Instituto Tecnológico Autónomo de México

Inferencia Bayesiana en Modelos de Mezcla en Media-Marianza

Tesis

que para obtener el título

Licenciado en Actuaría

presenta

David Edgardo Castillo Rodríguez

Asesor: Juan Carlos Martínez Ovando

Ciudad de México 2017

Índice general

Capítulo 1

Introducción

Con el fin de modelar algún fenómeno de interés de la natraleza se han propuesto modelos estocásticos, como la selección de un portafolio de mínima varianza, o la segmentación de individuos mediante un análisis de discriminante, por ejemplo. Estos modelos estocásticos suponen que los datos tienen alguna estructura de correlación dada, como es el caso del portafolio de mínima varianza que presupone una estructura de correlación normal; otros modelos requieren la estimación de alguna ley de probabilida, donde muchas veces se utiliza la distribución normal (o una mezcla finita normal), por su practicidad.

Dicho lo anterior, la presente tesis fue motivada por la siguiente pregunta ¿Qué hacer cuando hay evidencia de que un conjunto de datos no sigue una ley de probabilidad normal? Varios investigadores han dado respuesta a esta pregunta, mediante la distribución T, o la incorporación de un sesgo a la distribución normal (o a la distribución T), o mediante el uso de cópulas, o mediante el uso de la distribución hiperbólica generalizada, por ejemplo. De aquí que el tema central de esta tesis sea el cómo estimar una distribución hiperbólica (caso particular de la distribución hiperbólica generalizada).

El problema de cómo ajustar una distribución hiperbólica se realizará mediante distribuciones tipo mezcla normal en esperanza y varianza, ya que está distribución permite simplificar el problema, también para realizar el ajuste se utilizará un enfoque bayesiano, ya que se adecua de manera natural a las distribuciones tipo mezcla normal, y además permite explotar el conocimiento a priori de los datos para la obtención de los parámetros de la distribución a estimar.

La presente tesis se divide un 6 capítulos. En el capítulo 1 se introduce de manera breve el paradigma bayesiano de inferencia, el cual es necesario para el tipo de estimación que se hará. En el capítulo 2 se expone la notación a usar, así como las distribuciones necesarias para la estimación; también se habla brevemente de algunas características de las distribuciones mencionadas. En el capítulo 3 se introducen las distribuciones tipo mezcla normal, para así poder caracterízar a la distribución hiperbólica; también se habla brevemente de algunas características de estas distribuciones. En el capítulo 4 se construyen la función de verosimilitud, y densidades marginales completas necesarias para la estimación. En el capítulo 5 se hablará de los resultados y conclusiones. Y por último, en el capítulo 6 se incluye como anexo algunos resultados utilizados utilizados en la presente tesis, así como los respectivos códigos implementados, es importante mencionar que para la obtención de resultados se utilizó el software *R*.

Capítulo 2

Introducción al paradigma bayesiano de inferencia

En este capítulo se hablará brevemente sobre los antecedentes del paradigma bayesiano de inferencia, así como de algunas ideas principales que dieron origen a este. Y así, concluir en cómo es la estimación bayesiana, la cual es fundamental en este trabajo.

2.1. Antecedentes

En 1763, dos años después de la muerte de Thomas Bayes $(1702-1761)$, se publicó uno de sus ensayos, el cual consistió en resolver un problema de información inversa planteado por Jacob Bernoulli (An Essay towards solving a Problem in the Doctrine of Chances). El problema de Bernoulli consitia en obtener información sobre la realización de variables aleatorias independientes distribuidas Bernoulli; para ello, Bayes propuso en su ensayo un método que consistia en tener una suposición sobre la posibilidad de que un evento que tenga que ocurrir, algunas veces ocurra con éxito y otras veces no ocurra (o sea, un fracaso).

