UNIVERSIDAD CENTROCCIDENTAL
"LISANDRO ALVARADO"

Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

"INFERENCIA BAYESIANA PARA EL ANÁLISIS MULTIVARIADO"

Trabajo Especial de Grado presentado por

Br. José F. Querales

como requisito final
para obtener el título de Licenciado
en Ciencias Matemáticas
Área de Conocimiento: Estadistica.
Tutor: MSc Luz Rodríguez

Barquisimeto, Venezuela. Octubre de 2008

Universidad Centroccidental
"Lisandro Alvarado"
Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

ACTA
TRABAJO ESPECIAL DE GRADO

Los suscritos miembros del Jurado designado por el Jefe del Departamento de Matemáticas del Decanato de Ciencias y Tecnología de la Universidad Centroccidental "Lisandro Alvarado", para examinar y dictar el veredicto sobre el Trabajo Especial de Grado titulado:

"INFERENCIA BAYESIANA PARA EL ANÁLISIS MULTIVARIADO"

Presentado por el ciudadano BR. JOSÉ F. QUERALES titular de la Cédula de Identidad Nº 16.323.371. Con el propósito de cumplir con el requisito académico final para el otorgamiento del título de Licenciado en Ciencias Matemáticas.

Luego de realizada la Defensa y en los términos que imponen los Lineamientos para el Trabajo Especial de Grado de la Licenciatura en Ciencias Matemáticas, se procedió a discutirlo con el interesado habiéndose emitido el veredicto que a continuación se expresa:

[1] ____________________

Con una calificación de ____________________ puntos.

En fe de lo expuesto firmamos la presente Acta en la Ciudad de Barquisimeto a los _____ días del mes de ____________________ de __________.

TUTOR	FIRMA
PRINCIPAL	FIRMA
PRINCIPAL	FIRMA

OBSERVACIONES:

[1] Aprobado ó Reprobado

Gracias a Dios y a Nuestra Señora de Coromoto por darme ésta oportunidad

Agradecimientos

En primer lugar quiero dedicar este agradecimiento a mis padres Haidee y Francisco por ser mis guías durante mi vida y apoyarme en mi decisión de estudiar esta carrera. A mi familia en especial a mis hermanas, sobrina y a mi abuelita Rosa por estar pendiente de mí.

A las familias Lugo Briceño y Pérez Acosta. A mis amigos y compañeros Roberto, Anthony, Juan, Luiggy, Yankis, Borys, Efren, Javier, Marilyn, María, Gaby, Gladymar y Rafael. Y a todos que han estado a mi lado a lo largo de la carrera.

A mis profesores Jurancy, Edgar Guédez, Ismael Huerta, Edner Pineda, Wilfredo Ángulo, Neptalí Romero, Malón Mendoza, Hugo Lara, Jhonny Escalona y a mi tutora Luz Rodríguez quienes me han transmitido gran parte de sus conocimientos durante mi formación académica.

Por último quiero agradecer a la Familia Salcedo que en tan poco tiempo me han dado un gran apoyo en mi vida para seguir adelante, en especial a Litzy que ha sido y será de gran importancia en mi vida, ya que es la constante de Litchyz para converge a mis metas; le doy gracias a Dios por habermela puesto en mi vida.

“INFERENCIA BAYESIANA PARA EL ANÁLISIS MULTIVARIADO”

Resumen

El análisis multivariado es esa rama de la estadística dedicada al estudio de variables aleatorias correlacionadas entre sí. La esencia de la aplicación del análisis multivariado envuelve la motivación de resolver problemas y llegar a respuestas númerica, o generar grandes opiniones acerca de un fenómeno natural, así como también proveer resultados que pueden ser usados como base para tomar decisiones.

Por otra parte, la estadística Bayesiana es un término aplicado al cuerpo de las técnicas inferenciales que usan el teorema de Bayes para combinar la data observada con opiniones subjetivas o personales. La aproximación Bayesiana multivariada ha estado resolviendo con éxito problemas que han sido difíciles de tratar desde otros puntos de vista, por esta razón el mayor desarrollo que ha influido significativamente en las aplicaciones del análisis multivariado es el rápido desenvolvimiento de los resultados Bayesianos Multivariados. El objetivo de este proyecto es establecer formalmente la teoría correspondiente al Análisis Multivariado bajo el enfoque Bayesiano. Además, la distribución normal multivariada es introducida debido a la importancia y utilidad en diversos problemas de la vida real, así como también algunos ejemplos serán presentados para ilustrar la teoría.

Índice

Introducción

Para establecer un modelo probabilístico completo con todas las cantidades observables y las no observables, se debe establecer la distribución conjunta tanto de los parámetros, como de las cantidades observables (llamadas datos). Por ello, deben recordarse conceptos y resultados básicos de la teoría de probabilidad y algebra matricial necesarios para el desarrollo del trabajo.

En el Análisis Bayesiano se trata de encontrar métodos prácticos para hacer inferencia usando modelos de probabilidad tanto de las cantidades que se observan como de la cantidades no observables que suelen ser de interés. Se debe condicionar la función de distribución con respecto a las cantidades que han sido observadas. Ésta distribución condicional se conoce como a posteriori y nos provee de toda la información contenida en la muestra para inferir sobre los parámetros. Todos estos estudios se harán en los parámetros de la la distribución normal multivariada ; $\theta : p \times 1$ y $\Sigma : p \times p$; es decir, $N(\theta, \Sigma)$, en algunos caso estos parámetros son escalares.

Capítulo 1

Preliminares

§1.1. Traza de una matriz

Definición **1.1.1.** La suma de los elementos de la diagonal de una matriz cuadrada es llamada la traza, es decir; si $A = (a_{ij})$ con $i, j = 1, ..., p$

$$trA = \sum_{j=1}^{p} a_{jj}.$$

Esta definición será de gran utilidad, por ejemplo, para definir algunas distribuciones de probabilidad para el caso multivariado.

Mencionaremos algunas propiedades de la traza:

1. Supongamos que $A : p \times n, B : n \times p$, entonces

$$tr(AB) = tr(BA). \tag{1.1}$$

Por ejemplo; si x es un vector $p \times 1$,

$$tr(xx') = tr(x'x) = x'x$$

Este resultado se obtendrá porque el elemento ij de AB es $\sum_{\alpha=1}^{n} a_{i\alpha} b_{\alpha j}$ así,

$$tr(AB) = \sum_{i=1}^{p} \sum_{\alpha=1}^{n} a_{i\alpha} b_{\alpha j}.$$

Además, el elemento ij de BA es $\sum_{\alpha=1}^{p} b_{i\alpha} a_{\alpha j}$, así

$$tr(BA) = \sum_{i=1}^{n} \sum_{\alpha=1}^{p} b_{i\alpha} a_{\alpha j}.$$

2. Si $A : p \times n, B : p \times n$, entonces

$$tr(A + B) = tr(A) + tr(b). \tag{1.2}$$

3. Si α es un escalar y $A : p \times p$,

$$tr(\alpha A) = \alpha tr(A). \tag{1.3}$$

4. Si A es un escalar, $tr(A) = A$. Por ejemplo, si $x : p \times 1, A : p \times p, x'Ax$ es un escalar. Así

$$tr(Axx') = x'Ax$$

.

§1.2. Derivada de una función escalar de una matriz

La derivada de una función escalar f de una matriz $X = (x_{ij})$ con $i = 1, ..., p$ y $j = 1, n$ está definida como:

$$\frac{d}{dX} f(X) = \left(\frac{\partial f(X)}{\partial x_{ij}} \right), \qquad \begin{matrix} i = 1, ...p \\ j = 1, ..., n \end{matrix}$$

Algunas propiedades que serán utilizadas mas adelante son las siguientes:

1. Para $X : p \times p$, $|X| \neq 0$ y $X = X'$

$$\frac{d}{dX}|X| = 2|X|X^{-1} - diag(|X|X^{-1}) \tag{1.4}$$

2. Para $A' : p \times q$, $X : q \times p$

$$\frac{d}{dX} tr(A'X) = A \tag{1.5}$$

3. Si $x : p \times 1$, $A : p \times p$,

$$\frac{d}{dx}(x'Ax) = 2Ax \tag{1.6}$$

§1.3. Transformación de la matriz jacobiana

Definición **1.3.1.** Suponga que X, Y son matrices las cuales tienen el mismo número de elementos, r. Entonces si $X = f(Y)$. Esta función está determinada por r funciones reales: $y_1(x_1, ..., x_r), ..., y_r(x1, ..., x_r)$. Entonces La transformación jacobiana está definida como:

$$J(Y \longrightarrow X) = ||A|| \qquad A = \left(\frac{\partial y_i}{\partial x_i}\right), \qquad i, j = 1, ..., r$$

donde $||A||$ es el valor absoluto del determinante de A el cual es denotado por $|A|$, y $(x_1, ..., x_r)$ y $(y_1, ..., y_r)$ denota los distintos valores de X y Y, respectivamente.

Los siguientes Jacobianos son frecuentemente utilizados en el análisis multivariado. El valor absoluto será omitido para simplificar, pero debe darse por entendido

1. Si $Y : p \times p, X : p \times p$, X=X', y Y=AXA',

$$J(Y \rightarrow X) = |A|^{p+1}, \qquad |A| \neq 0 \tag{1.7}$$

2. Si $|A| \neq 0$, $dA^{-1} = -A^{-1}(dA)A^{-1}$ si $X = A^{-1}$,

$$J(A \rightarrow X) = |X|^{-(p+1)} \tag{1.8}$$

donde $X = X'$, $X : p \times p$.

Este resultado se obtiene de (1.7) dado que $J(X \rightarrow A) = J(dX \rightarrow dA) = J(dA^{-1} \rightarrow dA)$, y $J(A \rightarrow X) = [J(X \rightarrow A)]^{-1}$.

§1.4. Inversa de una matriz particionada por bloques

Sea $A : (q + r) \times (q + r)$ una matriz particionada por bloques arbitraria con $|A| \neq 0$ definida de la siguiente manera:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \tag{1.9}$$

donde $A_{11} : q \times q$, $A_{12} : q \times r$, $A_{21} : r \times q$ y $A_{22} : r \times r$.

Ahora, como $|A| \neq 0$ entonces existe A^{-1}, y la denotamos de la siguiente manera:

$$B = A^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \tag{1.10}$$

con $B_{11} : q \times q$,$B_{12} : q \times r$, $B_{21} : r \times q$ y $B_{22} : r \times r$. Calculemos B_{ij} en función de A_{ji} y viceversa con $i, j = 1, 2$

Como $B \cdot A = I_{(q+r)}$ entonces

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \cdot \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = I_{(q+r)}$$

Así, multiplicando las submatrices obtenemos:

$$\begin{aligned} B_{11} \cdot A_{11} + B_{12} \cdot A_{12} &= I_q & (1.11) \\ B_{11} \cdot A_{12} + B_{12} \cdot A_{22} &= 0_{q\times r} & (1.12) \\ B_{21} \cdot A_{11} + B_{22} \cdot A_{21} &= 0_{r\times q} & (1.13) \\ B_{21} \cdot A_{21} + B_{22} \cdot A_{22} &= I_r & (1.14) \end{aligned}$$

despejando B_{12} y B_{21} de (1.12) y (1.13) respectivamente tenemos:

$$\begin{aligned} B_{12} &= -B_{11} \cdot A_{12} \cdot (A_{22})^{-1} & (1.15) \\ B_{21} &= -B_{22} \cdot A_{21} \cdot (A_{11})^{-1} & (1.16) \end{aligned}$$

luego, (1.15) y (1.16) los sustituimos en (1.11) y (1.14) respectivamente

$$\begin{aligned} B_{11} \cdot (A_{11} - A_{12} \cdot (A_{22})^{-1} \cdot A_{21}) &= I_q \\ \Longrightarrow B_{11} &= (A_{11} - A_{12} \cdot (A_{22})^{-1} \cdot A_{21})^{-1} \\ &= A_{11,2}^{-1} & (1.17) \end{aligned}$$

y

$$\begin{aligned} B_{22} \cdot (A_{22} - A_{21} \cdot (A_{11})^{-1} \cdot A_{12}) &= I_r \\ \Longrightarrow B_{22} &= (A_{22} - A_{21} \cdot (A_{11})^{-1} \cdot A_{12})^{-1} \\ &= A_{22,1}^{-1} & (1.18) \end{aligned}$$

