UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

Instituto de Física y Matemáticas

Licenciatura en Matemáticas Aplicadas

Sobre inferencia bayesiana sesgada

TESIS

Para obtener el título de:

LICENCIADO EN MATEMÁTICAS APLICADAS

PRESENTA:

Brain de Jesús Salazar García

Director de Tesis

Dr. José del Carmen Jiménez Hernández

Huajuapan de León, Oaxaca, julio de 2024.

DEDICATORIA

A Mercedes Celia García Flores,
por ser mi modelo a seguir en la vida.

AGRADECIMIENTOS

Todo lo que he conseguido en esta vida, ha sido gracias al apoyo y la motivación que me brindan las personas que me rodean, me considero afortunado de tenerlos en mi vida.

A mi madre Mercedes Celia García Flores por ser tan fuerte y mostrarme que el cariño tiene distintas formas de expresarse. Gracias por tantas lecciones, buenos consejos, por todo el apoyo que me has dado durante toda mi vida y por guiarme para ser una buena persona.

A mis hermanas Melina y Betzhy por alegrarse y enorgullecerse de cada logro en mi vida, por mostrarme que hay distintas formas de disfrutar la vida y por ser mi motivo para salir adelante.

A mis tíos, primos y abuelos por estar al pendiente de mi, por todo su cariño y respeto y por ayudarme cuando más lo necesito.

A Laura Gabriela Morales Velasquez por ayudarme a enfrentar cada reto, gracias por estar conmigo en las buenas y en las malas, escucharme cuando lo necesitaba y por confiar tanto en mi.

A Ángel Bautista y Martha Alvarado por todo el apoyo y comprensión que me han dado, muchas gracias por toda la confianza y por tan buenos momentos compartidos.

Al Dr. José del Carmen Jiménez Hernández por confiar en mi para la elaboración de este proyecto, su paciencia a lo largo del mismo y por su disposición para enseñarme lo necesario y más. Mucha admiración y respeto para usted.

A mis sinodales Dra. Marisol López Cerino, M.C. Ana Delia Olvera Cervantes y M.M.M. Armando Alcalá Vallejo por el tiempo dedicado a la corrección de esta tesis, sus observaciones tan atinadas y por estar siempre al pendiente del desarrollo del trabajo.

Introducción

La inferencia bayesiana, es una teoría cuyo origen se remonta a hace más de 300 años (desde mediados del siglo XVII). Sin embargo, no fue hasta mediados de la década de los 50 que logró formalizarse y comenzó a tomar relevancia dentro del área de la inferencia. Su importancia ha ido creciendo con el tiempo, ya que se ha probado que es una herramienta que permite resolver problemas en los que la inferencia tradicional no proporciona buenos resultados. Tal es su utilidad que en los últimos años ha desempeñado un papel sobresaliente para lograr avances dentro de áreas como la psicología, ecología, física, aprendizaje-maquina y neurociencias; por mencionar algunas. El objetivo de la estadística, y en particular de la estadística bayesiana, es proporcionar una metodología para analizar adecuadamente la información con la que se cuenta, y decidir de manera razonable sobre la mejor forma de actuar (Rangel, 2015). En este sentido, una de las metas particulares de la inferencia bayesiana es encontrar una distribución de probabilidad para los parámetros de interés. No obstante, aunque en algunos casos se logra llegar de manera analítica a dicha distribución, también existen situaciones en las que esto no es posible, en este caso se recurren a los métodos de simulación. Dicha distribución de los parámetros se le conoce como distribución *a posteriori*, se obtiene a partir del teorema de Bayes el cual establece que esta es proporcional a la verosimilitud por la distribución previa de los parámetros. Una vez obtenida la distribución *a posteriori* se puede realizar la inferencia respecto a los parámetros de interés.

En 2018, Kaosu Matsumori, Yasuharu Koike y Kenji Matsumoto (Matsumori *et al.*, 2018) proponen que la distribución *a posteriori* sea proporcional a la verosimilitud elevada a una potencia β por la distribución *a priori* de los parámetros elevada a una potencia α. Ambos, β y α, les llaman sesgos cognitivos o sesgos exponenciales. El objetivo de esta tesis es encontrar la distribución a posteriori para los parámetros de interés aplicando la propuesta de Matsumori *et al.* (2018) al modelo normal multivariado conjugado. Uno de los trabajos que han abordado esta nueva propuesta es Montoya *et al.* (2024), en el cual se aborda el problema de la selección de portafolios con la incorporación de sesgos de comportamiento.

Índice general

1. Antecedentes teóricos **1**

1.1. Conceptos básicos 1

1.2. Teorema de Bayes 4

1.3. Variables aleatorias y sus distribuciones 6

1.3.1. Momentos, varianza y covarianza 9

1.3.2. Algunas distribuciones 11

1.4. Vectores aleatorios y sus distribuciones 12

1.4.1. Variables aleatorias independientes 14

1.4.2. Algunas distribuciones de vectores aleatorios 15

2. Estadística bayesiana **19**

2.1. Inferencia bayesiana 19

2.1.1. Teoría de la decisión 20

2.1.2. Estimación bayesiana 21

2.2. Distribución *a priori* 22

2.2.1. Distribuciones *a priori* no informativas 23

2.2.2. El principio de razón insuficiente 23

2.2.3. *A priori* no informativa de Jeffreys 25

2.2.4. Distribuciones *a priori* conjugadas 27

2.2.5. Otras alternativas para las distribuciones a priori 27

2.3. Modelo normal con media desconocida y varianza conocida 28

2.4. Modelo normal con media y varianza desconocidas 31

2.5. Modelo normal multivariado conjugado 38

3. Inferencia bayesiana sesgada **45**

3.1. Planteamiento 45

3.2. Modelo normal multivariado conjugado 47

4. Conclusiones **57**

Referencias **58**

◇

Lista de figuras

2.1. Modelo normal con media desconocida y varianza conocida. 31

3.1. Comportamiento de la distribución *a priori* y la función de verosimilitud para diferentes valores de α y β. Fuente: Recuperada de Matsumori *et al.* (2018) 47

Capítulo 1

Antecedentes teóricos

La probabilidad es el lenguaje matemático por medio del cual se cuantifica la incertidumbre. A pesar de que la palabra probabilidad se utiliza en la vida cotidiana de manera indiscriminada, sobre cualquier tipo de suceso, matemáticamente puede resultar bastante complejo asignarle una probabilidad a algún evento en particular. La teoría de la probabilidad a lo largo del tiempo a dado paso a distintas técnicas de inferencia. Por ello, en este primer capítulo se presentarán algunas definiciones y resultados que serán de utilidad para el mejor entendimiento del desarrollo de esta tesis. Cabe mencionar que dicha información fue recolectada de los siguientes textos, Gelman *et al.* (1995), Matsumori *et al.* (2018), Mood *et al.* (1978) y Mendoza y Regueiro (2011), Nuñez-Antonio y Castillo (2014), Hernández González (2005), Evans (2018).

1.1. Conceptos básicos

Dentro de la vida cotidiana, es muy común encontrarse con situaciones donde, la incertidumbre sobre algún evento será desvanecida solo cuando haya ocurrido dicho evento. Por ejemplo: el clima del día, el tráfico en la ciudad, el valor de las acciones o dónde se dejaron las llaves. La respuesta no se conoce con certeza. En su lugar lo que se hace es estimar una posible respuesta. La probabilidad es la ciencia de la incertidumbre Evans (2018). La probabilidad proporciona una estructura de razonamiento que permite trabajar con nuestro conocimiento y tomar decisiones a partir de lo que se sabe y lo que se desconoce. Para entender mejor la probabilidad es necesario conocer los siguientes conceptos.

Un **experimento** es el proceso por medio del cual se obtiene una observación o resultado. Los experimentos permiten recabar información acerca de un suceso. Existen dos tipos de experimentos:

Los **experimentos deterministas** son aquellos en los que se obtiene el mismo resultado de manera repetitiva, siempre que los experimentos se repitan bajo condiciones aproximadamente idénticas.

Los **experimentos aleatorios** son aquellos en los que los resultados no son esencialmente los mismos a pesar de que se realicen en condiciones aproximadamente idénticas.

En un experimento aleatorio se pueden presentar resultados distintos aún teniendo las mismas condiciones. Algunos ejemplos de experimentos aleatorios son: observar si cae cruz o cara al lanzar una moneda al aire, el número ganador de la lotería, entre otros. Se puede generalizar lo anterior con la siguiente definición.

Definición 1.1.1 *Los experimentos o fenómenos aleatorios son aquellos en los que no se puede predecir los resultados.*

La probabilidad se encarga de estudiar los experimentos aleatorios. Ya que no se sabe cuál será el resultado de un experimento aleatorio, es necesario agrupar al conjunto de todos los resultados posibles en un conjunto.

Definición 1.1.2 *El espacio muestral, o también llamado espacio muestra, de un experimento aleatorio ϵ es el conjunto de todos los resultados posibles del experimento y se denota por la letra griega Ω. A cada uno de los resultados en un espacio muestral se le llama elemento del espacio muestral, o simplemente punto muestral y se denota por ω.*

Para comprender mejor lo anterior se presenta el siguiente ejemplo.

Ejemplo 1.1.1 *Se realiza el experimento aleatorio de lanzar un dado. Interesa conocer el número de puntos en la cara superior. De aquí se tiene que, el espacio muestral de todos los resultados posibles es $\Omega = \{1, 2, 3, 4, 5, 6\}$. Por otro lado, supóngase que nos interesa saber si el número que cae es par o impar, en este caso nuestro espacio muestral cambia a $\Omega = \{par, impar\}$. En este ejemplo se observa que el espacio muestral de un experimento aleatorio no es único y depende del interés del observador.*

Un espacio muestral con un numero finito o a lo más numerable de elementos, se denomina **espacio muestral discreto**. Por otro lado, un espacio muestral con un numero infinito no numerable de elementos, es llamado **espacio muestral continuo.**
El Ejemplo (1.1.1) expone un espacio muestral discreto. Un ejemplo más elaborado de espacio muestral discreto es el siguiente:

Ejemplo 1.1.2 *Supóngase que se desea determinar el número de lanzamientos de una moneda que deberá hacerse hasta que aparezca la primera cara. Esta puede aparecer en*

lanzamiento $1, 2, 3, ...$ *De aquí el espacio muestral,* $\Omega = \{1, 2, 3, 4, ...\}$*, está formado por una infinidad numerable de puntos (tantos como números enteros positivos).*

Para representar el caso de un espacio muestral continuo se tiene el siguiente ejemplo.

Ejemplo 1.1.3 *Considérese el experimento aleatorio que consiste en seleccionar una muestra aleatoria de estudiantes de sexto grado en determinada ciudad y registrar su peso. El resultado puede ser cualquier número positivo. Cabe objetar que no habrá estudiantes que pesen menos de un kilogramo o más de 1000. Es cierto, pero no es objeción si se incluyen los resultados imposibles al enumerar cada resultado imaginable. Por tanto, este espacio muestral estará formado por todos los números positivos* $\Omega = \mathbb{R}^+$*.*

Definición 1.1.3 *Un evento es un subconjunto* A *del espacio muestral* Ω*, es decir, un conjunto de puntos muestrales.*

De aquí, se dice que el evento A a ocurrido, si el resultado de un experimento es un elemento de A. Un evento que consiste de un sólo punto muestral de Ω frecuentemente se llama un evento **elemental o simple**. En cambio, se llama **compuesto** cuando consta de más de un elemento del espacio muestral Ω. Los eventos o sucesos se denotan por las letras del alfabeto en mayúsculas $A, B, C, ...$ Los eventos triviales son el conjunto vacío $\emptyset$ y el mismo espacio muestral Ω, también conocidos como evento imposible y evento seguro respectivamente.

Ejemplo 1.1.4 *Un experimento aleatorio que consiste en lanzar dos dados y observar los números que salen. El espacio muestral consta de 36 puntos, que es* $\Omega = \{(1, 1), (1, 2), ..., (6, 6)\}$ *donde los pares ordenados de números representan los resultados respectivos del primero y del segundo dado. Definimos el evento A por la condición de que la suma de las dos puntuaciones sea igual a siete; tal evento está formado por seis puntos,* $A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$*.*

El concepto de probabilidad está dado en términos de eventos. Por ello, una vez que se definió el concepto de evento, se puede dar paso a analizar el concepto de probabilidad. La probabilidad se divide en **probabilidad clásica** y **probabilidad frecuentista**, ambas interpretaciones son estructuradas y justificadas a través de la **definición axiomática de la probabilidad**.

Sea Ω un espacio muestral y A cualquier evento de Ω. Es decir, A es cualquier subconjunto de Ω. Se dice que $\mathrm{P}(\cdot)$ es una *función de probabilidad* en el espacio muestral Ω si se satisfacen los tres axiomas siguientes.

Axioma 1.1.1 *P(A) es un número real tal que* $P(A) \geq 0$ *para todo evento* A *de* Ω

Axioma 1.1.2 *P(*Ω*) = 1.*

Axioma 1.1.3 *Si* $A_1, A_2, ...,$ *es una sucesión de eventos mutuamente excluyentes, es decir* $A_i \cap A_j = \emptyset$ *para* $i \neq j = 1, 2, ...,$ *entonces*

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots .$$

La definición axiomática de la probabilidad, formulada por Andrey Kolmogorov, proporciona una estructura formal y rigurosa para el estudio de la teoría de la probabilidad. Esto garantiza coherencia y proporciona un marco matemático preciso para el análisis y la inferencia probabilística.

1.2. Teorema de Bayes

Antes de enunciar y presentar el teorema de Bayes, se necesitan las siguientes definiciones.

Definición 1.2.1 *Sean* A *y* B *dos eventos de un espacio muestral* Ω *tal que* $P(B) > 0$. ***La probabilidad condicional*** *del evento* A *dado* B*, y que se denota por* $P(A|B)$*, es*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

El concepto de probabilidad condicional, da paso a la definición de independencia de eventos.

Definición 1.2.2 *Sean* A *y* B *dos eventos de un espacio muestral* Ω*. Se dice que* A *y* B *son independientes si se satisface cualquiera de las siguientes igualdades,*

a) $P(A|B) = P(A)$.

b) $P(B|A) = P(B)$.

c) $P(A \cap B) = P(A)P(B)$.

En otras palabras, la definición anterior significa que si $P(A|B)$ no depende del evento B entonces A y B son independientes.
Supóngase ahora que $\{B_1, B_2, ..., B_k\}$ es una partición[1] de Ω tal que $P(B_i) > 0$, para cada $i = 1, 2, ..., k$. Luego para cualquier evento A se tiene que

$$\begin{aligned} A &= A \cap \Omega \\ &= A \cap (B_1 \cup B_2 \cup \cdots \cup B_k) \\ &= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_k). \end{aligned}$$

Así, por la definición de eventos independientes,

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_k) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k) \\ &= \sum_{j=1}^{k} P(A|B_j)P(B_j). \end{aligned}$$

Por otro lado, nótese que por definición,

$$P(B_j|A) = \frac{P(A \cap B)}{P(A)},$$

si sustituye $P(A)$ por la igualdad encontrada anteriormente se tiene que,

$$P(B_j|A) = \frac{P(A \cap B)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)}.$$

La igualdad anterior con las hipótesis planteadas, conforman lo que se conoce como el **teorema de Bayes**. A continuación se analizará una de las muchas aplicaciones que tiene este resultado.

Ejemplo 1.2.1 *Un fusible electrónico es producido por cinco líneas de producción en una operación de manufactura. Los fusibles son costosos, sumamente confiables y se envían a proveedores en lotes de 100 unidades. Como la prueba es destructiva, la mayoría de los compradores de fusibles prueban sólo un número pequeño de ellos antes de decidirse a aceptar o rechazar lotes de fusibles que lleguen. Las cinco líneas de producción producen fusibles al mismo ritmo y normalmente producen sólo 2 % de fusibles defectuosos, que se dispersan al azar en la producción. Desafortunadamente, la línea 1 de producción sufrió problemas mecánicos y produjo 5 % de piezas defectuosas durante el mes de marzo. Esta situación llegó al conocimiento del fabricante después de que los fusibles ya habían sido enviados. Un cliente recibió un lote producido en marzo y probó tres fusibles. Uno falló. ¿Cuál es la probabilidad de que el lote se haya producido en la línea 1? ¿Cuál es la*

[1] Sean $\{B_1, B_2, ..., B_k\}$ tales que si $i \neq j$, $B_i \cap B_j = \emptyset$ y $\cup_{i=1}^{k} B_i = \Omega$.

probabilidad de que el lote haya provenido de una de las otras cuatro líneas?

Solución: Se denota con B al evento de que un fusible sea sacado de la línea 1 y con A el evento de que el fusible estuviera defectuoso. Se deduce entonces directamente que $P(B) = 0.2$ y $P(A|B) = 3(0.05)(0.95)^2 = 0.1353$.

Del mismo modo, $P(\bar{B}) = 0.8$ y $P(A|\bar{B}) = 3(0.02)(0.98)^2 = 0.057624$.

Observe que estas probabilidades condicionales fueron muy fáciles de calcular. Usando la ley de probabilidad total,

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \\ &= (0.135375)(0.2) + (0.057624)(0.8) \\ &= 0.0731. \end{aligned}$$

Por último, aplicando el teorema de Bayes

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{(0.135375)(0.2)}{0.0731742} = 0.37$$

y

$$P(\bar{B}|A) = 1 - P(B|A) = 1 - 0.37 = 0.63.$$

Por lo tanto la probabilidad de que el lote se haya producido en la linea 1 es 0.37 y la probabilidad de que el lote se haya producido en alguna otra linea es 0.63.

1.3. Variables aleatorias y sus distribuciones

Una variable aleatoria asigna un número real a cada elemento del espacio muestral. Esto da paso a estudiar el resultado de un experimento aleatorio en el espacio de los números reales mediante los valores que tome la variable aleatoria, (**v.a**).

Definición 1.3.1 *Sea Ω un espacio muestral en el que se define una función de probabilidad. Una variable aleatoria es una función X que va del espacio muestral Ω al conjunto de los números reales,*

$$X : \Omega \to \mathbb{R}$$

Si $\omega \in \Omega$, se denotará con $X(\omega)$ a la imagen de ω bajo X. A la imagen de X se le denomina recorrido o soporte de X y será denotado por $R(X)$.

$$R(X) = X(\Omega) = \{x \in \mathbb{R} : \exists\ \omega \in \Omega \text{ para el cual } X(\omega) = x\}.$$

$R(X)$ es nuevamente un espacio muestral.