En este punto de la historia, las ideas de Bayes carecian de claridad, pero diez años más tarde se esclarecieron debido a que Laplace las retomó, dandole así forma al paradigma Bayesiano de inferencia. Dichas ideas se consolidaron en su libro Theoriè analytique des probabilitès, 1812. Pero el trabajo de Laplace (concerniente al paradigma Bayesiano) aún carecia de formalidad teórica.

El paradigma bayesiano de inferencia cuntinuó eclipsado, hasta que Harold Jeffrey (Theory of probability, 1939), y Bruno De Finetti (La Prévision, ses Lois Logiques, ses Sources Subjectives, 1937) desarrollaron y dieron sustento teórico al paradigma bayesiano. Harold Jeffrey mantuvo una postura objetiva dentro del paradigma bayesiano, ya que el decía que la información inicial era capturada objetivamente mediante el uso de distribuciones no informativas. Por otro lado, De Finetti mantuvo una postura subjetiva, ya que en su obra se desarrolla la teoría de probabilidad refiriendose a esta como un grado un grado de creencia.

Hasta la fecha, el paradigma bayesiano sigue introduciendose en distintos campos de la ciencia, ya que es una gran herramienta en de inferencia o predicción, lo cuál es de vital interés en ciertos ámbitos de la vida humana, como en finanzas, econometría, actuaría, por ejemplo (en Savage (1972) o Lindley (1957) se puede encontrar una explicación más profunda del paradigma bayesiano de inferencia).

Ahora, teniendo un contexto sobre la evolución del paradigma bayesiano, veamos cuales son sus principios técnicos, y la idea que hay en estos.

2.2. Proceso de aprendizaje

Supongamos que tenemos un conjunto de eventos A_1, A_2, . . ., A_n, de los cuales nos gustaría hacer inferencia de acuerdo a una función de probabilidad, P, (referente a estos eventos), y con un nivel de información o evidencia disponible, el culal denotaremps por B. Con estos elementos

podemos definir nuestro estado de información actual sobre los eventos A_1, A_2, . . ., A_n, dada nuestra información inicial B, lo cual queda expresado, para cualquier A_i, como: $P(A_i|B)$, donde $\sum_{i=1}^{n} A_i = 1$.

Ahora, lo que nos interesa es actualizar nuestro nivel de información actual dada nueva información, digamos C, lo cual lo conseguimos de la siguiente manera utilizando la definición de probabilidad condicional Ross (1976),

$$\begin{aligned} P(A_i|C \wedge B) &= \frac{P(A_i \wedge C \wedge B)}{P(C \wedge B)} \\ &= \frac{P(A_i \wedge B)}{P(B)} \frac{P(B)}{P(C \wedge B)} \frac{P(C \wedge A_i \wedge B)}{P(A_i \wedge B)} \\ &= P(A_i|B)P(C|A_i \wedge B)/P(C|B) \end{aligned} \tag{2.1}$$

La ecuación anterior representa la idea de proceso de aprendizaje, o sea, dado el nivel de evidencia inicial denotado por B, mediante la incorporación de nueva información donotada por C, podemos actualizar nuestras creencias que ya teniamos con respecto al evento A_i, es decir, pasar de solamente tener $P(A_i|B)$ a $P(A_i|B \wedge C)$.

2.3. El proceso de aprendizaje en la inferencia estadística

Aplicando las ideas de la sección anterior, podemos incorporar nueva información en una distribución de probabilidad paramétrizada, cuyo parámetro es desconocido para nosotros. Este desconocimiento del verdadero valor del parámetro lo incorporamos en un distribución de probabilidad propia del parámetro desconocido. Por notación nos referimos a la distribución de probabilidad del parámetro como distribución inicial, y la

denotamos como : $\Pi(\theta)$, donde θ es el parámetro descocnocido el cual pertenece a un espacio parametral Θ.