Así

$$\begin{aligned} B_{12} &= -A_{11,2}^{-1} \cdot A_{12} \cdot (A_{22})^{-1} \qquad (1.19)\\ B_{21} &= -A_{22,1}^{-1} \cdot A_{21}(A_{11})^{-1} \qquad (1.20) \end{aligned}$$

Análogamente se puede probar que

$$\begin{aligned} A_{11} &= (B_{11} - B_{12} \cdot (B_{22})^{-1} \cdot B_{21})^{-1} = B_{11,2}{}^{-1}\\ A_{21} &= -B_{22} \cdot B_{21} \cdot (B_{11,2})^{-1}\\ A_{22} &= (B_{22} - B_{21} \cdot (B_{11})^{-1} \cdot B_{12})^{-1})^{-1} = B_{22,1}{}^{-1}\\ A_{12} &= -B_{11} \cdot B_{12} \cdot B_{22,1}^{-1} \end{aligned}$$

Otra forma de expresar B_{21} y B_{22} es la siguiente

$$\begin{aligned} B_{21} &= -(A_{22})^{-1} \cdot A_{21} \cdot B_{11} \qquad (1.21)\\ B_{22} &= A_{22}^{-1} + A_{22}^{-1} \cdot A_{21} \cdot B_{11} \cdot A_{12} \cdot (A_{22})^{-1} \qquad (1.22) \end{aligned}$$

En efecto, como las matrices A y B conmutan

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \cdot \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = I_{(q+r)}$$

obtenemos

$$\begin{aligned} A_{12} \cdot B_{11} + A_{22} \cdot B_{21} &= 0, \quad entonces\\ B_{21} &= -(A_{22})^{-1} \cdot A_{21} \cdot B_{11} \end{aligned}$$

y

$$\begin{aligned} A_{21} \cdot B_{12} + A_{22} \cdot B_{22} &= I, \quad \text{entonces reemplazamos } B_{12}(1{,}19)\ y\ A_{11,2} = B_{11},\ obtenemos\\ B_{22} &= A_{22}^{-1} + A_{22}^{-1} \cdot A_{21} \cdot B_{11} \cdot A_{12} \cdot A_{22}^{-1} \end{aligned}$$

§1.5. Determinante de una matriz particionada por bloques

DEFINICIÓN **1.5.1.** Sean $A' : (q+r) \times (q+r)$y $A'' : (q+r) \times (q+r)$ matrices particionadas por bloques dadas por, $A' = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$ y $A'' = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$, entonces:

$$\begin{aligned} det \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} &= det \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \\ &= |A_{11}| \cdot |A_{22}| \end{aligned} \tag{1.23}$$

donde $A_{11} : q \times q$, $A_{12} : q \times r$, $A_{21} : r \times q$ y $A_{22} : r \times r$

PROPOSICIÓN **1.5.1.** *Sea* $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ *una matriz particionada, con* $|A| \neq 0$

$$detA = |A_{11}| \cdot |A_{22} - A_{21} \cdot (A_{11})^{-1} \cdot A_{12}| \tag{1.24}$$

$$= |A_{22}| \cdot |A_{11} - A_{12} \cdot (A_{22})^{-1} \cdot A_{21}| \tag{1.25}$$

Demostración. Observamos que si multiplicamos las primeras q filas de A por $-A_{21}A_{11}^{-1}$ y se las sumamos a las r filas de A, entonces obtenemos

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21} \cdot (A_{11})^{-1} \cdot A_{12} \end{pmatrix}.$$

Luego, por propiedad de determinante y la definición 1.51

$$\begin{aligned} det \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} &= det \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21} \cdot (A_{11})^{-1} \cdot A_{12} \end{pmatrix} \\ &= |A_{11}| \cdot |A_{22} - A_{21} \cdot (A_{11})^{-1} \cdot A_{12}| \end{aligned}$$

análogamente se demuestra que

$$detA = |A_{22}| \cdot |A_{11} - A_{12} \cdot (A_{22})^{-1} \cdot A_{21}|$$

□

§1.6. Variables aleatorias

La relación entre los sucesos del espacio muestral y el valor numérico que se les asigna se establece a través de variables aleatorias.

Definición **1.6.1.** Una variable aleatoria es una función que asigna un valor numérico a cada suceso elemental del espacio muestral.

Es decir, una variable aleatoria es una variable cuyo valor numérico está determinado por el resultado del experimento aleatorio. La variable aleatoria la notaremos con letras mayúsculas X, Y, ... y con las letras minúscula x, y, ... sus valores.

La v.a. puede tomar un número numerable o no numerable de valores, dando lugar a dos tipos de v.a.: discretas y continuas.

Definición **1.6.2.** Se dice que una variable aleatoria X es discreta si puede tomar un número finito o infinito, pero numerable, de posibles valores.

Definición **1.6.3.** Se dice que una variable aleatoria X es continua si puede tomar un número infinito (no numerable) de valores, o bien, si puede tomar un número infinito de valores correspondientes a los puntos de uno o más intervalos de la recta real.

§1.7. Función de distribución acumulada (Fda)

Sean X, Y dos variables aleatorias definidas conjuntamente, es decir; X e Y tienen una distribución de probabilidad conjunta cuya Función de distribución acumulada (Fda) conjunta está dada por:

$$F(x, y) = P\{X \leq x, Y \leq y\}.$$

De manera general, cuando $X' = (X_1, ..., X_p)$ es un vector de variable aleatorias que son distribuidas conjuntamente la Fda está dada por:

$$F(x) = F(x_1, ..., x_p) = P\{X_1 \leq x_1, ..., X_p \leq x_p\}.$$

Toda Fda multivariada F satisface las siguientes propiedades:

1. F es monótona no-decreciente en cada componente de X

   Basta probar que $F(E) \geq 0$ con $E \subset X \subset \mathbb{R}$
   Sabemos que $P\{a \leq X \leq b\} = F(b) - F(a) = F(E)$
   definiendo $X = \Omega$ (el espacio muestral) tenemos que $E \subset X$ es un evento, luego por axioma de probabilidad

   $$P\{E\} \geq 0 \Rightarrow F(E) \geq 0$$

   Así se cumple lo que se quería probar.

2. $0 \leq F(x) \leq 1$

   Sea $S \subset X = \Omega(evento)$.

   Por lo anterior $F(S) \geq 0$. Falta probar que $F(S) \leq 1$ ahora $\Omega = S \cup S^c$.
   Entonces:

   $$\begin{aligned} P\{\Omega\} &= P\{S \cup S^c\} = P\{S\} + P\{S^c\} = 1 \\ \Longrightarrow &\quad P\{S\} = 1 - P(S^c) \leq 1 \qquad\qquad (ya\ que\ P\{S^c\} \geq 0) \end{aligned}$$

   luego, $0 \leq P\{S\} \leq 1$, y como S es un evento arbitrario, se cumple la propiedad.

3. $F(-\infty, x_2, ..., x_p) = F(x_1, -\infty, ..., x_p) = ... = F(x_1, x_2, ..., -\infty)$

   Sabemos que
   $F(x_1, x_2, ..., x_p) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} ... \int_{-\infty}^{x_p} f(x_1, x_2, ..., x_p) dx_p..dx_2 dx_1$ donde f es la función de densidad de $X = (X_1, X_2, ..., X_p)$, como f es continua

   $$\begin{aligned} F(-\infty, x_2, ..., x_p) &= \int_{-\infty}^{-\infty} \int_{-\infty}^{x_2} ... \int_{-\infty}^{x_p} f(x_1, x_2, ..., x_p) dx_p..dx_2 dx_1 \\ &= \int_{-\infty}^{x_p} \int_{-\infty}^{x_2} ... \int_{-\infty}^{-\infty} f(x_1, x_2, ..., x_p) dx_1..dx_2 dx_p \\ &= 0 \end{aligned}$$

Análogamente se prueba para $F(x_1, -\infty, ..., x_p)$

$$\therefore F(-\infty, x_2, ..., x_p) = F(x_1, -\infty, ..., x_p) = ... = F(x_1, x_2, ..., -\infty)$$

4. $F(\infty, \infty, ..., \infty) = 1$

$$\begin{aligned} F(\infty, \infty, ..., \infty) &= \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x_1, ..., x_p) dx_p ... d_1 \\ &= 1 \end{aligned}$$

ya que f es una función de densidad de X.

5. La probabilidad de un rectángulo de dimensión p es no-negativo. Probaremos para $p = 2$

$$\begin{aligned} P\{x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2\} &= F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \\ &\geq 0 \end{aligned}$$

Todas las propiedades son análogas al caso univariado excepto la última propiedad, existen funciones que cumplen las primeras 4 propiedades y la última no la cumple, así no son fdc.

Ejemplo **1.7.1.** Supongamos que

$$F(x_1, x_2) = \begin{cases} 0 \text{ , si } x_1 \leq 0 \text{ o } x_2 \leq 0 \; x_1 + x_2 \leq 1; \\ 1, \text{ en otro caso.} \end{cases}$$

esta función satisface las primeras 4 propiedades, lo cual es suficiente para una fdc en el caso univariado, pero dado que

$$F(1, 1) - F(1, \frac{1}{2}) - F(\frac{1}{2}, 1) + F(\frac{1}{2}, \frac{1}{2}) = -1$$

no es un fdc ya que no cumple la última propiedad, así $F(x_1, x_2)$ no puede ser una fdc bivariada.

Asumiremos que todas las funciones $F(x)$ serán continuas, en consecuencia ésta será expresada como la integral de una función $f(x)$ llamada densidad, es decir:

$$F(X) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} ... \int_{-\infty}^{x_p} f(x)dx \tag{1.26}$$

§1.8. Densidad

Supongamos que $F(X)$ es continua; entonces de (1.26) la función de densidad conjunta (fdc) de X es:

$$f(x) = f(x_1, ..., x_p) = \frac{\partial^p F(x)}{\partial x_1 ... \partial x_p} \tag{1.27}$$

Hay conjuntos donde los valores de x en (1.27) no existen. Análogo al caso univariado, esto es una relación para la probabilidad de un evento (o conjunto de valores en el espacio de dimensión p) en términos de la densidad conjunta para $X : p \times 1$

$$P\{X \subseteq R\} = \int_R ... \int f(x)dx \tag{1.28}$$

para una región R.

§1.9. Distribución marginal

En el análisis de datos multivariado , es típico comenzar con un vector con muchas componentes y; luego, encontrar posteriormente un subvector de interés. En tal caso, la distribución marginal de los subvectores es importante para la inferencia proporcional.

Sea $X' = (Y', Z')$, donde Y y Z son subvectores de $X : p \times 1$. [por ejemplo, $Y' \equiv (X_1, X_2), Z' \equiv (X_3, ..., X_p)$] entonces, si $g(y), h(z)$ denotan las densidades de Y, Z respectivamente, y si $f(x) = f(y, z)$ denota la densidad de X se tiene que:

$$g(y) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(y, z)dz \tag{1.29}$$

y

$$h(z) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(y, z)dy \tag{1.30}$$

donde todas las integrales son tomadas sobre $(\infty, -\infty), g(y)$ y $h(z)$ son llamadas las densidades marginales de Y y Z.

§1.10. Distribución condicional

La distribución condicional es de interés, y ocurre cuando un grupo de variables aleatorias están siendo estudiadas mientras un segundo grupo se mantiene fijo.
Sean A y B dos eventos que pueden ocurrir en un espacio de 2-dimensiones, entonces por definición, la probabilidad condicional de B dado A ésta dada por:

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

si $P(A) \neq O$. Si A es un evento donde la variable aleatoria X está en el intervalo $a \leq X \leq b$, y B es un evento donde la variable aleatoria Y esta en el intervalo $c \leq Y \leq d$, entonces

$$P\{c \leq Y \leq d | a \leq X \leq b\} = \frac{P\{a \leq X \leq b, c \leq Y \leq d\}}{P\{a \leq X \leq b\}}$$

y por (1.28)

$$P\{c \leq Y \leq d | a \leq X \leq b\} = \frac{\int_c^d \int_a^b f(x,y)dxdy}{\int_a^b g(x)dx}$$

donde $f(x, y)$ es la densidad conjunta de X, Y y $g(x)$ es la densidad marginal de X.
La densidad condicional de Y dado $X = x$ está definida como:

$$h(y|x) = \frac{f(x,y)}{g(x)}. \tag{1.31}$$

Así,

$$P\{c \leq Y \leq d | X = x\} = \int_c^d h(y|x)dy. \tag{1.32}$$

Generalizado a una dimensión p, sean $X' = (X_1, ..., X_p), Y' = (X_1, ..., X_k)$ y $Z' = (X_{k+1}, ..., X_p)$ los vectores aleatorios y con letra minúscula denotaremos los valores observados. La densidad condicional de Y dado Z está dada por:

$$g(Y|Z) = \frac{f(y,z)}{h(z)} = \frac{f(x)}{h(z)} \tag{1.33}$$

donde $f(x)$ denota la densidad del vector aleatorio X, y $h(z)$ denota la densidad marginal del vector Z.