Así X transforma todos los posibles resultados del espacio muestral en cantidades numéricas. Las letras mayúsculas V, W, X, Y y Z denotarán variables aleatorias, en caso de necesitar más de tres v.a., se usarán $Y_1, Y_2, Y_3, \ldots$.

Definición 1.3.2 *Una variable aleatoria X es discreta si $R(X)$ es un conjunto finito o infinito numerable.*

Definición 1.3.3 *Una variable aleatoria X es continua si $R(X)$ es un intervalo de la recta de los números reales.*

Dada una v.a. existen dos funciones que proveen de información acerca de las características de la v.a. Estas funciones, son la función de densidad de probabilidad (f.d.p.) y la función de distribución acumulada (f.d.a.) las cuales permiten representar al mismo tiempo tanto los valores que puede tomar la v.a. como las probabilidades de los distintos eventos de interés. Dado que $a \leq X \leq b$ es un evento, la f.d.p. proporciona un medio para calcular la probabilidad de que la variable X tome un valor dentro del intervalo $[a, b]$.
La f.d.p. de una v.a. continua X se define formalmente de la siguiente manera.

Definición 1.3.4 *Sea X una variable aleatoria continua. Se dice que la función integrable y no negativa $f_X : \mathbb{R} \to [0, \infty)$ es la función de densidad de probabilidad de X si para cualquier intervalo $(a, b) \subset \mathbb{R}$ se cumple la igualdad,*

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

Toda función de densidad de probabilidad $f_X(x)$ de una variable aleatoria continua cumple las siguientes propiedades:
a) $f_X(x) \geq 0$, para cualquier valor $x \in \mathbb{R}$.

b) $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Puesto que el área total bajo $f_X(x)$ es uno, la probabilidad del intervalo $a \leq X \leq b$ es el área acotada por la función de densidad y las rectas $X = a$ y$X = b$.
Para el caso discreto, se tiene la siguiente definición:

Definición 1.3.5 *Sea X una variable aleatoria discreta. Se dice que la función p es la función de densidad de probabilidad de X si para cualquier subconjunto A a lo más numerable de los números reales, se cumple que*

$$P(A) = \sum_{x \in A} p(x)$$

Respecto a la función de distribución acumulada se tiene la siguiente definición:

Definición 1.3.6 *Sea X una variable aleatoria. La función de distribución acumulada de la variable aleatoria X es la función $F_X : \mathbb{R} \to [0, 1]$, definida como*

$$F_X(x) = P(X \leq x), \;\; x \in \mathbb{R}.$$

En particular si, X es una v.a. continua, por definición la f.d.a. esta dada por,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

la función de distribución acumulada $F_X(x)$ es el área acotada por la función de densidad que se encuentra a la izquierda de la recta $X = x$. Además por el teorema fundamental del cálculo se cumple:

$$\frac{dF_X(x)}{dx} = F'_X(x) = f_X(x).$$

De este modo se puede determinar $f_X(x)$ a partir de $F_X(x)$. La f.d.a. $F_X(x)$, es una función que satisface ciertas propiedades, las cuales se enuncian en la siguiente proposición.

Proposición 1.3.1 *Toda función de distribución acumulada $F_X(x)$ cumple las siguientes propiedades:*

a) $\lim_{x \to -\infty} F_X(x) = 0$.

b) $\lim_{x \to \infty} F_X(x) = 1$.

c) Si $x_1 \leq x_2$, entonces $F_X(x_1) \leq F_X(x_2)$.

d) Si $x_1 \leq x_2$, entonces $P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1)$.

1.3.1. Momentos, varianza y covarianza

El valor esperado de una variable aleatoria o de cualquier función de una variable aleatoria se obtiene hallando el valor promedio de la función para todos los valores posibles de la variable. En realidad es una media teórica o ideal ya que realmente no se espera que la variable tome ese valor exacto pero si que se acerque a el.

Definición 1.3.7 *Sea X una variable aleatoria, el valor esperado de X, denotado por $E(X)$ se define como*

$$E(X) = \sum_X xp(x),$$

si X es discreta, y

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx,$$

si X es continua.

Estas cantidades se definen como el valor esperado solo si son absolutamente convergentes. Si no lo son, decimos que el valor esperado de X no existe.

Definición 1.3.8 *Sea X una variable aleatoria, el valor esperado de una función u de la variable aleatoria X es*

$$E[u(X)] = \sum_X u(x)p(x),$$

si X es discreta, y

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)dx,$$

si X es continua.

El valor esperado cumple con las propiedades de linealidad.
Los momentos de una distribución son los valores esperados de las potencias de la variable aleatoria. El momento r-ésimo de X se denota por μ'_r y está dado por:

$$\mu'_r = E\left(X^r\right).$$

Note que el primer momento μ'_1 es el valor esperado de X. Los momentos respecto a un

punto arbitrario cualquiera se definen como:

$$E[(X-a)]$$

y haciendo a igual a la media, se obtiene los momentos respecto a la media o momentos centrales, que suelen designarse por μ_r :

$$\mu_r = E\left[(X-\mu_1')^r\right].$$

Se tiene

$$\mu_1 = \mu_1' - \mu_1' = 0$$

y

$$\mu_2 = \mu_2' - (\mu_1')^2$$

Este segundo momento respecto a la media recibe el nombre de varianza de X. Otro término que proporciona información acerca del comportamiento entre variables aleatorias es la covarianza. En seguida se proporciona la definición de covarianza así como algunas de sus propiedades.

- La **covarianza** entre 2 variables aleatorias denotada por $\mathrm{Cov}(X,Y)$, se define como

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[X-E[X]][Y-E[Y]] \\ &= E[XY]-E[X]E[Y]\end{aligned}$$

Si $E[X]=0$ o $E[Y]=0$ entonces $\mathrm{Cov}(X,Y)=E[XY]$.

-La covarianza de X consigo misma es igual a la varianza de X: $\mathrm{Cov}(X,X)=\mathrm{Var}(X)$.

- $E[XY]$ es la correlación entre las dos variables aleatorias.

- Por la desigualdad de Cauchy-Schwartz

$$|E[XY]| \leq \sqrt{E\left[X^2\right]E\left[Y^2\right]}.$$

- X y Y se dice que son:

- Ortogonales si $E[XY]=0$.

- No correlacionadas si $\mathrm{Cov}(X,Y)=0$.

- El coeficiente de correlación se define como

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

donde $\sigma_X = \sqrt{\text{Var}(X)}$ y $\sigma_Y = \sqrt{\text{Var}(Y)}$ son las desviaciones estándar.

Propiedades:

- Acotado: $-1 \leq \rho_{X,Y} \leq 1$[2].

- Si $\rho_{X,Y} = 0$, X y Y son no correlacionadas.

- Si X e Y son independientes, entonces $\rho_{X,Y} = 0$.

La independencia implica no correlación, pero no al revés.

1.3.2. Algunas distribuciones

A continuación se muestran algunos ejemplos de funciones de densidad de probabilidad que serán de utilidad para el desarrollo de esta tesis.

Definición 1.3.9 *Sea X una v.a, se dice que X sigue una distribución* ***normal*** *con media μ y varianza σ^2 si su función de densidad de probabilidad está dada por*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

donde $\mu \in \mathbb{R}$, $\sigma^2 > 0$ y $x \in \mathbb{R}$ y se denota como $X \sim N(\mu, \sigma^2)$.

Por medio de la definición se puede mostrar que, X tiene valor esperado μ y varianza σ^2.

Definición 1.3.10 *Sea X una v.a. mayor que cero. Se dice que X sigue una distribución* ***gamma inversa*** *con parámetro de forma α y parámetro de escala β si su función de densidad de probabilidad está dada por*

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right), \tag{1.1}$$

con $\alpha > 0$ y $\beta > 0$ y se denota como $X \sim GI(\alpha, \beta)$.

X tiene valor esperado $\frac{\beta}{\alpha-1}$ y varianza $\beta^2/\left((\alpha-1)^2 * (\alpha-2)\right)$.

[2] $\rho_{X,Y} = 1$ significa que hay una relación lineal positiva entre las variables.
Por otro lado $\rho_{X,Y} = -1$ significa que hay una relación lineal negativa entre las variables.

Definición 1.3.11 *Sea X una v.a. que solo toma los valores 0 y 1.Se dice que X sigue una **distribución Bernoulli** con parámetro θ si su función de densidad de probabilidad está dada por*

$$p_X(x) = \theta^x(1-\theta)^{1-x}x, \tag{1.2}$$

donde $\theta \in [0,1]$.

X tiene valor esperado θ y varianza $\theta(\theta-1)$.

Definición 1.3.12 *Sea X una v.a. que toma valores en el intervalo $(0,1)$. Se dice que X sigue una **distribución beta** con parámetros α y β si su función de densidad de probabilidad está dada por*

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}, \tag{1.3}$$

donde $\alpha > 0$, $\beta > 0$ y $\mathrm{B}(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$.

X tiene valor esperado $\frac{\alpha}{\alpha+\beta}$ y varianza $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
De la definición anterior se sigue que el kernel[3] de una distribución beta esta dado por

$$x^{\alpha-1}(1-x)^{\beta-1}.$$

1.4. Vectores aleatorios y sus distribuciones

Ahora se extenderá el concepto de v.a. con valores reales a v.a. con valores en $\mathbb{R}^n$, a tales funciones se les llama **variables aleatorias multidimensionales** o también **vectores aleatorios**. También se estudiarán varios conceptos relacionados con estas funciones.

Definición 1.4.1 *Un vector aleatorio de dimensión n es una función $\boldsymbol{X} : \Omega \to \mathbb{R}^n$ dada por $\boldsymbol{X} = (X_1, X_2, ..., X_n)$, donde cada coordenada $X_1, X_2, ..., X_n$ es una variable aleatoria.*

Definición 1.4.2 *Se dice que el vector aleatorio $(X_1, X_2, ..., X_n)$ es discreto si cada coordenada que lo conforma es una variable aleatoria discreta, y se dice que es continuo en caso de que cada coordenada sea una variable aleatoria continua.*

A continuación se define la función de distribución conjunta de un vector aleatorio y la función de distribución acumulada (f.d.a.) de un vector aleatorio. Estos conceptos son completamente análogos al caso unidimensional.

[3]El kernel de una distribución de probabilidad es la parte de la f.d.p. que solo depende de la v.a.

Definición 1.4.3 *Sea* $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ *un vector aleatorio discreto. Se dice que la función* $p : \mathbb{R}^n \to [0, \infty)$ *es la función de densidad conjunta de las variables* $X_1, X_2, ..., X_n$*, si para* A *subconjunto del conjunto de valores que puede tomar el vector aleatorio se cumple,*

$$P\left[(x_1, x_2, \ldots, x_k) \in A\right] = \sum_{A} p_{X_1,..,X_n}\left(x_1, x_2, \ldots, \dot{x}_k\right)$$

donde $\sum_A p_{X_1,..,X_n}(x_1, x_2, \ldots, x_k)$ *significa súmese la función* p *para todos los puntos de* A*.*

La función de densidad p cumple las siguientes dos propiedades:

a) $p_{X_1,..,X_n}(x_1, ..., x_n) \geq 0, \;\; x_1, ..., x_n \in \mathbb{R}$.

b) $\sum_{X_1} ... \sum_{X_n} p_{X_1,...,X_n}(x_1, ..., x_n) = 1$.

Definición 1.4.4 *Sea* $\boldsymbol{X} = (X_1, ..., X_n)$ *un vector aleatorio continuo. Se dice que la función integrable y no negativa* $f_{\boldsymbol{X}}(x_1, ..., x_n) : \mathbb{R}^n \to [0, \infty)$ *es la función de densidad de probabilidad conjunta de las variables* $X_1, ..., X_n$ *si se cumple la igualdad,*

$$P(a_1 \leq X_1 \leq b_1, ..., a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} f_{X_1,...,X_n}(x_1, ..., x_n) dx_n ... dx_1,$$

con $a_i < b_i$*,* $i = 1, ..., n$*.*

La función de densidad f cumple las siguientes dos propiedades:

a) $f_{X_1,..,X_n}(x_1, ..., x_n) \geq 0, \;\; x_1, ..., x_n \in \mathbb{R}$.

b) $\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1,...,X_n}(x_1, ..., x_n) dx_n ... dx_1 = 1$.

En esta tesis a la función de densidad de probabilidad conjunta de $X_1, ..., X_n$, se llamará simplemente función de densidad conjunta de $X_1, ..., X_n$.

Definición 1.4.5 *Sea* $(X_1, ..., X_n)$ *un vector aleatorio discreto o continuo. La función de distribución acumulada del vector* $(X_1, ..., X_n)$ *es la función* $F_{X_1,...,X_n}(x_1, ..., x_n) : \mathbb{R} \to [0, 1]$*, dada por:*

$$F_{X_1,...,X_n}(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n), \;\; \forall (x_1, ..., x_n) \in \mathbb{R}^n.$$

En particular si $(X_1, ..., X_n)$ es un vector aleatorio continuo, la función de distribución

conjunta de $X_1, ..., X_n$, está dada por

$$F_{X_1,...,X_n}(x_1, ..., x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_{X_1,...,X_n}(t_1, ..., t_n) dt_n \cdots dt_1.$$

Por lo tanto, la función de densidad conjunta $f_{X_1,...,X_n}(x_1, ..., x_n)$ se encuentra diferenciando $F_{X_1,...,X_n}(x_1, ..., x_n)$ con respecto a $x_1, ..., x_n$ es decir,

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \frac{\partial^n F_{X_1,...,X_n}(x_1, ..., x_n)}{\partial x_1 ... x_n}.$$

Definición 1.4.6 *Sea $f_{X_1,...,X_n}(x_1, ..., x_n)$ la función de densidad de probabilidad conjunta del vector aleatorio continuo $(X_1, ..., X_n)$. Se define la función de densidad de probabilidad marginal de la variable aleatoria X_i como sigue*

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1,...,X_n}(t_1, ..., t_n) dt_1 \cdots dt_{i-1} dt_{i+1} \cdots dt_n.$$

Se puede verificar que las funciones de densidad marginales son efectivamente funciones de densidad univariadas.

Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ una variable aleatoria k-dimensional. El valor esperado de X_i para cualquier $i = 1, 2, \ldots, k$ se designa por $E(X_i)$ y se define por

$$E(X_i) = \sum_{X_1} \sum_{X_2} \cdots \sum_{X_k} X_i p_{X_1,...,X_k}(x_1, \ldots, x_k),$$

si $\boldsymbol{X}$ es discreta, y por

$$E(X_i) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} X_i f(X_1, \ldots, X_k) \, dX_1 dX_2 \ldots dX_k,$$

si $\boldsymbol{X}$ es continua.

1.4.1. Variables aleatorias independientes

En la sección (1.2) se habló acerca del concepto de eventos independientes. Si se quisiera extender la definición de independencia a variables aleatorias, intuitivamente se puede decir que, las variables aleatorias X y Y son independientes si cada acontecimiento que implica solamente a X es independiente de cada evento que involucra sólo a Y. A continuación se enuncia con precisión el concepto de independencia entre variables aleatorias continuas. Las siguientes definiciones son análogas para el concepto de independencia entre variables aleatorias discretas.

Definición 1.4.7 *Sea $(X_1, ..., X_n)$ un vector aleatorio continuo. Se dice que las variables aleatorias $X_1, ..., X_n$ son independientes si la función de densidad de probabilidad conjunta $f_{X_1,...,X_n}(x_1, ..., x_n)$ es igual al producto de las n funciones de densidad marginal correspondientes a cada variable X_i para $i = 1, ..., n$, esto es,*

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i),$$

para todo $(x_1, ..., x_n) \in \mathbb{R}^n$.

Una forma alterna de definir la independencia es mediante la función de distribución conjunta. Ambas definiciones son equivalentes, se dice que un conjunto infinito de variables aleatorias son independientes si cualquier subconjunto finito de ellas también lo es.

1.4.2. Algunas distribuciones de vectores aleatorios

La distribución normal multivariada es una distribución de probabilidad de un vector aleatorio continuo, dicha distribución será de gran utilidad para el desarrollo de esta tesis. A continuación se enuncian algunos conceptos previos a definir cuando un vector aleatorio tienen una distribución normal multivariada.

Definición 1.4.8 *Sea $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ un vector aleatorio continuo n-dimensional. Se dice que $\boldsymbol{X}$ tiene una distribución normal multivariada si su función densidad de probabilidad está dada por*

$$f_{\boldsymbol{X}}(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)},$$

donde $x = (x_1, ..., x_n)$, $\mu = (E(X_1), E(X_2), ..., E(X_n))$ y Σ es la matriz de varianzas y covarianzas, es decir $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$, y se denota como $\boldsymbol{X} \sim N(\mu, \Sigma)$.

Definición 1.4.9 *Sea $\boldsymbol{X} = (\theta, \sigma^2)$ un vector aleatorio tal que $\theta \in (-\infty, \infty)$ y $\sigma^2 \in (0, \infty)$. Se dice que $\boldsymbol{X}$ tienen una distribución conjunta normal gamma inversa con parámetro de ubicación μ, parámetro de forma α, parámetro de escala β y donde λ es un escalar si su función de densidad de probabilidad está dada por*

$$f\left(\theta, \sigma^2 \mid \alpha, \mu, \lambda, \beta\right) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \left(\sigma^2\right)^{-\alpha-1} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\lambda(\theta-\mu)^2 + \beta\right)\right\},$$

donde $\mu \in \mathbb{R}$, $\alpha > 0$, $\lambda > 0$ y $\beta > 0$, y se denota como $\boldsymbol{X} \sim NGI\left(\alpha, \mu, \lambda, \beta\right)$.

Se nota, de la definición anterior, que el kernel de una distribución normal-gamma inversa esta dado por

$$\left(\sigma^2\right)^{-\alpha-1} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\lambda(\theta-\mu)^2+\beta\right)\right\}.$$

Una distribución que será de utilidad para el desarrollo de esta tesis es la distribución Wishart-Inversa que se define a continuación.