Dicho lo anterior, dada una variable aleatoria X con distribución de densidad de probabilidad parametrizada por algún parámetro θ, tenemos la distribución $f_X(x|\theta)$ la cual se lee como distribución de X dado θ. Después, incorporando nuestro desconocimiento de θ, con $\Pi(\theta)$ podemos formular la siguiente expresión:

$$\begin{aligned} \Pi(\theta|X) &= \frac{f(x \wedge \theta)}{f(x)} \\ &= \frac{f(x \wedge \theta)}{f(x)} \frac{\Pi(\theta)}{\Pi(\theta)} \\ &= f(x|\theta) \frac{\Pi(\theta)}{f(X)} \end{aligned} \tag{2.2}$$

De la ecuación (2,2) podemos tener la siguiente expresión $\Pi(\theta|X) \propto f(x|\theta)\Pi(\theta)$. La parte derecha de la expresión anterior se conoce como distribución a posteriori de θ, esta se puede interpretar como la actualización del nivel de información que teniamos de θ (dicho nivel de información está considerado en $\Pi(\theta)$), al incluir nueva información proveniente de la realización (o realizaciones) de la variable aleatoria X.

La información proveniente de la variable aleatoria X queda capturada por su función de densiadad condicionada por θ, es decir, $f(x|\theta)$; también a esta densiadad condicionada la conocemos como función de verosimilitud.El símbolo $\propto$ nos indica que la expresión de la izquierda es proporcional a la de la derecha, es decir, sólo difieren por una constante que no depende del argumento de la parte izquierda.

La expresión (2,2) también nos da un algoritmo recursivo, el cual nos permite mejorar nuestra información del parámetro θ conforme más información obtengamos de la variable aleatoria X, es decir, si consideramos una

nueva ralización de $X = x*$, entonces nuestras nueva distribución apriori o inicial es $\Pi(\theta|x)$, mientras que la distribució a posteriori (o actualizada) queda dada como $\Pi(\theta|x \wedge x*)$, por lo que mediante un proceso análogo a (2,2) tendriamos la siguiente expresión:

$$\Pi(\theta|x \wedge x*) \propto f(x \wedge x* |\theta)\Pi(\theta|x \wedge x*) \tag{2.3}$$

2.4. Predicción

Hasta ahora hemos visto cómo incorporar nueva información de la variable aleatoria X para actualizar la información deisponible del parámetro θ. Pero también podemos formularnos la siguientes preguntas ¿Qué pasa con el próximo valor de X dados los valores x que ya fueron observados? ¿Hay forma de que la infromación previa de X sea incorporada para mejorar el nivel de información de futuras realizaciones? La respuesta a las preguntas anteriores tiene connotación positiva, y queda representada por la siguiente expresión:

$$\begin{aligned} f(y* |\overline{y}) &= \int_{S_\theta} f(y* \wedge \theta|\overline{y})d\theta \\ &= \int_{S_\theta} f(y* |\theta \wedge \overline{y})\Pi(\theta)d\theta, \end{aligned} \tag{2.4}$$

donde $y*$ representa la realización futura de Y, $\overline{y}$ representa un conjunto de realizaciones observadas de Y, y θ el parámetro de la distribución de probabilidad de Y.

Capítulo 3

Preliminares

En este capítulo se expone de manera breve la notación y algunas características de las distribuciones gaussiana, gaussiana p variada, gamma, gamma inversa, Wishart, Wishart inversa, gaussiana inversa generalizada, gaussiana inversa, las cuales se usarán en capítulos posteriores.

La notación a emplear adoptará la usada en el enfoque bayesiano, para ser consistente con el tipo de modelación que se realizará. Por lo que si la variable o vector aleatorio, X, tiene una función de densidad parametrizada por algún vector de parámetros, Θ, entonces la función de densidad de X se denotará por $f(X|\Theta)$, que se lee e interpreta como: X se distribuye $f(x)$ dado Θ. La función de densidad de Θ será $f(\Theta|\theta_0)$ para algún vector de parámetros θ_0. Esta función de densidad es conocida como distribución a priori, y se interpreta de la misma manera que la distribución de X, salvo que representa un grado subjetivo de creencia respecto a Θ. Para la función de densidad de Θ condicionada en X usaremos la notación $f(\Theta|X)$, la cual se conoce como densidad a posteriori (Bernardo & Smith, 2000).