§1.11. Independencia

Dos vectores aleatorios, Y, Z; se dicen que son independiente si una de las siguientes aplicaciones se cumplen:

$$f(y,z) = g(y)h(z) \tag{1.34}$$

ó

$$F(y,z) = G(y)H(z) \tag{1.35}$$

ó

$$P\{Y|Z\} = g(y) \tag{1.36}$$

donde $f(y,z), g(y), h(z)$ son la densidad de $X = (Y,Z), Y, Z$ respectivamente; F,G,H son la respectivas Fda, y $P(y|z)$ es la densidad condicional de $Y|Z$.

§1.12. Esperanza

Sea $X : p \times 1$ un vector columna con $X_i, i = 1, ..., p$ componentes aleatorias; donde $f(X) = f(x_1, ..., x_p)$ es la función de densidad conjunta
Cuando ésta existe, la esperanza de un vector X está definido como:

$$E(X) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix} \tag{1.37}$$

Análogamente, si $V : p \times n$, $E(V) = (E(V_{ij}))$, donde $V = (V_{ij})$

§1.12.1. Momento de segundo orden

La covarianza entre dos variables aleatorias Y y Z con momento de segundo orden finito está definido como:

$$Cov(Y, Z) = E[(Y - EY)(Z - EZ)] \tag{1.38}$$

ésto cuantitativamente puede ser positivo, negativo o cero; la covarianza matricial de un vector X está dada por la siguiente definición X:

$$\Sigma = (\sigma_{ij}) = E[(X - EX)(X - EX)'] \tag{1.39}$$

para $i, j = 1, ..., p$. Un elemento típico de Σ es $\sigma_{ij} = E(X_i - EX_i)(X_j - EX_j)$, $i, j = 1, .., p$ cuando $j = i$ los elementos están ubicados a lo largo de la diagonal de Σ y es llamada la varianza de X.
Recordemos que

$$VarX_i = E(X_i - EX_i)^2$$

si $i \neq j$, σ_{ij} es la covarianza de X_i y X_j. El coeficiente de correlación entre dos variables aleatorias escalares Y y Z con momento de segundo orden finito está definido

$$\rho = corr(Y, Z) = \frac{Cov(Y, Z)}{[(VarY)(VarZ)]^{1/2}}. \tag{1.40}$$

Esta es una medida de causa y efecto asociada con Y y Z. En general, $-1 \leq \rho \leq 1$, aunque en algunos casos, ρ es restringido a un intervalo más pequeño.

Una matriz de correlación es una matriz de coeficiente de correlación $R = (\rho_{ij})$ $i, j = 1, ..., p$. Esta matriz es útil para estudiar todas las asociaciones entre las componentes de un vector de variables aleatorias simultáneamente. La matriz de correlación es calculada en muchos modelos usados en análisis de datos multivariado ya que la matriz R frecuentemente provee un rápido entendimiento dentro de muchas relaciones insospechadas.

Los elementos de la diagonal, ρ_{jj}, de una matriz de correlación deberían ser todos uno, y los elementos fuera de la diagonal dados por:

$$\rho_{ij} = Corr(x_i, x_j) = \frac{cov(x_i, x_j)}{[Var(x_i)Var(x_j)]^{1/2}}, i \neq j$$

además, los ρ_{ij} deberían también satisfacer siempre la inecuación $-1 \leq \rho_{ij} \leq 1$ para $i, j = 1, ..., p$

§1.12.2. Transformaciones lineales

Sea $Y = AX + b$ una transformación lineal de una variable aleatoria $X : p \times 1$ con $A : k \times p$ y $b : k \times 1$ así Y es un vector $k \times 1$, $k \leq p$. Sean Σ_y y Σ_x la covarianzas de X y Y respectivamente entonces:

$$\Sigma_y = A\Sigma_x A',$$

en efecto

$$\begin{aligned}
Var(Y) &= E(Y - EY)(Y - EY)' \\
&= E[(AX + b) - E(AX + b)][(AX + b) - E(AX + b)]' \\
&= E[(AX + b) - AE(X) - b][(AX + b) - AE(X) - b)]' \\
&= E[A(X - E(X))][A(X - E(X))]' \\
&= E[A(X - E(X))][X - E(X)]'A' \\
&= AE[(X - EX)(X - EX)']A' \\
&= AVar(X)A'.
\end{aligned}$$

Ahora supongamos k=1 así A es un vector fila. Entonces Σ_y es un escalar y $Var(Y) = \Sigma_y$. Pero una varianza debería ser no-negativa. Dado que $A\Sigma_x A' \geq 0$ para todo A, Σ_x debería ser definida semi-positiva. Como Σ_x es arbitrario, el resultado puede cumplirse para toda matriz de covarianza.

§1.13. Transformaciones

Es útil examinar de nuevo un problema en términos de un conjunto de transformaciones de variables aleatorias. Si esa transformación existe, el siguiente teorema demuestra como la densidad es afectada.

Antes de enunciar el teorema, tenemos el siguiente lema el cual se usa para demostrar dicho teorema.

Lema 1. *Sea $g : A \to B$ un difeomorfismo de Abiertos de $\mathbb{R}^n$. Entonces para toda función $f : B \to \mathbb{R}$ que es integrable sobre B, la función $(f \circ g)|detDg|$ es integrable sobre A, y*

$$\int_B f = \int_A (f \circ g)|detDg| \quad \textit{donde } Dg \textit{ es la derivada de } g$$

Demostración. ver [9] □

Teorema **1.13.1.** *(Transformación) Sea* $Y_k = f_k(X_1, ..., X_p), k = 1, ..., p$ *una transformación 1-1 de* $Y = (Y_1, ..., Y_p)'$ *en* $X = (X_1, ..., X_p)$ *donde* $X_k = g_k(Y), k = 1, ..., p$ *denota la transformación inversa* $Y_k, k = 1, ..., p$*. Si* $p(x) = p(x_1, ..., x_p)$ *es la densidad de X, la densidad de* Y *está dada por*

$$g(y) = p[g_1(y), ..., g_p(y)]J(X \to Y)$$

donde $J(X \to Y)$ *denota el Jacobiano de la transformación*

Demostración. Por hipótesis Y_k y X_k son variables aleatorias diferenciables e inyectivas, $k = 1, 2, .., p$.

Observemos entonces que tanto Y_k como X_k $k = 1, 2, ..., p$ son de clase C^1. Luego, por el lema 1, tenemos:

$$\int_X P(x)dx = \int_Y p[g_1(y), ..., g_p(y)]J(X \to Y)dy$$

Así,

$$g(y) = p[g_1(y), ..., g_p(y)]J(X \to Y) \tag{1.41}$$

□

Nota: Si $Y_k = f_k(X)$, $k = 1, ..., p$ denota una transformación diferenciable con múltiples preimágenes, la densidad de Y puede ser obtenida aplicando (1.41) a cada solución de la trasformación inversa separadamente y entonces sumamos las densidades transformadas para cada solución.

Capítulo 2

Inferencia Bayesiana

§2.1. Teorema de Bayes, distribución a priori y posteriori

Sean X, Θ vectores aleatorios de dimensión p y k respectivamente que están distribuidos conjuntamente con densidad condicional de X dado Θ denotada por $f(x|\theta)$, y la densidad marginal de Θ dada por $g(\theta)$; los datos han sido generados al observar X, para algún Θ fijo no-observados. Nos gustaría hacer inferencia sobre Θ considerando tanto nuestro prejuicio (creencia priori), como también las observaciones de $(X|\Theta)$ que indirectamente relacionan a ésta. El teorema de bayes proporciona un mecanismo formal para llevar a cabo esto. En terminología bayesiana $g(\theta)$ es llamada la priori de Θ o densidad a priori Θ ya que ésta es la densidad de Θ priori a los datos observados.

Teorema **2.1.1.** *(Bayes)*

Sean $g_1(\theta)$ la densidad a priori de Θ, y $f(x|\theta)$ la densidad condicional de X dado θ, entonces la densidad de Θ dado X=x, es dada por:

$$h(\theta|x) = \frac{f(x|\theta)g_1(\theta)}{\int_{-\infty}^{+\infty} f(x|\theta)g_1(\theta)d\theta}. \tag{2.1}$$

Demostración. Sean z y g_2 las funciones de densidad de (X, θ) y de X respectivamente. Por la definición de la densidad condicional

$$f(x|\theta) = \frac{z(x,\theta)}{g_1(\theta)} \tag{2.2}$$

$$h(\theta|x) = \frac{z(x,\theta)}{g_2(x)} \tag{2.3}$$

despejando $z(x,\theta)$ de (2.2) y reemplazando en (2.3), obtenemos:

$$h(\theta|x) = \frac{f(x|\theta)g_1(\theta)}{g_2(x)} \tag{2.4}$$

por otra parte tenemos que:

$$\begin{aligned} g_2(x) &= \int_{-\infty}^{\infty} z(x,\theta)d\theta \\ &= \int_{-\infty}^{\infty} f(x|\theta)g_1(\theta)d\theta \end{aligned}$$

luego, sustituyendo $g_2(x)$ en (2.4)

$$h(\theta|x) = \frac{f(x|\theta)g_1(\theta)}{\int_{-\infty}^{+\infty} f(x|\theta)g_1(\theta)}d\theta.$$

□

Observemos que en (2.1) la integral depende solo de x y esté es fijo y conocido, así dicha integral es constante. Por lo tanto:

$$h(\theta|x) \propto f(x|\theta)g_1(\theta).$$

Una distribución para $\Theta : k \times 1$ posee todas la propiedades habituales de distribuciones de variables aleatorias observadas, excepto que Θ no es observada. Este tipo de distribuciones son llamadas distribuciones de probabilidad subjetiva.

EJEMPLO **2.1.1.** Se desea estimar la probabilidad, θ, de un evento, a parti del resultado de una sucesión de n *ensayos Bernoulli*, esto es, datos $x_1, x_2, ..., x_n$ que son iguales a uno si ocurre el evento (éxito) y cero si no ocurre. Sea x el número total de éxitos en la muestra de n ensayos. En este caso, el modelo muestral establece que:

$$f(x|\theta) = Bin(y|n,\theta) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}$$

Suponiendo que $g_1(\theta)$ es uniforme en el intervalo [0,1], se tiene que:

$$h(\theta|x) \propto \theta^x (1-\theta)^{n-x}$$

o lo que es lo mismo, la distribución no normalizada tiene un kernel equivalente a la distribución *Beta*, es decir:

$$\theta|x \sim Beta(x+1, n-x+1)$$

§2.2. Densidad general de la distribución normal multivariada

Sea $X : p \times 1$ un vector aleatorio con función de densidad $f(x)$. X tiene una distribución Normal multivariada (p variada) no-singular con vector de media $\theta : p \times 1$ y matriz covarianza $\Sigma : p \times p$ si

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\theta)'\Sigma^{-1}(x-\theta)\right] \tag{2.5}$$

para $\Sigma > 0$. Si $|\Sigma| = 0$, la distribución de X es llamada singular o normal degenerada y la densidad no existe.
Denotaremos esta distribución de X por:

$$\mathcal{L}(X) = N(\theta, \Sigma).$$

§2.2.1. Estandarización

Si $\mathcal{L}(X) = N(\theta, \Sigma)$, la distribución de X puede ser estandarizada por la transformación $Y = \Sigma^{-1/2}(X-\theta)$; es decir, $\mathcal{L}(Y) = N(0, I)$.

Observemos que la transformación $Y = \Sigma^{-1/2}(x-\theta)$ es inyectiva y diferenciable, la inversa existe, es diferenciable y es dada por $Y\Sigma^{1/2} + \theta = X$; además
$J(X \rightarrow Y) = |\Sigma|^{1/2}$.
Luego, por el teorema 1.13.1:

$$f(y) = \frac{1}{(2\pi)^{p/2}} exp\left[-\frac{1}{2}y'y\right]$$

$\therefore \mathcal{L}(Y) = N(0, I)$.

