posteriors inferidas a partir de tres cadenas (verde, naranja y azul).

Figura 4.3: Densidades de probabilidad aproximadas (izq.) a partir de las trazas del HMC (der.)

inferencia con 3 cadenas de 10,000 puntos a partir de las cuales se obtiene la estimación de los parámetros buscados. Este trabajo se ilustra en la figura 4.3, gráficos realizados con la librería Arviz que maneja los "*xarray*" - matrices de distintas dimensiones- en que se guardan los objetos de inferencia [31].

En el panel 4.3a muestra las distribuciones poco informativas a priori. A la izquierda se tiene la distribución formada a partir de una única traza (derecha) generada a partir de las formas funcionales definidas en el modelo. Se observa lo difusas que son las distribuciones, con $\sigma$ teniendo probabilidades similares en un amplio rango y $\mu$ teniendo una forma restringida, pero con gran desviación estándar. En el panel 4.3b se muestran las distribuciones posteriores de estas variables, luego de incluir los datos. A la derecha se ven tres trazas muy similares entre sí, por lo que apenas se distinguen unas de otras, y a la izquierda las distribuciones que éstas generan a partir de una estimación de densidad. Se observan dos cosas principales:

- para ambos parámetros, las tres cadenas convergieron a un mismo resultado y

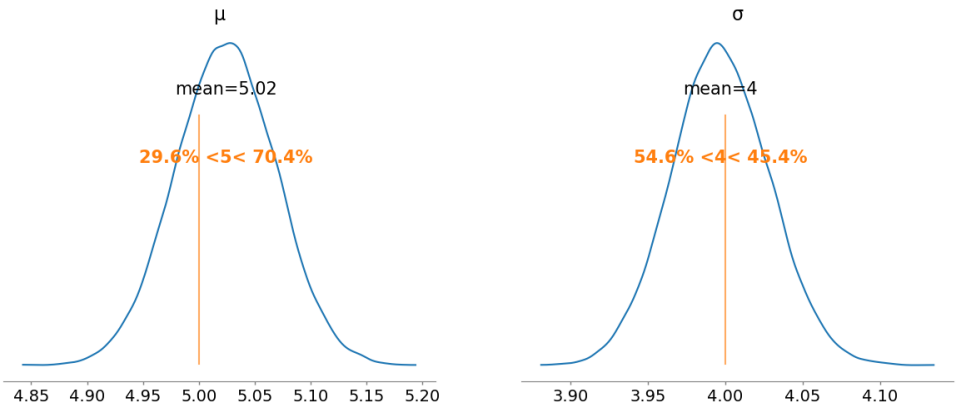



Figura 4.4: Densidades de probabilidad para los parámetros $\mu$ y $\sigma$ del modelo simple (azul), valores de referencia como rectas verticales (naranja) y valor medio estimado en negro.

- se reduce la incerteza de la distribución posterior. Al informarse con los datos, se disminuyen las anchas distribuciones a priori con las que comienza.

La figura 4.4 muestra las densidades totales a partir del conjunto de cadenas muestreadas. Se ve dónde están los valores a partir de los cuales se generaron los datos como rectas verticales y se informa el porcentaje de densidad que hay a cada lado de dicho valor en naranja. Las estimaciones para los parámetros son $\mu = 5{,}02 \pm 0{,}04$ y $\sigma = 4{,}0 \pm 0{,}03$.

### Influencia de los parámetros

Para la comprensión del modelado en este PPL se realizaron múltiples estudios preliminares sobre el modelo mostrado. En particular, se buscó comprender el efecto de los argumentos que alteran el sampleo. Algunas de las observaciones encontradas se detallan a continuación, mientras que los gráficos de soporte pueden hallarse en el apéndice B.

**Número de observaciones.** Como ya se explicó previamente, las observaciones ingresan al modelo no sólo al declarar las relaciones funcionales en las cual se basa el mismo, sino también como condicionales para la distribución de verosimilitud. A mayor número de observaciones ($N$), menor es la incerteza en las estimaciones de los valores de los parámetros.

**Trazas.** Las cadenas dependen de múltiples parámetros. A continuación se discuten los considerados en este trabajo. Cada cadena precisa un cierto número de puntos de entrenamiento, definidos como "*tuning steps*", donde el algoritmo define los parámetros de número de pasos y escala de los mismos entre cada muestra ($L$ y $\epsilon$ de la ecuación 4.6), colaborando a la convergencia final de la cadena. Estos puntos son descartados para la inferencia final. El número de cadenas "*chains*" determina cuántas trazas se correrán de manera independiente. Este argumento es

especialmente útil cuando se tiene un problema de alta dimensión ya que permite inicializar cada una en distintos puntos del espacio de parámetros para ayudar a explorar un área mayor del mismo. Las muestras de cada cadena pueden estar autocorrelacionadas, ya que cada punto tiene información de 'k' progenitores anteriores. Debido a esto, cada cadena puede tener un tamaño de muestra estimado menor que el pedido. El diagnóstico "*ess*" por sus siglas del inglés *estimated sample size* mide el número de muestras descocorrelacionadas y está aproximado como

$$Ess = \frac{CM}{\hat{\tau}},$$
$$\hat{\tau} = 1 + 2\sum_{t=1}^{2c+1} \hat{\rho}_t,$$

donde C es el número de cadenas, M el número de muestras o 'draws', y $\hat{\rho}_t$ es el estimador de autocorrelación de muestras con desfasaje $t$. Otro estadístico de diagnóstico es el coeficiente *r-hat*, $\hat{R}$, que compara la varianza entre cadenas con la varianza dentro de cada cadena.

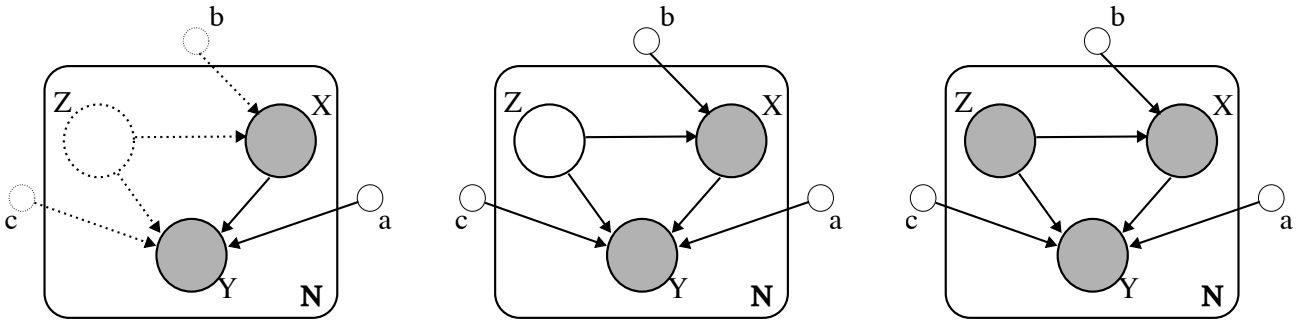
$$\hat{R} = \frac{\hat{V}}{W}$$

Acá $\hat{V}$ es el estimador de varianza posterior de las cadenas agrupadas, y $W$ la variación de cada cadena. En el límite tiende a 1, mientras que un valor mayor indica que una o más cadenas aún no convergieron. Realizar un "thinneado" no es estrictamente necesario para tener buenas estimaciones; al tener cadenas largas termina promediándose y cancelándose el ruido de la cadena [32].

Dado el gran número de dimensiones con las que se trabajará al manejar los datos de la UNSAM en el capítulo 6, si bien el método de muestreo default al usar "*sample*" es el *NUTS*, se usó "*sampling_jax.sample_numpyro_nuts*" que llama a Numpyro y JAX desde PyMC para paralelizar los muestreos y acelerar el proceso [33, 34, 35].

## 4.3. Camino al deconfounder: modelos de juguete

A modo de inmersión en las librerías se plantearon tres modelos sencillos sobre datos sintéticos generados de manera aleatoria con estructura de PPCA como se vió en la figura 2.3b y describió en la sección 3.3.1. Se generaron $N$ muestras de $D$ covariables y $K$ variables confundidoras que se combinaron a través de 3 matrices de coeficientes conocidos para obtener una




(a) *Unconfounded.* (b) *Confounded.* (c) *Oráculo.*

Figura 4.5: Representación gráfica de los modelos empleados. La variable $Y$ es el 'target' y $X$ representa las covariables que buscan explicarlo, ambas son observadas en todos los casos. La variable confundidora $Z$ codifica la cantidad de información presente en cada modelo, desde total desconocimiento en el 'unconfounded', inclusión sin observarla y finalmente inclusión con observación en el caso del modelo 'oráculo'.

variable objetivo o *target* de dimensión $T$.

Conociendo el verdadero método generativo, se realizaron tres modelizaciones sobre el mismo set de datos llamadas '*unconfounded*', '*confounded*' y '*oráculo*' que se exponen en la figura 4.5. Se usó la notación usual donde cada nodo sombreado representa una variable aleatoria observada y los no sombreados son variables latentes en 4.5c y 4.5b. Se agregó la notación de enlaces y nodos punteados en 4.5a para representar variables que no están presentes en el modelo de inferencia, pero sí fueron usadas para generar la variable objetivo. Al tener datos sintéticos, puede realizarse la inferencia suponiendo cada modelo y luego analizar la varianza y sesgo de las estimaciones respecto de los valores reales.

### 4.3.1. Generación de datos

El modelo generativo usado para cada $i = 1...N = 1200$ dato fue

$$X_i = bZ_i + \epsilon_{x,i},$$

$$Y_i = aX_i + cZ_i + \epsilon_{y,i},$$

donde cada $Z_i \in \mathbb{R}^K$ es el vector de las variables confundidoras, $X_i \in \mathbb{R}^D$ el vector de las covariables para cada observación $Y_i \in \mathbb{R}^T$, todos independiente e idénticamente distribuidos. Los errores $\epsilon_{x,y}$ para las covariables y objetivo respectivamente, son errores estándar de distribución $N(0, 1)$. Las matrices $b \in \mathbb{R}^{DxK}$, $a \in \mathbb{R}^{TxD}$ y $c \in \mathbb{R}^{TxK}$ se definieron una vez y fueron usadas

para todos los puntos. Con esta relación se generaron dos bases de datos:

1. $B_1$, donde $D = 3$, $K = 2$ y $T = 2$
2. $B_2$ con $D = 3$, $K = 2$ y $T = 1$

Con la base $B_2$ se generó un target alternativo de relación no lineal para evaluar qué modelo de inferencia capta mejor la discrepancia en caso de tener datos que no siguen las hipótesis del muestreado. Para esto, se cambió la generación de la variable objetivo por

$$Y_{i,NL} = aX_i + cZ_i^2 + \epsilon_{y,i}.$$

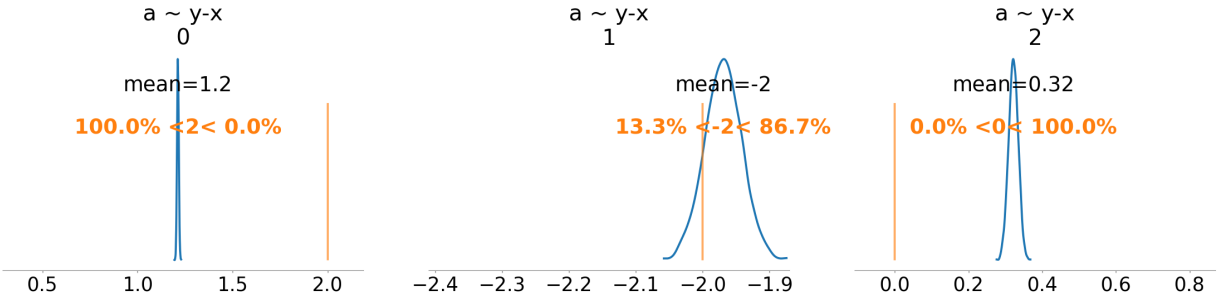
### 4.3.2. Resultados

Ponemos el interés en este momento en la matriz $a$ por dos motivos principales: es una variable compartida por los tres modelos mencionados, permitiendo una comparación directa, y además es la variable de interés en los estudios futuros, puesto que vamos a buscar caracterizar la relación entre las covariables y el objetivo, aún ante la presencia de posibles variables confundidoras. Todas las distribuciones de las tres modelizaciones planteadas se infirieron a partir de dos cadenas de 3000 muestras con 1800 puntos de tuneo, permitiendo tener análisis equivalentes.

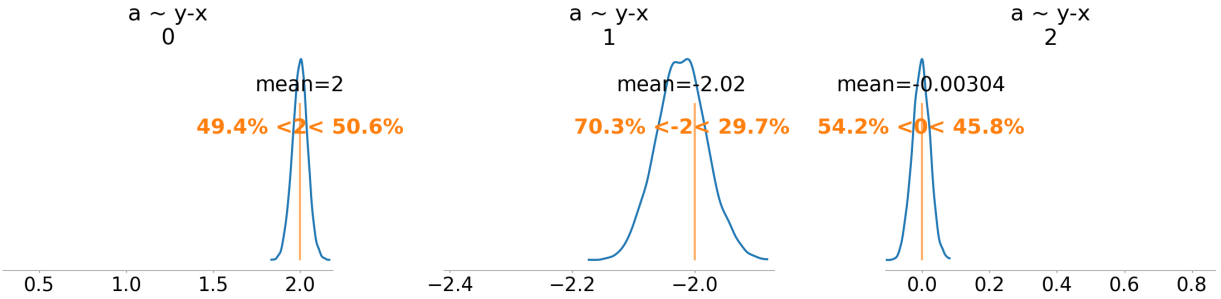
#### Base $B_1$

Dadas las dimensiones de la matriz $a$ en esta base, se separaron las densidades posteriores en dos figuras,4.6 y 4.7, una por cada fila de la matriz.

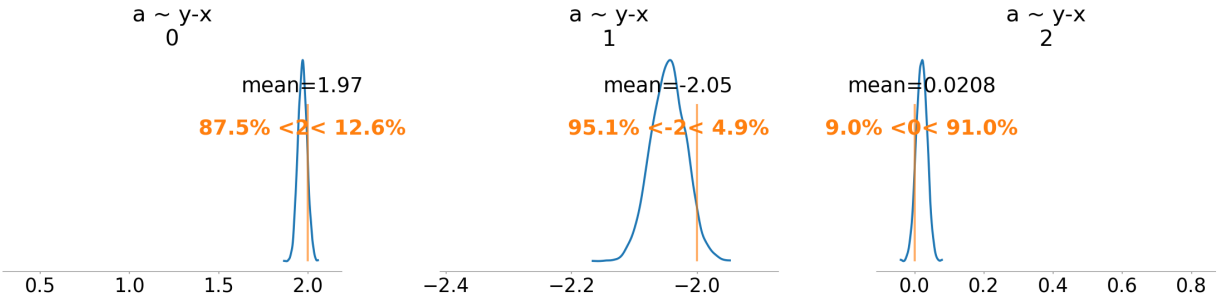
De manera general puede verse en 4.6a y 4.7a que el modelo 'unconfounded' predice valores sesgados de los coeficientes, sólo una de las coordenadas, $a_{0,1}$, captura el valor real. El modelo 'confounded' (4.6b y 4.7b) y 'oráculo' (4.6c y 4.7c) resultan en distribuciones que captan los valores reales, pero el oráculo presenta menos dispersión, como es de esperarse si se muestran todos los datos. En la tabla 4.1 se reportan los valores de sesgo cuadrático promedio y varianza promedio para esta base de datos y los tres modelos y se cuantifica lo aquí descripto. El modelo unconfounded tiene el máximo sesgo, con igual orden de magnitud en la dispersión que el oráculo, es decir, tiene distribuciones angostas alrededor de valores erróneos. El modelo confounded, donde consideramos la existencia de variables confundidoras sin observarlas, tiene el mismo sesgo que el oráculo, pero mayor dispersión. Es decir, logra captar los valores como con el oráculo, pero hay más incerteza en sus estimadores.




(a) *Unconfounded*.




(b) *Confounded*.




(c) *Oráculo*.

Figura 4.6: Distribuciones posteriores para los coeficientes de la primera fila de la matriz $a$ aplicada a la base $B_1$. Los coeficientes $j$ indican de qué coordenada $a_{0,j}$ de la matriz se trata.

**Base $B_2$**

Las figuras 4.8 y 4.9 muestran las distribuciones de los 3 coeficientes de la matriz $a$ para las bases de datos $B_2$ generados de manera lineal o cuadrática con la variables confundidora respectivamente. Sus valores de sesgo cuadrático medio y varianza también se muestran en la tabla 4.1.