3.1. Distribución gaussiana

Se dice que una variable aleatoria X, que toma valores en los números reales, se distribuye normal con media μ, y desviación estándar σ, si su función de densidad de probabilidad es de la siguiente manera,

$$f_x(x|\mu,\sigma) = \frac{1}{2\sqrt{\pi\sigma}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} I_{(-\infty,\infty)}(x). \tag{3.1}$$

En la distribución gaussiana, dos parámetros caracterizan la forma de la distribución, $\mu = E[X]$, y $\sigma^2 = Var[X]$, de aquí que el parámetro μ nos da información sobre la localización de la distribución, mientras que σ de la dispersión.

Para referirnos a que X se distribuye gaussiana con parámetros μ y σ, usaremos la notación, $X|\mu,\sigma$ se distribuye $N(x|\mu,\sigma)$.

Como se mencionó anteriormente, el parámetro μ es una medida de localización, mientras que el parámetro σ es una medida de dispersión de la distribución gaussiana, lo cual se ilustra en las siguientes gráficas. En la gráfica 3.1 se pueden observar cuatro distribuciones gaussianas, etiquetadas como a), b), c) y d), con el mismo parámetro de dispersión $\sigma = 1$, pero diferentes parámetros de localización, es decir, $\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 4$ y $\mu_4 = 6$ respectivamente.

Por otro lado, en la gráfica 3.2, se pueden observar cuatro distribuciones gaussianas, etiquetadas como a), b), c) y d), con parámetros de localización $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 0$ y $\mu_4 = 0$, y parámetros de dispersión $\sigma_1 = 0$, $\sigma_2 = 2$, $\sigma_3 = 4$ y $\sigma_4 = 6$, respectivamente; adviértase cómo incrementan las colas de la distribución gaussiana entre mayor es el parámetro de dispersión.

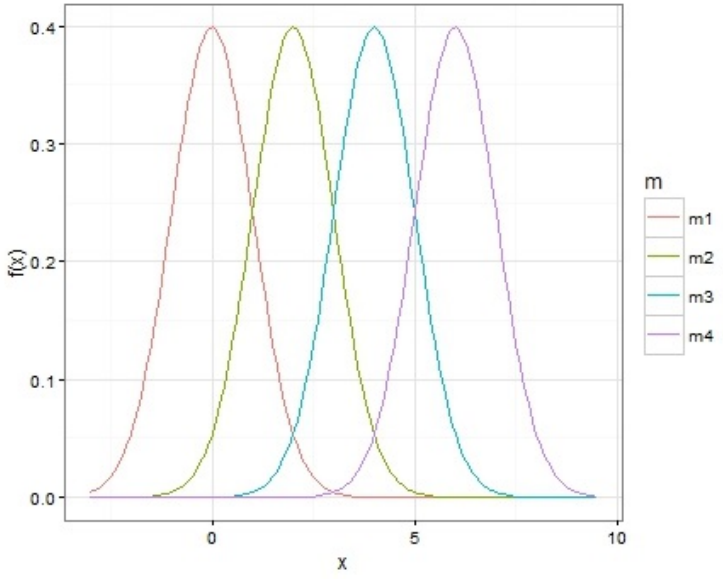

Figura 3.1: Distribuciones gaussianas con el mismo parámetro de dispersión, pero diferentes parámetros de localización.

Como caso particular de la distribución gaussiana tenemos la distribución gaussiana estándar, que se obtiene cuando el parámetro $\mu = 0$, y $\sigma = 1$. También podemos llegar a esta distribución mediante un proceso de normalización de la distribución (2,1), es decir, si X se distribuye $N(x|\mu, \sigma)$, y si consideramos la transformación $Y = \frac{X-\mu}{\sigma}$, entonces Y se distribuye $N(y|0, 1)$.