§2.2.2. Distribución normal bivariada

Sea $X : 2 \times 1$ un vector aleatorio bivariado con $\mathcal{L}(X) = N(\theta, \Sigma)$ y $\Sigma > 0$. Sean $\theta = (\theta_i)$y $\Sigma = (\sigma_{ij})$ i,j=1,2. Para simplificar tomemos $\sigma_{11} = \sigma_1^2, \sigma_{22} = \sigma_2^2$ y $\sigma_{12} = \rho\sigma_1\sigma_2$ donde ρ es el coeficiente de correlación entre X_1 y X_2. Si escribimos (2.5) para $p = 2$ tenemos que la densidad bivariada es dada por:

$$\begin{aligned} f(x) &= f(x_1, x_2) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \qquad (2.6) \\ &\quad exp\left\{-\frac{1}{2(1-\rho^2)}\left[(\frac{x_1-\theta_1}{\sigma_1})^2 - 2\rho(\frac{x_1-\theta_1}{\sigma_1})(\frac{x_2-\theta_2}{\sigma_2}) + (\frac{x_2-\theta_2}{\sigma_2})^2\right]\right\}. \end{aligned}$$

Acá

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\theta_2 & \sigma_2^2 \end{pmatrix} \qquad \Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2(1-\rho^2)} & -\frac{\rho}{\sigma_1\sigma_2(1-\rho^2)} \\ -\frac{\rho}{\sigma_1\sigma_2(1-\rho^2)} & \frac{1}{\sigma_2^2(1-\rho^2)} \end{pmatrix}.$$

La expresión entre corchetes de (2.6) controla la varianza de $f(x)$, es decir; si la expresión entre corchetes es constante entonces $f(x)$ es constante y viceversa. La referencia (2.6) muestra que si f(x) es constante entonces los puntos (x_1, x_2) están a lo largo de la elipse concéntrica con centro (θ_1, θ_2) con un menor y mayor ángulo respecto a los eje (x_1, x_2), los ángulos depende de los valores $(\rho, \sigma_1, \sigma_2)$. Si $\rho = 0$ las elipse son paralelas a los ejes coordenadas. Así, los entornos de la densidad de distribución normal bivariada son elipses (y más general las una distribución multivariada son hiperelipses). Los entornos son a menudo muy útiles para estudiar el comportamiento de una distribución en dos o tres dimensiones.

§2.2.3. Independencia

Sea $\mathcal{L}(X) = N(\theta, \Sigma)$ y $X : 2\times1$. Entonces, si $\rho = 0$ en (2.6), X_1 y X_2 no sólo están no-correlacionados, también son independientes. Es fácil ver que al sustituir $\rho = 0$ en (2.6) $f(x_1, x_2)$ se reduce al producto de una función de x_1 y una función de x_2. Por supuesto, lo contrario es también cierto, es decir; sí X_1 y X_2 son independientes, entonces X_1 y X_2 son no-correlacionados; en este sentido, el resultado se cumple para toda distribución bivariada (mientras que en el otro sentido, la falta de correlación generalmente no implica la independencia, aunque si para la distribución normal)).

§2.3. Distribución marginal de la Normal

Sea X un vector aleatorio y suponga $\mathcal{L}(X) = N(\theta, \Sigma)$. Entonces la distribución marginal de un subvector es también una distribución normal. Específicamente, sean

$$X = \begin{pmatrix} Y \\ Z \end{pmatrix} \quad , \quad \Theta = \begin{pmatrix} \theta_y \\ \theta_z \end{pmatrix} \quad y \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

donde $Y : q \times 1$, $Z : r \times 1$, $\theta_y : q \times 1$, $\theta_z : r \times 1$, $\Sigma_{11} : q \times q$, $\Sigma_{22} : r \times r$, $\Sigma_{12} : q \times r$ y $\Sigma_{21} : r \times q$ con $r + q = p$
Definamos la forma cuadrática:

$$Q(y, z) = [(y - \theta_y), (z - \theta_z)]' \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix} \begin{bmatrix} (y - \theta_y) \\ (z - \theta_z) \end{bmatrix}$$

donde

$$\Sigma^{-1} = \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix}$$

.

Ahora, sustituimos Q en (2.6)y obtenemos

$$f(y, z) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} exp \left[-\frac{1}{2} Q(y, z) \right]$$

usando el hecho de que Σ^{-1} es simétrica Q se puede escribir como:

$$Q(y, z) = (y - \theta_y)' \Sigma^{11} (y - \theta_y) + (z - \theta_z)' \Sigma^{22} (z - \theta_z) + 2(y - \theta)' \Sigma^{12} (z - \theta_z).$$

De acá, sí $g(y)$ denota la densidad marginal de Y, entonces

$$\begin{aligned} g(y) &= \int f(y, z) dz \\ &= \frac{H}{(2\pi)^{p/2} |\Sigma|^{1/2}} \\ &\quad \cdot exp \left\{ -\frac{1}{2} \left[(y - \theta_y)' \Sigma^{11} (y - \theta_y) + \theta_z' \Sigma^{22} \theta_z - 2(y - \theta_y)' \Sigma^{12} \theta_z \right] \right\} \end{aligned} \tag{2.7}$$

donde

$$H = \int exp\left\{-\frac{1}{2}\left[z'\Sigma^{22}z - 2(\theta_z'\Sigma^{22} + \theta_y'\Sigma^{12} - y'\Sigma^{12})z\right]\right\}dz$$

veamos que

$$H = (2\pi)^{r/2}|\Sigma^{22}|^{-1/2}exp\left(\frac{a'\Sigma^{22}a}{2}\right)$$

donde $a = \theta_z - (\Sigma^{22})^{-1}\Sigma^{21}(y-\theta_y)$.Definamos $K_1 = z'\Sigma^{22}z - 2(\theta_z'\Sigma^{22} + \theta_y'\Sigma^{12} - y'\Sigma^{12})z$. Estudiemos K_1 usando la simetría de Σ^{-1}

$$\begin{aligned}
K_1 &= z'\Sigma^{22}z - 2(\theta_z'\Sigma^{22} + \theta_y'\Sigma^{22} - y'\Sigma^{12})z \\
&= z'\Sigma^{22}z - \theta_z'\Sigma^{22}z - \theta_y'\Sigma^{12}z + y'\Sigma^{12}z - \theta_z'\Sigma^{22}z - \theta_y'\Sigma^{12}z + y'\Sigma^{12}z \\
&= z'\Sigma^{22}z - z'\Sigma^{22}\theta_z - z'\Sigma^{21}\theta_y + z'\Sigma^{21}y - \theta_z'\Sigma^{22}z - \theta_y'\Sigma^{12}z + y'\Sigma^{12}z - \theta_z'\Sigma^{22}z \\
&= z'\Sigma^{22}z - z'\Sigma^{22}\theta_z + z'\Sigma^{21}(y-\theta_y) + (y-\theta_y)'\Sigma^{12}z - \theta_z'\Sigma^{22}z \\
&= z'\Sigma^{22}z - z'\Sigma^{22}\theta_z + z'\Sigma^{21}(y-\theta_y) + (y-\theta_y)'\Sigma^{12}[\Sigma^{22}]^{-1}\Sigma^{22}z - \theta_z'\Sigma^{22}z \\
&= z'\Sigma^{22}z - z'\Sigma^{22}\theta_z + z'\Sigma^{21}(y-\theta_y) + \{[\Sigma^{22}]^{-1}\Sigma^{21}(y-\theta_y)\}'\Sigma^{22}z - \theta_z'\Sigma^{22}z \\
&= z'\Sigma^{22}z - z'\Sigma^{22}\theta_z + z'\Sigma^{21}(y-\theta_y) + \{[\Sigma^{22}]^{-1}\Sigma^{21}(y-\theta_y) - \theta_z\}'\Sigma^{22}z \\
&= z'\Sigma^{22}z + z'\Sigma^{22}\{[\Sigma^{22}]^{-1}\Sigma^{21}(y-\theta_y) - \theta_z\} + \{[\Sigma^{22}]^{-1}\Sigma^{21}(y-\theta_y) - \theta_z\}'\Sigma^{22}z \\
&= z'\Sigma^{22}z - z'\Sigma^{22}a - a'\Sigma^{22}z + a'\Sigma^{22}a - a'\Sigma^{22}a \\
&= (z-a)'\Sigma^{22}(z-a) - a'\Sigma^{22}a.
\end{aligned}$$

Así,

$$\begin{aligned}
H &= \int exp\left\{-\frac{1}{2}[(z-a)'\Sigma^{22}(z-a) - a'\Sigma^{22}a]\right\}dz \\
&= \frac{1}{(2\pi)^{r/2}|\Sigma^{22}|^{-1/2}}\int exp\left\{-\frac{1}{2}[(z-a)'\Sigma^{22}(z-a)\right\}dz(2\pi)^{r/2}|\Sigma^{22}|^{-1/2}exp\left(\frac{a'\Sigma^{22}a}{2}\right) \\
&= (2\pi)^{r/2}|\Sigma^{22}|^{-1/2}exp\left(\frac{a'\Sigma^{22}a}{2}\right) \qquad \textit{ya que es } z \sim N(a, (\Sigma^{22})^{-1})
\end{aligned}$$

ahora sustituimos H en (2.7)

$$\begin{aligned}
g(y) &= \frac{|\Sigma^{22}|^{-1/2}}{(2\pi)^{q/2}|\Sigma|^{1/2}} \\
&\cdot \exp\left\{-\frac{1}{2}[(y-\theta_y)'\Sigma^{11}(y-\theta_y) + \theta_z'\Sigma^{22}\theta_z - 2(y-\theta_y)'\Sigma^{12}\theta_z - a'\Sigma^{22}a]\right\}
\end{aligned} \quad (2.8)$$

Ahora estudiemos, $K_2 = (y-\theta_y)'\Sigma^{11}(y-\theta_y) + \theta_z'\Sigma^{22}\theta_z - 2(y-\theta_y)'\Sigma^{12}\theta_z - a'\Sigma^{22}a$

$$
\begin{aligned}
K_2 &= (y-\theta_y)'\Sigma^{11}(y-\theta_y) + \theta_z'\Sigma^{22}\theta_z - 2(y-\theta_y)'\Sigma^{12}\theta_z - a'\Sigma^{22}a \\
&= (y-\theta_y)'\Sigma^{11}(y-\theta_y) + \theta_z'\Sigma^{22}\theta_z - 2(y-\theta_y)'\Sigma^{12}\theta_z - \\
&\quad \{\theta_z - [\Sigma^{22}]^{-1}\Sigma^{21}(y-\theta_y\}'\Sigma^{22}\{\theta_z - [\Sigma^{22}]^{-1}\Sigma^{21}(y-\theta_y)\} \\
&= (y-\theta_y)'\Sigma^{11}(y-\theta_y) + \theta_z'\Sigma^{22}\theta_z - 2(y-\theta_y)'\Sigma^{12}\theta_z \\
&\quad -\{\theta_z' - (y-\theta_y)'\Sigma^{12}[\Sigma^{22}]^{-1}\}'\Sigma^{22}\{\theta_z - [\Sigma^{22}]^{-1}\Sigma^{21}(y-\theta_y)\} \\
&= (y-\theta_y)'\Sigma^{11}(y-\theta_y) + \theta_z'\Sigma^{22}\theta_z - 2(y-\theta_y)'\Sigma^{12}\theta_z - \theta_z'\Sigma^{22}\theta_z \\
&\quad +(y-\theta_y)'\Sigma^{12}\theta_z + \theta_z\Sigma^{21}(y-\theta_y) - (y-\theta_y)'\Sigma^{12}(\Sigma^{22})^{-1}\Sigma^{21}(y-\theta_y) \\
&= (y-\theta_y)'(\Sigma^{11} - \Sigma^{12}(\Sigma^{22})^{-1}\Sigma^{21})(y-\theta_y) \\
&= (y-\theta_y)'\Sigma_{11}^{-1}(y-\theta_y) \qquad\qquad \text{(por (1.17))}
\end{aligned}
$$