En el caso del modelado lineal, se repite el comportamiento reportado para $B_1$, con la salvedad que el modelo confounded tiene, bajo los mismos parámetros de muestreo que los otros modelos, distribuciones posteriores asimétricas como se ve en 4.8b. Con priors para los coeficientes poco informativos, definidos como $a_{ij} \sim N(0, 50)$, las 1200 observaciones pueden no ser suficientes para inferir correctamente las distribuciones de las matrices. Aumentar el número de observaciones atenúa este comportamiento pero el mismo no desaparece. El modelo 'confounded' es el más problemático a la hora de realizar inferencia ya que tiene un mayor grado de libertad, proponiendo la existencia de $K$ variables no observadas para cada observación dada, que sería lo que ocurre con los sistemas en la realidad. El modelo 'oráculo' tiene el mismo número de variables, pero se está ante la situación idílica de contar con las observaciones de las variables confundidoras. Esto permite que las distribuciones para los parámetros tengan mayor definición, como se ve en 4.8c y reporta el mínimo sesgo y varianza de los tres modelos.

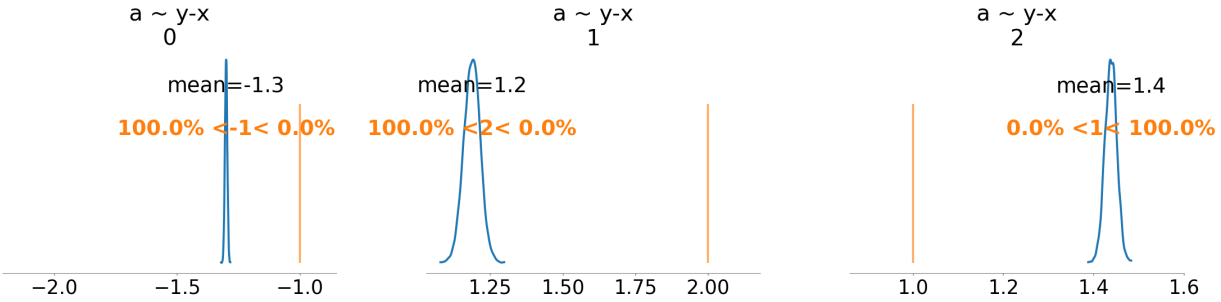
Finalmente, cuando los datos son generados de manera no lineal se obtuvieron las distribuciones ilustradas en en 4.9. El modelo 'confounded' haya distribuciones muy anchas que abarcan los valores verdaderos, pero tienen una forma aún más alejada de una gaussiana (4.9b). En este caso el modelo tiene más problemas para converger y presenta mayor dispersión ya que nos encontramos en el caso donde se plantea la existencia de variables confundidoras latentes, teniendo pocas observaciones para el gran número de variables libres, pero además el modelo planteado difiere del modelo generativo.

Llamativamente, el modelo unconfounded genera distribuciones de un orden menos de sesgo cuadrático promedio y el mismo orden de varianza que el oráculo, confirmando que si se presentan todas las variables pero no se indica el modelo correcto se pueden obtener inferencias sesgadas. Este hecho es relevante puesto que en situaciones reales se tiene conjeturas respecto del método generativo de los datos recibidos, y además se sabe que no se tiene todas las posibles covariables controladas y observadas. De esta forma, en cada investigación se parte bajo las condiciones del modelo 'confounded' presentado en este capítulo. Según el conocimiento que se tenga del caso a estudiar, los modelos propuestos estarán mas o menos alejados de la situación real, pero casi con absoluta certeza se tendrán variables no observadas que pueden afectar al resultado final y a otras variables. Se precisa contar con un método que permita introducir ob-

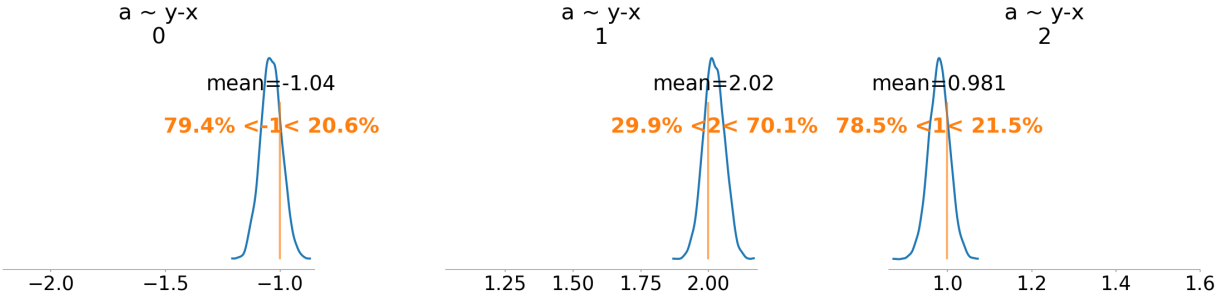
| | $B_1$ | | $B_2$ | | $B_2$ $(NL)$ | |
|---|---|---|---|---|---|---|
| **Modelo** | $\bar{S^2}$ | $\bar{\sigma^2}$ | $\bar{S^2}$ | $\bar{\sigma^2}$ | $\bar{S^2}$ | $\bar{\sigma^2}$ |
| **Unconfounded** | $2{,}8 \times 10^{-1}$ | $3{,}3 \times 10^{-4}$ | $1{,}0 \times 10^{-1}$ | $1{,}1 \times 10^{-3}$ | $5{,}3 \times 10^{-2}$ | $7{,}1 \times 10^{-2}$ |
| **Confounded** | $4{,}5 \times 10^{-4}$ | $1{,}4 \times 10^{-3}$ | $1{,}8 \times 10^{-3}$ | $5{,}5 \times 10^{-2}$ | $3{,}0 \times 10^{-1}$ | $1{,}6$ |
| **Oráculo** | $6{,}3 \times 10^{-4}$ | $5{,}5 \times 10^{-4}$ | $1{,}3 \times 10^{-4}$ | $5{,}9 \times 10^{-4}$ | $1{,}6 \times 10^{-1}$ | $1{,}1 \times 10^{-2}$ |

Tabla 4.1: Sesgo cuadrático promedio ($\bar{S^2}$) y varianza promedia ($\bar{\sigma^2}$) de los coeficientes de las matrices $a$ según la base usada para generar los datos y el modelo aplicado para inferir su distribución posterior.

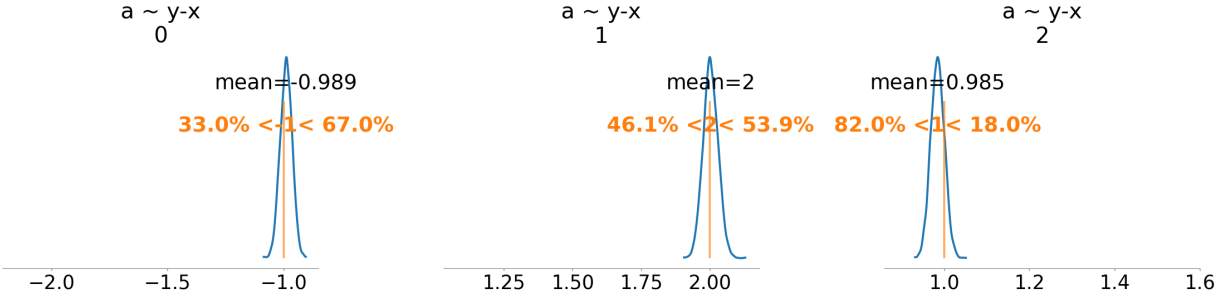
servaciones de estas variables confundidoras, como el del *deconfounder* propuesto en el capítulo 3, para colaborar con el muestreo de las cadenas y evitar caer en distribuciones con amplias desviaciones que aumenten el error en nuestras estimaciones.




(a) *Unconfounded*.

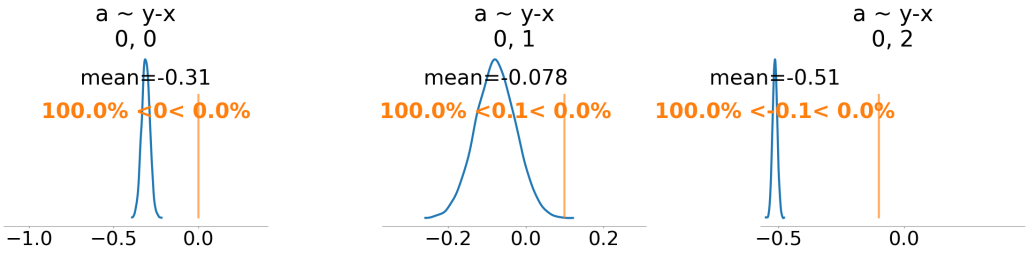



(b) *Confounded*.

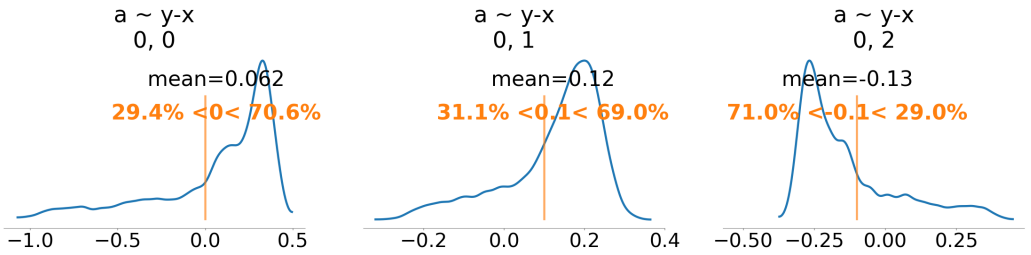



(c) *Oráculo*.

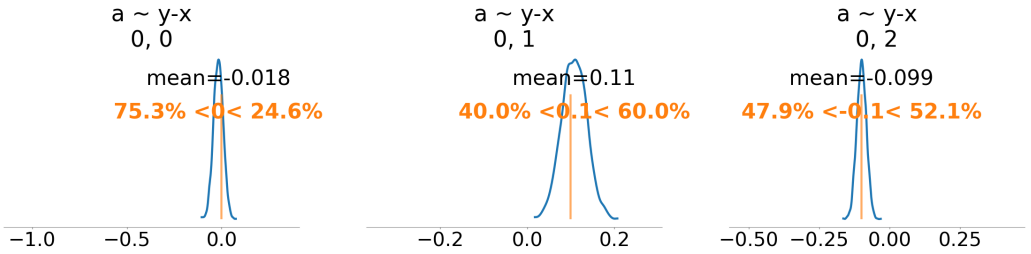
Figura 4.7: Distribuciones posteriores para los coeficientes de la segunda fila de la matriz $a$ aplicada a la base $B_1$. Los coeficientes $j$ indican de qué coordenada $a_{1,j}$ de la matriz se trata.




(a) *Unconfounded*.

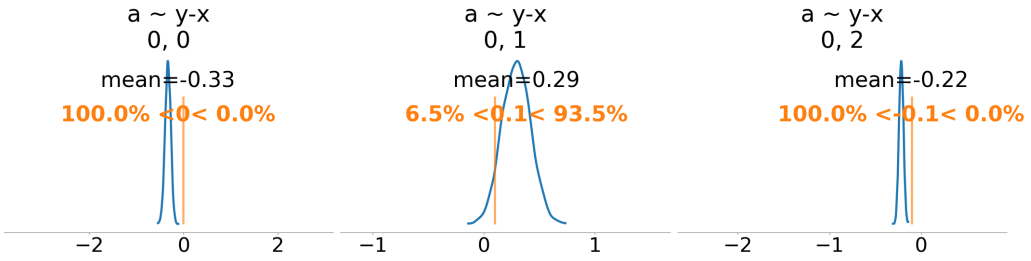



(b) *Confounded*.

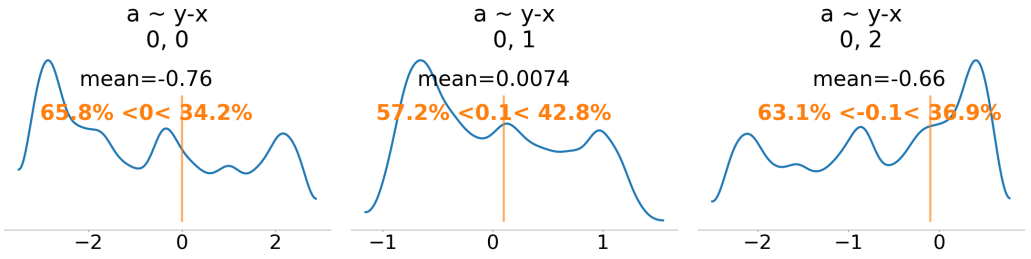



(c) *Oráculo*.

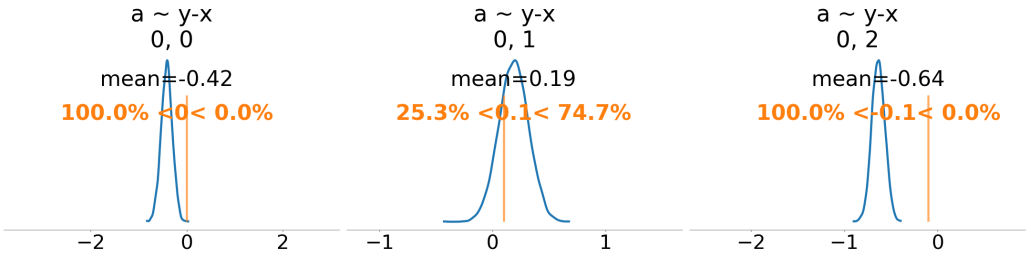
Figura 4.8: Distribuciones posteriores para los coeficientes de la matriz $a$ aplicada a la base $B_2$.




(a) *Unconfounded*.




(b) *Confounded*.




(c) *Oráculo*.

Figura 4.9: Distribuciones posteriores para los coeficientes de la matriz $a$ aplicada a la base $B_{NL}$.

# Capítulo 5

# Inferencia bayesiana sobre casos concretos ya estudiados

Habiendo motivado la necesidad de simular las variables confundidoras latentes consideradas en el modelo, este capítulo detalla cómo se aplicó el método de 'deconfounder' sobre dos ejemplos preexistentes para la validación del algoritmo. Se usaron bases de datos de fumadores y de cáncer de mama, y se adaptó el código de Python que los autores Wang y Blei pusieron a disposición en GitHub. El código de los autores hace uso de la librería Edwards con inferencia variacional, por lo que fue adaptado para usar PyMC [36].

## 5.1. Estudio de 'deconfounder' aplicado a base de datos de fumadores

La primera aplicación a datos reales se realizó sobre los resultados de una encuesta gastos médicos nacionales, *NMES* por sus siglas del ingles, realizada sobre una muestra representativa de la población de USA de 9708 individuos sobre hábitos de fumar, para quienes se cuenta además con sus gastos médicos [37]. De las 8 variables disponibles, las relevantes fueron la última edad de exposición al tabaco ('*last_age*'), el tiempo total de exposición al mismo ('*exposure*'), ambas medidas en años, y el estado marital ('*marital*'). Esta última variable toma valores discretos entre 1 y 5 según el estado civil: 1 para personas casadas, 2 en el caso de haber enviudado, 3 para personas divorciadas, 4 en caso de haberse separado y 5 en caso de estar soltero. Estas variables no son cardinales *a priori*, pero sí se estableció una relación ordinal entre las mismas.

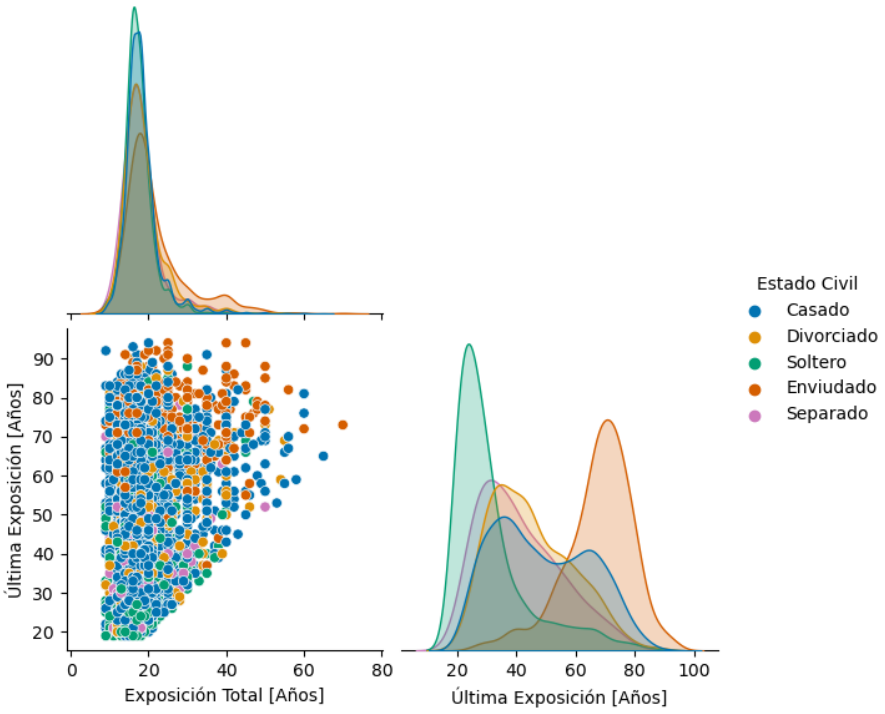



Figura 5.1: Relación entre los años de exposición total y la edad de la última exposición al tabaco. El estado marital se encuentra codificado con los colores. Las densidades en la diagonal se encuentran normalizadas en cada categoría.