Por último, la función generadora de momentos y característica de una distribución gaussiana es de la siguiente manera (Ross Shelder, Probability),

$$m_x(t) = \exp^{\sigma t + \frac{t\mu\sigma^2}{2}} \tag{3.2}$$

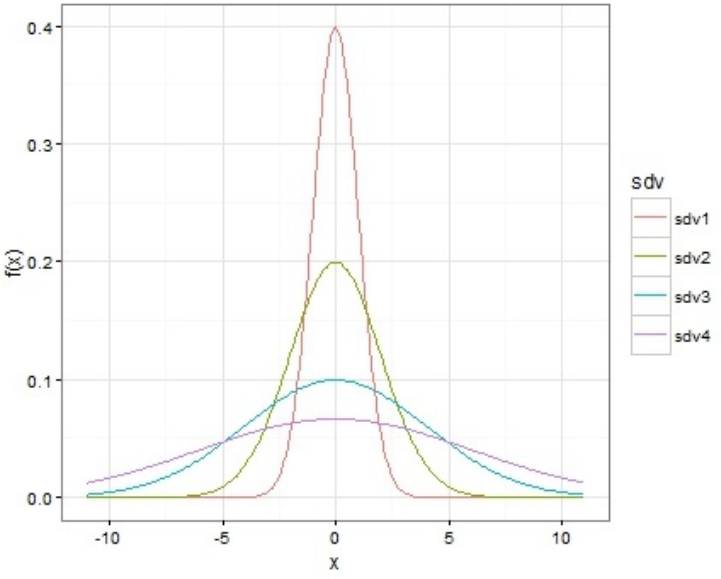

Figura 3.2: Distribuciones gaussianas con diferente parámetro de dispersión.

$$\Phi_x(it) = \exp\{\sigma it + \frac{it\mu\sigma^2}{2}\} \tag{3.3}$$

3.2. Distribución gaussiana p-variada

Se dice que un vector aleatorio X de dimensión p, con parámetros μ y Σ, que toma valores tanto positivos como negativos en cada entrada, tiene una distribución gaussiana p-variada si su función de densidad es de la siguiente forma:

$$f_x(X|\mu,\Sigma) = \frac{1}{(2\Pi)^{p/2}|\Sigma|^{1/2}} exp\{-1/2(X-\mu)'\Sigma^{-1}(X-\mu)\} I_{R^p}(x),$$

donde $\mu = E[X]$, de dimensión p, es el vector de medias, y $\Sigma = Cov[X]$ de dimensión $p \times p$, matriz simétetrica positiva definida, es la matriz de varianza-covarianza. Análogamente a la distribución gaussiana, el vector μ es un vector de posición, el cual indica donde está centrada la distribución del vector aleatorio X. Mientras que la matriz Σ está formada por las varianzas de cada X_i en las entradas diagonales, es decir en la (i, i) entrada, y por las covarianzas $Cov[X_{i,j}]$ en la (i, j) entrada de la matriz. De aquí que la matriz de varianzas-covarianzas indique la dispersión del vector aleatorio X alrededor del vector de medias μ.

En este caso, si X es un vector aleatorio de dimensión p, con vector de medias μ, y matriz ve varianza-covarianza Σ, usaremos la notación X se distribuye $N_p(x|\mu, \Sigma)$.

Como se mencionó anteriormente el vector μ es un parámetro de localización, de la distribución gaussiana p-variada, lo cual se ilustra a continuación para el caso bivariado. En la gráfica 3.3 se muestran tres distribuciones gaussianas a), b) y c); todas con la misma matriz de varianza-covarianza I, con $\sigma_1 = \sigma_2 = 1$ y $\sigma_{1,2} = 0$; pero con vectores de medias $\mu_1 = (-2, -2)$, $\mu_2 = (0, 0)$ y $\mu_3 = (2, 2)$, respectivamente.