Usando ésta última ecuación en (2.8) y la proposición 1.5.1, obtenemos que:

$$\mathcal{L}(Y) = N(\theta_y, \Sigma_{11})$$

análogamente se prueba que

$$\mathcal{L}(Z) = N(\theta_z, \Sigma_{22}) \tag{2.9}$$

§2.4. Distribución condicional normal

TEOREMA **2.4.1.** *Sea $X : p \times 1$ un vector aleatorio donde $\mathcal{L}(X) = N(\theta, \Sigma)$ y $X = (Y, Z)$. Entonces la distribución condicional de Y dado $Z = z$ es también normal donde el vector esperanza es una función lineal de z y la matriz de covarianza independiente de z, es decir; la distribución condicional Y dado $Z = z$ esta dada por:*

$$\mathcal{L}(Y|Z = z) = N(\theta_y + \Sigma_{12}(\Sigma^{22})^{-1}(z-\theta_z), \Sigma_{11,2})$$

donde

$$\Sigma_{11,2} = \Sigma_{11} - \Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{21}$$

Demostración. Sean $f(x)$ y $g(z)$ la función de densidad de X y Z dadas por (2.5) y (2.9) respectivamente. Por otra parte recordamos que la función condicional $H(Y|Z = Z)$ es dada por la ecuación (1.31). Así:

$$\begin{aligned}
h(Y|Z=z) &= \frac{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\Big\{ -\frac{1}{2}\,[(y-\theta_y),(z-\theta_z)]' \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix} \begin{bmatrix} (y-\theta_y) \\ (z-\theta_z) \end{bmatrix} \Big\}}{\frac{1}{(2\pi)^{r/2}|\Sigma_{22}|^{1/2}} exp\Big\{ -\frac{1}{2}[z-\theta_z]'{\Sigma_{22}}^{-1}[z-\theta] \Big\}} \\
&= \frac{|\Sigma_{22}|^{1/2}}{(2\pi)^{q/2}|\Sigma|^{1/2}} \\
& \quad exp\Big\{ -\frac{1}{2}\Big([(y-\theta_y).(z-\theta_z)]' \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix} \begin{bmatrix} (y-\theta_y) \\ (z-\theta_z) \end{bmatrix} - (z-\theta_z)'{\Sigma_{22}}^{-1}(z-\theta) \Big) \Big\}
\end{aligned}$$

Usemos la expresión (1.15),(1.21) y (1.22) para simplificar la siguiente ecuación:

$$\begin{aligned}
& [(y-\theta_y),(z-\theta_z)]' \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix} \begin{bmatrix} (y-\theta_y) \\ (z-\theta_z) \end{bmatrix} - (z-\theta_z)'{\Sigma_{22}}^{-1}(z-\theta) \\
= \; & (y-\theta_y)'\Sigma^{11}(y-\theta_y) + (z-\theta_z)'\Sigma^{21}(y-\theta_y) + (y-\theta_y)'\Sigma^{12}(z-\theta_z) \\
& +(z-\theta_z)'\Sigma^{22}(z-\theta_z) - (z-\theta_z)'{\Sigma_{22}}^{-1}(z-\theta) \\
= \; & (y-\theta_y)'\Sigma^{11}(y-\theta_y) - (z-\theta_z)'(\Sigma_{22})^{-1}\Sigma_{21}\Sigma^{11}(y-\theta_y) \\
& -(y-\theta_y)\Sigma^{11}\Sigma_{12}(\Sigma_{22})^{-1}(z-\theta_z) - (z-\theta_z)'{\Sigma_{22}}^{-1}(z-\theta) \\
& +(z-\theta_z)'{\Sigma_{22}}^{-1}(z-\theta) + (z-\theta_z)(\Sigma_{22})^{-1}\Sigma_{21}\Sigma^{11}\Sigma_{12}(\Sigma_{22})^{-1}(z-\theta_z) \\
= \; & (y-\theta_y)'\Sigma^{11}(y-\theta_y) - (z-\theta_z)'(\Sigma_{22})^{-1}\Sigma_{21}\Sigma^{11}(y-\theta_y) - (y-\theta_y)\Sigma^{11}\Sigma_{12}(\Sigma_{22})^{-1}(z-\theta_z) \\
& +(z-\theta_z){\Sigma_{22}}^{-1}\Sigma_{12}\Sigma^{11}\Sigma_{12}(\Sigma_{22})^{-1}(z-\theta_z) \\
= \; & [(y-\theta_y)' - (z-\theta_z)'(\Sigma_{22})^{-1}\Sigma_{21}]\Sigma^{11}(y-\theta_y) - \\
& [(y-\theta_y)' - (z-\theta_z)'(\Sigma_{22})^{-1}\Sigma_{21}]\Sigma^{11}\Sigma_{12}(\Sigma_{22})^{-1}(z-\theta_z) \\
= \; & [y-(\theta_y+\Sigma_{12}(\Sigma_{22})^{-1}z-\theta_z)]'\Sigma^{11}[y-(\theta_y+\Sigma_{12}(\Sigma_{22})^{-1}z-\theta_z)]
\end{aligned}$$

luego, por (1.17) y (1.25) obtenemos:

$$\mathcal{L}(Y|Z=z) = N(\theta_y + \Sigma_{12}(\Sigma^{22})^{-1}(z-\theta_z), \Sigma_{11,2})$$

□

§2.5. Principio de verosimilitud

En el caso en que y se conozca o se considere fijo, $p(y|\theta)$, como función de θ, se le conoce como la *función de verosimilitud*.

La inferencia Bayesiana obedece el principio de verosimilitud; es decir, para una muestra de datos, dos modelos $p(y|\theta)$ con la misma función de verosimilitud producen la misma inferencia sobre θ, o dicho de otra manera, que la función de verosimilitud contiene toda la información relevante que aportan los datos. Usar el principio de verosimilitud es aceptar que la inferencia está *condicionada* en los datos observados, pues la verosimilitud está parametrizada por los datos. Esto contrasta con la inferencia basada en las distribuciones de muestreo, donde se considera un estimador $\widehat{\theta} = f(y)$, el cual, de acuerdo con el tipo de experimento de muestreo, tiene una distribución muestral que resume las propiedades de estimador previo a la observación de los datos y por tanto, irrelevante para hacer inferencias luego que se han observado los datos.

§2.6. Distribución Wishart

§2.6.1. Densidad

Sea $V : p \times p$ una matriz simétrica y definida positiva, la matriz aleatoria V se dice que tiene una distribución Wishart no-singular p-dimensional con matriz escalar Σ, y n grados de libertad, $p \leq n$, si la distribución conjunta de los distintos elementos de V son continua con función de densidad:

$$p(V) = \frac{c|V|^{(n-p-1)/2}}{|\Sigma|^{n/2}} \exp\left(-\frac{1}{2}tr\Sigma^{-1}V\right) \qquad V > 0, \Sigma > 0 \tag{2.10}$$

y $p(V) = 0$, en otro caso; donde c es un número constante definido por:

$$c = \left[2^{np/2}\pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\left(\frac{n+1-j}{2}\right)\right]^{-1}.$$

Sí $n < p$, la distribución es singular y no es una densidad. Así, sí $V = v_{ij}$ y $\Sigma^{-1} = (\sigma^{ij})$, para $V > 0$

$$p(V) \propto \frac{|V|^{(n-p-1)/2}}{|\Sigma|^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} v_{ij}\sigma^{ij}\right).$$

Esta relación será expresada como:

$$\mathcal{L}(V) = W(\Sigma, p, n).$$

§2.7. Distribución inversa Wishart

§2.7.1. Densidad

Sea $U : p \times p$ una matriz aleatoria que tiene una distribución inversa Wishart con matriz escalar G definida positiva , y n grados de libertad; entonces para $2p < n$, la densidad de U está dada por:

$$p(U) = \frac{c_o |G|^{(n-p-1)/2}}{|U|^{n/2}} \exp\left(-\frac{1}{2} trU^{-1}G\right), \qquad U > 0$$

y $p(U) = 0$ en otro caso la relación será denotada por

$$\mathcal{L}(U) = W^{-1}(G, n, p).$$

La constante c_0 está dada por:

$$c_0^{-1} = 2^{(n-p-1)p/2} \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{n-p-1}{2}\right).$$

§2.8. Distribución a priori no-informativa

Supongamos que $\mathcal{L}(X|\Theta, \Sigma) = N(\theta, \Sigma)$. El problema de la subjetividad es para estimar una distribución a priori para (θ, Σ). Primero consideramos θ_j, una componente de θ *donde* $-\infty < \theta_j < \infty$, $j = 1, ..., p$.

Una manera de expresar la información vaga es usando la noción de que todos los valores de las variables aleatorias, pueden estar sobre la recta real, las cuales son igualmente probables. Así suponemos que la distribución de una variable aleatoria es uniforme en el intervalo (a, b), donde a y b son valores muy grandes negativo y positivo respectivamente. Sabemos que la distribución a posteriori $h(\theta|x) \propto f(x|\theta)g(\theta)$, es dada por el producto de la densidad a priori y la densidad condicional de la variable aleatoria observadas (función de verosimilitud) es importante que para determinar la distribución posteriori, solamente es necesario que el intervalo (a, b) extendido sobre la región en la cual la función de verosimilitud es apreciablemente distinta de cero.

Por ejemplo, cuando la función de verosimilitud está basada en la distribución Normal, una distribución a priori uniforme está definida en un rango mas o menos ubicado entre las 3 desviaciones estándar de la media en la distribución Normal. Fuera de este rango, las probabilidades a priori del parámetro no son importantes ya que éstas se obtienen multiplicando las colas de la función de densidad Normal, y por consiguiente no afecta significativamente la inferencia basada en la distribución posteriori.

El uso de la teoría de la estimación estable proporciona una función de densidad posteriori definida en un rango finito. Por lo tanto, para calcular las probabilidades a posteriori es necesario integrar la densidad posteriori sobre un rango finito. Sin embargo, frecuentemente es bastante engorroso matemáticamente trabajar con finito tales integrales propias cuando en realidad integrales impropias (aquellas con limites infinitos) serían mucho más fáciles de estudiar y evaluar. Por ésta razón, las densidades a priori uniformes son frecuentementes aproximada por densidades uniformes impropias; es decir, las funciones que son constantes en toda la recta real, y por tanto no puede ser densidades propias (ya que la integral no es 1). La aproximación usualmente no afecta significativamente la distribución posteriori (sobre la cual la inferencia será basada), y los cálculos siguientes serán simplificados. Cuando, expresamos opiniones confusas, la probabilidad a priori para una variable aleatoria θ_j es esparcida uniformemente sobre la recta real, y la densidad resultante es escrita como:

$$p(\theta_j) \propto constante \qquad\qquad -\infty < \theta_j < \infty$$

esta distribución es llamada difusa.

Es importante notar que este juicio de aproximación es sólo de interés cuando, a pesar de que $p(\theta_j)$ no es propia, el resultado de la distribución posteriori es propia.

Si θ es un vector con componentes cuyo rango están en la recta real y se supone que pueden ser distribuidas independientemente, una densidad a priori difusa para el vector de variables aleatorias está dada por $p(\theta) \propto constante$.

Así, si $\mathcal{L}(X|\Theta, \Sigma) = N(\Theta, \Sigma)$, difusa sobre $\Theta : p \times 1$ asumimos sus componentes distribuidas independientemente, y la densidad es dada por:

$$p(\Theta) \propto constante \qquad\qquad -\infty < \theta_j < \infty \tag{2.11}$$

Ahora consideremos una densidad a priori impropia para una variable $\sigma^2, 0 < \sigma^2 < \infty$. Luego, aplicando logaritmo a σ^2, el problema se reduce al estudio de una densidad a priori para la nueva variable $log(\sigma^2)$, donde $-\infty < log(\sigma^2) < \infty$, entonces tenemos el mismo caso de la densidad a priori impropia para el vector aleatorio θ. Esto es, para ser difusa sobre $log(\sigma^2)$, tomamos

$$p(log(\sigma^2)) \propto constante \qquad\qquad 0 < \sigma^2 < \infty$$

Usando el teorema 1.13.1, obtenemos la densidad a priori impropia

$$\begin{aligned} p(\sigma^2) &= p(log(\sigma^2))\left|\frac{dlog(\sigma^2)}{d\sigma^2}\right| \propto \left|\frac{dlog(\sigma^2)}{d\sigma^2}\right| = \frac{1}{\sigma^2} \\ \Longrightarrow & \quad p(\sigma^2) \propto \frac{1}{\sigma^2} \qquad o' \qquad p(\sigma) \propto \frac{1}{\sigma} \end{aligned} \tag{2.12}$$

según la parametrización que sea necesaria.