En la figura 5.1 se tienen los perfiles de distribución de la exposición total y la edad de última exposición en función de la categoría de estado civil a la cual pertenecen. Si bien la exposición total tiene un máximo alrededor de 20 años para todas las categorías, puede verse cómo difieren los perfiles de última exposición; quienes están casados presentan dos máximos cercanos a los de las personas divorciadas y enviudadas. Mientras que los solteros tienen un pico a temprana edad, quienes enviudaron tienen una última exposición a mayor edad, en promedio.

A fin de poder contrastar los resultados obtenidos de la inferencia con algún valor de referencia, se generaron datos semi-sintéticos de las expensas definiendo un único vector de coeficientes $\boldsymbol{\beta} = (-0{,}7; 0{,}01; -0{,}95)$, cada uno de ellos fijados aleatoriamente a partir del muestreo de una distribución $N(0, 1)$. Las tres covariables son estandarizadas, restando la media y dividiendo por su desviación, $\tilde{x_i} = \frac{x_i - \mu}{\sigma}$ para que los valores sean muestras distribuidas de forma $\tilde{x_i} \sim N(0, 1)$. El target de expensas médicas de cada individuo se generó de la forma

$$y_n = \boldsymbol{\beta} \boldsymbol{x}_n + \epsilon_n,$$

donde

$$\boldsymbol{x_n} = (x_{n,marital}, x_{n,exposure}, x_{n,last\_age})$$

es el vector de covariables de cada unidad, y $\epsilon_n \sim N(0,1)$ su término de error. Si bien las tres covariables fueron usadas para generar el gasto médico, en el modelado se consideró como posibles causas a las variables de exposición total al tabaco y al estado marital, mientras que la edad de última exposición se dejó como confundidora latente; la última exposición tiene un impacto en los gastos médicos, y además influye en la exposición total y el estado civil, según la publicación de Wang y Blei.

Con este conjunto de datos de causas estandarizadas y target generado se proponen entonces dos modelos de factores, llamados lineal y cuadrático según la potencia de la variable confundidora latente, con priors

$$\begin{aligned} &z_n \sim N(0,1), \\ &W_{k,d}, W_{2\ k,d} \sim N(0, \sigma_W)\ y \\ &\sigma_W \sim HN(\sigma = 1), \end{aligned} \tag{5.1}$$

y verosimilitudes

$$\begin{aligned} &\boldsymbol{x_{lin}} \sim N\left(x|\mu_x = \mathbf{zW}, \sigma_x = 0{,}1\right), \\ &\boldsymbol{x_{cuad}} \sim N\left(x|\mu_x = \mathbf{zW} + \mathbf{z}^2\mathbf{W}_2, \sigma_x = 0{,}1\right). \end{aligned} \tag{5.2}$$

Se corrieron 5 cadenas con 4000 puntos de entrenamiento y 8000 puntos de traza final. Dado que las cadenas individuales se bloquean en mínimos locales disjuntos, se realizó una selección según la log-probabilidad de cada cadena, manteniendo únicamente la de mayor valor.

Para la selección de uno u otro modelo se realizó el chequeo predictivo llevando a cabo el modelo de factores con un 20 % de los datos enmascarados. A partir de 50 repeticiones de obtener muestras de los datos enmascarados a partir de la inferencia realizada sobre el resto del conjunto de datos, y contrastarlos con los valores verdaderos, se construyó el gráfico de la figura 5.2, donde los valores medios para el modelo de factores lineal fue $p - score_{lin} = 0{,}093 \pm 0{,}005$ y para el modelo cuadrático $p - score_{quad} = 0{,}18 \pm 0{,}04$. Se aprecia que el modelo de factores que supera el chequeo predictivo es el cuadrático. Los chequeos predictivos realizados en este trabajo superan, para ambos modelos, los valores reportados en la publicación de Wang y Blei. Hay que recordar que se realizó el estudio con una única cadena de máxima log-probabilidad, cabe la posibilidad de que nuestra técnica de muestreo esté produciendo muestras de un mínimo particular.

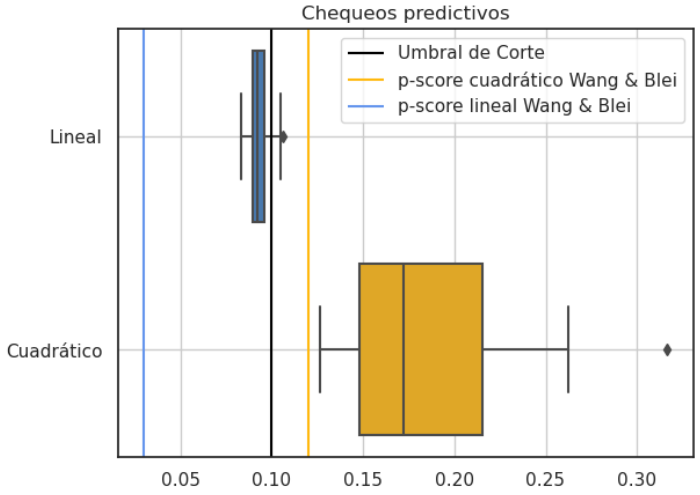



Figura 5.2: Gráficos de caja para los chequeos predictivos del modelo de factores lineal y cuadrático. Sus valores medios fueron $p - score_{lin} = 0{,}09 \pm 0{,}005$ y $p - score_{quad} = 0{,}18 \pm 0{,}04$. Las líneas verticales marcan los valores obtenidos en el paper de Wang y Blei (*azul* y *amarilla*), y el umbral de corte (*negro*) allí propuesto.

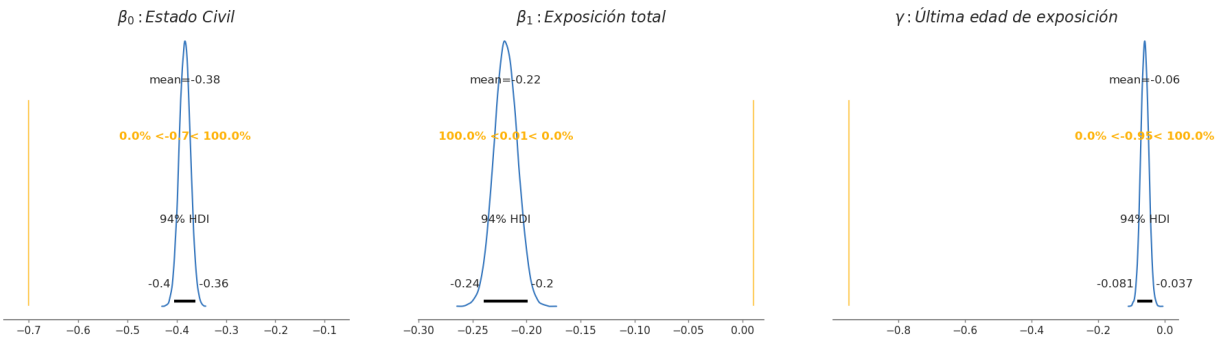
Se prosiguió a realizar la inferencia del modelo de factores cuadráticos sobre el conjunto de datos completos, con los mismos parámetros de muestreo ya mencionados. Se selecciona nuevamente la cadena de mayor probabilidad logarítmica, y con ella se obtiene, a partir de los valores medios de las densidades de las confundidoras, valores sustitutos para usar en el modelo de resultados.

El modelo de resultados establece la relación del target (los gastos médicos generados) con las causas mediante el vector $\beta$ y con las confundidoras sustitutas mediante el coeficiente $\gamma$, en el caso del modelo del *deconfounder*. Las distribuciones prior y la verosimilitud propuesta en cada modelo fueron

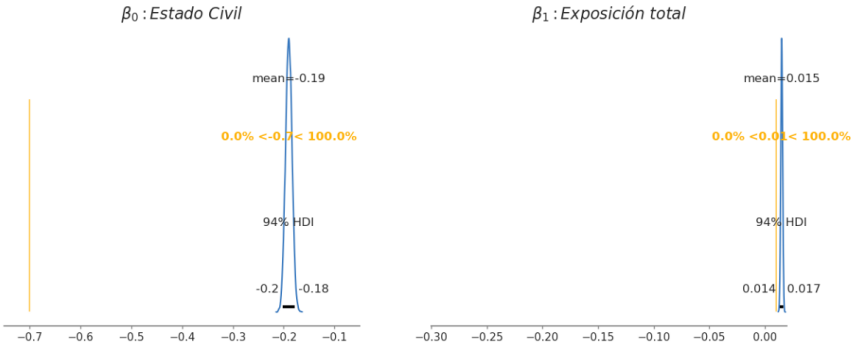
$$\begin{aligned} &\beta_i = N\left(\mu = 0, \sigma = 1\right), \ i = 1, 2, \\ &\gamma = N\left(\mu = 0, \sigma = 1\right), \\ &y_n^{deconfounder} = N\left(\mu = \boldsymbol{\beta}\boldsymbol{X}_n + \gamma\hat{\boldsymbol{Z}}_n, \sigma = 1\right), \\ &y_n^{naive} = N\left(\mu = \boldsymbol{\beta}\boldsymbol{X}_n, \sigma = 1\right). \end{aligned} \tag{5.3}$$

En el modelo *naive*, se establece que la variable confundidora no tiene influencia sobre el resultado, es decir que se impone $\boldsymbol{\gamma} = 0$.

La inferencia bajo estas condiciones generó distribuciones posteriores de los parámetros muy angostas y sesgadas, puesto que los valores reales con los cuales se generaron los datos




(a) Modelo *deconfounder*.




(b) Modelo *naive*.

Figura 5.3: Distribuciones posteriores de los parámetros $\beta$ y $\gamma$ con el modelado de datos bajo método del deconfounder y de los parámetros $\beta$ para el modelado naive. Las rectas verticales marcan los valores verdaderos de cada coeficiente.

observados no eran abarcados, como se ve en la figura 5.3. El sesgo cuadrático para el vector $\boldsymbol{\beta}$ obtenido con el modelado naive fue de $S^2_{naive} = 0{,}259 \pm 0{,}006$. El modelo deconfounder tuvo una leve mejora con sesgo cuadrático del vector $\beta$ de $S^2_{deconf.} = 0{,}15 \pm 0{,}01$, y el del coeficiente $\gamma$ de $S^2_\gamma = 0{,}79 \pm 0{,}02$.

Dado que se esperaba captar los coeficientes reales con mayor exactitud, se realizó el mismo estudio de inferencia bayesiana con un modelo *oráculo*. Es decir, teniendo como observadas el total de las covariables usadas para definir el target, por lo que el vector $\boldsymbol{\beta}$ y los vectores de covariables $\boldsymbol{X}_n$ tienen ahora tres dimensiones e incluyen las variables estado marital, exposición total y última edad de exposición. El modelo oráculo queda definido a través de

$$\beta_i = N\left(\mu = 0, \sigma = 1\right), \; i = 1, 2, 3,$$
$$y_n^{oraculo} = N\left(\mu = \boldsymbol{\beta}\boldsymbol{X}_n, \sigma = 1\right).$$

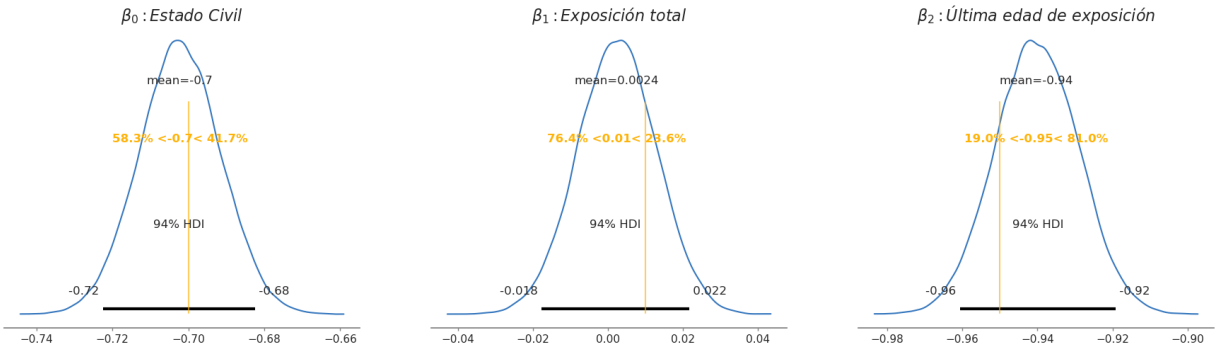



Figura 5.4: Distribuciones posteriores para los tres coeficientes del parámetro $\beta$ bajo el modelo *oráculo*.

| Modelo | $S^2$ | $S^2_{ref}$ |
|---|---|---|
| **Oráculo** | $1 \times 10^{-4}$ | $5{,}06 \times 10^{-2}$ |
| **Naive** | $2{,}59 \times 10^{-1}$ | $2{,}42 \times 10^{-1}$ |
| **Deconfounder** | $1{,}50 \times 10^{-1}$ | $1{,}78 \times 10^{-1}$ |

Tabla 5.1: Tabla con los valores de sesgo cuadrático para los coeficientes $\beta$ relacionados con el estado civil y exposición al tabaco. Se reportan los valores obtenidos en este trabajo y los de referencia del trabajo publicado en el paper del deconfounder [2].

Las distribuciones posteriores para los parámetros pueden verse en la figura 5.4, y el sesgo cuadrático en estos casos es del orden $O(10^{-4})$.

En la tabla 5.1 se muestran los valores de sesgo cuadrático total para las causas de estado civil y exposición total obtenidos en este trabajo, comparados con los valores de referencia presentados en el artículo de referencia. Como puede verse, si bien nuestros valores están sesgados y los gráficos de las distribuciones posteriores mostraban que no se llegaban a capturar los valores reales de los coeficientes, el sesgo de nuestros estimadores está en el orden del trabajo publicado. Puede ser que esta base de datos no sea la adecuada para probar un método para múltiples causas, puesto que estamos considerando sólo dos variables y usando una única variable confundidora latente. Además estamos asumiendo, de la misma manera que en el trabajo publicado, que la variable seleccionada como confundidora afecta las dos causas, pero no es algo que hayamos determinado ni cuantificado.

## 5.2. Estudio de 'deconfounder' aplicado a base de datos cáncer de mamas

Se replicó el estudio con el código adaptado sobre una base de datos de pacientes diagnosticados en base a estudios de cáncer de mama en Wisconsin, del año 1995 [38]. La misma cuenta con mediciones de atributos relevantes para la clasificación de tumores a partir del análisis de propiedades de núcleos celulares en las muestras histológicas de aspirado en punción de tejidos mamarios, y está disponible de manera libre en Scikit-Learn [39].

Dicha base de datos cuenta con 569 muestras reales con 32 atributos sin valores faltantes. Del total de los atributos computados, los relevantes en nuestro análisis son las 10 variables de valores promedio (números reales) de cada núcleo que se presentan en la tabla 5.2. Como *target*, se tiene el diagnóstico del tumor, maligno o benigno codificado con 1 o 0 respectivamente, los cuales se distribuyen en 212 unidades malignas y 357 benignas. Las variables descartadas del análisis son los errores estándar de cada categoría, y los llamados 'peores valores' que serían los máximos encontrados, o el promedio de los tres mayores, en toda la muestra histológica. Un análisis exploratorio de las variables muestra que el radio promedio del núcleo tiene gran

| # | Causa | # | Causa |
|---|---|---|---|
| **0** | Radio medio | **5** | Compacticidad media |
| **1** | Perímetro medio (*) | **6** | Concavidad media |
| **2** | Área medio (*) | **7** | Puntos cóncavos medios |
| **3** | Textura media | **8** | Simetría media |
| **4** | Suavidad media | **9** | Dimensión fractal media |

Tabla 5.2: Tabla de variables seleccionadas como causas del conjunto de datos del cáncer de mamas. Las causas con '(*)' indican que fueron removidas del estudio, debido a su correlación con el radio medio.

correlación con el perímetro promedio ($r_{pearson} = 0{,}998$) y con el área ($r_{pearson} = 0{,}987$). Para disminuir la varianza de los estimadores hacia el final del estudio y fortalecer la hipótesis de superposición detallada en la ecuación 3.14, se eliminan el área y el perímetro promedio como variables causales del análisis. Dada la relación entre el radio, perímetro y área del núcleo se consideró que mantener una única covariable no genera una gran pérdida de información en el estudio. El preprocesado de los datos entonces nos deja con 8 covariables de interés como causas asignadas. Las mismas son estandarizadas restando su valor medio y dividiendo por su desviación estándar para poder tener magnitudes comparables, de la misma forma que se hizo

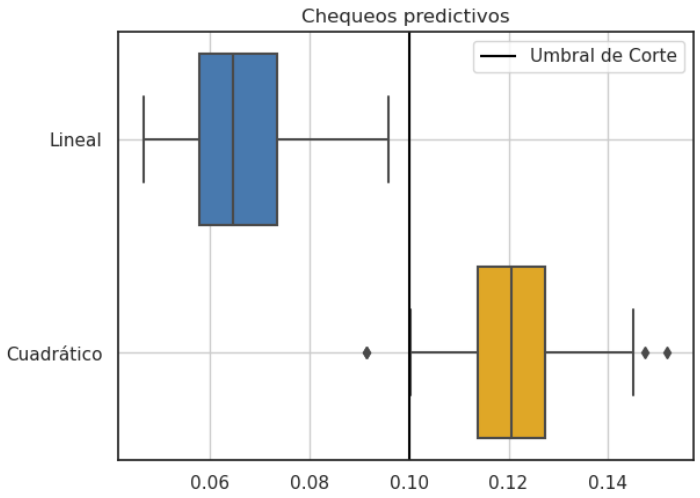



Figura 5.5: Gráficos de caja para los chequeos predictivos del modelo de factores lineal y cuadrático. Sus valores medios fueron $p - score_{lin} = 0{,}07 \pm 0{,}01$ y $p - score_{quad} = 0{,}12 \pm 0{,}01$.

en la sección anterior al tener atributos con unidades distintas.