Definición 1.4.10 *Sea* $\boldsymbol{X} = (\boldsymbol{X_1}, ..., \boldsymbol{X_n})$ *donde cada* $\boldsymbol{X_i}$ *es un vector aleatorio de dimensión* n*, con esto* $\boldsymbol{X}$ *es una matriz de dimensión* $n \times n$ *definida positiva. Se dice que* $\boldsymbol{X}$ *sigue una distribución Wishart inversa con matriz de escala* τ *y* γ *grados de libertad si su función de densidad de probabilidad está dada por*

$$\frac{|\tau|^{\gamma/2}}{2^{\gamma n/2}\Gamma_n\left(\frac{\gamma}{2}\right)}|\boldsymbol{X}|^{-\frac{1}{2}(\gamma+n+1)}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\tau\boldsymbol{X}^{-1}\right)\right\} \tag{1.4}$$

donde $\gamma > n-1$ *y* τ *es una matriz de* $n \times n$ *definida positiva,* $\operatorname{tr}(\cdot)$ *es la traza de una matriz,* Γ_n *es la función gamma multivariante y se denota como,* $\boldsymbol{X} \sim WI(\tau, \gamma)$*.*

Mediante el uso de la definición de valor esperado se puede probar que,

$$E(\boldsymbol{X}) = \frac{\tau}{\gamma - n - 1}.$$

Definición 1.4.11 *Sea* $\boldsymbol{X} = (X_1, X_2, ..., X_p)$ *un vector aleatorio continuo p-dimensional. Se dice que* $\boldsymbol{X}$ *tiene una distribución t de Student multivariante si su función densidad de probabilidad está dada por,*

$$\frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\left[1+\frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\nu+p)/2},$$

donde el parámetro de ubicación $\boldsymbol{\mu}$ *es un vector de valores reales,* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$*, la matriz de escala* $\boldsymbol{\Sigma}$ *es una matriz real definida positiva de* $p \times p$ *y* $\nu > 0$ *es un número real, representa los grados de libertad*[4]*. Se denota como* $\boldsymbol{X} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*.*

Definición 1.4.12 *Sea* $\boldsymbol{\mu}$ *un vector aleatorio que sigue una distribución normal multivariada con media* $\boldsymbol{\mu}_0$ *y matriz de varianzas y covarianzas* $\frac{1}{\lambda}\boldsymbol{\Sigma}$*, donde* $\boldsymbol{\Sigma}$ *tiene una distribución Wishart inversa con parametros* $(\boldsymbol{\Psi}, \nu)$*.* $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *tienen una distribución conjunta normal Wishart inversa si su función de densidad conjunta está dada por*

$$\frac{\lambda^{D/2}|\boldsymbol{\Psi}|^{\nu/2}|\boldsymbol{\Sigma}|^{-\frac{\nu+D+2}{2}}}{(2\pi)^{D/2}2^{\frac{\nu D}{2}}\Gamma_D\left(\frac{\nu}{2}\right)}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right)-\frac{\lambda}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\right\}.$$

[4]Los grados de libertad se definen como el número de observaciones independientes menos el número de parámetros estimados o restricciones impuestas en el modelo.

Se denota como $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim NWI(\boldsymbol{\mu_0}, \lambda, \boldsymbol{\Psi}, \nu)$.

Capítulo 2

Estadística bayesiana

La estadística bayesiana es una metodología sustentada en un conjunto de axiomas, que permite realizar inferencias coherentes acerca de variables que presentan aleatoriedad. Está basada principalmente en la interpretación subjetiva de la probabilidad y en el teorema de Bayes (Nuñez-Antonio y Castillo, 2014). Su objetivo es brindar una herramienta para el correcto análisis de la información y con esto, poder tomar decisiones razonables sobre la mejor forma de actuar. A lo largo del siguiente capítulo, se explica cuáles son las interrogantes a las que da respuesta la inferencia bayesiana, así como los ingredientes necesarios para lograr esto y las distintas opciones que nos proporciona el enfoque bayesiano para modelar de una mejor manera nuestra información. Los conceptos de este capítulo se tomaron de Correa Morales y Barrera Causil (2018), Gelman *et al.* (1995), Gutiérrez y Zhang (2021), Nuñez-Antonio y Castillo (2014), Rangel (2015), Mendoza y Regueiro (2011).

2.1. Inferencia bayesiana

La inferencia bayesiana tiene como objetivo obtener información acerca de los valores de k cantidades desconocidas

$$\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k),$$

con k mayor o igual que uno. Para lograr esto se debe tener información *a priori*, la cual deberá estar expresada en términos de una función de densidad de probabilidad

$$p_{\boldsymbol{\theta}}(\boldsymbol{\theta}).$$

De la misma manera, se deberá contar con n observaciones

$$\boldsymbol{y} = (y_1, y_2, ..., y_n),$$

del vector aleatorio $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, cuya distribución de probabilidad tiene como parámetros a las k cantidades desconocidas. Las componentes de $\mathbf{Y}$ son variables aleatorias, luego la dependencia de $\mathbf{Y}$ puede ser expresada en términos de una función de densidad de probabilidad

$$p_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}).$$

Se busca una manera de expresar el estado de información acerca de $\boldsymbol{\theta}$ teniendo en cuenta la información *a priori* y la distribución de probabilidad de los datos observados. Para lograr esto, se utiliza el teorema de Bayes del cual se tiene,

$$p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p_{(\boldsymbol{\theta},\boldsymbol{Y})}(\boldsymbol{\theta},\boldsymbol{y})}{p_{\boldsymbol{Y}}(\boldsymbol{y})} = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\theta})p_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta})}{p_{\boldsymbol{Y}}(\boldsymbol{y})},$$

donde $p_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{\theta})p_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$. Nótese que $p_{\boldsymbol{Y}}(\boldsymbol{y})$ no depende de $\boldsymbol{\theta}$ cuando se fija a $\boldsymbol{y}$, por lo que puede ser considerado una constante respecto a $\boldsymbol{\theta}$. Esto produce que la ecuación anterior se escriba como,

$$p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y}) \propto p_{\boldsymbol{\theta}}(\boldsymbol{\theta})p_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}).$$

Nótese que cuando se ve a $p_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta})$ como función de $\boldsymbol{\theta}$, se obtiene la función de verosimilitud

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = p_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}),$$

a partir de la cual se obtiene la función *log-verosimilitud*

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = log(L(\boldsymbol{\theta}|\boldsymbol{y})).$$

Así, definiendo a $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ como la función de densidad de probabilidad *a priori* de $\boldsymbol{\theta}$ o distribución *a priori* y a $p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y})$ como la función de densidad de probabilidad *a posteriori* de $\boldsymbol{\theta}$ o distribución *a posteriori*, se puede pensar en el teorema de Bayes de una manera más sencilla de recordar

distribución *a posteriori* $\propto$ **distribución *a priori*** $\times$ **función de verosimilitud**.

La notación utilizada hasta aquí es general en el sentido de que no hace distinción entre variables aleatorias continuas o discretas. Sin embargo, de aquí en adelante se usará la notación respectiva para v.a continuas y v.a. discretas. Además se hará referencia a $\boldsymbol{\theta}$ como el vector de parámetros de interés.

2.1.1. Teoría de la decisión

En el ámbito bayesiano, la estimación del valor del parámetro $\boldsymbol{\theta}$ depende de las observaciones de $\boldsymbol{Y}$, en este sentido se puede pensar en la estimación del parámetro

como una función que depende de las observaciones ($A(\boldsymbol{Y})$). Dentro de la teoría de la decisión a esta función $A(\boldsymbol{Y})$ se le conoce como regla de decisión.
Ya que el parámetro posee cierta incertidumbre, el echo de proporcionar una estimación para este, trae consigo un grado de falla. Para formalizar esto, se define la función de pérdida L, la cual depende de la estimación y del parámetro estimado, $L(A, \boldsymbol{\theta})$. Dicha función de pérdida le asigna a cada pareja $(A, \boldsymbol{\theta})$ un número real positivo, el cual cuantifica la pérdida que se tiene de la estimación respecto al parámetro.
En general, se busca una estimación que minimice la pérdida. Sin embargo, esta acción no se puede realizar de manera directa ya que la función de pérdida depende del parámetro $\boldsymbol{\theta}$ el cual se puede pensar como una variable aleatoria. Por ello, la herramienta fundamental que se tiene para obtener información acerca del parámetro $\boldsymbol{\theta}$ es su función de densidad *a posteriori* $p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y})$. Esto da paso a que se defina el riesgo R como,

$$R = E(L) = \int L(A, \boldsymbol{\theta}) p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}.$$

Esto lleva a que se busque la estimación que minimice el riesgo R (Gutiérrez y Zhang, 2021).
Supóngase que se desea estimar un parámetro de interés $\boldsymbol{\theta}$ en el contexto de la estadística bayesiana y se denota la función de densidad *a posteriori* de $\boldsymbol{\theta}$ como $p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y})$, entonces si se utiliza la función de pérdida cuadrática [1], el riesgo asociado será

$$\begin{aligned} R = E(L(A, \boldsymbol{\theta})) =& E\left[(A - \boldsymbol{\theta})^2\right] = E\left[A^2 - 2A\boldsymbol{\theta} + \boldsymbol{\theta}^2\right] = A^2 - 2AE[\boldsymbol{\theta}] + E\left[\boldsymbol{\theta}^2\right] \\ &= A^2 - 2AE[\boldsymbol{\theta}] + E\left[\boldsymbol{\theta}^2\right] - (E[\boldsymbol{\theta}])^2 + (E[\boldsymbol{\theta}])^2 \\ &= A^2 - 2AE[\boldsymbol{\theta}] + (E[\boldsymbol{\theta}])^2 + \mathrm{Var}(\boldsymbol{\theta}) \\ &= (A - E[\boldsymbol{\theta}])^2 + \mathrm{Var}(\boldsymbol{\theta}), \end{aligned}$$

que se minimiza si $A = E(\boldsymbol{\theta})$. Es decir, la mejor acción para estimar $\boldsymbol{\theta}$ es utilizar su valor esperado tomado con respecto a la distribución posterior $p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y})$.

2.1.2. Estimación bayesiana

El enfoque bayesiano, además de especificar un modelo para los datos observados $\boldsymbol{y} = (y_1, \ldots, y_n)$ dado un vector de parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, usualmente en forma de densidad condicional $p_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{\theta})$, supone que $\boldsymbol{\theta}$ es una variable aleatoria y que tiene una función de densidad *a priori* $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, que depende de parámetros fijos, los cuales se conocen como hiperparámetros. De esta forma, la inferencia concerniente a $\boldsymbol{\theta}$ se basa en una función de densidad posterior o *a posteriori* $p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y})$ (Gutiérrez y Zhang, 2021).

La estimación bayesiana se puede separar en dos etapas:

[1] La función de pérdida cuadrática se define como $(A - \boldsymbol{\theta})^2$.

- Antes de la recolección de los datos, en donde el usuario propone, basado en su conocimiento, experiencia o fuentes externas, una distribución de probabilidad previa para el parámetro de interés. Con esta distribución es posible calcular estimaciones puntuales y por intervalos con el fin de confirmar que la distribución *a priori* propuesta se ajusta al problema de estudio.

- Después de la recolección de los datos. Siguiendo el teorema de Bayes, el usuario actualiza su conocimiento acerca del comportamiento probabilístico del parámetro de interés mediante la distribución *a posteriori* de éste. Con esta distribución es posible calcular estimaciones puntuales y por intervalos justo como en el enfoque frecuentista.

Inferencia *a posteriori*

Una vez recolectados los datos, se actualiza el estado de informacion acerca de $\boldsymbol{\theta}$. Se puede calcular la probabilidad *a posteriori* de que $\boldsymbol{\theta}$ esté en la región G dados los datos observados como

$$\mathrm{P}(\boldsymbol{\theta} \in G \mid \boldsymbol{y}) = \int_G p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}.$$

También es posible calcular la estimación puntual para el vector $\boldsymbol{\theta}$ dados los datos observados. La cual está dada por alguna medida de tendencia central para la distribución $p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y})$. En particular, si se toma una función de pérdida cuadrática, entonces se tiene que un buen estimador puntual es,

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int_\Theta \boldsymbol{\theta} p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}.$$

La región C de $100 \times (1-\alpha)$ de credibilidad es tal que

$$1 - \alpha \leq \Pr(\boldsymbol{\theta} \in C \mid \boldsymbol{y}) = \int_C p_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}.$$

2.2. Distribución *a priori*

La elección de la distribución *a priori* es un paso crucial dentro del análisis bayesiano, ya que esta afecta directamente a la distribución *a posteriori* tal y como se puede observar en el teorema de Bayes. La distribución *a priori* es la encargada de describir de manera adecuada el estado de información que se tiene de los parámetros de interés. Por ejemplo, si se cree que un parámetro toma valores cercanos a 74 entonces la distribución *a priori* debe de tomar valores cercanos a 74, como podría ser una distribución normal centrada

en ese valor. Por otro lado, también se debe considerar que al momento de realizar los cálculos de las distribuciones *a posteriori*, existen distribuciones *a priori* más sencillas de manejar. A continuación se presentan algunos tipos de distribuciones *a priori*.

2.2.1. Distribuciones *a priori* no informativas

Existen situaciones en las que no se quiere introducir información en la distribución *a priori*, porque

- No se sabe nada sobre el comportamiento de los parámetros de interés.
- Se quiere ser objetivo.

En estas situaciones se tienen que elegir distribuciones *a priori* no informativas (Correa Morales y Barrera Causil, 2018). El objetivo de las distribuciones *a priori* no informativas es que la distribución *a priori* tenga un impacto mínimo sobre la distribución *a posteriori* de los parámetros de interés. Esto se traduce a que la distribución *a priori* sea relativamente plana respecto a la función de verosimilitud. De esta manera, los datos tendrán una mayor influencia sobre la distribución *a posteriori*, y por tanto, en todas las inferencias que de ella se obtengan.

Estas distribuciones no informativas se clasifican en dos grupos:

- **Propias:** cuando la distribución de probabilidad integra a una constante finita, se dice que es propia.
- **Impropias:** una distribución $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ se dice impropia si

$$\int_{\theta} p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \infty.$$

Nótese, que el uso de distribuciones *a priori* impropias puede conducir a distribuciones *a posteriori* impropias. En ese caso se imposibilita la opción de hacer inferencia.

2.2.2. El principio de razón insuficiente

Este principio, que es muy intuitivo, establece que si no hay información para diferenciar entre valores diferentes de $\boldsymbol{\theta}$, se debe dar la misma probabilidad a todos los valores.

Esto implica considerar una distribución *a priori* uniforme para $\boldsymbol{\theta}$. Si el soporte de $\boldsymbol{\theta}$ es infinito, la distribución *a priori* impropia, resulta ser,

$$p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto 1.$$

Ejemplo 2.2.1 *Suponga que se tienen* $\boldsymbol{y} = (y_1, \ldots, y_n)$ *observaciones de* $\boldsymbol{Y} = (Y_1, ..., Y_n)$*, donde cada* Y_i *es una v.a. independiente y que sigue una distribución normal con media* θ *y varianza conocida* σ^2*, es decir,* $Y_i \overset{iid}{\sim} N(\theta, \sigma^2)$*. Supóngase que* $f_\theta(\theta) \propto 1$*, que es la distribución a priori uniforme (impropia) sobre los números reales. Ya que las observaciones son independientes, se tiene que la función de densidad de probabilidad conjunta está dada por:*

$$\begin{aligned} f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta) &= f_{Y_1|\theta}(y_1|\theta) \times \cdots \times f_{Y_n|\theta}(y_n|\theta) \\ &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\}. \end{aligned}$$

Usando el teorema de Bayes se tiene que la distribución a posteriori de θ *es de la forma:*

$$\begin{aligned} f_{\theta|\boldsymbol{Y}}(\theta \mid \boldsymbol{y}) &\propto f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid \theta)\, f_\theta(\theta) \\ &\propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\} \cdot 1. \\ &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \theta}{\sigma}\right)^2\right\}. \end{aligned}$$

Nótese que

$$\sum_{i=1}^{n}(y_i - \theta)^2 = \sum_{i=1}^{n} y_i^2 - 2\theta n\bar{y} + n\theta^2$$

y además, ya que

$$S^2 = \frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2),$$

resulta que

$$\sum_{i=1}^{n}(y_i - \theta)^2 = S^2(n-1) + n\bar{y}^2 - 2\theta n\bar{y} + n\theta^2 = S^2(n-1) + n(\bar{y} - \theta)^2. \tag{2.1}$$

De aquí

$$f_{\theta|\boldsymbol{Y}}(\theta \mid \boldsymbol{y}) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(S^2(n-1)+n(\bar{y}-\theta)^2\right)\right\}$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(n(\bar{y}-\theta)^2\right)\right\}$$

y, por lo tanto la distribución a posteriori es:

$$\theta|\boldsymbol{Y}=\boldsymbol{y} \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right),$$

la cual es una distribución propia. De aquí, la media a posteriori de θ es $\bar{y}$ y su varianza a posteriori es $\frac{\sigma^2}{n}$.

2.2.3. *A priori* no informativa de Jeffreys

La distribución *a priori* de Jeffreys satisface la propiedad local de uniformidad para distribuciones *a priori* no informativas, es decir, es uniforme sobre el espacio muestral de los parámetros. Esta *a priori* está basada en la matriz de información de Fisher.

Definición 2.2.1 *Sea $f_{X|\theta}(x \mid \theta)$ la función de densidad de X dado θ. La información de Fisher, denotada por $\mathcal{I}(\theta)$, se define como,*

$$\mathcal{I}(\theta) = E_X\left[\left(\frac{\partial}{\partial\theta}\log f_{X|\theta}(x \mid \theta)\right)^2\right] = -E_X\left[\frac{\partial^2}{\partial\theta^2}\log f_{X|\theta}(x \mid \theta)\right].$$

Si $\boldsymbol{\theta}$ es un vector de p componentes, entonces,

$$\mathcal{I}(\boldsymbol{\theta}) = -E_X\left[\frac{\partial^2 \log(f(x \mid \boldsymbol{\theta}))}{\partial\theta_i \partial\theta_j}\right]_{p\times p},$$

con $i, j = 1, ..., p$, y entonces $\mathcal{I}(\boldsymbol{\theta})$ será una matriz de dimensión $p \times p$.

Definición 2.2.2 *La distribución a priori de Jeffreys para el parámetro $\boldsymbol{\theta}$, se define como:*

$$p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto |\mathcal{I}(\boldsymbol{\theta})|^{1/2}.$$

La distribución *a priori* de Jeffreys es localmente uniforme y por lo tanto no informativa, esta propiedad es útil ya que proporciona un esquema automatizado para

hallar distribuciones *a priori* no informativas para cualquier modelo paramétrico, además es impropia para muchos modelos.

Ejemplo 2.2.2 *Supóngase que* $\boldsymbol{y} = (y_1, ..., y_n)$ *son observaciones de* $\boldsymbol{Y} = (Y_1, ..., Y_n)$, *donde cada* Y_i *es una variable aleatoria distribuida normal e independientemente con media* θ *y varianza conocida* σ^2, *esto es,* $Y_i \overset{iid}{\sim} N(\theta, \sigma^2)$.