Veremos que éste resultado es el mismo que calculó Jeffrey en el principio de la invarianza. Caso más general, si $\mathcal{L}(X|D) = N(0, D)$ donde $D = diag({\sigma_1}^2, ..., {\sigma_p}^2)$, y si los $\sigma_j' s$ son independientes, una densidad a priori impropia para los p elementos está dado por:

$$\begin{aligned} p(D) &= p({\sigma_1}^2, ..., {\sigma_p}^2) = \prod_{j=1}^{p} p({\sigma_j}^2) \propto \frac{1}{{\sigma_1}^2 \cdots {\sigma_p}^2} \\ & p(D) \propto \frac{1}{|D|} \end{aligned} \tag{2.13}$$

La noción de las priori impropias es desarrollada ahora para una matriz de covarianza completa, Σ. Primero notamos que $\Sigma = \Sigma'$, estos son p distintos elementos en la primera fila, $p-1$ elementos distintos en la segunda fila, $p-2$ elementos distintos en la tercera fila, y así sucesivamente obtenemos que la matriz Σ tiene un total de

$$\sum_{i=1}^{p}(p-(i-1)) = p^2 - \frac{p(p+1)}{2} + p = \frac{p(p+1)}{2}$$

elementos distintos; o tiene $(p+1)/2$ grupos de p elementos distintos, y el resultado para cada grupo de p elementos es dado por (2.13). Una Generalización multivariada de (2.12) y (2.13) está en términos de la varianza generalizada $|\Sigma|$, la impresición acerca de la distribución de Σ es representada por la afirmación de que Σ sigue la densidad a priori (impropia)

$$p(\Sigma) \propto \frac{1}{|\Sigma|^{(p+1)/2}}, \quad \Sigma > 0 \tag{2.14}$$

En consecuencia, los términos ”priori difusa” o ”priori impropia” serán usados en el sentido de (2.11) y (2.14).

La priori en (2.14) fue propuesta por primera vez por Jeffrey, 1961, para p=1,2. Él usó el argumento de la invarianza el cual será discutido más adelante.

Ahora, estudiemos la a priori de Σ^{-1} y veamos que coinciden con (2.14).

Sean $x_1, ..., x_p$ p-vectores independientes observados con distribución $N(0, \Sigma)$ y sea $X = (x_1, ..., x_p)$. La función de verosimilitud está dada por:

$$\begin{aligned} L(x_1, ..., x_p|\Sigma^{-1}) &\propto \frac{1}{|\Sigma|^{p/2}} exp\left\{-\frac{1}{2}\sum_{j=1}^{p} x_j'\Sigma^{-1}x_j\right\} \\ &\propto \frac{1}{|\Sigma|^{p/2}} exp\left\{-\frac{1}{2}tar(XX'\Sigma^{-1})\right\}. \end{aligned}$$

Sea $y_j = \Sigma^{-1/2}x_j$, $j = 1, .., p$ y sea $Y = (y_1, ..., y_p) = \Sigma^{-1/2}X$. Entonces por teorema 1.13.1, cada y_j tiene una distribución $N(0, I)$, luego si $W = Y'Y = X'\Sigma^{-1}X$, W tiene una distribución Wishart , con densidad proporcional a

$$|W|^{-1/2} exp\left\{-\frac{1}{2}trW\right\}, \qquad W > 0$$

Dado que la distribución de W no depende de Σ^{-1} asumimos la distribución posteriori de $W = X'\Sigma^{-1}X$ (veamos a Σ^{-1} como desconocida y a X como fija y conocida) es la misma distribución Wishart. Esto es,

$$p(X'\Sigma^{-1}X|X) \propto |X'\Sigma^{-1}X|^{-1/2} exp\left\{-\frac{1}{2}tr(X'\Sigma^{-1}X)\right\}.$$

Transformando W a Σ^{-1} y usando la ecuación (1.7), por el teorema 1.13.1 obtenemos:

$$\begin{aligned} p(\Sigma^{-1}|X) &= p(W|X)J(\Sigma^{-1} \longrightarrow X) \\ &\propto |X'\Sigma^{-1}X|^{-1/2} exp\left(-\frac{1}{2}tr(X'\Sigma^{-1}X)\right)|X|^{p+1} \\ &\propto |\Sigma^{-1}|^{-1/2} exp\left(-\frac{1}{2}tr(X'\Sigma^{-1}X)\right) \\ &= |\Sigma^{-1}|^{-1/2} exp\left(-\frac{1}{2}tr(XX'\Sigma^{-1})\right). \end{aligned}$$

Por lo tanto

$$p(\Sigma^{-1}|X) \propto |\Sigma^{-1}|^{-1/2} exp\left(-\frac{1}{2}tr(XX'\Sigma^{-1})\right)$$

luego, por consecuencia del teorema de Bayes

$$p(\Sigma^{-1}|X) \propto L(X_1, ..., X_p|\Sigma^{-1})p(\Sigma^{-1})$$

así, obtemos que

$$p(\Sigma^{-1}) \propto \frac{1}{|\Sigma^{-1}|^{(p+1)/2}}.$$

§2.9. Información de Fisher

A continuación veremos que el entendimiento y la medición de la información es uno de los aspectos clave de la estadística.

Definición **2.9.1.** Sea X un vector aleatorio con función de probabilidad $p(x|\theta)$. La información de Fisher θ es la medida de fisher de X y ésta definida por:

$$I(\theta) = E_{X|\theta}\left[\frac{-\partial^2 \log p(x|\theta)}{\partial\theta\partial\theta'}\right]$$

con elementos $I_{ij}(\theta)$ dada por:

$$I_{ij}(\theta) = E_{X|\theta}\left[-\frac{\partial^2 \log p(x|\theta)}{\partial\theta_i\partial\theta_i}\right], \qquad i, j = 1, ..., p$$

La medida de la información de θ que se define de ésta forma es la relación del valor medio de la curvatura de la probabilidad. La información de Fisher observada corresponde a la segunda derivada respecto a θ del log de la verosimilitud multiplicada por (-1); es decir

$$J_x = -\frac{\partial^2 \log p(x|\theta)}{\partial\theta\partial\theta'}$$

y es interpretada como una medida local del contenido de la información, mientras el valor esperado de la información de Fisher es una medida global.

Hay muchas propiedades que se pueden obtener de la información de Fisher. Una de las más útiles es la siguiente:

Lema 2. *Sea $X = (X_1, ..., X_p)$ una colección de variables aleatorias independientes con distribuciones $p_i(x, \theta)$, $i = 1, ..., p$.*

Sean J_x y J_{x_i} las medidas de la información observada obtenidas a través de X y X_i; $i = 1, ..., p$ respectivamente. Sean I y I_i la información de Fisher obtenida a través de X y X_i , $i = 1, ..., p$ respectivamente. Entonces

$$J_x(\theta) = \sum_{i=1}^{p} J_{x_i}(\theta) \qquad I(\theta) = \sum_{i=1}^{p} I_i(\theta).$$

Demostración. Primero por la independencia de los X_i

$$p(x|\theta) = \prod_{i=1}^{p} p_i(x_i|\theta)$$

entonces

$$\log p(x|\theta) = \sum_{i=1}^{p} \log p_i(x_i|\theta)$$

Ahora, derivamos dos veces respecto a θ y multiplicamos por (-1) para obtener:

$$-\frac{\partial^2 \log p(x|\theta)}{\partial\theta\partial\theta'} = -\sum_{i=1}^{p} \frac{\partial^2 \log p_i(x_i|\theta)}{\partial\theta\partial\theta'} \tag{2.15}$$

Cuando vamos a probar el resultado a cerca de la información observada, tomamos la esperanza respecto a $X|\theta$ en ambos lados de (2.15). Entonces

$$\begin{aligned} E_{X|\theta}\left[-\frac{\partial^2 \log p(x|\theta)}{\partial\theta\partial\theta'}\right] &= E_{x|\theta}\left[-\sum_{i=1}^{p} \frac{\partial^2 \log P_i(x_i|\theta)}{\partial\theta\partial\theta'}|\theta\right] \\ &= -\sum_{i=1}^{p} E_{x|\theta}\left[\frac{\partial^2 \log P_i(x_i|\theta)}{\partial\theta\partial\theta'}|\theta\right] \quad (\textit{propiedades de } E_{X|\theta}) \\ &= \sum_{i=1}^{p} I_i(\theta). \end{aligned} \tag{2.16}$$

Así de (2.15) y (2.16) se cumple:

$$J_x(\theta) = \sum_{i=1}^{p} J_{x_i}(\theta) \qquad I(\theta) = \sum_{i=1}^{p} I_i(\theta).$$

$\square$

§2.10. Invariancia

La densidad a priori dada en (2.14) fue alcanzada sobre la base de una opinión confusa, o teoría de estimación estable, aproximada. Otra aproximación, debido a Jeffrey (1961), está basada sobre la noción de que las afirmaciones de la probabilidad hechas acerca de la variable aleatoria observada debería permanecer invariante bajo los cambios en la parametrización del problema. Para sastifacer este criterio, Jeffrey muestra que un vector de parámetro θ está definido de la siguiente manera:

Definición **2.10.1.** Consideremos una observación X donde la función de densidad es dada por $p(x|\theta)$; la priori de Jeffrey no-informativa tiene una densidad dada por:

$$p(\theta) \propto [I(\theta)]^{1/2}, \theta \in \Theta \qquad (Caso\, escalar)$$

y

$$p(\theta) \propto |I(\theta)|^{1/2} \qquad (caso\, multivariado) \tag{2.17}$$

Lema 3. *La priori de Jeffrey $p(\theta) \propto [I(\theta)]^{1/2}$ es invariante bajo las transformaciones 1-1, es decir; si $\phi = \phi(\theta)$ es una transformación 1-1 de θ, entonces la priori de Jeffrey para ϕ es $p(\theta) \propto [I(\phi)]^{1/2}$*

Demostración. Sea $\phi = \phi(\theta)$ una transformación de θ 1-1.

Derivamos $\log p(x|\phi)$ con respecto a ϕ, entonces

$$\frac{\partial \log p(x|\phi)}{\partial \phi} = \frac{\partial \log p(x|\phi(\theta))}{\partial \theta}\frac{\partial \theta}{\partial \phi}$$

donde $\theta = \theta(\phi)$ es la transformación de ϕ. Para obtener el parámetro de información de Fisher, necesitamos la segunda derivada de el logaritmo de la función de verosimilitud, entonces

$$\frac{\partial^2 \log p(x|\phi)}{\partial \phi^2} = \frac{\partial \log p(x|\phi(\theta))}{\partial \theta}\frac{\partial^2 \theta}{\partial \phi^2} + \frac{\partial^2 \log p(x|\phi(\theta))}{\partial \theta^2}\left(\frac{\partial \theta}{\partial \phi}\right)^2$$

Multiplicamos ambos lados por (-1) y calculamos el valor esperado respecto a $p(x|\theta)$. Así

$$I(\phi) = -E_{x|\theta}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\right]\frac{\partial^2 \theta}{\partial \phi} + I(\theta)\left(\frac{\partial \theta}{\partial \phi}\right)^2.$$

Veamos que $E_{x|\theta}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\right] = 0$.