Se realizó el mismo proceso para el modelo de factores que en la sección anterior con los datos de los fumadores, actualizando las dimensiones en 5.1 y 5.2 ahora que se consideran $D = 8$ causas y proponen $K = 2$ variables confundidoras latentes. La realización de 50 chequeos predictivos a partir de la distribución posterior generada por la cadena de máxima probabilidad posterior se presenta en la figura 5.5. Nuevamente el modelo cuadrático es el único que queda por sobre el umbral de corte de 0,1.

Se realizó la inferencia sobre el conjunto total de covariables observadas con el modelo de factores cuadráticos y seleccionó nuevamente la cadena de máxima probabilidad posterior. A partir de los valores medios de esta traza se obtuvieron los estimadores de $\boldsymbol{Z}$ a usar como confundidores sustitutos. En el modelo de resultados se usan los mismos priors que en 5.3, pero la verosimilitud para el target binario es ahora

$$y_n = Bern\left(p = \mathbb{S}\left(\boldsymbol{\beta}\boldsymbol{X}_n + \gamma\hat{\boldsymbol{Z}}_n\right)\right), \tag{5.4}$$

donde $\mathbb{S}$ es la función sigmoidea.

Las distribuciones posteriores de todos los coeficientes obtenidas a partir de las 5 cadenas de 8000 puntos de traza después de 4000 puntos de tuneo se muestran en la figura 5.6. En este caso no podemos establecer valores 'verdaderos' de referencia con los cuales contrastar los resultados al tratarse de un conjunto de datos con target real, dado por la asignación de tumor maligno o benigno según las calificaciones médicas. Los valores ilustrados en la figura son entonces los

valores medios de nuestras densidades posteriores, en negro, y los valores medios obtenidos en el tutorial de Wang y Blei, marcados mediante la línea vertical naranja, y una desviación estándar de distancia a cada lado como la región sombreada en gris. Para tener en cuenta el error reportado por el tutorial, se realizó el cálculo del p-valor sobre la diferencia $d$ entre las dos estimaciones, de distribución $N\left(d|\mu = \mu_{tesis} - \mu_{tutorial}, \sigma^2 = \sigma^2_{tesis} + \sigma^2_{tutorial}\right)$. En este caso sólo es significativa ($p_{val} = 0$) la diferencia en el radio medio, puesto que la suavidad media y confundidora 2 presentan p-valores de 0,174 y 0,113 respectivamente.

## 5.3. Conclusiones

En este capítulo se ilustró la manera en que fue empleado el método del deconfounder en dos bases de datos. En el caso del dataset de fumadores, se generaron datos semi-sintéticos para poder contrastar los estimadores generados por la inferencia con los valores reales para determinar el sesgo del método. Si bien el método no logra capturar el comportamiento real, el sesgo en esta instancia es igual o ligeramente mejor que el presentado en el trabajo previamente publicado.

En la base de datos de cáncer de mamas no se tienen valores reales con los cuales contrastar los resultados, pero sí se consiguieron distribuciones posteriores de los parámetros de manera que los valores reportados por el tutorial de Wang y Blei fueran mayoritariamente consistentes con las mismas.

Teniendo en cuenta estos resultados, pareciera que el modelo propuesto en esta tesis logra replicar los resultados propuestos por los autores que proponen el método del deconfounder. Por ello, en el próximo capítulo se expone la base de datos de interés y motivación del trabajo sobre deserción universitaria y se aplican los modelos acá discutidos con las adecuaciones dimensionales correspondientes.

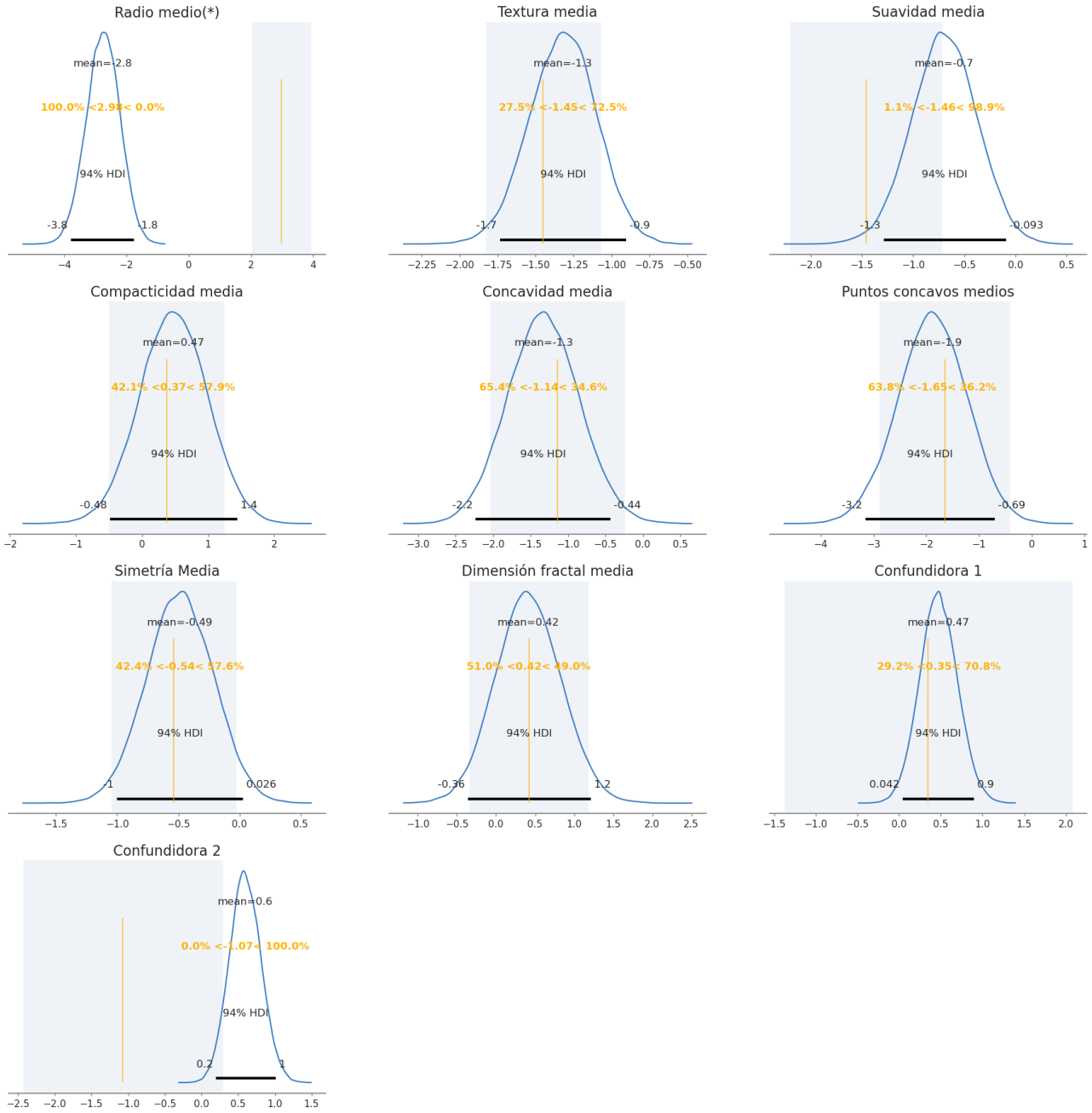



Figura 5.6: Distribuciones posteriores de los parámetros $\beta$ que acompañan a cada causa y $\gamma$ con las dos variables confundidoras bajo consideración. Las líneas verticales y el área sombreada muestran los valores medios y la dispersión de dichos coeficientes reportados en el código tutorial por los autores Wang y Blei. La diferencia $d$ de la variable con (*) fue la única de p-valor significativo, existiendo una baja probabilidad de observar los valores de Wang y Blei dadas las distribuciones reportadas en este trabajo.

# Capítulo 6

# Inferencia causal para la deserción universitaria de UNSAM

La continuidad de la educación es un problema social vigente a nivel mundial. Cada vez se requieren más especializaciones para insertarse en el mundo laboral de manera exitosa, extendiendo la brecha de oportunidades entre quienes se forman a nivel universitario y quienes cuentan con nivel secundario o menor. Teniendo universidades públicas y de calidad en el país, es de interés mejorar la retención universitaria y disminuir la fuga de recursos empleados para formar y capacitar a un gran número de personas. Las instituciones cuentan con bases de datos digitalizadas con información académica y social de sus inscriptos, lo que permite un análisis de datos con herramientas como la inferencia causal.

El trabajo de este capítulo tiene como objetivo buscar comprender las causas de la deserción estudiantil universitaria, entendiendo que si se estudian y describen las causas, se pueden diseñar políticas de retención adaptadas a las causas de mayor relevancia e impacto. La Universidad Nacional del General San Martín (UNSAM) cuenta con una base de datos de los trabajos prácticos regularizados y finales de las materias por cuatrimestre, por lo que se puede medir el desempeño de cada persona en el tiempo. Se va a limitar el análisis al primer año de estudio universitario y la deserción (o no) a esa altura de la carrera.

En este capítulo se presentará el estudio de la base de datos de 3034 estudiantes de la UNSAM pertenencientes a las unidades académicas de la Escuela de Ciencia y Tecnología (ECyT), Escuela de Humanidades (EH), y del Instituto de Ciencias de la Rehabilitación y el Movimiento (ICRM) entre los años 2017 y 2021, sobre las cuales se hace una selección de las unidades de EH.

## 6.1. Antecedentes

En la misma línea de trabajo, el licenciado Pablo Aguila realizó un estudio de predicción de deserción con estos datos aplicando métodos de aprendizaje automático supervisado como *random forest*, *redes neuronales* y *regresión lineal* [40].

En su trabajo definió una base de datos a partir de la unificación de tres fuentes de información: una base con datos académicos de las materias regularizadas y finalizadas durante el primer semestre y primer año y sus notas, la encuesta de índole socio-económico respondida por cada estudiante al momento del ingreso a la universidad, y la información geográfica del radio censal de cada persona a partir del cotejo con la el reporte del 2010 del censo del INDEC.

Con los datos académicos de la UNSAM, sólo mantuvo las unidades donde podía definirse la 'fecha de inicio de actividad académica' como el día de primera regularización de materia o primer final rendido, pero excluyendo las instancias donde las materias fueran dadas por equivalencia o no correspondiesen al primer año de alguna de las carreras de la UNSAM. Asignó seis variables o *features* académicos a cada unidad: el número de materias regularizadas, el número de finales rendidos y el promedio de notas de estos finales al fin del primer semestre y para el primer año en su totalidad. Definió la variable objetivo o *target* de manera binaria con un 0 o un 1 para las personas que continuaron con sus estudios y aquellas que desertaron, respectivamente. Se consideró como 'desertora' a toda aquella persona que no presentó actividad de ningún tipo en un tiempo mayor que 365 días de su fecha de inicio de actividad académica. Cabe destacar que con esta definición se puede estar asignando de manera errónea la etiqueta de desertora a personas que cursen materias pero no logren regularizarlas por otros motivos. La encuesta a cada ingresante permitió definir variables sobre la conformación familiar de la persona y las tareas por fuera de la actividad académica. Dentro de las mismas están el número de hijos, el número de personas a cargo, la cantidad de horas semanales de trabajo o búsqueda del mismo, el máximo nivel educativo alcanzado por el padre y la madre y la dirección de residencia. Aguila contrastó esta dirección con la API de Google Maps y obtuvo el radio censal de cada persona, asociando entonces 78 variables censales de dicho radio a cada persona. Así se tiene información del tipo de vivienda, condiciones como presencia de computadora e internet, la distancia al campus, la distancia al barrio popular más cercano entre otras. Además construyó la variable "SES", el nivel socio económico según las denominaciones de la metodología del trabajo de estratificación social [41].

El licenciado empleó los métodos mencionados con el objetivo de *predecir* la deserción estudiantil y analizar la eficacia de dicha predicción con distintas medidas temporales de antelación.

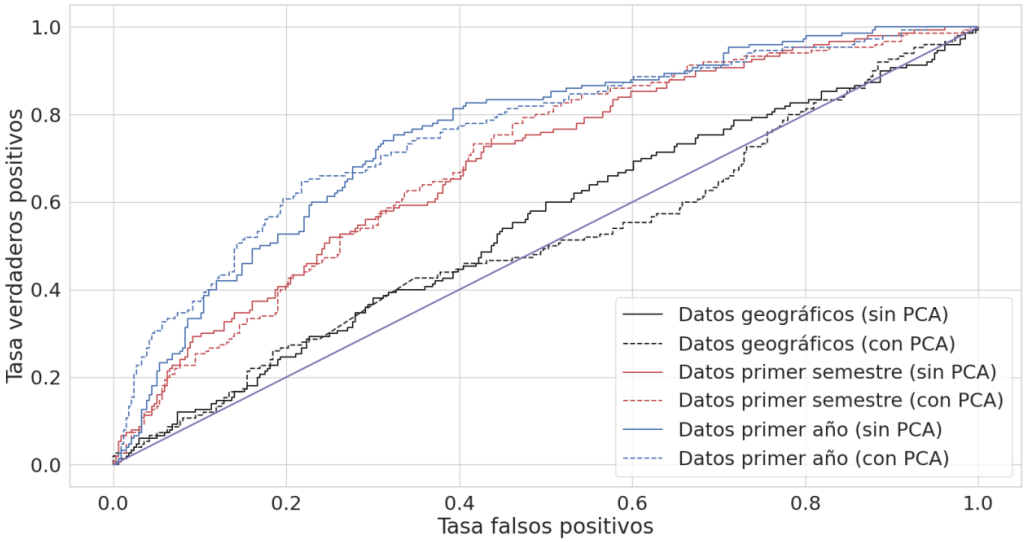



Figura 6.1: Curvas ROC de las redes neuronales para el dataset completo con y sin realizar PCA a las variables numéricas. Figura de Aguila[40].

La figura 6.1 muestra el rendimiento del modelo de clasificación por redes neuronales según el umbral de corte, comparando los verdaderos positivos (personas que se clasifican como desertoras y dejan) con los falsos positivos (quienes se clasifican como desertoras pero siguen en la universidad). Se tomaron tres sub-conjuntos de entrenemiento del set de datos completos: los datos geográficos, para predecir la deserción en cuanto ingresa a la universidad y los datos académicos del primer semestre y primer año para predecir con seis o un mes de antelación, respectivamente. Las predicciones a partir de los datos geográficos no se diferencian de manera satisfactoria respecto de asignaciones al azar, mientras que los datos del primer semestre y el año completo sí mejoran la tasa de predicciones positivas verdaderas. Se reportó un desempeño similar para los casos sin y con PCA sobre las varuiables numéricas, los datos del primer semestre obtienen una tasa de 0,80 de verdaderos positivos ante 0,50 y 0,58 falsos positivos para los datos sin y con PCA respectivamente.

Las predicciones obtenidas con la base de datos segmentada por unidad académica tienen mayor rendimiento que con la base de datos total. Se descubrió que las mejores predicciones son logradas sobre las unidades de la EH. La figura 6.2 muestra las curvas ROC y el área bajo la curva para la regresión logística, el Random Forest y la red neuronal.