La función de densidad de probabilidad conjunta está dada por

$$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid \theta) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \theta\right)^2\right\},$$

tomando logarítmo natural, resulta

$$\ln\left(f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid \theta)\right) = \ln\left(\left(2\pi\sigma^2\right)^{-\frac{n}{2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \theta\right)^2.$$

Luego

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln\left(f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid \theta)\right) &= \frac{-1}{2\sigma^2}\sum_{i=1}^{n} -2\left(y_i - \theta\right) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(-2y_i + 2\theta\right)\end{aligned}$$

y

$$\frac{\partial^2}{\partial\theta^2}\ln\left(f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid \theta)\right) = -\frac{n}{\sigma^2} \Rightarrow E_{\boldsymbol{Y}}\left[\frac{n}{\sigma^2}\right] = -\frac{n}{\sigma^2}.$$

Así, por definición de a priori de Jeffreys se tiene que,

$$f_\theta(\theta) \propto \sqrt{\frac{n}{\sigma^2}} \propto 1.$$

Que coincide con la distribución a priori del Ejemplo (2.2.1), por lo que

$$\theta|\mathbf{Y} = \boldsymbol{y} \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right).$$

2.2.4. Distribuciones *a priori* conjugadas

Cuando se tiene el problema de hacer inferencia acerca de un conjunto de parámetros $\boldsymbol{\theta}$, la distribución *a priori* juega un papel importante sobre las conclusiones a obtenerse. En ese sentido, el enfoque bayesiano ofrece la posibilidad de que dicha distribución *a priori* no sea única y que algunas elecciones resulten más ventajosas que otras al momento de hacer los cálculos involucrados y de expresar la información respecto a $\boldsymbol{\theta}$. Esta idea es la que da paso a la siguiente definición.

Definición 2.2.3 *Sea Y una v.a con función de densidad de probabilidad $f_{Y|\boldsymbol{\theta}}(y \mid \boldsymbol{\theta})$, entonces la familia de distribuciones $\mathcal{F} = \{p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\}$ se dice que es cerrada o equivalentemente conjugada si cuando la distribución a priori para $\boldsymbol{\theta}$ pertenece a $\mathcal{F}$, la distribución a posteriori correspondiente a cualquier muestra aleatoria de Y también pertenece a $\mathcal{F}$ (Mendoza y Regueiro, 2011).*

La conjugación proporciona la selección de una clase de distribuciones *a priori* limitada, si la clase es lo suficientemente grande esto puede no ser un gran problema. En particular cuando $f_{\boldsymbol{Y}|\boldsymbol{\theta}}(y \mid \boldsymbol{\theta})$ resulta ser una distribución normal, la familia exponencial es conjugada para esta (Correa Morales y Barrera Causil, 2018).

2.2.5. Otras alternativas para las distribuciones a priori

Dos alternativas para las distribuciones *a priori* son la distribución *a priori* de máxima entropía y la distribución *a priori* indiferente. A continuación se presenta la definición de distribución *a priori* de máxima entropía.

Definición 2.2.4 ***Distribución a priori de máxima entropía.*** *Cuando θ es univariable y puede tomar cualquier valor sobre la recta real, y la media y la varianza a priori están especificadas, la distribución a priori de máxima entropía es la Normal con la media y la varianza especificadas.*

En caso de soporte finito la distribución de máxima entropía es la uniforme, sin embargo, cuando existe alguna información previa, por ejemplo una media, entonces se construye esta *a priori* teniendo en cuenta esta restricción (Correa Morales y Barrera Causil, 2018). Considere el siguiente ejemplo, donde el soporte es finito, digamos $\theta_1, \ldots, \theta_k$, con $E[\theta_i] = C$, entonces se debe maximizar

$$H = -\sum_{i=1}^{k} p_\theta\left(\theta_i\right) \log\left(p_\theta\left(\theta_i\right)\right),$$

sujeto a la restricción,

$$\sum_{i=1}^{k} \theta_i p_\theta(\theta_i) = C,$$

y que,

$$\sum_{i=1}^{k} p_\theta(\theta_i) = 1.$$

Los $p_\theta(\theta_i)$ que resuelvan este problema forman la distribución de máxima entropía.

Por otro lado, se tiene la distribución *a priori* indiferente, la cual se puede pensar como un caso especial de la familia conjugada de una distribución. A continuación se enuncia la definición formal.

Definición 2.2.5 ***Distribución a priori indiferente.*** *Se define que una distribución a priori es indiferente si identificando una clase de distribuciones conjugadas se selecciona una distribución a priori de esta clase que satisfaga:*

- *La distribución a priori debe ser impropia.*
- *Una «muestra mínima necesaria» debe inducir una distribución a posteriori propia.*

2.3. Modelo normal con media desconocida y varianza conocida

Considere datos que pueden modelarse adecuadamente por medio de una distribución normal, en este primer caso se asume que la varianza, σ^2, es conocida y el objetivo es estimar la media, θ. Suponga que $\boldsymbol{Y}=(Y_1, ..., Y_n)$ es un vector aleatorio en donde cada Y_i es una v.a. que sigue una distribución normal con media θ y varianza σ^2 y son independientes entre si. Esto es $Y_i \overset{iid}{\sim} N(\theta, \sigma^2)$. Además suponga que la distribución *a priori* para θ es una distribución normal con media μ_0 y varianza σ_0^2, es decir,

$$\theta \sim N\left(\mu_0, \sigma_0^2\right),$$

esto significa que la f.d.p de θ es,

$$f_\theta(\theta) = (2\pi\sigma_0^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\theta - \mu_0)^2\right\}$$

$$\propto \left(\sigma_0^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right\}.$$

Por otro lado, ya que las observaciones son independientes entre sí, la función de densidad

de probabilidad conjunta está dada por el producto de las distribuciones marginales.

$$f_{\boldsymbol{Y}|\theta}\left(\boldsymbol{y}|\theta\right) = f_{Y_1|\theta}\left(y_1|\theta\right) \cdot ... \cdot f_{Y_n|\theta}\left(y_n|\theta\right)$$

$$= \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\theta)^2\right\}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}.$$

Haciendo uso del teorema de Bayes la distribución *a posteriori* de θ es,

$$f_{\theta|\boldsymbol{Y}}\left(\theta \mid \boldsymbol{y}\right) = \frac{f_{(\boldsymbol{Y},\theta)}\left(\boldsymbol{y},\theta\right)}{f_{\boldsymbol{Y}}\left(\boldsymbol{y}\right)} = \frac{f_{\boldsymbol{Y}|\boldsymbol{\theta}}\left(\boldsymbol{y} \mid \theta\right) f_{\theta}(\theta)}{f_{\boldsymbol{Y}}\left(\boldsymbol{y}\right)} \propto f_{\boldsymbol{Y}|\theta}\left(\boldsymbol{y} \mid \theta\right) f_{\theta}(\theta)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} \cdot \left(\sigma_0^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right\}.$$

$$\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\theta}{\sigma}\right)^2\right\} \cdot \exp\left\{-\frac{1}{2}\left(\frac{\theta-\mu_0}{\sigma_0}\right)^2\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\frac{1}{\sigma^2}\left(y_i^2 - 2y_i\theta + \theta^2\right) + \frac{1}{\sigma_0^2}\left(\theta^2 - 2\theta\mu_0 + \mu_0^2\right)\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\frac{1}{\sigma^2}\left(-2y_i\theta + \theta^2\right) + \frac{1}{\sigma_0^2}\left(\theta^2 - 2\theta\mu_0\right)\right]\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\left(-2\theta\sum_{i=1}^{n}y_i + \theta^2\sum_{i=1}^{n}1\right) + \frac{\theta^2}{\sigma_0^2} - \frac{2\theta\mu_0}{\sigma_0^2}\right]\right\}$$

$$= \exp\left\{-\frac{1}{2}\left(\theta^2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right) - 2\theta\left(\frac{\sum_{i=1}^{n}y_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\right)\right\}$$

$$= \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\theta^2 - 2\theta\left(\frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)\right)\right\},$$

el cual es el kernel de una distribución normal con parámetros

$$\frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} \quad \text{y} \quad \frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}.$$

Por lo tanto

$$\theta \mid \boldsymbol{Y} = \boldsymbol{y} \sim N\left(\left(\frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right), \frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}\right).$$

Así, θ tiene una media *a posteriori*

$$\frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

y una varianza

$$\frac{1}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}.$$

De aquí, una estimación puntual para θ es su media *a posteriori*.

Observando la forma de la media *a posteriori*, se observa que este es una combinación convexa entre el estimador clásico de máxima verosimilitud $\hat{\theta}_C = \bar{y}$ y el estimador previo $\hat{\theta}_P = \mu_0$, puesto que:

$$\begin{aligned}
\hat{\theta}_B =& \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \\
=& \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\bar{y} + \frac{\frac{1}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\mu_0 \\
=& \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\hat{\theta}_C + \frac{\frac{1}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\hat{\theta}_P.
\end{aligned}$$

De donde se puede concluir que, para una distribución fija, entre mayor sea el tamaño muestral n, más peso tendrá el estimador clásico $\hat{\theta}_C$ en la estimación bayesiana. De la misma forma, para un conjunto fijo de datos $\boldsymbol{y}$, entre menor sea la varianza previa, σ_0^2, más certeza se tiene sobre la información previa y por consiguiente la estimación bayesiana $\hat{\theta}_B$ se acercará más a la estimación.
En la Figura 2.1 se muestra el comportamiento de la distribución *a priori*, la función de verosimilitud y la distribución *a posteriori* fijando los parámetros $\mu_0 = 5, \sigma_0^2 = 0.01, \bar{y} = 2, \sigma^2 = 1$ y variando a n con $n = 5, 10, 50, 200$. Se puede observar que a medida que el tamaño muestral n aumente, la función de verosimilitud (vista como la función del parámetro θ) se vuelve más concentrada alrededor del valor de $\bar{y}$, y en consecuencia, la función de densidad posterior de θ se sitúa más cercana a la función de verosimilitud, y la estimación bayesiana se acerca más a la estimación clásica $\bar{y}$ (Gutiérrez y Zhang, 2021).

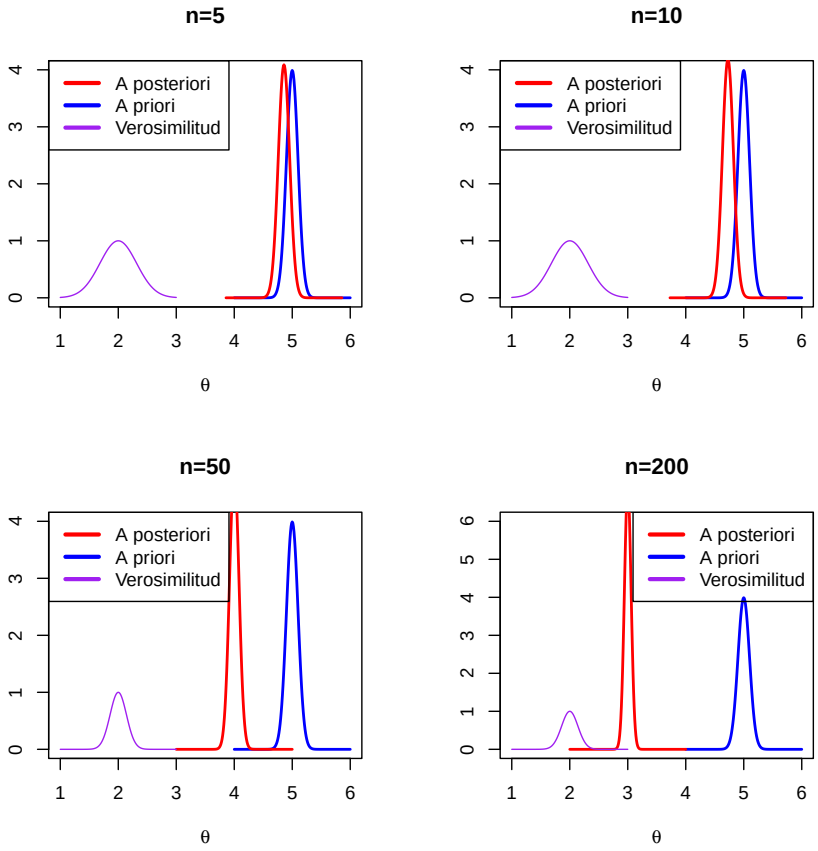

Figura 2.1: Modelo normal con media desconocida y varianza conocida.

2.4. Modelo normal con media y varianza desconocidas

Considere datos que pueden modelarse adecuadamente por medio de una distribución normal, en este caso se asume que la media y la varianza son desconocidas y el objetivo es estimar a ambos parámetros. Suponga que $\boldsymbol{Y} = (Y_1, ..., Y_n)$ es un vector aleatorio en donde cada Y_i es una v.a que sigue una distribución normal con media θ, varianza σ^2 y son independientes entre sí. Esto es $Y_i \overset{iid}{\sim} N(\theta, \sigma^2)$. Además suponga que la distribución *a priori* conjunta para (θ, σ^2) está dada por,

$$f_{\theta,\sigma^2}(\theta, \sigma^2) = f_{\theta|\sigma^2}(\theta|\sigma^2) f_{\sigma^2}(\sigma^2), \tag{2.2}$$

donde

$$\theta|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \qquad \text{y} \qquad \sigma^2 \sim GI\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right),$$

con $\kappa_0 \in \mathbb{R}\backslash\{0\}$.

La distribución muestral conjunta está dada por el producto de las distribuciones marginales, esto es,

$$\begin{aligned} f_{\boldsymbol{Y}|\theta,\sigma^2}\left(\boldsymbol{y}|\theta,\sigma^2\right) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(y_i-\theta)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}. \end{aligned}$$

Usando la ecuación (2.1),

$$\begin{aligned} f_{\boldsymbol{Y}|\theta,\sigma^2}(\boldsymbol{y}|\theta,\sigma^2) &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left(S^2(n-1)+n(\bar{y}-\theta)^2\right)\right\} \\ &\propto \left(\sigma^2\right)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left(S^2(n-1)+n(\bar{y}-\theta)^2\right)\right\}. \end{aligned}$$

Por otra parte, por el teorema de Bayes,

$$f_{\theta,\sigma^2|\boldsymbol{Y}}\left(\theta,\sigma^2|\boldsymbol{y}\right) \propto f_{\boldsymbol{Y}|\theta,\sigma^2}\left(\boldsymbol{y}|\theta,\sigma^2\right) f_{\theta,\sigma^2}\left(\theta,\sigma^2\right).$$

Sustituyendo la distribución *a priori* conjunta dada en la ecuación (2.2), se tiene

$$f_{\theta,\sigma^2|\boldsymbol{Y}}(\theta,\sigma^2|\boldsymbol{y}) \propto f_{\boldsymbol{Y}|\theta,\sigma^2}\left(\boldsymbol{y}|\theta,\sigma^2\right) f_{\theta|\sigma^2}\left(\theta|\sigma^2\right) f_{\sigma^2}\left(\sigma^2\right).$$

Las tres distribuciones de las que depende la distribución *a posteriori* son conocidas. Por ello, para encontrar la distribución *a posteriori* basta realizar el producto de las tres distribuciones que aparecen en la ecuación anterior. Para realizar esto, se comenzará analizando el producto $f_{\theta|\sigma^2}\left(\theta|\sigma^2\right) f_{\sigma^2}\left(\sigma^2\right)$. Nótese primero que

$$\theta|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right),$$

esto significa que su función de densidad de probabilidad está dada por

$$f_{\theta|\sigma^2}\left(\theta \mid \sigma^2\right) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{\kappa_0}{2\sigma^2}(\theta-\mu_0)^2\right\}$$

$$\propto \left(\sigma^2\right)^{-1/2} \exp\left\{-\frac{\kappa_0}{2\sigma^2}(\theta-\mu_0)^2\right\}.$$

De la misma manera, se tiene

$$\sigma^2 \sim GI\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right),$$

esto significa que su función de densidad de probabilidad está dada por

$$f_{\sigma^2}(\sigma^2) = \frac{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\nu_0/2}}{\Gamma\left(\frac{\nu_0}{2}\right)} (\sigma^2)^{-\frac{\nu_0}{2}-1} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}$$

$$\propto (\sigma^2)^{-\frac{\nu_0}{2}-1} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}.$$

Por lo que

$$f_{\theta|\sigma^2}\left(\theta|\sigma^2\right) f_{\sigma^2}\left(\sigma^2\right) \propto \left(\sigma^2\right)^{-1/2} \exp\left\{-\frac{\kappa_0}{2\sigma^2}(\theta-\mu_0)^2\right\} \cdot \left(\sigma^2\right)^{-\frac{\nu_0}{2}-1} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}$$

$$= \left(\sigma^2\right)^{-\frac{\nu_0+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[\kappa_0(\theta-\mu_0)^2 + \nu_0\sigma_0^2\right]\right\}.$$

Lo cual es el kernel de una distribución normal gamma inversa con parámetros $\left(\frac{\nu_0+1}{2}, \kappa_0, \mu_0, \nu_0\sigma_0^2\right)$.