En efecto

$$\begin{aligned}
\int p(x|\theta)dx &= 1 \\
\Longrightarrow 0 &= \int \frac{\partial p(x|\theta)}{\partial\theta}dx \\
&= \int \frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial\theta}p(x|\theta)dx \\
&= \int \frac{\partial \log p(x|\theta)}{\partial\theta}p(x|\theta)dx \\
&= E_{x|\theta}\left[\frac{\partial \log p(x|\theta)}{\partial\theta}\right].
\end{aligned}$$

Luego,

$$\begin{aligned}
I(\phi) &= I(\theta)\left(\frac{\partial\theta}{\partial\phi}\right)^2 \\
\Longrightarrow [I(\phi)]^{1/2} &= [I(\theta)]^{1/2}\left|\frac{\partial\theta}{\partial\phi}\right|.
\end{aligned}$$

Por la regla de la probabilidad si la densidad de θ es proporcional a $[I(\theta)]^{1/2}$ entonces ϕ tiene la densidad

$$p(\phi) \propto [I(\phi)]^{1/2} = [I(\theta(\phi))]^{1/2}\left|\frac{\partial\theta}{\partial\phi}\right| = p(\theta).$$

Así, se cumple que las transformaciones 1-1 son invariantes. □

El mismo resultado es cierto para el caso multivariado, es decir;

Corolario 1. *La priori de Jeffrey es invariante bajo la transformación 1-1 en el caso multivariado.*

Demostración. Esto sólo es una transformación Φ de θ la cual satisface la regla de invariancia y tiene densidad constante. Esta transformación es fácil de obtener al tomar

$$p(\Phi) \propto |I(\theta)|^{1/2}\left|\frac{\partial\theta}{\partial\Phi}\right| \propto K$$

o'

$$\left|\frac{\partial\theta}{\partial\Phi}\right| \propto |I(\theta)|^{-1/2} \Rightarrow \left|\frac{\partial\Phi}{\partial\theta}\right| \propto |I(\theta)|^{1/2} \Rightarrow \Phi \propto \int_\theta |I(u)|^{1/2}du$$

□

Por lo tanto, para el caso multivariado basta calcular una función inversamente proporcional.

Jeffrey calculó que una densidad a priori no es necesariamente cierta cuando ésta tiene más de un parámetro en el problema (tal como en la distribución Normal, con 2 parámetros). Lo correcto para tal caso, se recomienda trabajar con un parámetro escalar, y luego con el otro parámetro, es decir; restringir el conjunto.

Ejemplo **2.10.1.** Suponga que X es un escalar y $\mathcal{L}(X;\theta,\sigma^2) = N(\theta,\sigma^2)$, entonces la función de verosimilitud es proporcional a:

$$p(x|\theta,\sigma^2) \propto \frac{1}{\sigma}\exp\left[-\frac{1}{2\sigma^2}(x-\theta)^2\right] \qquad -\infty < x < \infty,$$

si aplicamos log en ambos lados ,entonces

$$L \equiv \log p(x|\theta,\sigma^2) = -\log\sigma - \frac{1}{2\sigma^2}(x-\theta)^2.$$

Ahora, calculamos I_{ij} con $i,j = 1,2$ para hallar el valor de I

Caso 1: $i = j = 1$

$$\begin{aligned}
\frac{\partial \log p(x|\theta,\sigma^2)}{\partial\theta} &= \frac{(x-\theta)}{\sigma^2} \\
\Rightarrow \frac{\partial^2 \log p(x|\theta,\sigma^2)}{\partial\theta^2} &= -\frac{1}{\sigma^2} \\
\Rightarrow E\left[-\frac{\partial^2 \log p(x|\theta,\sigma^2)}{\partial\theta^2}\right] &= \frac{1}{\sigma^2} \\
\Rightarrow I_{11} &= \frac{1}{\sigma^2}
\end{aligned}$$

Caso 2: $i = j = 2$

$$\begin{aligned}
\frac{\partial \log p(x|\theta,\sigma^2)}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(x-\theta)^2 \\
\Rightarrow \frac{\partial^2 \log p(x|\theta,\sigma^2)}{\partial (\sigma^2)^2} &= \frac{1}{2(\sigma^2)^2} - \frac{1}{2(\sigma^2)^2}(x-\theta)^2 \\
\Rightarrow E\left[-\frac{\partial^2 \log p(x|\theta,\sigma^2)}{\partial (\sigma^2)^2}\right] &= -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}E[(x-\theta)^2] \\
&= -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\sigma^2 \\
&= \frac{1}{2\sigma^4} \\
\Rightarrow I_{22} = \frac{1}{2\sigma^4} &
\end{aligned}$$

Caso 3: $i = 1 y j = 2$

$$\begin{aligned}
\frac{\partial^2 \log p(x|\theta,\sigma^2)}{\partial \theta \partial \sigma^2} &= -\frac{1}{\sigma^2}(x-\theta) \\
\Rightarrow \frac{1}{\sigma^4}E(x-\theta) &= \frac{1}{\sigma^4}(E(x)-\theta) = 0
\end{aligned}$$

Luego, como $\log p(x|\theta,\sigma^2)$ es C^2 entonces $I_{12} = I_{21}$.
Usando la definición 2.11.1 y los casos anteriores, sin tomar los parámetros por separados obtenemos:

$$p(\theta,\sigma^2) \propto \frac{1}{\sigma^3}.$$

Ahora, supongamos que σ^2 es una constante conocida y θ es desconocida. Entonces:

$$\begin{aligned}
I &= I_{11} \\
\Longrightarrow p(\theta) &\propto \textit{constante}.
\end{aligned}$$

Después supongamos que θ es constante conocida mientras σ^2 es desconocida. Entonces

$$\begin{aligned}
I &= I_{22} = \frac{1}{2\sigma^4} \\
\Longrightarrow p(\sigma^2) &\propto \frac{1}{\sigma^2}.
\end{aligned}$$

Ahora, combinando términos por cada parámetro, asumiendo que θ y σ^2 son independientes

$$p(\theta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Así, el argumento de la varianza de Jeffrey se aplica a un parámetro a la vez y da el mismo resultado que en el caso a priori no-informativo.

EJEMPLO **2.10.2.** Supongamos $X : p \times 1$ y $\mathcal{L}(x|\theta, \Lambda) = N(\theta, \Lambda), \Lambda > 0$, con $\Lambda = \Sigma^{-1}$ la función de verosimilitud es proporcional a:

$$p(x|\theta) \propto |\Lambda|^{1/2} \exp\left\{-\frac{1}{2}(x-\theta)'\Lambda(x-\theta)\right\}$$

Ahora, aplicamos log en ambos lados, entonces

$$L \equiv \log p(x|\theta, \Lambda) = \frac{1}{2}\log|\Lambda| + c - \frac{1}{2}(x-\theta)'\Lambda(x-\theta),$$

Para alguna constante c. En este problema nosotros desarrollaremos la invarianza de Jeffrey por la aplicación de una matriz de parámetros en un tiempo. Entonces

$$\begin{array}{rcl} L & = & \log|\Lambda| + x'\Lambda x - 2\theta'\Lambda x + \theta'\Lambda\theta \quad \text{(Por Simetría de } \Lambda) \\ \dfrac{\partial L}{\partial \theta} & = & -\dfrac{1}{2}[2\Lambda\theta - 2\Lambda x] \qquad (por\ (1{,}21)) \\ \Rightarrow \dfrac{\partial^2 L}{\partial\theta\partial\theta'} & = & -\Lambda \\ \Rightarrow E\left(-\dfrac{\partial^2 L}{\partial\theta\partial\theta'}\right) & = & \Lambda = constante. \end{array}$$

Así, la matriz de información de Fisher es constante

$$p(\theta) \propto constante \quad .[\text{por } (2.17)]$$

Ahora, asumimos que θ es constante. Entonces, si L es escrita de una forma más conveniente

$$L = \frac{1}{2}\log|\Lambda| + c - \frac{1}{2}tr(x-\theta)(x-\theta)'\Lambda$$

y usando (1.4) y (1.5) obtenemos

$$\begin{aligned}\frac{\partial L}{\partial \Lambda} &= \frac{1}{2}\left[\frac{1}{|\Lambda|}(2|\Lambda|\Lambda^{-1} - diag(|\Lambda|\Lambda^{-1}))\right] - \frac{1}{2}(x-\theta)(x-\theta)' \\ &= \Lambda^{-1} - diag(\Lambda^{-1}) - \frac{1}{2}(x-\theta)(x-\theta)'\end{aligned}$$

luego por (1.8) y tomando $\Lambda \equiv (\lambda_{ij})$, tenemos

$$\frac{\partial L}{\partial \lambda_{ij}\partial \lambda_{kl}} \propto |\Lambda|^{-(p+1)}$$

Tomando la esperanza con respecto a x el resultado no cambia, es decir; si I denota la matriz de información de Fisher

$$|I| \propto |\Lambda|^{-(p+1)}$$

Así, la priori invariante para Λ es dada

$$P(\Lambda) \propto \frac{1}{|\Lambda|^{(p+1)/2}}$$

Dado el Jacobiano de Λ a Σ es $|\Sigma|^{-(p+1)}$, la densidad a priori invariante implícita para Σ es

$$P(\Sigma) \propto \frac{1}{|\Sigma|^{(p+1)/2}}$$

Finalmente, si θ y Σ son independientes, la densidad a priori es dada por:

$$P(\theta, \Sigma) \propto \frac{1}{|\Sigma|^{(p+1)/2}} \quad \text{(Por (2.17))}$$

§2.11. Conjugada a priori natural

Si la densidad a priori es fácil de calcular, ésta puede ser hallada por medio del problema de una densidad a priori informativa. Una forma, para hallar la densidad a priori es usando la conjugada natural.

La idea básica es escribir la densidad como la función de verosimilitud para las variables aleatorias observadas, entonces intercambiando los roles de las variables aleatorias observadas con los parámetros, asumimos que éste último es aleatorio y el anterior es fijo, modificando la proporcionalidad constante apropiadamente para que la nueva "densidad" al integrarla sea igual a uno. En resumen. Sea $L(x|\theta)$ la verosimilitud obtenida de la familia F. Si existe una constante k de tal manera que la función sea una función de densidad, $p(\theta) = kL(x|\theta)$, entonces la familia F de las densidades p es la conjugada natural de la familia con respecto a la distribución de muestreo de verosimilitud L.

Ventajas de usar la densidad de la conjugada a priori natural

1. La distribución posteriori pertenece a la misma familia que la de la priori.

2. La distribución posteriori es fácil de manejar matemáticamente.

Ejemplo **2.11.1.** Sea $X : p \times 1$ un vector aleatorio; supongamos que tiene una distribución $\mathcal{L}(X|\theta) = N(\theta, A_0)$, donde A_0 es una matriz $p \times p$ no-singular fija y conocida. La densidad de ésta observación es dada por:

$$p(x|\theta, A_0) \propto \exp\left\{-\frac{1}{2}(x-\theta)'{A_0}^{-1}(x-\theta)\right\}.$$

Intercambiamos los roles de x y θ; la densidad a priori nos da:

$$p(\theta) \propto \exp\left\{-\frac{1}{2}(\theta-x)'{A_0}^{-1}(\theta-x)\right\}.$$

Esta expresión, cuando la observamos como una densidad en θ, tiene la forma de una densidad Normal con media x y matriz de covarianza A_0. Aunque, la familia de la densidad puede ser obtenida por un vector medio a arbitrario, y una matriz de covarianza A arbitraria (kernel de la conjugada natural). Entonces, la densidad de la familia de la conjugada natural a priori es de la forma:

$$p(\theta) \propto \exp\left\{-\frac{1}{2}(\theta-a)'A^{-1}(\theta-a)\right\}$$

ó

$$\mathcal{L}(\theta) = N(a, A)$$

§2.12. Conjugada a priori natural para la distribución normal

Ahora consideramos el caso general de $X : p \times 1$ con $\mathcal{L}(X|\theta, \Sigma) = N(\theta, \Sigma)$, y Σ desconocida.