Se encuentra que el método con mayor área es el método de Random Forest, con $AUC_{anual} = 0{,}81$; las chances de que la probabilidad de asignación de una persona de target 1 tomada al azar sea mayor que la probabilidad de una unidad aleatoria de target 0 es de 0,81, por lo que

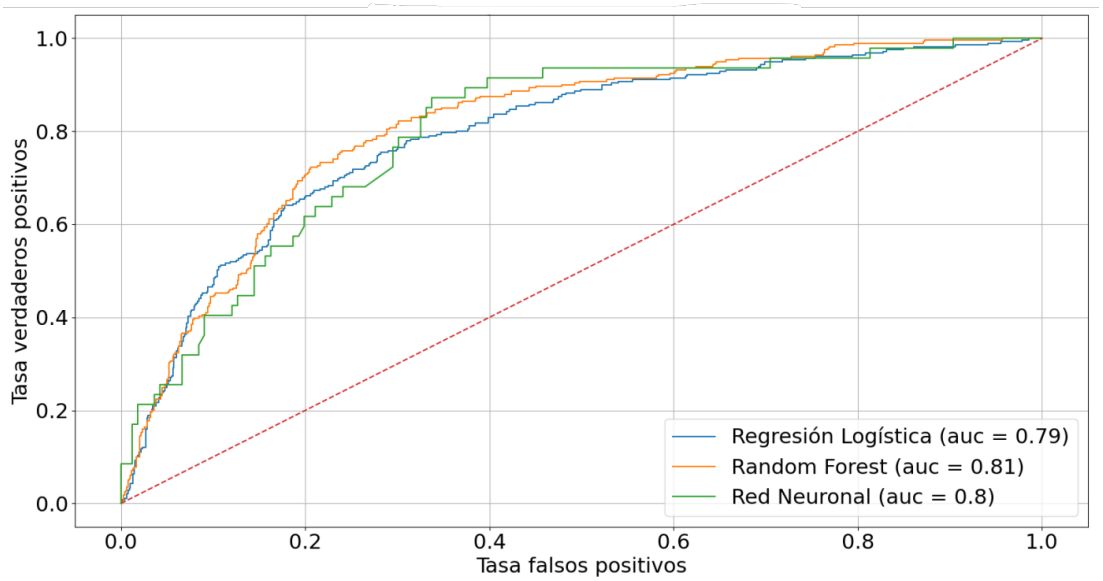



Figura 6.2: Curvas ROC con areas AUC para los tres métodos de aprendizaje aplicados al dataset de la Escuela de Humanidades, considerando las features geográficas y las académicas del primer año entero. Figura de Aguila[40].

tiene un buen poder de separación entre ambos targets. Si la etiqueta se asigna al finalizar el primer semestre o en cuanto comienza los estudios, el método de Random Forest reportó $AUC_{semestral} = 0{,}72$ y $AUC_{inicio} = 0{,}55$ respectivamente. El mismo estudio sobre la unidad académica ECyT presentó un AUC máximo de 0,66 con la información de todo el primer año. **En este trabajo de tesis se usó la información del primer año de las unidades de la EH.**

## 6.2. Preparación de la base de datos

De las 3034 unidades de la base de datos de Aguila, sólo 1330 corresponden a estudiantes de la EH. Las mismas tienen 147 covariables medidas. Estas observaciones se separaron en un conjunto de entrenamiento y otro de testeo de 976 y 354 unidades, respectivamente. Se realizó un pre-procesamiento por separado para asegurar que la información contenida en el conjunto de evaluación quede fuera del análisis de entrenamiento.

En primer lugar se descartaron los datos censales puros y tomaron las primeras cuatro componentes principales, PCA0, PCA1, PCA2 y PCA3. Explican un $\sim 65\,\%$ de la variabilidad de los datos y la primera componente principal, PCA0, está correlacionada con "SES", con un coeficiente de Pearson de $\rho = -0{,}92$, por lo que a la hora de seleccionar las covariables a

estudiar como causas 'SES' será excluida. El resto del dataset consta de variables numéricas y categóricas. Uno de los pasos del pre-procesado fue codificar las variables categóricas ordinales (aquellas cuyas categorías presentan un orden intrínseco al ser variables cardinales) como la cantidad de hijos, personas a cargo, el nivel de educación alcanzado por los padres y la cantidad de horas de trabajo o actividad según la tabla 6.1. Los datos faltantes y las respuestas "sin información" se completaron con 0, así estas unidades fueron tratadas asumiendo que no le correspondía la pregunta.

| Valor | Horas de actividad | Trabajo | Hijos | Personas a cargo | Máxima educación alcanzada por la madre/el padre |
|---|---|---|---|---|---|
| 0 | No trabajó y no buscó trabajo | Sin información | No tiene | No tiene | Sin información |
| 1 | No trabajó y buscó trabajo | Hasta 10hs | Uno | Uno | Escuela primaria incompleta |
| 2 | Trabajó al menos una hora | Mas de 10 y hasta 20hs | Dos | Dos | Escuela primaria completa |
| 3 | - | Más de 20 y hasta 35hs | Más de dos | Más de dos | Colegio secundario incompleto |
| 4 | - | 35hs o más | - | - | Colegio secundario completo |
| 5 | - | - | - | - | Estudios superiores completos |
| 6 | - | - | - | - | Estudios universitarios incompletos |
| 7 | - | - | - | - | Estudios universitarios completos |
| 8 | - | - | - | - | Estudios de postgrado |

Tabla 6.1: Codificación de las variables categóricas relevantes en la base de datos de UNSAM.

Se creó la variable 'educación máxima total' sumando los valores de educación máxima alcanzada de cada madre y padre para estratificar por la influencia total en una única escala. Con la misma idea, las variables 'horas de actividad' y 'trabajo' se combinaron en una única variable denominada 'horas de ocupación', haciendo referencia al tiempo ocupado en tareas extracurriculares. Todas las personas con un valor $\leq 2$ en esta variable no tienen trabajo. Finalmente, si bien las variables 'hijos' y 'personas a cargo' hacen referencia al número de dependientes que cada persona evaluada tiene bajo su cuidado, se dejaron dos variables distintas para permitir la posibilidad de políticas dirigidas a cada necesidad (un jardín de infantes puede servir para unidades con hijos, pero no necesariamente con personas a cargo de la tercera edad).

De los datos académicos de la UNSAM, se manipularon las variables numéricas de materias regularizadas y finales rendidos debido a su alta correlación. Para cumplir con la hipótesis de solapamiento pero preservar la información de ambas variables, se construyó una tasa anual como $tasa = \frac{finales^2}{regularizadas}$. El promedio anual se mantuvo separado.

Finalmente, las covariables seleccionadas como causas pueden separarse en tres tipos:

1. Causas de responsabilidades externas: la cantidad de hijos (hijos), la cantidad de familiares a cargo (familiares a cargo) y las horas ocupadas con trabajo o actividad (horas ocupadas),

2. causas académicas: la educación total de los padres (educación total), la tasa anual y el promedio anual,

3. causas geográficas y censales: la distancia a barrios populares (distancia BP), la distancia al campus Miguelete de UNSAM, donde está ubicada la EH (distancia campus), la densidad de población por kilómetro cuadrado (densidad de población) y las primeras cuatro componentes principales de los datos censales (PCA0, PCA1, PCA2 y PCA3).

La matriz de correlación de la figura 6.3 muestra el valor absoluto del coeficiente de correlación entre cada causa. La máxima correlación se encuentra entre el número de hijos y el número de personas a cargo, con un coeficiente de $\rho = 0{,}60$. La causa PCA0, que informa de manera indirecta e inversa el nivel socio-económico, está correlacionada con la distancia al campus ($\rho = 0{,}47$) y la distancia a los barrios populares ($\rho = -0{,}60$). El resto de las variables tiene coeficientes de correlación de Pearson $\rho < 0{,}35$.

Un primer modelo sencillo para explicar la deserción a partir de las causas seleccionadas sería una regresión lineal simple. La tabla 6.2 presenta los coeficientes de correlación entre dichas causas y el target (discreto), a modo de 'peso' de relevancia de cada una de las causas en la deserción. Los máximos coeficientes de correlación son los de la tasa y promedio anual y educación total de los padres, de manera negativa. Esto tiene sentido puesto que las primeras dos causas miden de forma numérica el compromiso de cada estudiante con sus estudios: a mejor rendimiento académico, menos tendencia de dejar la carrera. La educación de los padres parece tener entonces influencia positiva en cada estudiante: cuanto mayor nivel de educación en cada familia, menor la probabilidad de deserción. Las horas ocupadas presentan el siguiente máximo de correlación pero con signo opuesto, indicando que la ocupación fuera de la universidad se asocia con una mayor deserción. Todos estos coeficientes tienen valores $\rho < 0{,}3$, y la mayoría menores que 0,08. Parece ser que ninguna de estas causas puede explicar por sí misma la situación de deserción. Hay que recordar que el target en este caso es discreto y que los coeficientes de correlación pierden la medida de estructuras no lineales que pueden estar presentes en estos datos.

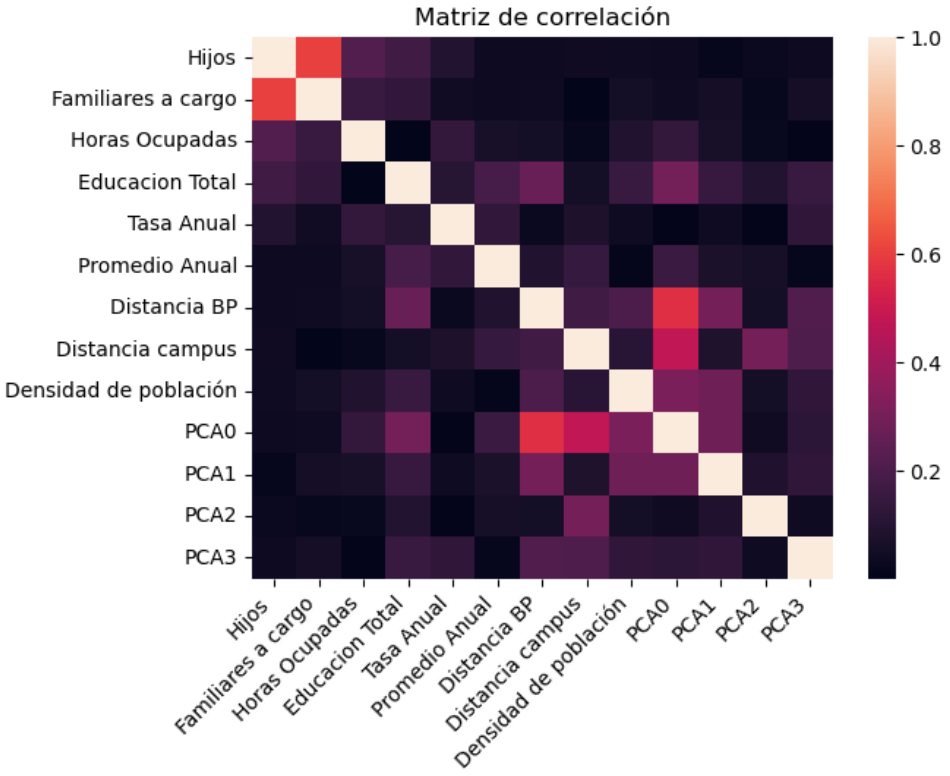



Figura 6.3: Matriz de correlación entre las covariables seleccionadas como causas en el análisis. Se presentan los valores absolutos de los coeficientes de correlación en escala de colores. La mayoría de las causas se encuentran descorrelacionadas, con $\rho < 0{,}35$.

| # | Causa | Coef. de correlación |
|---|---|---|
| 0 | Hijos | -0.03 |
| 1 | Familiares a cargo | -0.01 |
| 2 | Horas Ocupadas | 0.08 |
| 3 | Educacion Total | -0.09 |
| 4 | Tasa Anual | -0.29 |
| 5 | Promedio Anual | -0.11 |
| 6 | Distancia BP | -0.00 |
| 7 | Distancia campus | -0.04 |
| 8 | Densidad de población | 0.03 |
| 9 | PCA0 | -0.03 |
| 10 | PCA1 | -0.03 |
| 11 | PCA2 | 0.03 |
| 12 | PCA3 | -0.02 |

Tabla 6.2: Coeficiente de correlación de Pearson entre cada causa con el target discreto. Todos los valores reportados con valor absoluto menor que 0,3

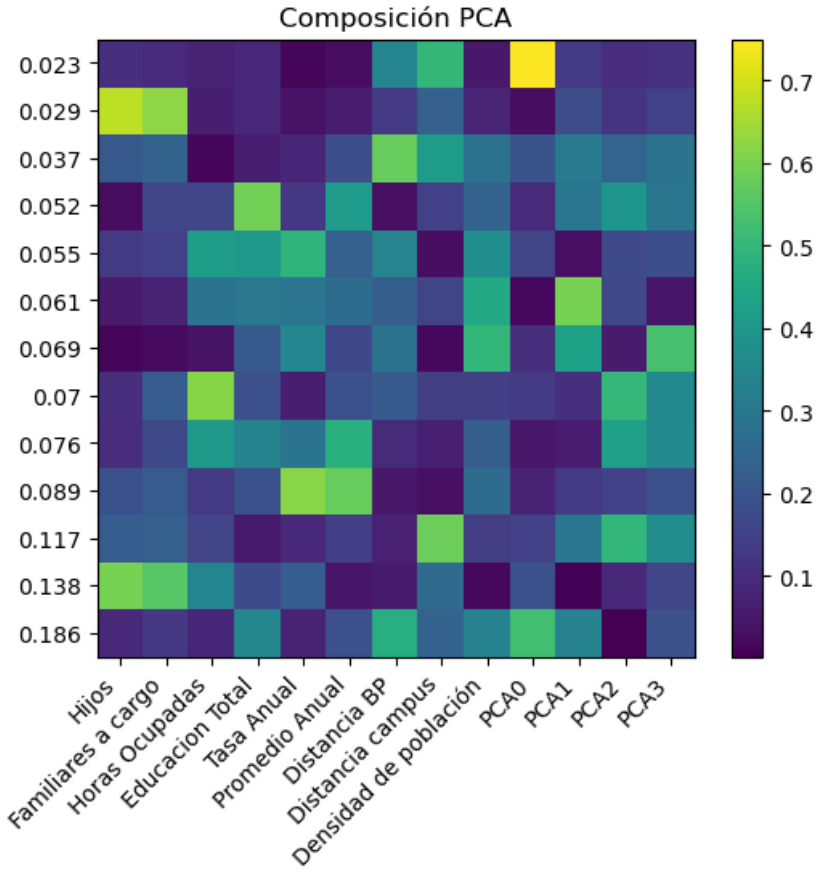



Figura 6.4: Descomposición de las causas seleccionadas en sus componentes principales. El mapa de colores ilustra el peso, en valor absoluto, de cada causa para la componente del PCA. La variabilidad explicada por cada componente está reportada en el eje vertical, y a la derecha la escala de valores.

## 6.3. Aplicación del método del deconfounder

Como se detalló en el capítulo 3, el método del deconfounder requiere de la definición de un modelo de factores y otro modelo de resultados. Para el modelo de factores se tienen $D = 13$ causas a explicar con un número $K$ de variables confundidoras latentes que planteamos definir a partir de un análisis de componentes principales sobre las causas. El mapa de calor de la figura 6.4 muestra una fila por componente principal. En el eje vertical está el valor de la fracción de la variabilidad total explicada por esa componente, con la primera componente en la fila inferior. El color en cada coordenada indica el peso de la causa en el eje horizontal para esa componente principal, con los valores máximos tendiendo al amarillo.

De abajo para arriba, podría decirse que la primera componente es del tipo geográfico, con

máximo peso de las causas de distancia a barrios populares y PCA0. La segunda componente engloba más bien las responsabilidades externas, con influencia del número de hijos, familiares y horas ocupadas. La tercera puede ser el costo de trasladarse a la facultad al resaltar el peso de la distancia al campus, pero también mezcla un poco las primeras dos componentes. Es llamativo cómo las causas académicas de tasa y promedio anual no aparecen hasta la cuarta componente principal. A partir de este análisis se define el número de confundidoras como $K = 4$. Como el PCA es lineal, cualquier relación entre causas que se aleje de una línea recta se perderá en este análisis. Las variables confundidoras no tienen por qué ser cuatro, y tampoco deben ser estas cuatro descritas, pero se entiende que es una aproximación usada para informar el número de variables confundidoras que de otra forma sería una selección arbitraria.