Luego la distribución *a posteriori* conjunta de (θ, σ^2) está dada por :

$$\begin{aligned} f_{\theta,\sigma^2|\boldsymbol{Y}}\left(\theta,\sigma^2|\boldsymbol{y}\right) &\propto \left(\sigma^2\right)^{-\frac{\nu_0+1}{2}-1}\exp\left\{-\frac{1}{2\sigma^2}(\kappa_0(\theta-\mu_0)^2+\nu_0\sigma_0^2)\right\} \\ &\quad\cdot\left(\sigma^2\right)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}(S^2(n-1)+n(\bar{y}-\theta)^2)\right\} \\ &= \left(\sigma^2\right)^{-\frac{\nu_0+1+n}{2}-1}\exp\left\{-\frac{1}{2\sigma^2}\left[\kappa_0(\theta-\mu_0)^2+\nu_0\sigma_0^2+S^2(n-1)+n(\bar{y}-\theta)^2\right]\right\}. \end{aligned} \tag{2.3}$$

En seguida, se desarrollarán las operaciones que se encuentran dentro de los corchetes.

$$\begin{aligned} &\kappa_0\left(\theta-\mu_0\right)^2+\nu_0\sigma_0^2+S^2(n-1)+n(\bar{y}-\theta)^2 \\ &= \kappa_0\left(\theta^2-2\theta\mu_0+\mu_0^2\right)+n\left(\bar{y}^2-2\bar{y}\theta+\theta^2\right)+\nu_0\sigma_0^2+S^2(n-1) \\ &= \theta^2\left(\kappa_0+n\right)-2\theta\left(\kappa_0\mu_0+n\bar{y}\right)+\kappa_0\mu_0^2+n\bar{y}^2+\nu_0\sigma_0^2+S^2(n-1) \\ &= \left(\kappa_0+n\right)\left[\theta^2-2\theta\frac{\left(\kappa_0\mu_0+n\bar{y}\right)}{\left(\kappa_0+n\right)}\right]+\kappa_0{\mu_0}^2+n\bar{y}^2+\nu_0{\sigma_0}^2+S^2(n-1) \\ &= \left(\kappa_0+n\right)\left[\theta^2-2\theta\frac{\left(\kappa_0\mu_0+n\bar{y}\right)}{\left(\kappa_0+n\right)}+\left(\frac{\left(\kappa_0\mu_0+n\bar{y}\right)}{\left(\kappa_0+n\right)}\right)^2-\left(\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}\right)^2\right] \\ &+\kappa_0\mu_0^2+n\bar{y}^2+\nu_0\sigma_0^2+S^2(n-1) \\ &= \left(\kappa_0+n\right)\left[\left(\theta-\left(\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}\right)\right)^2-\left(\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}\right)^2\right]+\kappa_0\mu_0^2+n\bar{y}^2+\nu_0\sigma_0^2 \\ &+S^2(n-1) \\ &= \left(\kappa_0+n\right)\left[\left(\theta-\left(\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}\right)\right)^2\right]-\left(\kappa_0+n\right)\left(\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}\right)^2+\kappa_0\mu_0^2+n\bar{y}^2+\nu_0\sigma_0^2 \\ &+S^2(n-1). \end{aligned}$$

Sustituyendo en la ecuación (2.3) se obtiene,

$$\begin{aligned} &f_{\theta,\sigma^2|\boldsymbol{Y}}(\theta,\sigma^2|\boldsymbol{y}) \\ &\propto \left(\sigma^2\right)^{-\frac{\nu_0+1+n}{2}-1}\exp\left\{-\frac{1}{2\sigma^2}\left[\kappa_0(\theta-\mu_0)^2+\nu_0\sigma_0^2+S^2(n-1)+n(\bar{y}-\theta)^2\right]\right\} \\ &= \left(\sigma^2\right)^{-\frac{\nu_0+1+n}{2}-1}\exp\left\{-\frac{1}{2\sigma^2}\left[\left(\kappa_0+n\right)\left(\theta-\left(\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}\right)\right)^2-\left(\kappa_0+n\right)\left(\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}\right)^2\right.\right. \\ &\left.\left.+\kappa_0\mu_0^2+n\bar{y}^2+\nu_0\sigma_0^2+S^2(n-1)\right]\right\}, \end{aligned}$$

se tiene que θ y σ^2 siguen una distribución conjunta *a posteriori* normal gamma inversa

con parámetros $(\nu_1, \kappa_1, \mu_1, \sigma_1^2)$ es decir,

$$f_{\theta,\sigma^2|\boldsymbol{Y}}\left(\theta, \sigma^2 \mid \boldsymbol{y}\right) = \left(\sigma^2\right)^{-\nu_1-1} \exp\left\{-\frac{1}{2\sigma^2}\left(\kappa_1\left(\theta-\mu_1\right)^2 + \kappa_1\sigma_1^2\right)\right\},$$

donde

$$\begin{aligned}
\nu_1 &= \frac{\nu_0 + n + 1}{2} \\
\kappa_1 &= \kappa_0 + n \\
\mu_1 &= \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n} \\
\sigma_1^2 &= \frac{-\left(\kappa_0 + n\right)\left(\frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0+n}\right)^2 + \kappa_0\mu_0^2 + n\bar{y}^2 + \nu_0\sigma_0^2 + S^2(n-1)}{\kappa_0 + n}.
\end{aligned}$$

Para realizar la inferencia *a posteriori* de cada parámetro se puede proceder de la siguiente forma:

1. Para hallar la distribución *a posteriori* condicional de θ, dada por $f_{\theta|\sigma^2,\mathbf{Y}}\left(\theta \mid \sigma^2, \mathbf{y}\right)$, se debe considerar a σ^2 como constante fija y conocida, tal como en la sección (2.3). Basado en lo anterior, es posible utilizar la siguiente regla de probabilidad

$$f_{\theta|\sigma^2,\mathbf{Y}}\left(\theta \mid \sigma^2, \boldsymbol{y}\right) = \frac{f_{\theta,\sigma^2|\mathbf{Y}}\left(\theta, \sigma^2 \mid \boldsymbol{y}\right)}{f_{(\sigma^2,\mathbf{Y})}\left(\sigma^2, \boldsymbol{y}\right)} f_{\boldsymbol{Y}}(\boldsymbol{y}) \propto f_{\theta,\sigma^2|\mathbf{Y}}\left(\theta, \sigma^2 \mid \boldsymbol{y}\right). \tag{2.4}$$

La ecuación (2.4) sugiere que la distribución condicional *a posteriori* de $\theta, f_{\theta|\sigma^2,\mathbf{Y}}\left(\theta \mid \sigma^2, \boldsymbol{y}\right)$, se encuentra utilizando la distribución *a posteriori* conjunta, $f_{\theta,\sigma^2|\mathbf{Y}}\left(\theta, \sigma^2 \mid \boldsymbol{y}\right)$, suponiendo que todas las expresiones que involucren al valor σ^2 se pueden incluir en la constante de proporcionalidad (Gutiérrez y Zhang, 2021).

Aplicando este razonamiento se puede concluir que la distribución posterior de θ condicional a $(\sigma^2, \mathbf{Y})$ está dada por

$$\theta \mid \sigma^2, \mathbf{Y} = \boldsymbol{y} \sim \text{Normal}\left(\mu_1, \sigma^2/\left(n + \kappa_0\right)\right),$$

con $\mu_1 = \frac{n\bar{y} + \kappa_0\mu_0}{n+\kappa_0}$.

2. Dado que σ^2 no depende de ningún otro parámetro entonces, utilizando la distribución *a posteriori* conjunta, es posible encontrar su distribución marginal *a posteriori* de la siguiente forma

$$f_{(\sigma^2|\mathbf{Y})}\left(\sigma^2 \mid \boldsymbol{y}\right) = \int f_{(\theta,\sigma^2|\mathbf{Y})}\left(\theta, \sigma^2 \mid \boldsymbol{y}\right) d\theta,$$

además, nótese que

$$\frac{\kappa_0 n}{\kappa_0+n}\left(\mu_0-\bar{y}\right)^2=-\left(\kappa_0 n\right)\left(\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}\right)^2+\kappa_0\mu_0^2+n\bar{y}^2.$$

De aquí,

$$\begin{aligned}
f_{\sigma^2|\mathbf{Y}}\left(\sigma^2\mid\mathbf{Y}\right)&=\int f_{\theta,\sigma^2|\mathbf{Y}}\left(\theta,\sigma^2\mid\mathbf{Y}\right)d\theta\\
&\propto\left(\sigma^2\right)^{-\frac{\nu_0+n+1}{2}-1}\exp\left\{-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2+(n-1)S^2+\frac{\kappa_0 n}{\kappa_0+n}(\mu_0-\bar{y})^2\right]\right\}\\
&\times\int_{-\infty}^{\infty}\exp\left\{-\frac{n+\kappa_0}{2\sigma^2}\left(\theta-\mu_1\right)^2\right\}d\theta\\
&\propto\left(\sigma^2\right)^{-\frac{\nu_0+n}{2}-1}\exp\left\{-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2+(n-1)S^2+\frac{\kappa_0 n}{\kappa_0+n}(\mu_0-\bar{y})^2\right]\right\}\\
&\times\int_{-\infty}^{\infty}\frac{\sqrt{n+\kappa_0}}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{n+\kappa_0}{2\sigma^2}\left(\theta-\mu_1\right)^2\right\}d\theta\\
&\propto\left(\sigma^2\right)^{-\frac{\nu_0+n}{2}-1}\exp\left\{-\frac{\left(n+\nu_0\right)\sigma_1^2}{2\sigma^2}\right\}.
\end{aligned}$$

Donde $\left(n+\nu_0\right)\sigma_1^2=\nu_0\sigma_0^2+(n-1)S^2+\frac{\kappa_0 n}{\kappa_0+n}(\mu_0-\bar{y})^2$. La expresión anterior corresponde a la función de densidad de una distribución gamma inversa con parámetros $\left(\frac{n+n_0}{2},\frac{(n+\nu_0)\sigma_1^2}{2}\right)$.

3. Un razonamiento similar se puede formular para el parámetro θ; utilizando la distribución *a posteriori* conjunta, es posible encontrar su distribución marginal *a posteriori* de la siguiente forma

$$f_{\theta|\mathbf{Y}}(\theta\mid\boldsymbol{y})=\int f_{\theta,\sigma^2|\mathbf{Y}}\left(\theta,\sigma^2\mid\boldsymbol{y}\right)d\sigma^2.$$

Note que:

$$\begin{aligned}
f_{\theta|\mathbf{Y}}(\theta\mid\boldsymbol{y})&=\int_0^{\infty}f_{\theta,\sigma^2|\mathbf{Y}}\left(\theta,\sigma^2\mid\boldsymbol{y}\right)d\sigma^2\\
&\propto\int_0^{\infty}\left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0+n+1}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}\left[\left(\nu_0+n\right)\sigma_1^2+\left(\kappa_0+n\right)\left(\theta-\mu_1\right)^2\right]\right\}d\sigma^2.
\end{aligned}$$

Haciendo un cambio de variable tal que

$$z=\frac{A}{2\sigma^2},\quad\text{donde}\quad A=\left(\nu_0+n\right)\sigma_1^2+\left(\kappa_0+n\right)\left(\theta-\mu_1\right)^2,$$

por tanto

$$d\sigma^2 = -\frac{A}{2z^2}dz.$$

Así

$$\begin{aligned}
f_{\theta|\mathbf{Y}}(\theta \mid \boldsymbol{y}) &\propto \left(\frac{1}{A}\right)^{\frac{\nu_0+n+1}{2}+1} \int_{\infty}^{0} \frac{-A}{2z^2}(2z)^{\frac{\nu_0+n+1}{2}+1} e^{-z} dz \\
&\propto A^{-\frac{\nu_0+n+1}{2}} \underbrace{\int_0^{\infty} z^{\frac{\nu_0+n+1}{2}-1}}_{\text{Gamma}\left(\frac{\nu_0+n+1}{2},1\right)} e^{-z} dz \\
&\propto A^{-\frac{\nu_0+n+1}{2}} \\
&= \left[(\nu_0+n)\,\sigma_1^2 + (\kappa_0+n)\,(\theta-\mu_1)^2\right]^{-\frac{\nu_0+n+1}{2}} \\
&\propto \left[1 + \frac{(\kappa_0+n)\,(\theta-\mu_1)^2}{(\nu_0+n)\,\sigma_1^2}\right]^{-\frac{\nu_0+n+1}{2}} \\
&= \left[1 + \frac{1}{\nu_0+n}\left(\frac{\theta-\mu_1}{\sigma_1/\sqrt{\kappa_0+n}}\right)^2\right]^{-\frac{\nu_0+n+1}{2}},
\end{aligned}$$

que corresponde al kernel de una distribución t de Student con $\nu_0 + n$ grados de libertad, parámetro de localización μ_1 y parámetro de escala $\frac{\sigma_1}{\sqrt{\kappa_0+n}}$, esto es,

$$\theta \mid \mathbf{Y} = \boldsymbol{y} \sim t_{n+\nu_0}\left(\mu_1, \frac{\sigma_1^2}{\kappa_0+n}\right).$$

Las distribuciones encontradas permiten estimar directamente los parámetros θ y σ^2 usando las esperanzas teóricas de las distribuciones marginales *a posteriori*. Por ende, las estimaciones puntuales son:

$$\hat{\theta} = \mu_1 = \frac{n\bar{y} + \kappa_0\mu_0}{n+\kappa_0}$$

$$\hat{\sigma}^2 = \frac{(n+\nu_0)\,\sigma_1^2/2}{[(n+\nu_0)\,/2] - 1} = \frac{(n+\nu_0)\,\sigma_1^2}{n+\nu_0-2} \approx \sigma_1^2 = \frac{\nu_0\sigma_0^2 + (n-1)S^2 + \frac{\kappa_0 n}{\kappa_0+n}(\mu_0-\bar{y})^2}{n+\nu_0}.$$

Los intervalos del $(1-\alpha)\times 100\,\%$ de credibilidad para θ y σ^2 se construyen usando los percentiles $\alpha/2$ y $1-\alpha/2$ de las respectivas distribuciones posteriores dadas en los resultados.

2.5. Modelo normal multivariado conjugado

Considere datos que pueden modelarse de manera adecuada por medio de una distribución normal multivariada, en este caso se asume que el vector de medias $\boldsymbol{\theta}$ y la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ son desconocidas y el objetivo es estimar sus valores. Suponga que $\boldsymbol{Y} = (\boldsymbol{Y_1}, ..., \boldsymbol{Y_n})$ está conformado por n vectores aleatorios de dimensión m cada uno, donde cada $\boldsymbol{Y_i}$ sigue una distribución normal multivariada con vector de medias $\boldsymbol{\theta}$, matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ y son independientes entre sí. Es decir $\boldsymbol{Y_i} \overset{iid}{\sim} N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Además suponga que la distribución *a priori* conjunta de $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ está dada por

$$f_{\boldsymbol{\theta},\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}\right) = f_{\boldsymbol{\theta}|\boldsymbol{\Sigma}}\left(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}\right) f_{\boldsymbol{\Sigma}}\left(\boldsymbol{\Sigma}\right),$$

donde

$$\boldsymbol{\theta} \mid \boldsymbol{\Sigma} \sim N\left(\theta_0, \frac{\boldsymbol{\Sigma}}{\lambda}\right) \quad \text{y} \quad \boldsymbol{\Sigma} \sim WI\left(\tau_0, \gamma_0\right),$$

con $\lambda \in \mathbb{R} \backslash \{0\}$. De aquí se tiene que:

$$f_{\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma}}\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}\right) = (2\pi)^{-m \cdot n/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{\sum_{i=1}^{n} -\frac{1}{2}\left(y_i - \boldsymbol{\theta}\right)^{\top} \boldsymbol{\Sigma}^{-1}\left(y_i - \boldsymbol{\theta}\right)\right\}$$

$$f_{\boldsymbol{\theta}|\boldsymbol{\Sigma}}\left(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}\right) = (2\pi)^{-m/2} \left|\frac{\boldsymbol{\Sigma}}{\lambda}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta} - \theta_0\right)^{\top} \boldsymbol{\Sigma}^{-1} \lambda \left(\boldsymbol{\theta} - \theta_0\right)\right\}$$

$$f_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) = \frac{|\tau_0|^{\gamma_0/2}}{2^{\gamma_0 m/2} \Gamma_p\left(\frac{\gamma_0}{2}\right)} |\Sigma|^{-(\gamma_0 + m + 1) \cdot \frac{1}{2}} \exp\left\{-\frac{1}{2} \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right)\right\}.$$

Basta escribir el producto $f_{\boldsymbol{\theta}|\boldsymbol{\Sigma}}\left(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}\right) f_{\boldsymbol{\Sigma}}\left(\boldsymbol{\Sigma}\right)$ para darse cuenta que

$$(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \sim NWI\left(\theta_0, \lambda, \tau_0, \gamma_0\right).$$

Por otro lado, se sabe que la distribución conjunta *a posteriori* está dada por:
$f_{\boldsymbol{\theta},\boldsymbol{\Sigma}|\boldsymbol{Y}}\left(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \boldsymbol{y}\right) = f_{\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma}}\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}\right) \times f_{\boldsymbol{\theta},\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}\right)$
$= f_{\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma}}\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}\right) \times f_{\boldsymbol{\theta}|\boldsymbol{\Sigma}}\left(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}\right) \times f_{\boldsymbol{\Sigma}}\left(\boldsymbol{\Sigma}\right)$

$$= (2\pi)^{-m \cdot n/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{\sum_{i=1}^{n} -\frac{1}{2}\left(y_i - \boldsymbol{\theta}\right)^{\top} \boldsymbol{\Sigma}^{-1}\left(y_i - \boldsymbol{\theta}\right)\right\} \times$$

$$(2\pi)^{-m/2} \left|\frac{\boldsymbol{\Sigma}}{\lambda}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta} - \theta_0\right)^{\top} \boldsymbol{\Sigma}^{-1} \lambda \left(\boldsymbol{\theta} - \theta_0\right)\right\} \times$$

$$\frac{|\tau_0|^{\gamma_0/2}}{2^{\gamma_0 m/2} \Gamma_p\left(\frac{\gamma_0}{2}\right)} |\Sigma|^{-(\gamma_0 + m + 1) \cdot \frac{1}{2}} \exp\left\{-\frac{1}{2} \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right)\right\}$$

$$\propto |\boldsymbol{\Sigma}|^{-(\gamma_0+m+n+2)} \exp\left\{ \sum_{i=1}^{n} -\frac{1}{2} (y_i - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1} (y_i - \boldsymbol{\theta}) - \frac{\lambda}{2} (\boldsymbol{\theta} - \theta_0)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \theta_0) \right.$$
$$\left. -\frac{1}{2} \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right) \right\}.$$