Supongamos que $X_1, ..., X_n$ son n observaciones independientes e idénticamente distribuida. El problema está en hallar la densidad de la conjugada a priori natural de θ y Σ, y esto puede ser realizado ya que existen suficientes distribuciones. La densidad conjunta de las observaciones es dada por:

$$p(x_1, ..., x_n|\theta, \Sigma) \propto \frac{1}{|\Sigma|^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} [(x_i - \theta)'\Sigma^{-1}(x_i - \theta)] \right\}.$$

A veces es conveniente reparametrizar los términos de $(\theta, \Lambda) = (\theta, \Sigma^{-1})$ Λ es llamada la matriz de precisión (cuanto más pequeña la varianza mayor sería la precisión). La densidad reescrita es:

$$\begin{aligned}
p(x_1, ..., x_p|\theta, \Lambda) &\propto |\Lambda|^{n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} [(x_i - \theta)'\Lambda(x_i - \theta)] \right\} \\
&= |\Lambda|^{n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{p} [x_i'\Lambda x_i - \theta'\Lambda x_i + \theta'\Lambda\theta - x_i'\Lambda\theta] \right\} \\
&= |\Lambda|^{n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{p} [x_i'\Lambda x_i - 2\theta'\Lambda x_i + \theta'\Lambda\theta] \right\} \qquad (\Lambda = \Lambda') \\
&= |\Lambda|^{n/2} \exp\left\{ -\frac{n}{2} \left[\frac{1}{n} \sum_{i=1}^{n} x_i'\Lambda x_i - 2\theta'\Lambda\overline{x} + \theta'\Lambda\theta \right] \right\}.
\end{aligned}$$

Así.

$$p(x_1, ..., x_p|\theta, \Lambda) \propto |\Lambda|^{n/2} \exp\left\{ -\frac{n}{2} \left[\frac{1}{n} \sum_{i=1}^{n} x_i'\Lambda x_i - 2\theta'\Lambda\overline{x} + \theta'\Lambda\theta \right] \right\}$$

donde $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ es la usual expresión para la media muestral, ahora completamos cuadrado con respecto a θ

$$
\begin{aligned}
& |\Lambda|^{n/2} \exp\left\{-\frac{n}{2}\left[\frac{1}{n}\sum_{i=1}^{n} x_i'\Lambda x_i - 2\theta'\Lambda\overline{x} + \overline{x}\Lambda\overline{x} + \theta'\Lambda\theta - \overline{x}\Lambda\overline{x}\right]\right\} \\
= \;& |\Lambda|^{n/2} \exp\left\{-\frac{n}{2}\left[(\theta-\overline{x})'\Lambda(\theta-\overline{x}) - \overline{x}\Lambda\overline{x} + \frac{1}{n}\sum_{i=1}^{n} x_i'\Lambda x_i\right]\right\}.
\end{aligned}
$$

Luego

$$
p(x_1,...,x_p|\theta,\Lambda) \propto |\Lambda|^{n/2} \exp\left\{-\frac{n}{2}\left[(\theta-\overline{x})'\Lambda(\theta-\overline{x}) - \overline{x}\Lambda\overline{x} + \frac{1}{n}\sum_{i=1}^{n} x_i'\Lambda x_i\right]\right\} \tag{2.18}
$$

Entonces, reescribimos la expresión (2.18) como el producto de 2 factores

$$
\begin{aligned}
p(x_1,...,x_p|\theta,\Lambda) \;\propto\; & \left[|\Lambda|^{1/2} \exp\left\{-\frac{n}{2}(\theta-\overline{x})'\Lambda(\theta-\overline{x})\right\}\right] \qquad (2.19)\\
& \cdot \left[|\Lambda|^{(n-1)/2} \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n} x_i'\Lambda x_i + n\overline{x}\Lambda\overline{x}\right)\right\}\right]
\end{aligned}
$$

Note que sólo el primer factor contiene θ, así usando el ejemplo 2.12.1 obtenemos una familia conjugada natural para la distribución a priori condicional para θ en la forma:

$$
\mathcal{L}(\theta|\Lambda) = N(a,(\Lambda b)') \tag{2.20}
$$

donde a es un p-vector arbitrario obtenido con los datos de $\overline{x}$ y b es un escalar positivo arbitrario obtenido con n. Sin embargo, lo que realmente es importante de la solución es la densidad a priori conjunta $p(\theta,\Lambda)$, dada:

$$
p(\theta,\Lambda) = p(\theta|\Lambda)p(\Lambda). \tag{2.21}
$$

Sólo es necesario establecer $p(\Lambda)$ de la segunda expresión entre corchetes de (2.19). Dado que la traza de un escalar es igual a un escalar, entonces la expresión puede ser reescrita como:

$$
\begin{aligned}
& |\Lambda|^{(n-1)/2} \exp\left\{-\frac{1}{2} tr[\sum_{i=1}^{n} x_i' \Lambda x_i - n\overline{x}' \Lambda \overline{x}]\right\} \\
= & |\Lambda|^{(n-1)/2} \exp\left\{-\frac{1}{2}\left(tr[\sum_{i=1}^{n} x_i' \Lambda x_i] - tr[n\overline{x}' \Lambda \overline{x}]\right)\right\} \\
= & |\Lambda|^{(n-1)/2} \exp\left\{-\frac{1}{2}\left(tr[\sum_{i=1}^{n} \Lambda x_i x_i'] - tr[n\Lambda \overline{x}\overline{x}']\right)\right\} \\
= & |\Lambda|^{(n-1)/2} \exp\left\{-\frac{1}{2} tr[\sum_{i=1}^{n} \Lambda x_i x_i' - n\Lambda \overline{x}\overline{x}']\right\} \\
= & |\Lambda|^{(n-1)/2} \exp\left\{-\frac{1}{2} tr[\Lambda(\sum_{i=1}^{n} x_i x_i' - n\overline{x}\overline{x}')]\right\} \\
= & |\Lambda|^{(n-1)/2} \exp\left\{-\frac{1}{2} tr[\Lambda \sum_{i=1}^{n} (x_i x_i' - \overline{x}\overline{x}')]\right\} \\
\propto & |\Lambda|^{(n-1)/2} \exp\left\{-\frac{1}{2} tr[\Lambda \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})']\right\} \qquad (2.22)
\end{aligned}
$$

Hagamos $V = \sum_{i=1}^{n}(x_i - \overline{x})'(x_i - \overline{x})$ donde es la matriz de covarianza muestral. Por otra parte notemos que la expresión (2.22) es exactamente la ecuación (2.10) ya que n y V son arbitrario, así la distribución a priori para Λ es dada por:

$$
\mathcal{L}(\Lambda) = W(G, p, m) \tag{2.23}
$$

donde $G > 0$ es una matriz escalar (dimensión-p) y $m \geq p$.

Ahora, sustituimos (2.20) y (2.23) en (2.21) para obtener la distribución a priori conjugada natural de θ y Λ, la cual es dada por la distribución Normal-Wishart

$$
p(\theta, \Lambda) \propto |\Lambda|^{(m-p)/2} \exp\left\{-\frac{1}{2}[(\theta - a)'\Lambda(\theta - a)b + tr\Lambda G^{-1}]\right\} \tag{2.24}
$$

Ésta densidad a priori conjuga natural si es usada en futuros análisis solo es necesario los valores m, a, b y G basada en los estudios o historias evidentes.

Para obtener una priori conjugada natural para (θ, Σ) de (θ, Λ) es necesario transformar las variables en (2.23). El resultado será una distribución Normal-Inversa

Wishart. Retomando a la función de verosimilitud (2.19) y parametrizando esto en términos de (θ, Σ) para cambiar variables, tenemos:

$$\begin{aligned} p(x_1, ..., x_p|\theta, \Sigma) \quad \propto \quad & \left[|\Sigma|^{-1/2} \exp\left\{-\frac{n}{2}(\theta - \overline{x})'\Sigma^{-1}(\theta - \overline{x})\right\}\right] \\ & \cdot \left[|\Sigma^{-1}|^{(n-1)/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i'\Sigma^{-1}x_i + n\overline{x}\Sigma^{-1}\overline{x}\right\}\right]. \end{aligned} \tag{2.25}$$

Usando de nuevo la condicional a priori $\mathcal{L}(\theta|\Sigma) = N(a, \Sigma)$, y usamos la relación de (2.21), obtenemos que:

$$p(\theta, \Sigma) = p(\theta|\Sigma)p(\Sigma) \tag{2.26}$$

el segundo factor de (2.25) se puede escribir como la ecuación (2.22)

$$|\Sigma^{-1}|^{(n-1)/2} \exp\left\{-\frac{1}{2}tr\Sigma^{-1}V\right\}. \tag{2.27}$$

Sea $W = V^{-1}$, usando (2.24) y (1.8) obtenemos la densidad Inversa de Wishart

$$p(W|G) \propto |W^{-1}|^{(m+p+1)/2}|G^{-1}|^{m/2} \exp\left(-\frac{1}{2}trW^{-1}G^{-1}\right) \tag{2.28}$$

más aún , W cumple con la distribución Inversa-Wishart y es escrita como:

$$\mathcal{L}(W) = W^{-1}(H, p, v).$$

Así, la densidad de W puede ser expresada como:

$$p(W|H) \propto |W|^{v/2} \exp\left\{-\frac{1}{2}trW^{-1}H\right\} \qquad W > 0, H > 0, v \geq p.$$

La comparación de (2.27) y (2.28) demuestra que Σ tiene una distribución Inversa-Wishart. Así, esto es consistente con (2.23) ya que, la distribución marginal de la conjugada a priori natural para Σ es:

$$\mathcal{L}(\Sigma) = W^{-1}(H, p, v) \qquad H > 0, \Sigma > 0, v = m + p + 1 \tag{2.29}$$

donde $m > p - 1$ para una distribución propia.

Nota: Esto puede verse de (2.29) y la definición de la densidad que el limite $H \to 0$ y $m \to p-1$, el resultado de la densidad impropia toma la forma:

$$p(\Sigma) \propto \frac{1}{|\Sigma|^{(m+p+1)/2}} \to \frac{1}{|\Sigma|^{p}}.$$

Además, para p=1 el resultado es el mismo como para la densidad invariante de $p(\Sigma) \propto |\Sigma|^{-(p+1)/2}$. Sin embargo, para $p > 1$ el resultado diverge.

La conjugada a priori natural es la densidad Normal-Inversa de Wishart y es obtenidas de (2.20),(2.26) y (2.29), para $v > 2p$ como:

$$p(\theta, \Sigma) \propto |\Sigma|^{(v+1)/2} \exp\left\{-\frac{1}{2}[(\theta - a)'\Sigma^{-1}(\theta - a)b + tr\Sigma^{-1}H]\right\} \tag{2.30}$$

Observación:La distribución Normal a priori genera una Normal a posteriori; La Normal-Wishart a priori genera una Normal-Wishart posteriori, y así sucesivamente.

Ejemplo **2.12.1.** Examinemos el caso de $X : p \times 1$, $\mathcal{L}(X|\theta, \Sigma) = N(\theta, \Sigma)$ donde la densidad de la conjugada priori natural para (θ, Σ) es la densidad Normal-Wishart por (2.30). Así, de (2.25) y (2.30), la densidad a posteriori de (θ, Σ) dada la muestra es definida por $p(\theta, \Sigma|x_1, ..., x_p) \propto p(x_1, ..., x_p|\theta, \Sigma)p(\theta, \Sigma)$, o

$$\begin{aligned} p(\theta, \Sigma|x_1, ..., x_p) \quad \propto \quad & \frac{1}{|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2}tr[n(\theta - \overline{x})(\theta - \overline{x})' + V]\Sigma^{-1}\right\} \\ & \frac{1}{|\Sigma|^{(v+1)/2}} \exp\left\{-\frac{1}{2}tr[b(\theta - a)(\theta - a)' + H]\Sigma^{-1}\right\} \end{aligned}$$

combinando los términos obtenemos:

$$\begin{aligned} p(\theta, \Sigma|x_1, ..., x_p) \quad = \quad & \frac{1}{|\Sigma|^{(n+v+1)/2}} \\ \cdot \quad & \exp\Bigg\{ -\frac{1}{2}\Big[(n+b)\left(\theta - \frac{n\overline{x} + ab}{n+b}\right)'\Sigma^{-1}\left(\theta - \frac{n\overline{x} + ab}{n+b}\right) \\ & + tr(V + H)\Sigma^{-1} + \frac{nb}{n+b}(\overline{x} - a)'\Sigma^{-1}(\overline{x} - a)\Big]\Bigg\} \end{aligned} \tag{2.31}$$

Comparando (2.30) y (2.31) demostramos que la priori y la posteriori son de la familia de la Distribución Normal Inversa- Wishart.

Referencias Bibliográficas

[1] Anderson, T.W.(1958). An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons.

[2] Box, G.E.P y Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley Publishing Co.

[3] Bradley, P.C. y Thomas, A.L. (1996). Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall.

[4] Casella, G. y Berger, R.L. (1990). Statistical Inference. Thomson.

[5] Congdon Peter (2001). Statistical Modelling. New York: John Wiley and Sons.

[6] Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

[7] Press, James S. (1982). Applied Multivariate Analysis: using Bayesian and Frequentist Methods of Inference. Second Edition. Malabar, Florida.

[8] West, M. and Harrison, J. (1997). Bayesian Forecasting and Dynamic Models, 2nd.ed. Springer, New York.

[9] James R. Munkes.(1993).Analysis on Manifolds,Massachusetts Institute of tecnology.

[10] lelys B. de Guenni, Isabel L. Salvador, María E. Pérez (2008). Análisis de datos con técnicas bayesianas. XXI Escuela Venezolana de Matemáticas.