Para el modelo de factores se plantearon los modelos lineal y cuadrático con la estructura entre variables dada por las ecuaciones 3.8, con priors definidos como

$$\begin{aligned} z_{n,k} &\sim N(0,1), \\ W_{k,d}, W_{2k,d} &\sim N(0,\sigma_W),\ y \\ \sigma_W &\sim HN(\sigma = 1), \end{aligned}$$

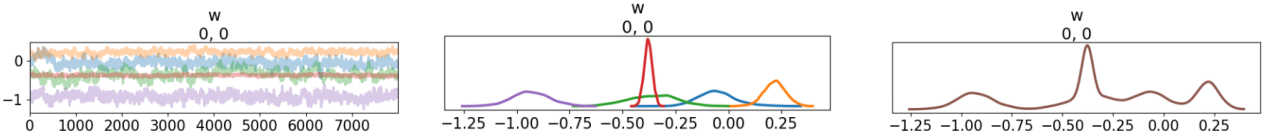
donde $\sigma_W$ se usa para tener un hiperparámetro de la desviación de coeficientes de las $\boldsymbol{W}$'s, de los cuales no tenemos información. Las siglas $HN$ indican la distribución de media normal. La verosimilitud de las causas observadas viene dada por

$$\boldsymbol{x} \sim N\left(x|\mu_x, \sigma_x = 0{,}1\right),$$

con valores medios $\mu_x$ definidos según 3.9:

$$\begin{aligned} \mu_x^{lin} &= \mathbf{zW}, \\ \mu_x^{cuad} &= \mathbf{zW} + \mathbf{z}^2\mathbf{W}_2. \end{aligned}$$

La inferencia de este capítulo se realizó con 5 cadenas de 4000 pasos de ajuste y 8000 muestras. Al realizar esta inferencia se encontró una dificultad agregada; la gran dimensionalidad de los datos y la simetría de rotación que hay al considerar variables latentes en los modelos lineales generan que los muestreos del espacio de parámetros puedan ser conflictivos, como se comenzaba a esbozar en el capítulo 5. En este caso, se realizaron modificaciones en los argumentos del muestreo con Numpyro y JAX de manera que las cadenas tengan mayor profundidad




(a) Trazas para $W_{00}$. (b) Distribución por cadena. (c) Distribución total.

Figura 6.5: Muestreo de la coordenada [0,0] de la matriz $\boldsymbol{W}$. De izquierda a derecha: las trazas de las cadenas, la distribución de cada cadena, y la distribución total generada a partir de todas las cadenas.

en las capas de árbol (con *max_tree_depth*), además de tener una matriz de densidad completa (con *dense_mass*), ya que por defecto es una matriz de masas diagonal [42]. Aún así, las cadenas tienden a tener poca mezcla y gran autocorrelación, por lo que los diagnósticos como *r_hat* y *ess* empeoran respecto de los ejemplos sencillos y los modelos de juguete previamente ilustrados en este trabajo. Esta observación es relevante por dos motivos centrales. En primer lugar, un mal muestreo del modelo de factores implica que las trazas obtenidas no reconstruyen las distribuciones propias del problema. En segundo lugar, al tener un muestreo acotado, puede ser que las cadenas no converjan, o lo hagan en mínimos locales, y las representaciones posteriores obtenidas para las variables confundidoras no se asemejen a una distribución normal. En los peores casos de muestreo las distribuciones posteriores de algunas variables confundidoras eran multi-modales, con cada cadena centrada en un valor distinto, como se puede ver en la figura 6.5. Las trazas de cada cadena están representadas en 6.5a y la distribución que generan en 6.5b. La distribución estimada a partir del total de las cadenas está en 6.5c. Allí, tomar un único valor medio como confundidora sustituta genera que se seleccione incorrectamente un valor de muy baja probabilidad para continuar con el análisis del modelo de resultados.

Para disminuir las probabilidades de caer en estos casos patológicos se analizan los valores logarítmicos de probabilidad de cada una de las 5 cadenas, representados en la figura 6.6. Se obtiene un ranking de las mismas y mantiene la traza de máxima probabilidad. Los resultados expuestos a continuación están basados en una única cadena de $logp = 233940 \pm 70$. El resto de las cadenas tienen valores de logp menores a distancia $\delta \in [4080, 18030]$. Se destaca que para afianzar estos resultados requeriríamos más tiempo de corrida o cambiar la estrategia de muestreo.

### 6.3.1. Resultados

Como se presentó en el capítulo 3, primero se plantearon los modelos de factores lineal y cuadrático sobre una base de datos enmascarada, aplicando una matríz que retiene el 20 % de las observaciones de manera aleatoria. Se obtuvieron las distribuciones de las matrices $\boldsymbol{W}$ y $\boldsymbol{W_2}$ y de las 4 variables confundidoras para cada unidad observada. Con las mismas, se realizó

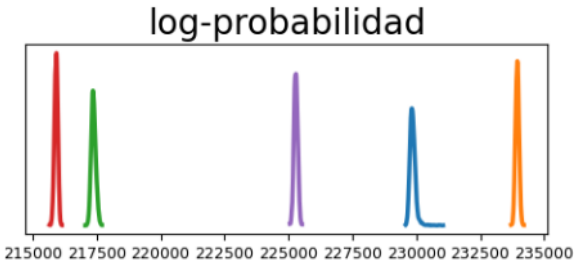



Figura 6.6: Log-probabilidad de cada cadena según las distribuciones posteriores de todas las variables del modelo.

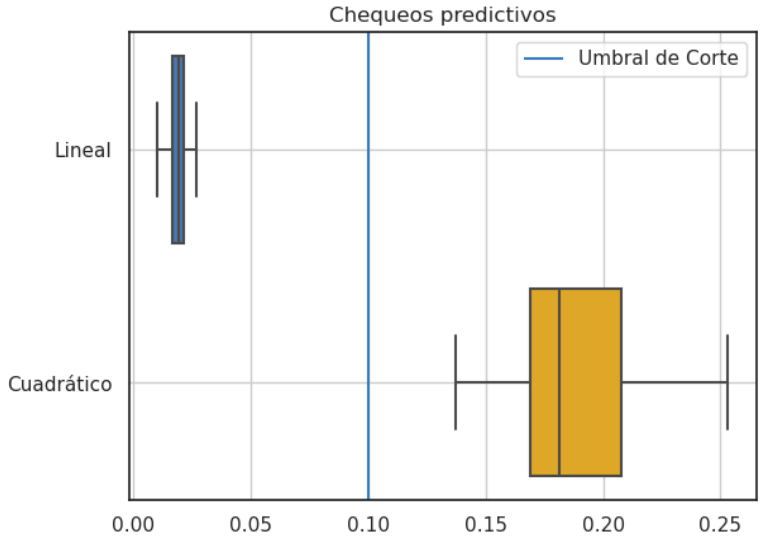



Figura 6.7: Box-plot de los p-valores para 50 chequeos predictivos del modelo lineal en azul ($p - score_{lin} = 0{,}019 \pm 0{,}004$) y cuadrático en amarillo ($p - score_{cuad} = 0{,}19 \pm 0{,}03$). El umbral de corte está representado por la línea celeste.

el chequeo predictivo para evaluar la concordancia entre los datos generados por la inferencia y los que habían sido retenidos del proceso de inferencia. La figura 6.7 muestra los diagramas de caja o box-plots resultantes de haber realizado el chequeo predictivo 50 veces, generando replicas de los datos enmascarados a través de muestras de las matrices $\boldsymbol{W}$ y $\boldsymbol{W_2}$ y $\boldsymbol{Z}$ cada vez. Si bien el modelo cuadrático tiene mayor amplitud de valores posibles de p-valor, supera el umbral de corte de $p > 0{,}1$ en todas las instancias y su valor es diez veces el del modelo lineal. Por este motivo, el resto del análisis se continua con el modelo cuadrático.

Se realizó la inferencia sobre el conjunto completo de observaciones con el modelo de factores cuadrático, seleccionando la cadena de mayor probabilidad logarítmica. Los estimadores para las confundidoras se construyen con el valor medio de la distribución para cada $z_{n,k}$. La figura

6.8 presenta el resultado de ambas matrices, donde los coeficientes de cada coordenada son los valores medios. El sombreado corresponde con el valor absoluto en sus unidades de error, calculado como $\frac{|\boldsymbol{W_{ij}}|}{\sigma_{ij}}$.

Los únicos valores menores que dos sigmas son los de color azul intenso (4 coordenadas en la matriz $\boldsymbol{W}$ y 7 de $\boldsymbol{W_2}$). Los valores más significativos son los de tono rojizo. Si bien no podemos afirmar cuáles son las variables confundidoras, notamos ciertas estructuras predominantes que se mantienen en ambas matrices. Las figuras sugieren que $z_0$ se asocia principalmente a las causas de tipo geográfico. Hay una relación en las causas de horas ocupadas y educación total de los padres, pero los coeficientes más relevantes son los de la distancia a barrios populares, la densidad poblacional, y las primeras tres componentes principales de la información censal. De las causas geográficas, resta la distancia al campus, que se relaciona principalmente con la última variable confundidora, $z_3$, junto con PCA0 y PCA2. Por su parte, la confundidora $z_1$ tiene mayor relación en las causas de responsabilidades externas: el número de hijos, los familiares a cargo y las horas ocupadas. Finalmente, las causas académicas de educación familiar, la tasa anual de materias y el promedio anual guardan relación principalmente con la variable $z_2$, pero la misma se vincula con peso similar a las causas geográficas. Es la variable confundidora que se asocia más fuertemente a la causa de tasa anual, y además la que menor contraste tiene en su composición. El resto de los coeficientes tienen valores significativos en su error pero son cercanos a cero. La relación asociativa entre las causas y confundidoras es de muy bajo peso.

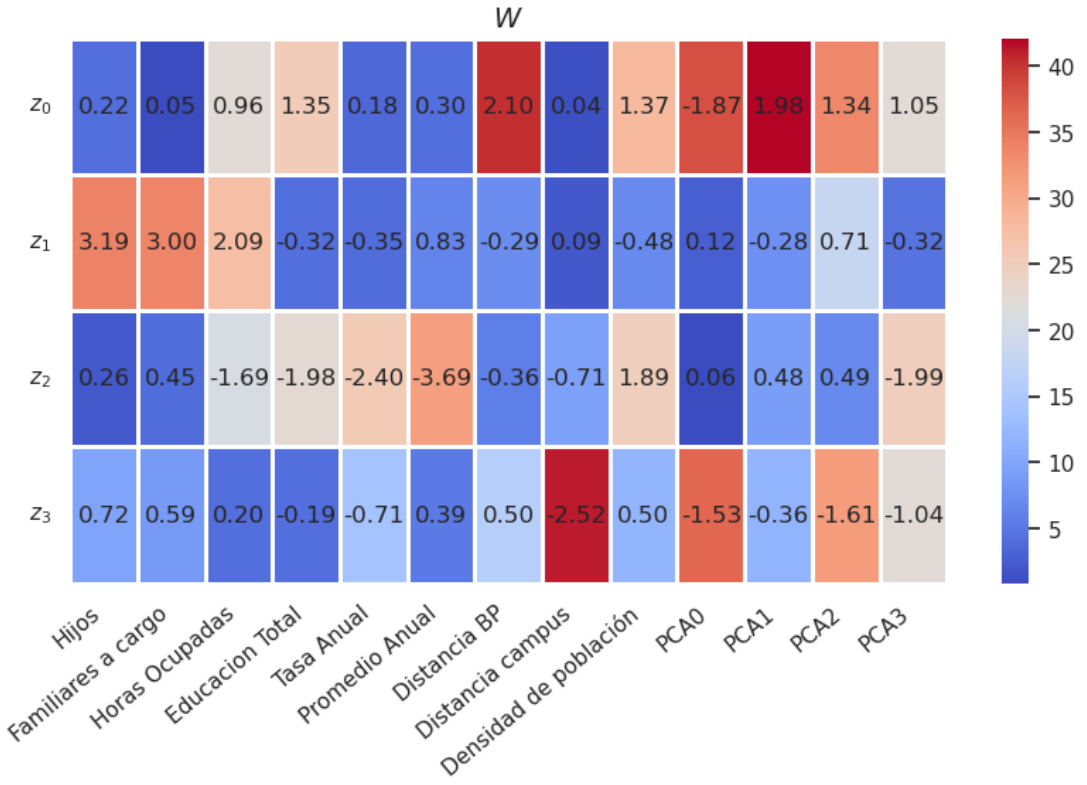
Continuamos con el método del deconfounder usando las estimaciones de confundidoras a partir de valores medios como confundidoras sustitutas en el modelo de resultados. En este modelo empleamos distribuciones prior del tipo

$$\beta_d \sim N(0,1),$$
$$\gamma_k \sim N(0,1),$$

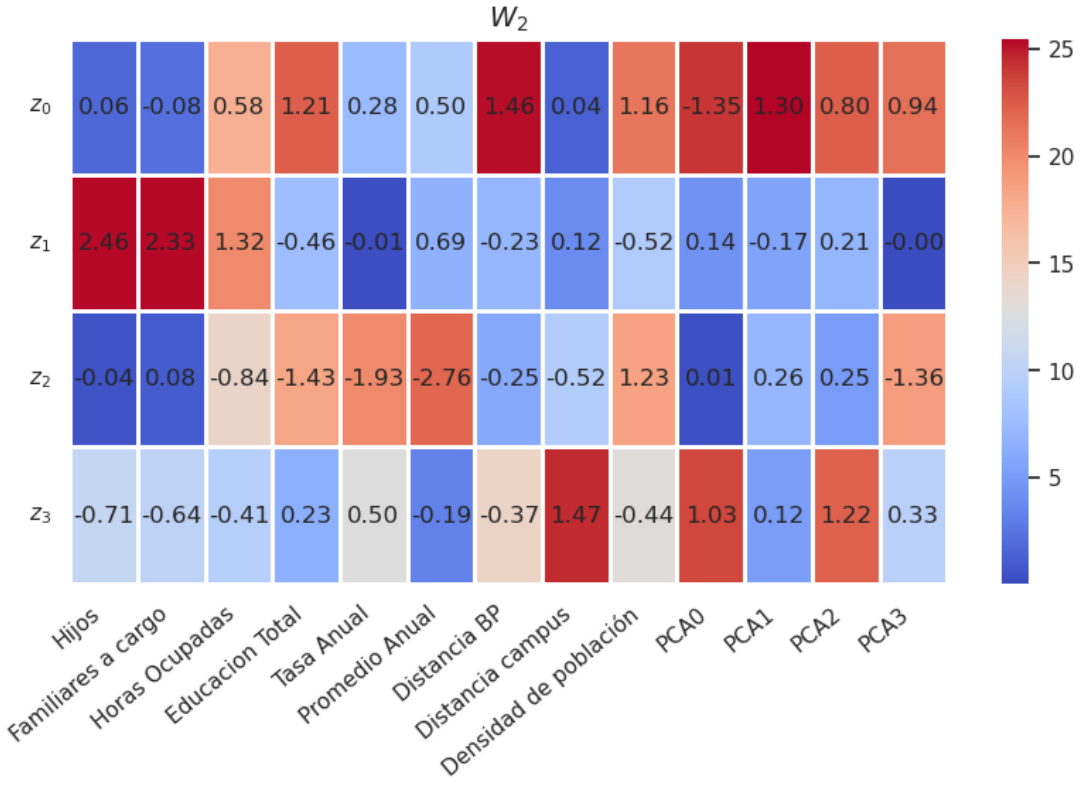
y verosimilitud de la forma

$$\boldsymbol{y} \sim Bern\left(\boldsymbol{y}|p = \mathbb{S}\left(\boldsymbol{X\beta} + \hat{\boldsymbol{Z}}\boldsymbol{\gamma}\right)\right),$$

donde $\mathbb{S}$ representa la función sigmoidea, y $\hat{\boldsymbol{Z}}$ son los estimadores de las variables confundidoras. De la inferencia se obtienen distribuciones de los vectores $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ que relacionan las causas y las variables confundidoras con la deserción estudiantil. A modo de comparación, se realizó también inferencia con un modelo *naive*, con el mismo prior para $\boldsymbol{\beta}$ pero sin confundidoras sustitutas ni




(a) Matriz $\mathbf{W}$.




(b) Matriz $\boldsymbol{W_2}$.

Figura 6.8: Reconstrucción de matrices $\boldsymbol{W}$ y $\boldsymbol{W_2}$ a partir de los valores medios de la distribución de *1 de las trazas de MCMC*. El sombreado indica su valor en unidades absolutas de error, con la escala de colores a la derecha de cada matriz.

$\boldsymbol{\gamma}$. La relación entre el target y sus causas es

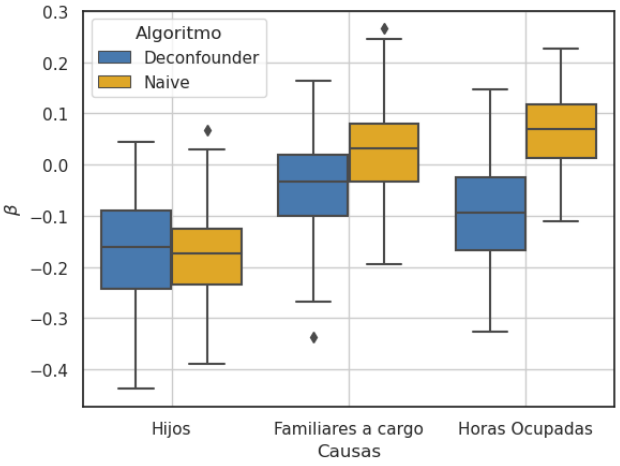
$$\boldsymbol{y} \sim Bern\left(\boldsymbol{y}|p = \mathbb{S}\left(\boldsymbol{X}\boldsymbol{\beta}_{naive}\right)\right),$$

forzando a ignorar la presencia de variables confundidoras. En ambos casos se utilizan todos los datos $y_{obs}$. En la figura 6.9 se presentan las estimaciones de los vectores resultantes, tomados a partir de los valores medios.