A continuación se analizará únicamente la parte dentro de la exponencial:

$$\sum_{i=1}^{n} \left[-\frac{1}{2} (y_i - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1} (y_i - \boldsymbol{\theta}) \right] - \frac{\lambda}{2} \left[(\boldsymbol{\theta} - \theta_0)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \theta_0) \right] - \frac{1}{2} \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right)$$
$$= -\frac{1}{2} \left[\sum_{i=1}^{n} \left(y_i^\top \boldsymbol{\Sigma}^{-1} y_i - 2\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} y_i + \boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} \right) + \lambda \left(\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} - 2\theta_0^\top \boldsymbol{\Sigma} \boldsymbol{\theta} + \theta_0^\top \boldsymbol{\Sigma}^{-1} \theta_0 \right) \right.$$
$$\left. + \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right) \right]$$
$$= -\frac{1}{2} \left[\sum_{i=1}^{n} y_i^\top \boldsymbol{\Sigma}^{-1} y_i - 2n\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \bar{y} + n\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} - 2\lambda \theta_0^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} + \lambda \theta_0^\top \boldsymbol{\Sigma}^{-1} \theta_0 \right.$$
$$\left. + \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right) \right]$$
$$= -\frac{1}{2} \left[(n+\lambda) \left(\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} \right) - 2 \left(n\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \bar{y} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \theta_0 \right) + \sum_{i=1}^{n} y_i^\top \boldsymbol{\Sigma}^{-1} y_i + \lambda \theta_0^\top \boldsymbol{\Sigma}^{-1} \theta_0 \right.$$
$$\left. + \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right) \right]$$

$$= -\frac{1}{2} \left[(n+\lambda) \left(\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \left(\frac{n\bar{y} + \lambda\theta_0}{n+\lambda} \right) \right) + \sum_{i=1}^{n} y_i^\top \boldsymbol{\Sigma}^{-1} y_i - 2n\bar{y} \boldsymbol{\Sigma}^{-1} \bar{y} \right.$$
$$\left. + 2n\bar{y}^\top \boldsymbol{\Sigma}^{-1} \bar{y} + \lambda \theta_0^\top \boldsymbol{\Sigma}^{-1} \theta_0 + \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right) \right]$$
$$= -\frac{1}{2} \left[(n+\lambda) \left(\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1} \left(\frac{n\bar{y} + \lambda\theta_0}{n+\lambda} \right) + \left(\frac{n\bar{y} + 1\theta_0}{n+\lambda} \right)^\top \boldsymbol{\Sigma}^{-1} \left(\frac{n\bar{y} + \lambda\theta_0}{n+\lambda} \right) \right.\right.$$
$$\left.\left. - \left(\frac{n\bar{y} + 1\theta_0}{n+\lambda} \right)^\top \boldsymbol{\Sigma}^{-1} \left(\frac{n\bar{y} + \lambda\theta_0}{n+\lambda} \right) \right) + \sum_{i=1}^{n} y_i^\top \boldsymbol{\Sigma}^{-1} y_i - 2\sum_{i=1}^{n} \bar{y}^\top \boldsymbol{\Sigma}^{-1} y_i + n\bar{y}^\top \boldsymbol{\Sigma}^{-1} \bar{y} \right.$$
$$\left. + n\bar{y}^\top \boldsymbol{\Sigma}^{-1} \bar{y} + \lambda \theta_0^\top \boldsymbol{\Sigma}^{-1} \theta_0 + \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right) \right]$$
$$= -\frac{1}{2} \left[(n+\lambda) \left(\left(\boldsymbol{\theta} - \frac{n\bar{y} + \lambda\theta_0}{n+\lambda} \right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\theta} - \frac{n\bar{y} + \lambda\theta_0}{n+\lambda} \right) \right) - (n\bar{y} + \lambda\theta)^\top \boldsymbol{\Sigma}^{-1} \left(\frac{n\bar{y} + \lambda\theta_0}{n+\lambda} \right) \right.$$
$$\left. + \sum_{i=1}^{n} (y_i - \bar{y})^\top \boldsymbol{\Sigma}^{-1} (y_i - \bar{y}) + n\bar{y}^\top \boldsymbol{\Sigma}^{-1} \bar{y} + \lambda \theta_0^\top \boldsymbol{\Sigma}^{-} \theta_0 + \operatorname{tr}\left(\tau_0 \boldsymbol{\Sigma}^{-1}\right) \right]$$

$$
\begin{aligned}
&= -\frac{1}{2}\left[(n+\lambda)\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)+\operatorname{tr}\left[-\left(n\bar{y}+\lambda\theta_0\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\right.\right.\\
&\quad\left.\left.+\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(y_i-\bar{y}\right)+n\bar{y}^{\top}\boldsymbol{\Sigma}^{-1}\bar{y}+\lambda\theta_0^{\top}\boldsymbol{\Sigma}^{-1}\theta_0+\tau_0\boldsymbol{\Sigma}^{-1}\right]\right]\\
&= -\frac{1}{2}\left[(n+\lambda)\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\right.\\
&\quad\left.+\operatorname{tr}\left[\left(-\left(\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\left(n\bar{y}+\lambda\theta_0\right)^{\top}+\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(y_i-\bar{y}\right)^{\top}+n\bar{y}\bar{y}^{\top}+\lambda\theta_0\theta_0^{\top}+\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right]\\
&= -\frac{1}{2}\left[(n+\lambda)\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\right.\\
&\quad+\operatorname{tr}\left[\left(-\frac{n^2\bar{y}\bar{y}^{\top}-2\lambda n\bar{y}\theta_0^{\top}-\lambda^2\theta_0\theta_0^{\top}}{n+\lambda}+\frac{(n+\lambda)\left(n\bar{y}\bar{y}^{\top}\right)}{n+\lambda}+\frac{(n+\lambda)\left(\lambda\theta_0\theta_0^{\top}\right)}{n+\lambda}\right.\right.\\
&\quad\left.\left.\left.+\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(y_i-\bar{y}\right)^{\top}+\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right]\\
&= -\frac{1}{2}\left[(n+\lambda)\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\right.\\
&\quad\left.+\operatorname{tr}\left[\left(\frac{\lambda n\bar{y}\bar{y}^{\top}-2\lambda n\bar{y}\theta_0^{\top}+\lambda n\theta_0\theta_0^{\top}}{\lambda+n}+\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(y_i-\bar{y}\right)^{\top}+\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right]\\
&= -\frac{1}{2}\left[(n+\lambda)\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\right.\\
&\quad\left.+\operatorname{tr}\left[\left(\frac{\lambda n}{\lambda+n}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^{\top}+\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(y_i-\bar{y}\right)^{\top}+\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right].
\end{aligned}
$$

Si se define

$$(n-1)S=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(y_i-\bar{y}\right)^{\top},$$

se tiene que

$$
\begin{aligned}
f_{\boldsymbol{\theta},\boldsymbol{\Sigma}|\boldsymbol{Y}}\left(\boldsymbol{\theta},\boldsymbol{\Sigma}\mid\boldsymbol{y}\right)\propto|\boldsymbol{\Sigma}|^{-(\gamma_0+m+n+2)/2}\exp\left\{-\frac{(n+\lambda)}{2}\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\right.\\
\left.-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\lambda n}{\lambda+n}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^{\top}+(n-1)S+\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\}.
\end{aligned}
$$

Por lo tanto, definiendo

$$\begin{aligned}\theta_n =& \frac{n\bar{y} + \lambda\theta_0}{n+\lambda}\\ \lambda_n =& n + \lambda\\ \gamma_n =& \gamma_0 + n\\ \tau_n =& \frac{\lambda n}{\lambda + n}\left(\bar{y} - \theta_0\right)\left(\bar{y} - \theta_0\right)^\top + (n-1)S + \tau_0,\end{aligned}$$

se tiene

$$\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \boldsymbol{Y} = \boldsymbol{y} \sim NWI(\theta_n, \lambda_n, \gamma_n, \tau_n).$$

A partir de la distribución *a posteriori* conjunta se pueden obtener las distribuciones marginales *a posteriori* de cada parámetro, la distribución marginal *a posteriori* de sigma está dada por:

$$\begin{aligned}
&f_{\boldsymbol{\Sigma}|\boldsymbol{Y}}(\boldsymbol{\Sigma} \mid \boldsymbol{y}) = \int_\Theta f_{\boldsymbol{\theta},\boldsymbol{\Sigma}|\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \boldsymbol{y}) d\theta\\
&\propto \int_\Theta |\boldsymbol{\Sigma}|^{-(\gamma_0+m+n+2)/2} \exp\left\{-\frac{(n+\lambda)}{2}\left(\boldsymbol{\theta} - \frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\right.\\
&\left. - \frac{1}{2}\operatorname{tr}\left[\left(\frac{\lambda n}{\lambda+n}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^\top + (n-1)S + \tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\}\\
&\propto \int_\Theta |\boldsymbol{\Sigma}|^{-(\gamma_0+m+n+2)/2} \exp\left\{-\frac{n+\lambda}{2}\left(\boldsymbol{\theta} - \frac{n\bar{y}+\lambda\theta_1}{n+\lambda}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\right\} d\theta\\
&\cdot \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\lambda n}{\lambda+n}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^\top + (n-1)S + \tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\} d\theta\\
&\propto \int_\Theta \left|\frac{\boldsymbol{\Sigma}}{\lambda+n}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta} - \frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)^\top \left(\frac{\boldsymbol{\Sigma}}{n+\lambda}\right)^{-1}\left(\boldsymbol{\theta} - \frac{n\bar{y}+\lambda\theta_0}{n+\lambda}\right)\right\} d\theta\\
&- \left|\frac{\boldsymbol{\Sigma}}{\lambda+n}\right|^{1/2} |\boldsymbol{\Sigma}|^{-(\gamma_0+m+n+2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\lambda n}{\lambda+n}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^\top + (n-1)S + \tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\}\\
&\propto 1 \cdot \left|\frac{\boldsymbol{\Sigma}}{\lambda+n}\right|^{1/2} |\boldsymbol{\Sigma}|^{-(\gamma_0+m+n+2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\lambda n}{\lambda+n}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^\top + (n-1)S + \tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\}\\
&\propto \boldsymbol{\Sigma}^{-(\gamma_0+m+n+1)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\lambda n}{\lambda+n}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^\top + (n-1)S + \tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\}.
\end{aligned}$$

De lo anterior, se puede concluir que la distribución marginal *a posteriori* de la matriz $\boldsymbol{\Sigma}$ es

$$\boldsymbol{\Sigma} \mid \mathbf{Y} = \boldsymbol{y} \sim WI\left(\tau_n, \gamma_n\right).$$

Donde

$$\tau_n = \frac{\lambda n}{\lambda + n} (\bar{y} - \theta_0) (\bar{y} - \theta_0)^\top + (n-1)S + \tau_0$$
$$\gamma_n = \gamma_0 + n.$$

Luego, una buena estimación para $\mathbf{\Sigma}$ es el valor esperado de la distribución marginal *a posteriori* dado por

$$\frac{\tau_n}{\gamma_n - m - 1} = \frac{\frac{\lambda n}{\lambda+n} (\bar{y} - \theta_0) (\bar{y} - \theta_0)^\top + (n-1)S + \tau_0}{\gamma_0 + n - m - 1}.$$

Teniendo en cuenta que la estimación previa de $\mathbf{\Sigma}$ viene dada por $\hat{\mathbf{\Sigma}}_{\text{pre}} = \frac{\tau_0}{\gamma_0 - m - 1}$, podemos ver que la estimación bayesiana de $\mathbf{\Sigma}$ está conformada por tres componentes: la estimación previa $\hat{\mathbf{\Sigma}}_{\text{pre}}$, la estimación clásica $\mathbf{S}$ y una medida de discrepancia entre la estimación previa y la clásica de $\boldsymbol{\theta}$. Para encontrar correctas formas de escoger los parámetros previos de $\mathbf{\Sigma}$, por ahora ignoramos el último componente, y vemos que la estimación previa $\hat{\mathbf{\Sigma}}_{\text{pre y}}$ y la estimación clásica $\mathbf{S}$ entran al cómputo de la estimación bayesiana con los pesos de $\gamma_0 - m - 1$ y $n - 1$, de esta forma, podemos escoger γ_0 tal que $\gamma_0 - m$ represente el número de la información previa, y el valor de τ_0 se puede calcular a partir de γ_0 y $\hat{\mathbf{\Sigma}}_{\text{pre}}$ (Gutiérrez y Zhang, 2021).
Por otro lado, la distribución marginal *a posteriori* de $\boldsymbol{\theta}$ es,

$$f_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int_{R^m \times R^m} f_{\boldsymbol{\theta},\mathbf{\Sigma}|\mathbf{Y}}(\boldsymbol{\theta}, \mathbf{\Sigma} \mid \boldsymbol{y}) d\mathbf{\Sigma}$$
$$= \int_{R^m \times R^m} |\mathbf{\Sigma}|^{-(\gamma_0+m+n)/2-1} \exp\left\{ -\frac{1}{2} \left[\operatorname{tr}\left(\tau_0 \mathbf{\Sigma}^{-1}\right) + \lambda(\boldsymbol{\theta} - \theta_0)^\top \mathbf{\Sigma}^{-1} (\boldsymbol{\theta} - \theta_0) + \right.\right.$$
$$\left.\left. \sum_{i=1}^{n} (y_i - \boldsymbol{\theta})^\top \mathbf{\Sigma}^{-1} (y_i - \boldsymbol{\theta}) \right] \right\} d\mathbf{\Sigma}$$
$$= \int_{R^m \times R^m} |\mathbf{\Sigma}|^{-(\gamma_0+m+n+2)/2} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left[\tau_0 + \lambda(\boldsymbol{\theta} - \theta_0)(\boldsymbol{\theta} - \theta_0)^\top + \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}) (y_i - \boldsymbol{\theta})^\top \right] \mathbf{\Sigma}^{-1} \right\} d\mathbf{\Sigma}$$
$$\propto \left| \tau_0 + \lambda(\boldsymbol{\theta} - \theta_0)(\boldsymbol{\theta} - \theta_0)^\top + \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}) (y_i - \boldsymbol{\theta})^\top \right|^{-\frac{\gamma_0+n+1}{2}}$$
$$= \left| \tau_0 + \lambda(\boldsymbol{\theta} - \theta_0)(\boldsymbol{\theta} - \theta_0)^\top + \sum_{i=1}^{n} (y_i - \overline{y}) (y_i - \overline{y})^\top + n(\overline{y} - \boldsymbol{\theta})(\overline{y} - \boldsymbol{\theta})^\top \right|^{-\frac{\gamma_0+n+1}{2}}$$
$$= \left| \tau_0 + (n-1)S + \frac{\lambda n}{\lambda + n} (\theta_0 - \overline{y})(\theta_0 - \overline{y})^\top + (\lambda + n) (\boldsymbol{\theta} - \theta_n) (\boldsymbol{\theta} - \theta_n)^\top \right|^{-\frac{\gamma_0+n+1}{2}}$$

$$
\begin{aligned}
&= \left| \tau_n + (\lambda + n)\,(\boldsymbol{\theta} - \theta_n)\,(\boldsymbol{\theta} - \theta_n)^\top \right|^{-\frac{\gamma_0+n+1}{2}} \\
&\propto \left| I_m + (\lambda + n)\,\tau_n^{-1}\,(\boldsymbol{\theta} - \theta_n)\,(\boldsymbol{\theta} - \theta_n)^\top \right|^{-\frac{\gamma_0+n+1}{2}} \\
&= \left| 1 + (\lambda + n)\,(\boldsymbol{\theta} - \theta_n)^\top\,\tau_n^{-1}\,(\boldsymbol{\theta} - \theta_n) \right|^{-\frac{\gamma_0+n+1}{2}} \\
&= \left| 1 + \frac{1}{n + \gamma_0 - m + 1}\,(\boldsymbol{\theta} - \theta_n)^\top \left(\frac{\tau_n}{(\lambda + n)\,(n + \gamma_0 - m + 1)} \right) (\boldsymbol{\theta} - \theta_n) \right|^{-\frac{\gamma_0+n+1}{2}} .
\end{aligned}
$$

Esto significa que la distribución marginal *a posteriori* del parámetro $\boldsymbol{\theta}$ es la distribución t de Student multivariante con

$$
\boldsymbol{\theta} \,\middle|\, \mathbf{Y} = \boldsymbol{y} \sim t_{n+\gamma_0-m+1} \left(\theta_n, \frac{\tau_n}{(\lambda + n)\,(n + \gamma_0 - m + 1)} \right)
$$

con $\theta_n = \frac{\lambda\theta_0 + n\bar{y}}{\lambda+n}$ y τ_n dado en la ecuación (2.5).

El anterior resultado indica que la estimación bayesiana del parámetro $\boldsymbol{\theta}$ está dada por

$$
\hat{\theta} = \theta_n = \frac{n\bar{y} + \lambda\theta_0}{n + \lambda} = \frac{n}{n + \lambda}\bar{y} + \frac{\lambda}{n + \lambda}\theta_0
$$

donde se puede observar que $\hat{\theta}$ se acercará a la estimación clásica $\bar{y}$ cuando n es grande comparado a λ, de lo contrario se acercará a la estimación previa θ_0. La varianza posterior para el i-ésimo componente de $\boldsymbol{\theta}$ está dada por

$$
\operatorname{var}(\theta_i \mid \mathbf{Y}) = \frac{\tau_{ii}}{(\lambda + n)\,(n + \gamma_0 - m + 1)} \frac{n + \lambda - m + 1}{n + \lambda - m - 1} \approx \frac{\tau_{ii}}{(\lambda + n)\,(n + \gamma_0 - m + 1)}
$$

donde τ_{ii} denota el i-ésimo elemento en la diagonal de la matriz τ_n.

Capítulo 3

Inferencia bayesiana sesgada

La inferencia bayesiana sesgada es una propuesta alterna al método bayesiano tradicional. Su objetivo sigue siendo el mismo que el del método bayesiano tradicional, sin embargo, cuenta con una variante, la introducción de sesgos en forma de potencias en la distribución *a priori* y en la función de verosimilitud. Esta variante permite describir de una manera más acertada los comportamientos de algunos fenómenos que se observan en la realidad. En este capítulo se muestra de una forma más detallada la propuesta que ofrece la inferencia bayesiana sesgada, así como el análisis del modelo normal multivariado conjugado bajo este nuevo enfoque se obtiene de manera explícita la distribución *a posteriori* para este modelo así como las distribuciones marginales *a posteriori* de los parámetros del modelo.

3.1. Planteamiento

El método bayesiano es una herramienta que permite hacer inferencias para la toma de decisiones incluso con una cantidad limitada de datos, y tiene éxito donde los métodos estadísticos tradicionales no han logrado capturar la dinámica inherente. Este método se ha aplicado en muchos campos de investigación, como la genética, la lingüística, el procesamiento de imágenes, la cosmología, la ecología, el aprendizaje automático, la psicología y la neurociencia (Matsumori *et al.*, 2018).

Sin embargo, según Matsumori *et al.* (2018) en muchas situaciones, el comportamiento real presenta desviaciones respecto a lo obtenido con el método bayesiano. Matsumori *et al.* (2018) propone la introducción de sesgos exponenciales para la explicación de dichas desviaciones. Dichos sesgos exponenciales son insertados en forma de dos potencias que afectan a una distribución única, es decir, se inserta una potencia β que afecta únicamente a la verosimilitud y una potencia α que afecta únicamente a la distribución *a priori*. De

esta manera la distribución *a posteriori* se encontrará de la forma,

$$p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y}) \propto [p_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta})]^{\beta}[p_{\boldsymbol{\theta}}(\boldsymbol{\theta})]^{\alpha},$$

donde α y β son positivos.
Cabe notar que, si a la expresión anterior se le aplica logaritmo se obtiene

$$\ln\left(p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y})\right) = \beta \ln\left(p_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta})\right) + \alpha \ln\left(p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right) + cte.$$

donde β es el peso del logaritmo de la función de verosimilitud, α es el peso del logaritmo de la distribución *a priori* y *cte.* es el término de normalización (Matsumori *et al.*, 2018). Cuando se le asignan distintos valores a α y a β se tiene los siguientes casos,

1. Si $\alpha = \beta = 1$ se tiene la inferencia bayesiana estándar (usual), como la mostrada en el capítulo anterior.

2. Cuando $0 \leq \alpha < 1$ la distribución *a priori* es más plana que lo normal, la distribución *a priori* va teniendo menor influencia sobre la distribución *a posteriori* conforme se actualizan los datos muestrales, la información más remota se vuelve menos relevante, a este caso se le conoce como *olvido*.