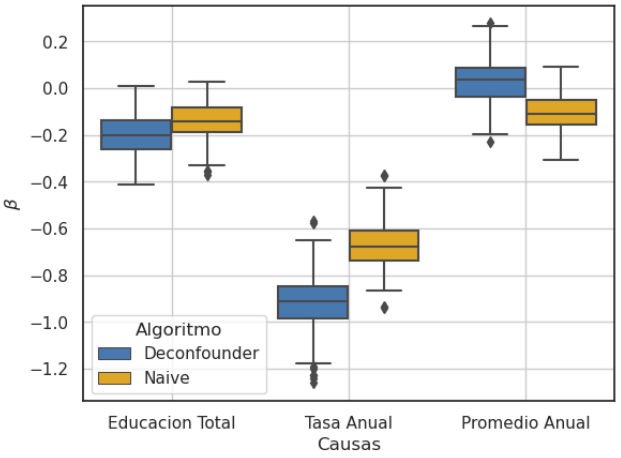
Las figuras 6.9a a 6.9d presentan los gráficos de caja de las muestras para cada parámetro identificando la causa con la que se relaciona, coloreadas según si provienen del modelo deconfounder o del modelo naive y agrupados según los tipos de causas. Se evaluó la posibilidad de que las muestras obtenidas en ambos modelos provinieran de la misma distribución mediante test-T y test-Z para evaluar las medias, test de Mood para evaluar la mediana y un test de Kolmogórov-Smirnov de dos colas sobre las distribuciones posteriores de $\boldsymbol{\beta}$. Aunque las cadenas generadas por el muestreo HMC tenían naturalmente baja autocorrelación, presentando un *ess* del orden del número de muestras, se realizó una selección de 1 punto por cada 100. Los p-valores de los tests estadísticos fueron considerados altamente significativos ($p < 0{,}001$) para todos los coeficientes evaluados, exceptuando el coeficiente asociado a la causa del número de hijos, donde el resultado fue no significativo ($p_{medias} \sim 0{,}3$, $p_{mediana} \sim 0{,}2$ y $p_{KS} \sim 0{,}07$). En 6.9a se nota que estas distribuciones se asemejan más entre sí que el resto de las causas del mismo grupo. La tabla 6.3 reporta los valores del estadístico del test Kolmogórov-Smirnov. De todas las causas, los coeficientes con mayor disparidad entre los modelos fueron las horas ocupadas, la tasa y el promedio anual, la distancia al campus, y el PCA1, con valor del estadístico $\tau > 0{,}5$.

Se nota que el modelo del deconfounder presenta una corrección al caso naive. La tasa anual es la causa que mayor incidencia tiene en la probabilidad de deserción de ambos modelos, el deconfounder aumenta su relevancia. Se modifica el signo de los coeficientes que acompañan a las variables de familiares a cargo, horas ocupadas y el promedio anual, invirtiendo su relación con la deserción. En general, el deconfounder amplifica los pesos relativos. Al pasar parte de la relación a las variables confundidoras, se logra un mayor contraste y se empieza a dilucidar la estructura de las causas. Disminuye el peso de las últimas dos componentes censales.

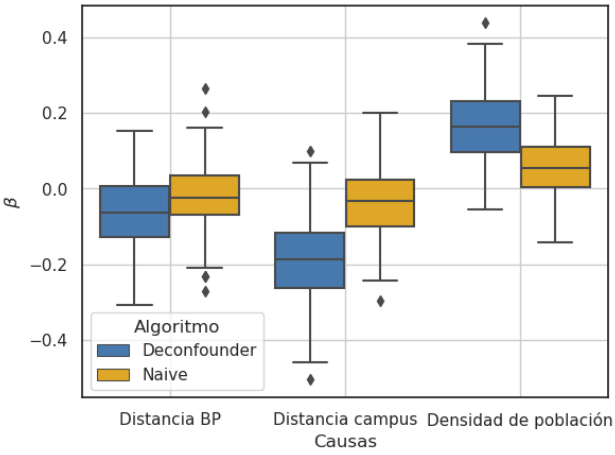
El gráfico 6.9e presenta los valores medios de los coeficientes de ambos vectores para todas las causas y variables confundidoras, usando las desviaciones estándar como barras de error. Es llamativa la manera en que las variables confundidoras tienen más peso relativo en el resultado que las causas observadas. Dada la amplitud de los valores del vector $\boldsymbol{\gamma}$, casi el total del efecto de las causas quedan codificadas dentro de las variables confundidoras. Sus primeros tres



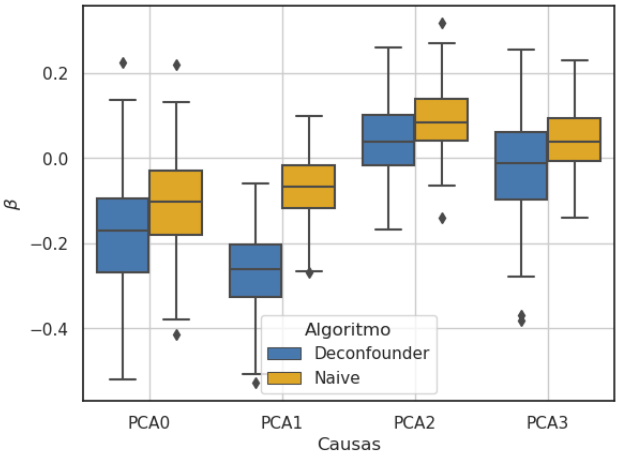

(a) Causas externas.



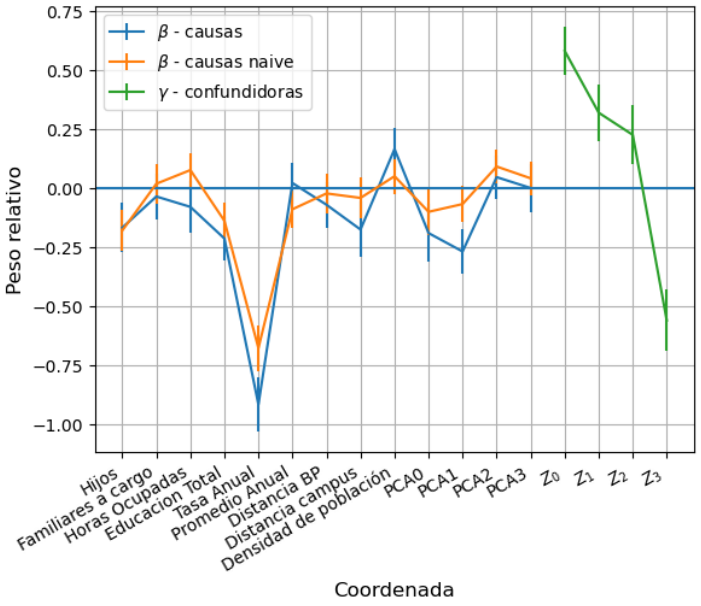

(b) Causas académicas.




(c) Causas censales geográficas.




(d) Causas censales PCA.




(e) Coeficientes de los vectores $\boldsymbol{\beta}$ (azul) y $\boldsymbol{\gamma}$ (verde) del modelo de resultados comparados con los valores de realizar inferencia de manera naive (amarillo).

Figura 6.9: Valores medios y dispersión de los coeficientes $\boldsymbol{\beta}$ (azul), $\boldsymbol{\beta}_{naive}$ (amarillo) y $\boldsymbol{\gamma}$ (verde).

| | Causa | Estadístico KS ($\tau$) |
|---|---|---|
| 0 | Hijos | 0.130 |
| 1 | Familiares a cargo | 0.325 |
| 2 | Horas Ocupadas | 0.665 |
| 3 | Educacion Total | 0.340 |
| 4 | Tasa Anual | 0.785 |
| 5 | Promedio Anual | 0.585 |
| 6 | Distancia BP | 0.235 |
| 7 | Distancia campus | 0.570 |
| 8 | Densidad de población | 0.465 |
| 9 | PCA0 | 0.255 |
| 10 | PCA1 | 0.740 |
| 11 | PCA2 | 0.300 |
| 12 | PCA3 | 0.285 |

Tabla 6.3: Tabla con los valores del estadístico del test Kolmogórov-Smirnov representando la máxima diferencia entre las funciones de distribución cumulativa para las muestras posteriores de $\boldsymbol{\beta}$ provenientes del modelo de resultados naive y con el deconfounder.

coeficientes presentan una relación positiva con el target, mientras que la última coordenada tiene signo negativo. En el análisis de las matrices $W$'s se notó que esta variable tenía influencia sobre las causas geográficas. En particular, es la variable confundidora con mayor impacto sobre la causa de distancia al campus. Cuanto mayor $z_3$, menor probabilidad de deserción, y mayor efecto lineal negativo sobre la distancia al campus.

Lamentablemente, no todos los resultados obtenidos son significativos al tener en cuenta los errores. Para disminuir el efecto del problema de muestreo, una de las posibles vías de ataque sería plantear un modelo de factores con las matrices $\boldsymbol{W}$ y $\boldsymbol{W_2}$ no-centradas. Estudios preliminares parecen indicar que esto puede mejorar la mezcla entre cadenas de la inferencia del modelo de factores, y la autocorrelación de las mismas. Como se mencionó más arriba, precisaríamos de más puntos de inferencia para aumentar la confianza en el muestreo.

Desde un principio se planteó que el propósito de este trabajo no es predecir sino que busca explicar las relaciones causales intrínsecas presentes entre las variables aleatorias de estudio. En los casos de análisis sobre datasets semi-sintéticos o presentes en estudios previos se tiene un valor objetivo con el cual contrastar la exactitud y rendimiento de un nuevo algoritmo, pero en el estudio de deserción estudiantil no contamos con estos valores. Por este motivo, se realizó una evaluación del rendimiento en base a la capacidad predictiva del método deconfounder sobre el conjunto de validación de la base de datos de la UNSAM que se había separado con este

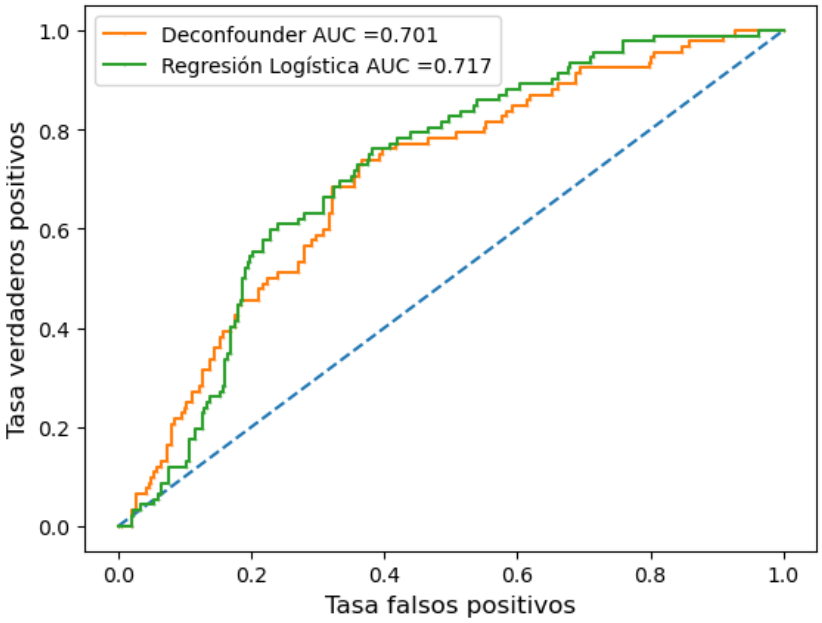



Figura 6.10: Curvas ROC.

Figura 6.11: Evaluación de las predicciones del método deconfounder (amarillo) comparado con una regresión logística simple (verde) sobre los datos de validación.

propósito.

Se aplicó una regresión logística del target en función de las causas sobre la base de datos de entrenamiento y se conservaron los coeficientes de cada variable. Con estos coeficientes y las observaciones del conjunto de validación se reconstruyeron predicciones del target de deserción.

Para replicar de manera estadística la probabilidad asignada por el método del deconfounder, se plantean matrices $\boldsymbol{W}$ y $\boldsymbol{W_2}$ y los vectores $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ a partir de un muestreo aleatorio de distribución normal con parámetros iguales al de los valores medios y varianza establecidos por el método del deconfounder como se describió anteriormente. Dadas las matrices, se realiza inferencia sobre los valores de las variables confundidoras condicionadas sobre las observaciones de causas del conjunto de validación. Se estiman los valores de dichas variables con sus valores medios. Así, se tienen las matrices y los vectores del método del deconfounder, las causas observadas del conjunto de validación, y los estimadores de las confundidoras. Se establece la probabilidad de deserción aplicando una función sigmoidea para asginar un target 0 ó 1 a cada unidad del conjunto.

Se contrastaron las tasas de falsos y verdaderos positivos para generar las curvas ROC de cada método y calcular el área debajo de las mismas, como se observa en la figura 6.10. El

método de deconfounder tiene $AUC_{ent} = 0{,}737$ sobre la base de datos en la que fue entrenada y dicho valor se modificó a $AUC_{val} = 0{,}701$ al evaluarlo sobre el conjunto de validación. El método simple de regresión logística, sesgado para reproducir predicciones, presenta un $AUC_{RL-val} = 0{,}717$.

Recordando que los antecedentes de mejor rendimiento sobre los mismos datos fueron los del método de Random Forest con un área de $AUC = 0{,}81$, el método del deconfounder ofrece una alternativa más simple de modelo con menor número de parámetros. Tenemos mayor explicabilidad sin disminuir drásticamente la predictibilidad.

# Capítulo 7

# Conclusiones generales y posibles acciones a futuro

En los capítulos 2 y 3 se profundizaron los conceptos y conocimientos sobre la inferencia bayesiana y causal. Se trabajó sobre los modelos gráficos de redes bayesianas y planteó una manera de sortear el problema fundamental de Rubin en la inferencia causal. Para aplicar los métodos computacionales planteados en el capítulo 4 se generaron datos y modelos probabilísticos de juguete. De esta manera, se afianzaron las metodologías, argumentos y parámetros relevantes a la hora de tomar las muestras de MCMC que se usaron en los procesos de inferencia. Se adaptó el código abierto de Wang y Blei, escrito en Python empleando la librería Edwards con inferencia variacional, a un código empleando la librería de PyMC, que continua ofreciendo actualizaciones y mejoras, además de ofrecer un gran soporte en su *discourse*. Al momento de escribir esta tesis, anunciaron un paquete accesorio, CausalPy, que parece ofrecer herramientas alternativas para el análisis aquí planteado.

Las primeras aplicaciones del método del deconfounder sobre bases de datos externas fueron sobre casos ya estudiados, utilizando observaciones de gastos médicos en la población fumadora de Estados Unidos y la clasificación de tumores como malignos o benignos de un estudio médico en Winsconsin. De esta manera se tuvo una instancia de familiarización con el método donde se pudieron comparar los resultados obtenidos con valores de referencia, ya sean semi-sintéticos para los gastos médicos o resultados de otro análisis causal para los tumores. En el primer caso, el sesgo obtenido es de orden similar al publicado en el artículo que introduce el método. En el último caso, se tuvo una mayoría de concordancia en los parámetros inferidos, donde sólo uno de los parámetros de las causas reportados no proviene, estadísticamente, de las distribuciones que inferimos.

Para el caso de mayor interés en este trabajo se aplicó el método del deconfounder sobre una base de datos generada durante el trabajo del Licenciado Pablo Aguila. El mismo juntó dos fuentes de datos de la universidad (académica y socio-económica) y la cotejó con el censo para armar una base de datos de estudiantes de la UNSAM entre los años 2017 y 2021. Se realizó una exploración de los datos para familiarizarse con la información disponible, además de manipular variables con el fin de preservar información pero cumplir con las hipótesis de superposición e independencia del método. En esta instancia se trabajó también con análisis de componentes principales para informar el número de variables confundidoras a tener en cuenta. En base a la selección de 13 causas y considerando 4 variables confundidoras se realizó un estudio causal planteando modelos de factores lineal y cuadrático, realizando un chequeo predictivo para confirmar el uso del modelo de factores cuadrático, del cual se obtienen los valores sustitutos de las confundidoras para usar en el modelo de resultados. Del mismo se obtuvieron los valores de impacto causal de cada variable bajo consideración, pudiendo contrastarlas con los valores estimados en caso de tener un modelo sin variables confundidoras. Se obtuvo una corrección a los mismos, además de notar que quedan abarcados por la amplitud de los parámetros de las variables confundidoras, de mayor impacto en la deserción estudiantil. Puede ser que no se estén observando las verdaderas causas que afectan a cada estudiante a la hora de dejar los estudios.