3. Si $\alpha > 1$ la distribución *a priori* es más nítida que lo usual, la influencia sobre la distribución *a posteriori* es mayor, el predominio de la información más antigua va creciendo conforme la información se actualiza, a este caso se le llama *estereotipo*

4. Cuando $0 \leq \beta < 1$ la distribución muestral es más plana que lo usual, su influencia sobre la distribución *a posteriori* es menor, a este tipo de sesgo se le llama *rigidez*, ya que los datos observados afectan poco. Cuando los datos observados tiene datos atípicos, la distribución *a posteriori* puede verse muy afectada por estos datos. La *rigidez* disminuye la influencia de estos datos.

5. Si $\beta > 1$ la influencia de la función de verosimilitud sobre la distribución de probabilidad *a posteriori* es mayor. A este tipo de sesgo lo llamamos *flexibilidad* ya que los datos observados influyen fuertemente en la probabilidad posterior.

La Figura 3.1 ilustra estos casos, cuando β se acerca a ∞, la distribución *a posteriori* converge a la moda de la función de verosimilitud [el mismo resultado que la estimación de máxima verosimilitud (ML)]. Por tanto, la flexibilidad proporciona un método intermedio entre las estimaciones bayesianas y ML. La región roja ($\alpha < \beta$) enfatiza la función de verosimilitud más que la distribución *a priori*, a la inferencia de esta región se le llama inferencia bayesiana tipo ML. La región azul ($\alpha > \beta$), sin embargo, enfatiza más la distribución *a priori* que la función de verosimilitud, a la inferencia de esta región se le llama inferencia bayesiana robusta. En la región donde α, pero no β, está cerca de 0, la forma de la distribución *a posteriori* está determinada principalmente solo por la función de verosimilitud. En la región donde β (pero no α), está cerca de 0, la forma de

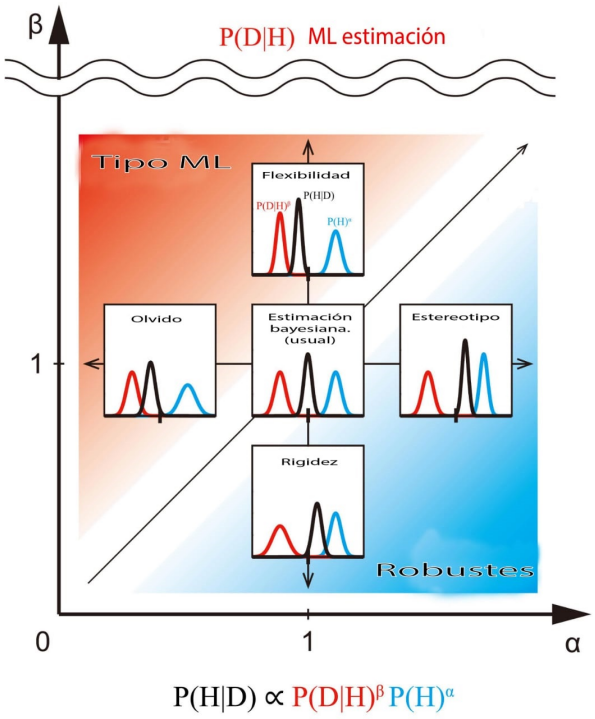

Figura 3.1: Comportamiento de la distribución *a priori* y la función de verosimilitud para diferentes valores de α y β.
Fuente: Recuperada de Matsumori *et al.* (2018)

la distribución *a posteriori* está determinada principalmente solo por la distribución *a priori*. En la región donde tanto α como β están cerca de 0, la forma de la distribución *a posteriori* es casi plana (Matsumori *et al.*, 2018).

3.2. Modelo normal multivariado conjugado

En esta sección se analiza el modelo normal multivariado conjugado bajo el enfoque de la inferencia bayesiana sesgada.
Suponga que se tienen datos que pueden modelarse de manera adecuada por medio de una distribución normal multivariada, en este caso se asume que el vector de medias $\boldsymbol{\theta}$ y la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ son desconocidas. Suponga que $\boldsymbol{Y} = (\boldsymbol{Y_1}, ..., \boldsymbol{Y_n})$ está conformado por vectores aleatorios de dimensión m, donde cada $\boldsymbol{Y_i}$ sigue una distribución normal multivariada con vector de medias $\boldsymbol{\theta}$, matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ y son independientes entre sí. Es decir $\boldsymbol{Y_i} \overset{iid}{\sim} N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Además suponga que la distribución *a priori* conjunta de $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$ está dada por,

$$f_{\boldsymbol{\theta},\boldsymbol{\Sigma}}(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = f_{\boldsymbol{\theta}|\boldsymbol{\Sigma}}(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})\, f_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}), \tag{3.1}$$

donde,

$$\boldsymbol{\theta} \mid \boldsymbol{\Sigma} \sim N\left(\theta_0, \frac{\boldsymbol{\Sigma}}{\lambda}\right) \quad \text{y} \quad \boldsymbol{\Sigma} \sim WI\left(\tau_0, \gamma_0\right). \tag{3.2}$$

Teorema 3.2.1 *Bajo el enfoque de inferencia bayesiana sesgada con los supuestos dados en las ecuaciones (3.1) y (3.2), la distribución a posteriori conjunta de los parámetros* $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ *es la distribución normal Wishart inverza con parámetros,*

$$\theta_n = \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}$$

$$\lambda_n = n\beta + \alpha\lambda$$

$$\gamma_n = \beta n + \alpha\left(\gamma_0 + m + 2\right) - m - 2$$

$$\tau_n = \frac{\alpha\lambda n\beta}{\alpha\lambda + n\beta}\left(\bar{y} - \theta_0\right)\left(\bar{y} - \theta_0\right)^{\top} + \beta(n-1)S + \alpha\tau_0.$$

Demostración.
Bajo el enfoque de la inferencia bayesiana sesgada se sabe que la distribución a posteriori está dada por,

$$\begin{aligned}
f_{\boldsymbol{\theta},\boldsymbol{\Sigma}|\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \boldsymbol{y}) =& \left[f_{\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma}}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma})\right]^{\beta} \times \left[f_{\boldsymbol{\theta},\boldsymbol{\Sigma}}(\boldsymbol{\theta}, \boldsymbol{\Sigma})\right]^{\alpha} \\
=& \left[f_{\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\Sigma}}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma})\right]^{\beta} \times \left[f_{\boldsymbol{\theta}|\boldsymbol{\Sigma}}\left(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}\right)\right]^{\alpha}\left[f_{\boldsymbol{\Sigma}}\left(\boldsymbol{\Sigma}\right)\right]^{\alpha} \\
=& \left[(2\pi)^{-m\cdot n/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\left\{\sum_{i=1}^{n} -\frac{1}{2}\left(y_i - \boldsymbol{\theta}\right)^T \boldsymbol{\Sigma}^{-1}\left(y_i - \boldsymbol{\theta}\right)\right\}\right]^{\beta} \times \\
& \left[(2\pi)^{-m/2}\left|\frac{\boldsymbol{\Sigma}}{\lambda}\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta} - \theta_0\right)^T \boldsymbol{\Sigma}^{-1}\lambda\left(\boldsymbol{\theta} - \theta_0\right)\right\}\right]^{\alpha} \times \\
& \left[\frac{|\tau_0|^{\gamma_0/2}}{2^{\gamma_0 m/2}\Gamma_p\left(\frac{\gamma_0}{2}\right)}|\Sigma|^{-(\gamma_0+m+1)\frac{1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\tau_0\boldsymbol{\Sigma}^{-1}\right)\right\}\right]^{\alpha} \\
& \propto |\ \boldsymbol{\Sigma}\ |^{-(\beta n + \alpha(1+\gamma_0+m+1))/2} \\
& \exp\left\{\beta\sum_{i=1}^{n} -\frac{1}{2}\left(y_i - \boldsymbol{\theta}\right)^T \boldsymbol{\Sigma}^{-1}\left(y_i - \boldsymbol{\theta}\right) - \frac{\alpha}{2}\left(\boldsymbol{\theta} - \theta_0\right)^T \boldsymbol{\Sigma}^{-1}\lambda\left(\boldsymbol{\theta} - \theta_0\right)\right. \\
& \left. - \frac{\alpha}{2}\operatorname{tr}\left(\tau_0\boldsymbol{\Sigma}^{-1}\right)\right\}
\end{aligned}$$

Definiendo a $\lambda' = \alpha\lambda \quad$ *y* $\tau_0' = \alpha\tau_0$*, el argumento de la exponencial se puede reescribir como,*

$$\beta\sum_{i=1}^{n} -\frac{1}{2}\left(y_i-\boldsymbol{\theta}\right)^T\boldsymbol{\Sigma}^{-1}\left(y_i-\boldsymbol{\theta}\right) - \frac{\lambda'}{2}\left(\boldsymbol{\theta}-\theta_0\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\theta_0\right) - \frac{1}{2}\operatorname{tr}\left(\tau_0'\boldsymbol{\Sigma}^{-1}\right)$$

$$= -\frac{1}{2}\Bigg[\beta\sum_{i=1}^{n}\left(y_i^\top\boldsymbol{\Sigma}^{-1}y_i - 2\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}y_i + \boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right) + \lambda'\left(\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\theta_0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \theta_0^\top\boldsymbol{\Sigma}^{-1}\theta_0\right)$$

$$+\operatorname{tr}\left(\tau_0'\boldsymbol{\Sigma}^{-1}\right)\Bigg]$$

$$= -\frac{1}{2}\Bigg[\beta\sum_{i=1}^{n} y_i^\top\boldsymbol{\Sigma}^{-1}y_i - 2n\beta\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\bar{y} + n\beta\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \lambda'\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\lambda'\theta_0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}$$

$$+\lambda'\theta_0^\top\boldsymbol{\Sigma}^{-1}\theta_0 + \operatorname{tr}\left(\tau_0'\boldsymbol{\Sigma}^{-1}\right)\Bigg]$$

$$= -\frac{1}{2}\Bigg[\left(n\beta+\lambda'\right)\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\left(n\beta\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\bar{y} + \lambda'\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\theta_0\right) + \beta\sum_{i=1}^{n} y_i^\top\boldsymbol{\Sigma}^{-1}y_i$$

$$+\lambda'\theta_0^\top\boldsymbol{\Sigma}^{-1}\theta_0 + \operatorname{tr}\left(\tau_0'\boldsymbol{\Sigma}^{-1}\right)\Bigg]$$

$$= -\frac{1}{2}\Bigg[\left(n\beta+\lambda'\right)\left(\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\left(\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)\right) + \beta\sum_{i=1}^{n} y_i^\top\boldsymbol{\Sigma}^{-1}y_i$$

$$- 2n\beta\bar{y}^\top\boldsymbol{\Sigma}^{-1}\bar{y} + 2n\beta\bar{y}^\top\boldsymbol{\Sigma}^{-1}\bar{y} + \lambda'\theta_0^\top\boldsymbol{\Sigma}^{-1}\theta_0 + \operatorname{tr}\left(\tau_0'\boldsymbol{\Sigma}^{-1}\right)\Bigg]$$

$$= -\frac{1}{2}\Bigg[\left(n\beta+\lambda'\right)\left(\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\theta^\top\boldsymbol{\Sigma}^{-1}\left(\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right) + \right.$$

$$\left.\left(\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right) - \left(\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)\right)$$

$$+\beta\sum_{i=1}^{n} y_i^\top\boldsymbol{\Sigma}^{-1}y_i - 2\beta\sum_{i=1}^{n}\bar{y}^\top\boldsymbol{\Sigma}^{-1}y_i + n\beta\bar{y}^\top\boldsymbol{\Sigma}^{-1}\bar{y} + n\beta\bar{y}^\top\boldsymbol{\Sigma}^{-1}\bar{y} + \lambda'\theta_0^\top\boldsymbol{\Sigma}^{-1}\theta_0 +$$

$$\operatorname{tr}\left(\tau_0'\boldsymbol{\Sigma}^{-1}\right)\Bigg]$$

$$= -\frac{1}{2}\Bigg[\left(n\beta+\lambda'\right)\left(\left(\boldsymbol{\theta} - \frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)\right) -$$

$$\left(n\beta\bar{y}+\lambda'\theta_0\right)^\top\boldsymbol{\Sigma}^{-1}\left(\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right) + \beta\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^\top\boldsymbol{\Sigma}^{-1}\left(y_i-\bar{y}\right) + n\beta\bar{y}^\top\boldsymbol{\Sigma}^{-1}\bar{y} +$$

$$\lambda'\theta_0^\top\boldsymbol{\Sigma}^{-1}\theta_0 + \operatorname{tr}\left(\tau_0'\boldsymbol{\Sigma}^{-1}\right)\Bigg]$$

$$
\begin{aligned}
&= -\frac{1}{2}\Bigg[(n\beta+\lambda')\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)\\
&+\operatorname{tr}\Bigg[-(n\beta\bar{y}+\lambda'\theta_0)^{\top}\boldsymbol{\Sigma}^{-1}\left(\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)\\
&+\beta\sum_{i=1}^{n}(y_i-\bar{y})^{\top}\boldsymbol{\Sigma}^{-1}(y_i-\bar{y})+n\beta\bar{y}^{\top}\boldsymbol{\Sigma}^{-1}\bar{y}+\lambda'\theta_0^{\top}\boldsymbol{\Sigma}^{-1}\theta_0+\tau_0'\boldsymbol{\Sigma}^{-1}\Bigg]\Bigg]\\
&= -\frac{1}{2}\Bigg[(n\beta+\lambda')\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)\\
&+\operatorname{tr}\Bigg[\Bigg(-\left(\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)(n\beta\bar{y}+\lambda'\theta_0)^{\top}+n\beta\bar{y}\bar{y}^{\top}+\lambda'\theta_0\theta_0^{\top}+\\
&\beta\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^{\top}+\tau_0'\Bigg)\boldsymbol{\Sigma}^{-1}\Bigg]\Bigg]
\end{aligned}
$$

Manipulando algebraicamente el argumento de la traza, se obtiene:

$$
\begin{aligned}
&\operatorname{tr}\Bigg[\Bigg(\frac{-(n\beta)^2\bar{y}\bar{y}^{\top}-2\lambda' n\beta\bar{y}\theta_0^{\top}-\lambda'^2\theta_0\theta_0^{\top}}{n\beta+\lambda'}+\frac{(n\beta+\lambda')(n\beta\bar{y}\bar{y}^{\top})}{n\beta+\lambda'}+\frac{(n\beta'+\lambda')\left(\lambda'\theta_0\theta_0^{\top}\right)}{n\beta+\lambda'}\\
&\quad+\beta\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^{\top}+\tau_0'\Bigg)\boldsymbol{\Sigma}^{-1}\Bigg]\\
&=\operatorname{tr}\left[\left(\frac{\lambda' n\beta}{\lambda'+n\beta}(\bar{y}-\theta_0)(\bar{y}-\theta_0)^{\top}+\beta\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^{\top}+\tau_0'\right)\boldsymbol{\Sigma}^{-1}\right].
\end{aligned}
$$

Definiendo $(n-1)S=\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^{\top}$ *y sustituyendo en la distribución a posteriori conjunta, resulta:*

$$
\begin{aligned}
f_{\boldsymbol{\theta},\boldsymbol{\Sigma}|\boldsymbol{Y}}(\boldsymbol{\theta},\boldsymbol{\Sigma}\mid\boldsymbol{y})\propto|\boldsymbol{\Sigma}|^{-(\beta n+\alpha(1+\gamma_0+m+1))/2}&\\
\exp\Bigg\{-\frac{(n\beta+\lambda')}{2}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\lambda'\theta_0}{n\beta+\lambda'}\right)&\\
-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\lambda' n\beta}{\lambda'+n\beta}(\bar{y}-\theta_0)(\bar{y}-\theta_0)^{\top}+\beta(n-1)S+\tau_0'\right)\boldsymbol{\Sigma}^{-1}\right]\Bigg\}.&
\end{aligned}
$$

Si se hace,

$$\theta_n = \frac{n\beta\bar{y} + \lambda'\theta_0}{n\beta + \lambda'} = \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}$$

$$\lambda_n = n\beta + \lambda' = n\beta + \alpha\lambda$$

$$\gamma_n = \beta n + \alpha\left(\gamma_0 + m + 2\right) - m - 2$$

$$\begin{aligned}\tau_n &= \frac{\lambda' n\beta}{\lambda' + n\beta}\left(\bar{y} - \theta_0\right)\left(\bar{y} - \theta_0\right)^\top + \beta(n-1)S + \tau_0' \\ &= \frac{\alpha\lambda n\beta}{\alpha\lambda + n\beta}\left(\bar{y} - \theta_0\right)\left(\bar{y} - \theta_0\right)^\top + \beta(n-1)S + \alpha\tau_0,\end{aligned}$$

se tiene el resultado buscado, por lo tanto,

$$\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \boldsymbol{Y} = \boldsymbol{y} \sim NWI\left(\theta_n, \lambda_n, \gamma_n, \tau_n\right). \blacksquare$$

A partir de la distribución *a posteriori* conjunta se pueden obtener las distribuciones marginales *a posteriori* de cada parámetro. El siguiente teorema muestra como conseguir la distribución *a posteriori* marginal de $\boldsymbol{\Sigma}$ a partir de la distribución *a posteriori* conjunta de $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$.

Teorema 3.2.2 *Con las hipótesis del Teorema (3.2.1), la distribución a posteriori marginal de* $\boldsymbol{\Sigma}$ *es la distribución Wishart inversa con parámetros* (τ_n, γ_n).

Demostración.
Por definición, se sabe que la distribución marginal está dada por,

$$\begin{aligned}&f_{\boldsymbol{\Sigma}|\boldsymbol{Y}}(\boldsymbol{\Sigma} \mid \boldsymbol{y}) = \int_\Theta f_{\boldsymbol{\theta},\boldsymbol{\Sigma}|\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \boldsymbol{y}) d\theta \\ &= \int_\Theta |\boldsymbol{\Sigma}|^{-(\beta n + \alpha(1+\gamma_0+m+1))/2} \exp\left\{-\frac{(n\beta+\alpha\lambda)}{2}\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\right. \\ &\left. -\frac{1}{2}\operatorname{tr}\left[\left(\frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^\top + \beta(n-1)S + \alpha\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\} d\theta \\ &\propto \int_\Theta |\boldsymbol{\Sigma}|^{-(\gamma_n+m+2)/2} \exp\left\{-\frac{\beta n + \alpha\lambda}{2}\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \lambda\alpha\theta_0}{n\beta+\lambda\alpha}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \lambda\alpha\theta_0}{n\beta+\lambda\alpha}\right)\right\} d\theta \\ &\cdot \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^\top + \beta(n-1)S + \alpha\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\}\end{aligned}$$

$$\propto \int_{\Theta} \left|\frac{\boldsymbol{\Sigma}}{n\beta+\alpha\lambda}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\lambda\alpha\theta_0}{n\beta+\lambda\alpha}\right)^{\top}\left(\frac{\boldsymbol{\Sigma}}{n\beta+\lambda\alpha}\right)^{-1}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\lambda\alpha\theta_0}{n\beta+\lambda\alpha}\right)\right\} d\theta$$

$$\left|\frac{\boldsymbol{\Sigma}}{n\beta+\lambda\alpha}\right|^{1/2} |\boldsymbol{\Sigma}|^{-(\gamma_n+m+2)/2}$$

$$\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^{\top}+\beta(n-1)S+\alpha\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\}$$

$$\propto 1\cdot\left|\frac{\boldsymbol{\Sigma}}{\lambda+n}\right|^{1/2} |\boldsymbol{\Sigma}|^{-(\gamma_n+m+2)/2}$$

$$\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^{\top}+\beta(n-1)S+\alpha\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\}$$

$$\propto \boldsymbol{\Sigma}^{-(\gamma_n+m+1)/2}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^{\top}+\beta(n-1)S+\alpha\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\}.$$

La expresión anterior corresponde a la función de densidad de una distribución Wishart inversa con parametros (τ_n, γ_n)*, esto es,*

$$\boldsymbol{\Sigma} \mid \mathbf{Y} = \boldsymbol{y} \sim WI\left(\tau_n, \gamma_n\right).$$

A partir de aquí, se sabe que una buena estimación puntual de la matriz de varianzas y covarianzas es el valor esperado de la distribución marginal *a posteriori* dado por

$$\frac{\tau_n}{\gamma_n-m-1} = \frac{\frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^{\top}+\beta(n-1)S+\alpha\tau_0}{\beta n+\alpha\left(\gamma_0+m+2\right)-2m-3}.$$

Teniendo en cuenta que la estimación previa de $\boldsymbol{\Sigma}$ viene dada por $\hat{\boldsymbol{\Sigma}}_{\text{pre}} = \frac{\tau_0}{\gamma_0-m-1}$, podemos ver que la estimación bayesiana sesgada de $\boldsymbol{\Sigma}$ está conformada por tres componentes: la estimación previa $\hat{\boldsymbol{\Sigma}}_{\text{pre}}$, la estimación clásica S y una medida de discrepancia entre la estimación previa y la clásica de $\boldsymbol{\theta}$. Este último termino se incorpora con un peso de $\frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}$, la estimación previa $\hat{\boldsymbol{\Sigma}}_{\text{pre y}}$ y la estimación clásica S se añaden al cómputo de la estimación bayesiana sesgada con los pesos de $\alpha(\gamma_0-m-1)$ y $\beta(n-1)$. Esto significa que si se le da un mayor peso a la función de verosimilitud, la estimación clásica tendrá mayor influencia sobre la estimación bayesiana sesgada, por otro lado, si se le da mayor importancia a la distribución *a priori*, la estimación previa tendrá mayor influencia sobre la estimación bayesiana sesgada. Sin embargo, a medida que crecen α y β la discrepancia entre la estimación previa y la estimación clásica de $\boldsymbol{\theta}$ también va creciendo. Suponga que se fija el valor de α y se hace crecer a β, el valor de $\frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}$ va a converger a $\alpha\lambda$. Si por el contrario, se fija el valor β y se hace crecer a α, el valor de $\frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}$ converge a $n\beta$.

Por otro lado, a partir de la distribución *a posteriori* conjunta, se puede mostrar que la distribución marginal *a posteriori* de $\boldsymbol{\theta}$ es la distribución t de Student multivariante, esto

se muestra en el siguiente teorema.

Teorema 3.2.3 *Con las hipótesis del Teorema (3.2.1), la distribución a posteriori marginal de* $\boldsymbol{\theta}$ *es la distribución t de Student multivariante con* $\nu = \beta n + \alpha\left(\gamma_0 + m + 2\right) - 2m - 1$ *grados de libertad y parámetros*

$$\theta_n \quad y \quad \frac{\tau_n}{(n\beta + \alpha\lambda)(\beta n + \alpha\left(\gamma_0 + m + 2\right) - 2m - 1)}.$$

Demostración.
Por definición, la distribución marginal de $\boldsymbol{\theta}$ *está dada por,*

$$f_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int_{R^m \times R^m} f_{\boldsymbol{\theta},\boldsymbol{\Sigma}|\mathbf{Y}}(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \boldsymbol{y}) d\boldsymbol{\Sigma}$$

$$= \int_{R^m \times R^m} |\boldsymbol{\Sigma}|^{-(\beta n + \alpha(\gamma_0 + m + 2))/2}$$
$$\exp\left\{-\frac{(n\beta + \alpha\lambda)}{2}\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)\right.$$
$$\left. -\frac{1}{2}\operatorname{tr}\left[\left(\frac{\alpha\lambda n\beta}{\alpha\lambda + n\beta}\left(\bar{y} - \theta_0\right)\left(\bar{y} - \theta_0\right)^\top + \beta(n-1)S + \alpha\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\} d\boldsymbol{\Sigma}$$

$$= \int_{R^m \times R^m} |\boldsymbol{\Sigma}|^{-(\beta n + \alpha(\gamma_0 + m + 2))/2}$$
$$\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left((n\beta + \alpha\lambda)\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)^\top\right.\right.\right.$$
$$\left.\left.\left. + \frac{\alpha\lambda n\beta}{\alpha\lambda + n\beta}\left(\bar{y} - \theta_0\right)\left(\bar{y} - \theta_0\right)^\top + \beta(n-1)S + \alpha\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\} d\boldsymbol{\Sigma}.$$

$$= \int_{R^m \times R^m} \left|(n\beta + \alpha\lambda)\left(\theta - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)\left(\theta - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)^\top + \right.$$
$$\left.\frac{\alpha\lambda n\beta}{\alpha\lambda + n\beta}\left(\bar{y} - \theta_0\right)\left(\bar{y} - \theta_0\right)^\top + \beta(n-1)S + \alpha\tau_0\right|^{(\beta n + \alpha(\gamma_0 + m + 2) - m - 1)/2} |\boldsymbol{\Sigma}|^{-(\beta n + \alpha(\gamma_0 + m + 2))/2}$$
$$\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left((n\beta + \alpha\lambda)\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)\left(\boldsymbol{\theta} - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)^\top\right.\right.\right.$$
$$\left.\left.\left. + \frac{\alpha\lambda n\beta}{\alpha\lambda + n\beta}\left(\bar{y} - \theta_0\right)\left(\bar{y} - \theta_0\right)^\top + \beta(n-1)S + \alpha\tau_0\right)\boldsymbol{\Sigma}^{-1}\right]\right\} d\boldsymbol{\Sigma}.$$
$$\left|(n\beta + \alpha\lambda)\left(\theta - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)\left(\theta - \frac{n\beta\bar{y} + \alpha\lambda\theta_0}{n\beta + \alpha\lambda}\right)^\top\right.$$
$$\left. + \frac{\alpha\lambda n\beta}{\alpha\lambda + n\beta}\left(\bar{y} - \theta_0\right)\left(\bar{y} - \theta_0\right)^\top + \beta(n-1)S + \alpha\tau_0\right|^{-(\beta_n + \alpha(\gamma_0 + m + 2) - m - 1)/2}$$

$$
\begin{aligned}
&= \int_{R^m \times R^m} 2^{-(\beta_n+\alpha(\phi_0+m+2)-m-1)m/2} \\
&\left| (n\beta+\alpha\lambda)\left(\theta - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\theta - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top + \frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}(\bar{y}-\theta_0)(\bar{y}-\theta_0)^\top \right. \\
&\left. + \beta(n-1)S + \alpha\tau_0 \right|^{(\beta n+\alpha(\gamma_0+m+2)-m-1)/2} |\boldsymbol{\Sigma}|^{-(\beta n+\alpha(\gamma_0+m+2))/2} \\
&\exp\left\{ -\frac{1}{2}\operatorname{tr}\left[\left((n\beta+\alpha\lambda)\left(\boldsymbol{\theta} - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\boldsymbol{\theta} - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top \right.\right.\right. \\
&\left.\left.\left. + \frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}(\bar{y}-\theta_0)(\bar{y}-\theta_0)^\top + \beta(n-1)S + \alpha\tau_0 \right)\boldsymbol{\Sigma}^{-1} \right] \right\} d\boldsymbol{\Sigma}. \\
&\left| (n\beta+\alpha\lambda)\left(\theta - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\theta - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top \right. \\
&\left. + \frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}(\bar{y}-\theta_0)(\bar{y}-\theta_0)^\top + \beta(n-1)S + \alpha\tau_0 \right|^{-(\beta_n+\alpha(\gamma_0+m+2)-m-1)/2} \\
&2^{(\beta_n+\alpha(\phi_0+m+2)-m-1)m/2} \\
&= \int_{R^m \times R^m} \left(\Gamma_m\left(\frac{\beta_n+\alpha(\gamma_0+m+2)-m-1}{2} \right) \right)^{-1} 2^{-(\beta_n+\alpha(\phi_0+m+2)-m-1)m/2} \\
&\left| (n\beta+\alpha\lambda)\left(\theta - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\theta - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top + \frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}(\bar{y}-\theta_0)(\bar{y}-\theta_0)^\top \right. \\
&\left. + \beta(n-1)S + \alpha\tau_0 \right|^{(\beta n+\alpha(\gamma_0+m+2)-m-1)/2} |\boldsymbol{\Sigma}|^{-(\beta n+\alpha(\gamma_0+m+2))/2} \\
&\exp\left\{ -\frac{1}{2}\operatorname{tr}\left[\left((n\beta+\alpha\lambda)\left(\boldsymbol{\theta} - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\boldsymbol{\theta} - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top \right.\right.\right. \\
&\left.\left.\left. + \frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}(\bar{y}-\theta_0)(\bar{y}-\theta_0)^\top + \beta(n-1)S + \alpha\tau_0 \right)\boldsymbol{\Sigma}^{-1} \right] \right\} d\boldsymbol{\Sigma}. \\
&\left| (n\beta+\alpha\lambda)\left(\theta - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\theta - \frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top \right. \\
&\left. + \frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}(\bar{y}-\theta_0)(\bar{y}-\theta_0)^\top + \beta(n-1)S + \alpha\tau_0 \right|^{-(\beta_n+\alpha(\gamma_0+m+2)-m-1)/2} \\
&\Gamma_m\left(\frac{\beta_n+\alpha(\gamma_0+m+2)-m-1}{2} \right) 2^{(\beta_n+\alpha(\phi_0+m+2)-m-1)m/2}
\end{aligned}
$$

La expresión antes del diferencial de $\boldsymbol{\Sigma}$ es la función de densidad de una distribución Wishart inversa cuya integral es 1. La expresión después del diferencial de $\boldsymbol{\Sigma}$ es

proporcional a,

$$\begin{aligned}
&\propto \left| (n\beta+\alpha\lambda)\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top \right. \\
&\quad \left. + \frac{\alpha\lambda n\beta}{\alpha\lambda+n\beta}\left(\bar{y}-\theta_0\right)\left(\bar{y}-\theta_0\right)^\top + \beta(n-1)S + \alpha\tau_0 \right|^{-\frac{\beta n+\alpha(\gamma_0+m+2)-m-1}{2}} \\
&= \left| (n\beta+\alpha\lambda)\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top + \tau_n \right|^{-\frac{\beta n+\alpha(\gamma_0+m+2)-m-1}{2}} \\
&= |\tau_n| \left| (n\beta+\alpha\lambda)\tau_n^{-1}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top + I_m \right|^{-\frac{\beta n+\alpha(\gamma_0+m+2)-m-1}{2}} \\
&\propto \left| (n\beta+\alpha\lambda)\tau_n^{-1}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top + I_m \right|^{-\frac{\beta n+\alpha(\gamma_0+m+2)-m-1}{2}},
\end{aligned}$$

en la expresión anterior el término

$$\tau_n^{-1}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right),$$

es un vector de dimensión $m\times 1$, por lo que se puede aplicar la propiedad $|I+uv^\top| = 1+u^\top v$, así la expresión anterior es igual a,

$$\begin{aligned}
&= \left| 1 + (n\beta+\alpha\lambda)\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top \tau_n^{-1}\left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right) \right|^{-\frac{\beta n+\alpha(\gamma_0+m+2)-m-1}{2}} \\
&= \left| 1 + \frac{1}{\beta n+\alpha\left(\gamma_0+m+2\right)-2m-1}\times \right. \\
&\quad \left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right)^\top \left(\frac{\tau_n}{(n\beta+\alpha\lambda)(\beta n+\alpha\left(\gamma_0+m+2\right)-2m-1)}\right)^{-1} \\
&\quad \left. \left(\boldsymbol{\theta}-\frac{n\beta\bar{y}+\alpha\lambda\theta_0}{n\beta+\alpha\lambda}\right) \right|^{-\frac{\beta n+\alpha(\gamma_0+m+2)-m-1}{2}}.
\end{aligned}$$

Por lo tanto,

$$\boldsymbol{\theta}\,\Big|\,\mathbf{Y}=\boldsymbol{y} \sim t_\nu\left(\theta_n, \frac{\tau_n}{(n\beta+\alpha\lambda)(\beta n+\alpha\left(\gamma_0+m+2\right)-2m-1)}\right). \blacksquare$$

El anterior resultado indica que la estimación bayesiana sesgada del parámetro $\boldsymbol{\theta}$ está dada por

$$\hat{\boldsymbol{\theta}} = \theta_n = \frac{n\beta\bar{y} + \lambda\alpha\theta_0}{n\beta + \lambda\alpha} = \frac{n\beta}{n\beta + \lambda\alpha}\bar{y} + \frac{\lambda\alpha}{n\beta + \lambda\alpha}\theta_0,$$

en donde se puede observar que la nueva estimación está dada por un promedio del estimador clásico $\bar{y}$ y la estimación previa θ_0. Note que el sesgo que afecta a la verosimilitud aparece como un peso del estimador clásico y el sesgo que afecta a la distribución *a priori* aparece como un peso de la estimación previa. En este caso si se fija el valor de α y se hace crecer a β la nueva estimación va a converger a la estimación clásica, por otro lado si se fija el valor de β y se hace crecer el valor de α, el valor de la nueva estimación converge a la estimación previa.

Capítulo 4

Conclusiones

En esta tesis se realizó el análisis de distintos modelos estadísticos bajo el enfoque bayesiano, con el fin de familiarizarse y entender el proceso de inferencia dentro de la estadística bayesiana. En específico se obtuvo, la distribución *a posteriori* marginal del parámetro θ en el modelo normal con media desconocida y varianza conocida; las distribuciones *a posteriori* conjunta y marginales para los parámetros θ y σ^2 en el modelo normal conjugado y las distribuciones *a posteriori* conjunta y marginales para los parámetros $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$ en el modelo normal multivariado conjugado.

Posteriormente, se analizó la propuesta hecha por Matsumori *et al.* (2018) estudiando los distintos casos posibles de la distribución *a posteriori* de acuerdo a los valores de los sesgos exponenciales y se aplicó al modelo normal multivariado conjugado mediante la introducción de sesgos exponenciales en la función de verosimilitud y en la distribución a priori. Así, se logró encontrar de manera explícita la distribución *a posteriori* conjunta de los parámetros $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$, la cual resultó ser una distribución normal Wishart inversa. Es aquí donde se tiene el éxito de este trabajo ya que con esto, se logra cumplir el objetivo planteado para el desarrollo de esta tesis y permite verificar que la hipótesis planteada es verdadera. Además como un aporte extra a el resultado obtenido, se encontraron las distribuciones marginales *a posteriori* de los parámetros $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$ que resultaron ser una distribución t de Student multivariante y una distribución Wishart inversa, respectivamente.

Uno de los principales aportes de esta tesis es proporcionar de manera detallada los cálculos para poder encontrar las distribuciones *a posteriori* conjuntas y marginales de los parámetros, así como proporcionar un análisis teórico acerca del comportamiento de las estimaciones puntuales de cada parámetro.

Como trabajo a futuro se pueden abordar otros modelos a partir del enfoque bayesiano sesgado además de mostrar algunas aplicaciones de estos en algunas áreas de interés.

Referencias

Correa Morales, J. C. y Barrera Causil, C. J. (2018). Introducción a la estadística Bayesiana. *Textos Académicos*.

Evans, M. J. (2018). *Probabilidad y estadística: la ciencia de la incertidumbre*. Reverté.

Gelman, A., Carlin, J. B., Stern, H. S. y Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gutiérrez, A. y Zhang, H. (2021). *Modelos Bayesianos con R y STAN*.

Hernández González, S. (2005). Historia de la estadística.

Matsumori, K., Koike, Y. y Matsumoto, K. (2018). A biased Bayesian inference for decision-making and cognitive control. *Frontiers in Neuroscience*, 12, 734.

Mendoza, M. y Regueiro, P. (2011). Estadística Bayesiana. *Instituto Teconológico Autónomo de México*.

Montoya, L., Vargas, S. y Jiménez-Hernández, J. (2024). Selección de Cartera: Un enfoque de Sesgos de Comportamiento. *Revista Mexicana de Economía y Finanzas. ISSN: 24486795 Aceptado para su publicación.*.

Mood, A. M., Graybill, F. *et al.* (1978). *Introducción a la teoría de la estadística*. Madrid: Aguilar,.

Nuñez-Antonio, G. y Castillo, A. (2014). *Introducción a la estadística bayesiana: Un enfoque aplicado*. Universidad Autónoma Metropolitana.

Rangel, E. N. (2015). ¿ Qué es la Estadística Bayesiana? *Journal of Basic Sciences*, 1, 1.