En un futuro, se planea continuar con el estudio de esta situación con algunas de estas consideraciones:

- cambiar de estrategia o método de muestreo, especialmente usar una reparametrización no centrada, que, en base a intentos preliminares, parecen resolver parte del problema de cadenas de baja mezcla y alta autocorrelación en el análisis de modelo de factores,
- analizar la relación entre los valores del chequeo predictivo y el número de variables confundidoras consideradas, especialmente en el modelo de los datos de la UNSAM,
- empelar el paquete CausalPy, que se lanzó al finalizar este trabajo por lo que no se incursionó en el mismo,
- realizar un estudio del efecto causal sumando causas de a una.

# Bibliografía


[1] Bob Rehder. Independence and dependence in human causal reasoning. *Cognitive Psychology*, 72:54–107, 2014.

[2] Yixin Wang and David M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[3] George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[4] Mark A Pitt and In Jae Myung. When a good fit can be bad. *Trends in cognitive sciences*, 6(10):421–425, 2002.

[5] Emilia Musso, Facundo Brizuela del Moral, Antonella Paola Di Naranjo, German Leandro Pereno, and Sabrina Sánchez. Deserción universitaria y rendimiento académico en estudiantes trabajadores y/o con hijos a cargo. 2021.

[6] Vanina L Celada. Acerca de las causas de deserción universitaria en argentina a principios del siglo xxi, de las políticas implementadas y nuevas propuestas de retención de población estudiantil. 2020.

[7] Andrés Santos Sharpe. Un análisis histórico del abordaje sobre el abandono universitario en argentina. *Historia de la educación-anuario*, 17(2):0–0, 2016.

[8] Ministerio de Educación. *Síntesis de Información Universitaria 2020-2021*. https://www.argentina.gob.ar/educacion/universidades/informacion/publicaciones/sintesis [Revisado: diciembre 2022].

[9] Phil Gregory. *Bayesian logical data analysis for the physical sciences: a comparative approach with mathematica® support*. Cambridge University Press, 2005.

[10] Christopher M. Bishop and Nasser M. Nasrabadi. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York, 2016.

[11] Sharon Bertsch McGrayne. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of C*. Yale University Press, 2011.

[12] Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, and Christopher Yau. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26, 2021.

[13] George E. P. Box. Sampling and bayes inference in statistical modelling (with discussion)'. *Journal of the Royal Statistical Society A*, 143:383–430, 1980.

[14] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

[15] Xi Jiang, Guanghua Xiao, and Qiwei Li. A bayesian modified ising model for identifying spatially variable genes from spatial transcriptomics data. *Statistics in Medicine*, 41(23):4647–4665, 2022.

[16] Matthew T. Moores, Geoff K. Nicholls, Anthony N. Pettitt, and Kerrie Mengersen. Scalable bayesian inference for the inverse temperature of a hidden potts model, 2015.

[17] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[18] Susan Athey Dustin Tran, Francisco J. R. Ruiz and David M.Blei. Model criticism for bayesian causal inference. *arXiv: Methodology*, 2016.

[19] Judea Pearl. *Causality*. Cambridge university press, 2009.

[20] Consortium W. T. C. C. et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[21] Donald B Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

[22] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[23] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970.

[24] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[25] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[26] Osvaldo A. Martin, Ravin Kumar, and Junpeng Lao. *Bayesian Modeling and Computation in Python*. Chapman & Hall, Boca Raton, December 2021.

[27] Thomas V. Wiecki John Salvatier and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2016:e55, 4 2016.

[28] Cam Davidson-Pilon. *Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference*. Addison-Wesley Data & Analytics Series. Pearson Education, 2015.

[29] *Introductory Overview of PyMC — PyMC dev documentation*. https://docs.pymc.io/en/latest/learn/core_notebooks/pymc_overview.html[Revisado: enero 2023].

[30] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[31] Ravin Kumar, Colin Carroll, Ari Hartikainen, and Osvaldo Martin. Arviz a unified library for exploratory analysis of bayesian models in python. *Journal of Open Source Software*, 4(33):1143, 2019.

[32] William A. Link and Mitchell J. Eaton. On thinning of chains in mcmc. *Methods in Ecology and Evolution*, 3(1):112–115, 2012.

[33] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. http://github.com/google/jax, Version= 0.3.13.

[34] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

[35] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

[36] Yixin Wang and David M. Blei. The blessings of multiple causes: A tutorial. https://github.com/blei-lab/deconfounder_tutorial [Revisado: diciembre 2022].

[37] *National Medical Expenditure Survey (NMES), US Department of Health and Human Services Public Health service*, 1987. Extraído de https://github.com/blei-lab/deconfounder_public/tree/master/smoking_R/dat[Revisado: febrero 2023].

[38] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. University of California, Irvine, School of Information and Computer Sciences http://archive.ics.uci.edu/ml [Revisado: diciembre 2022].

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[40] Pablo Felipe Aguila Suarez. Aprendizaje automático aplicado a problemas sociales. 2022. [Tesis de licenciatura, DF, FCEN, UBA: octubre 2022].

[41] Agustín De Grande, Pablo y Salvia. Estratificación y desigualdad social (total país), 2010, 2021. https://mapa.poblaciones.org/map/97801 [Recuperado: 10 de mayo 2022].

[42] Bad posterior geometry and how to deal with it, 2021. https://num.pyro.ai/en/stable/tutorials/bad_posterior_geometry.html [Recuperado: enero 2023].

# Apéndice A

# Dinámica Hamiltoniana

## A.1. Propiedades

### Preserva volumen

Para que preserve el volumen, el determinante del Jacobiano de $T$ debe ser 1 ($|\nabla T| = 1$). Por el teorema de Liouville, alcanza con probar que la divergencia del campo vectorial es nula. Se plantea esta divergencia y usan las ecuaciones Hamiltonianas definidas en la ecuación 4.2 para demostrarlo.

$$\begin{aligned}\nabla \cdot \left(\frac{dq}{dt}, \frac{dp}{dt}\right) &= \sum_{d=1}^{D} \left(\frac{\partial}{\partial q_d}\frac{dq_d}{dt} + \frac{\partial}{\partial p_d}\frac{dp_d}{dt}\right) \\ &= \sum_{d=1}^{D} \left(\frac{\partial}{\partial q_d}\frac{\partial H}{\partial p_d} - \frac{\partial}{\partial p_d}\frac{\partial H}{\partial q_d}\right) = 0\end{aligned}$$

### Conservación de la energía

Resta probar que la energía global, el Hamiltoniano $H$ no cambia con el tiempo. Es decir, $\frac{dH}{dt} = 0$.

$$\begin{aligned}\frac{dH}{dt} &= \sum_{d=1}^{D} \left(\frac{\partial H}{\partial q_d}\frac{dq_d}{dt} + \frac{\partial H}{\partial p_d}\frac{dp_d}{dt}\right) \\ &= \sum_{d=1}^{D} \left(\frac{\partial H}{\partial q_d}\frac{\partial H}{\partial p_d} - \frac{\partial H}{\partial p_d}\frac{\partial H}{\partial q_d}\right) = 0\end{aligned}$$

## A.2. Tasa de aceptación de HMC

Con la tasa de aceptación definida según 4.5, recordando la definición del Hamiltoniano 4.4 y que se preserva la energía, por lo que

$$H(\boldsymbol{q}_0, \boldsymbol{p}_0) = H(\boldsymbol{q}_t, \boldsymbol{p}_t)$$

$$\begin{aligned}
A(\boldsymbol{q}_t, \boldsymbol{p}_t, \boldsymbol{q}_0, \boldsymbol{p}_0) &= min\left(1, \frac{\pi(\boldsymbol{q}_t)\phi_M(\boldsymbol{p}_t)}{\pi(\boldsymbol{q}_0)\phi_M(\boldsymbol{p}_0)}|\nabla T|\right), \\
A(\boldsymbol{q}_t, \boldsymbol{p}_t, \boldsymbol{q}_0, \boldsymbol{p}_0) &= min\left(1, \exp\left(log\left(\frac{\pi(\boldsymbol{q}_t)\phi_M(\boldsymbol{p}_t)}{\pi(\boldsymbol{q}_0)\phi_M(\boldsymbol{p}_0)}\right)\right) \cdot 1\right), \\
A(\boldsymbol{q}_t, \boldsymbol{p}_t, \boldsymbol{q}_0, \boldsymbol{p}_0) &= min\left(1, \exp\left(H(\boldsymbol{q}_0, \boldsymbol{p}_0) - H(\boldsymbol{q}_t, \boldsymbol{p}_t)\right)\right), \\
A(\boldsymbol{q}_t, \boldsymbol{p}_t, \boldsymbol{q}_0, \boldsymbol{p}_0) &= min\left(1, \exp(0)\right)
\end{aligned}$$

## A.3. Propiedades de integración numérica mediante el método *leapfrog*

De las ecuaciones de actualización de parámetros en 4.6, se desprende que los Jacobianos son

$$\begin{aligned}
\nabla\hat{T}_{\boldsymbol{p}}(\boldsymbol{q}, \boldsymbol{p}) &= \begin{bmatrix} 1 & 0 \\ \frac{\epsilon}{2}\nabla^2 log\ \pi(\boldsymbol{q}) & 1 \end{bmatrix}, \\
\nabla\hat{T}_{\boldsymbol{q}}(\boldsymbol{q}, \boldsymbol{p}) &= \begin{bmatrix} 1 & \epsilon M^{-1} \\ 0 & 1 \end{bmatrix},
\end{aligned}$$

y cumplen con la condición $|\nabla T| = 1$, por lo que se preserva el volumen.

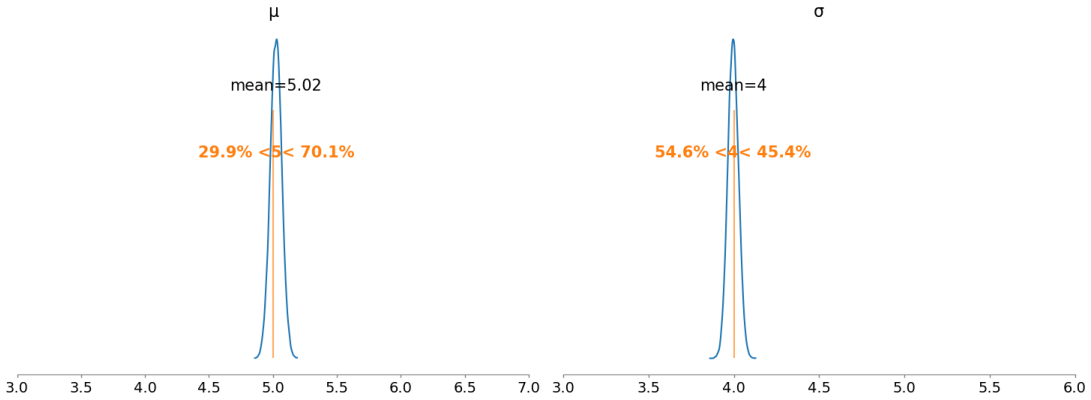
# Apéndice B

# Gráficos complementarios al capítulo 4

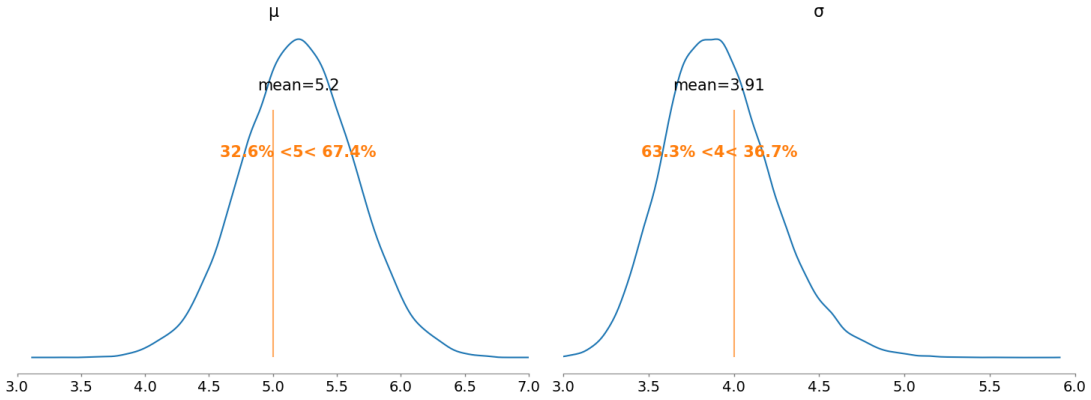
A continuación se presenta la misma densidad de probabilidad que la figura 4.4 pero con los ejes ampliados para facilitar su contraste con la distribución de realizar el mismo procedimiento de muestreo reduciendo la cantidad de datos a un 1 %, usando $N = 80$. Los resultados para el nuevo caso son $\mu = 5{,}2 \pm 0{,}4$ y $\sigma = 3{,}9 \pm 0{,}3$ (recordamos que con $N = 8000$ eran $\mu = 5{,}02 \pm 0{,}04$ y $\sigma = 4{,}0 \pm 0{,}03$), por lo que se tiene 10 veces la incertidumbre de antes. Este efecto se nota en la figura, ya que en B.1b se ve una distribución gaussiana centrada en la misma zona, pero más ancha que B.1a.

La figura B.2 ilustra el efecto del argumento *tune* en la función de sampleo. A la izquierda se tienen las densidades posteriores, mientras que a la derecha se ven sólo los primeros 2000 puntos del muestreo, para que se llegue a ver el comportamiento de arrastre o *drifting* de las tres cadenas en el caso de B.2b que no sucede en B.2a. Los pasos de entrenamiento son necesarios para asegurarse que los puntos empleados para hacer la inferencia sean del sector de cadena que ya convergió. El efecto que tienen estas cadenas pueden verse en la distribución posterior con colas pesadas en valores extremos, por lo que las estimaciones de los parámetros informan valores de $\mu = 5{,}0 \pm 0{,}5$ y $\sigma = 4 \pm 3$, pero no es correcto ya que éstos implicarían una distribución normal con esos parámetros, y la asimetría observada no coincide con este reporte.

Cambiar el número de cadenas o de muestreo no genera gran impacto en la inferencia en cuanto a su forma funcional y estimadores, pero si cambian diagnósticos como el tamaño de muestra estimada 'ess', que pasa del orden de $\sim$ 20,000 a $\sim$ 900 para una cadena y $\sim$ 2,000 para 3 cadenas de 1000 puntos.

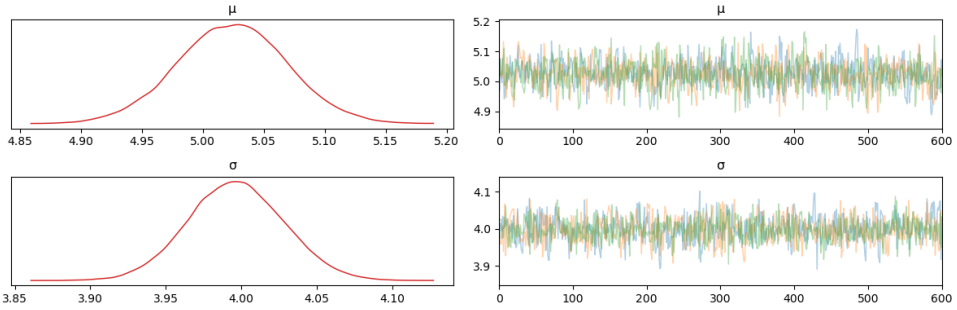



(a) N = 8000

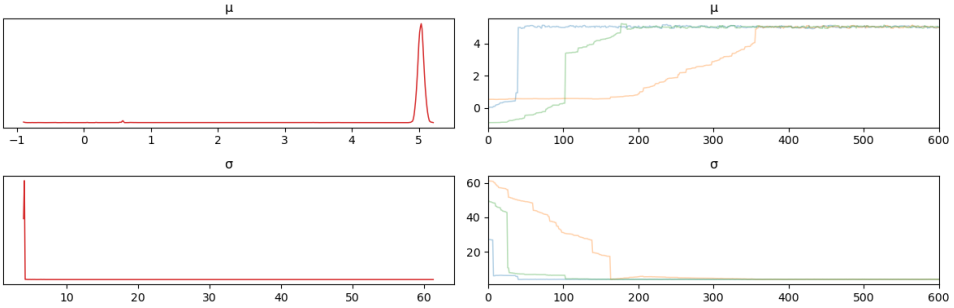



(b) N = 80

Figura B.1: Densidades posteriores para los parámetros $\mu$ y $\sigma$ según el número $N$ de datos observados. Parámetros de *draw*, *tune* y *chains* idénticos entre sí.




(a) *Tune* = 1500




(b) *Tune* = 10

Figura B.2: Densidades posteriores y trazas recortadas para los parámetros $\mu$ y $\sigma$ con $N = 8000$ datos observados, mismo número de cadenas y puntos de muestreo, efecto del parámetro *tune*.