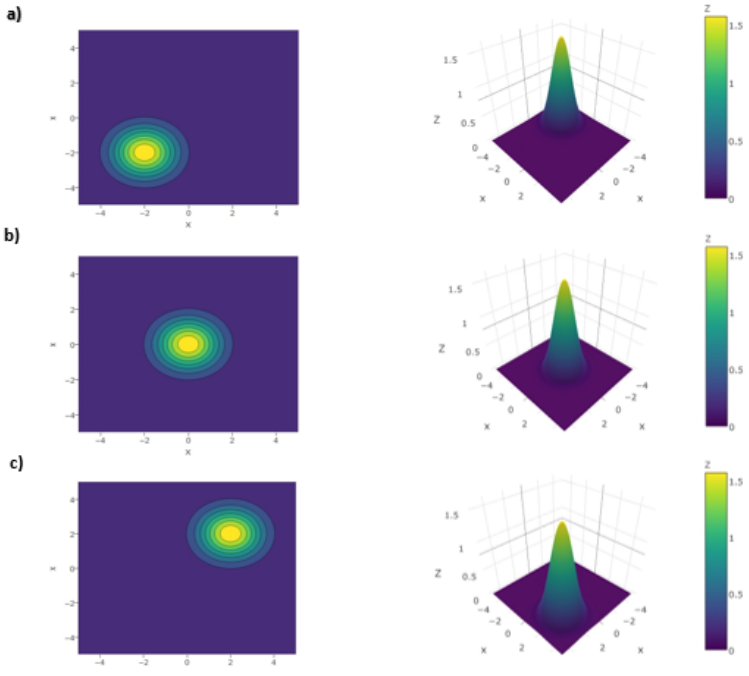

Figura 3.3: Distribución gaussiana p-varaida con diferentes vectores de medias μ.

Ahora si nos enfocamos en la matriz de varianza-covarianza Σ, entre mayor sea la varianza del elemento x_i del vector aleatorio X, mayor será la dispersión de la distribución de X en el eje coordenado correspondiente a x_i.

Por otro lado, entre mayor o menor sea la covarianza entre algún par de elementos x_i y x_j pertenecientes a X, la distribución de X tendrá una notable forma elíptica en los ejes cordenados i y j, la cual está ligada a la interpretación de la correlación lineal entre x_i y x_j, pues si $Cov[x_i, x_j]$ es mayor a cero, la distribución de X tendrá una foma elíptica centrada en

el vector de medias μ, pero rotada entre 0 y 90 grados, lo que indica que hay una dependencia lineal positiva entre el i-ésimo y j-ésimo elemento de X. Mientras que, sí la covarianza entre el i-ésimo y j-ésimo elemento de X es menor a cero, la distribución de X en los ejes coordenados i y j tendrá una forma elíptica rotada entre 90 y 180 grados. Los dos puntos anteriores se ilustran a cuntinuación para el caso bivariado.

En la gráfica 3.4 a) se observa una distribución gaussiana bivariada con vector de medias $\mu = (0, 0)$, y con matriz de varianza-covarianza, Σ, con las entradas $Var[X] = 1$, $Var[Y] = 5$ y $Cov[X, Y] = 0$. En b) con vector de medias $\mu = (0, 0)$, $Var[X] = 1$, $Var[Y] = 1$ y $Cov[X, Y] = 0$. Y en c) con vector de medias $\mu = (0, 0)$, $Var[X] = 5$, $Var[Y] = 1$ y $Cov[X, Y] = 0$.

En la gráfica 3.5 a) se observa una distribución gaussiana bivariada con vector de medias $\mu = (0, 0)$, y con matriz de varianza-covarianza, Σ, con las entradas $Var[X] = 2$, $Var[Y] = 2$ y $Cov[X, Y] = 0{,}75$. En b) con vector de medias $\mu = (0, 0)$, $Var[X] = 5$, $Var[Y] = 5$ y $Cov[X, Y] = -0{,}75$.

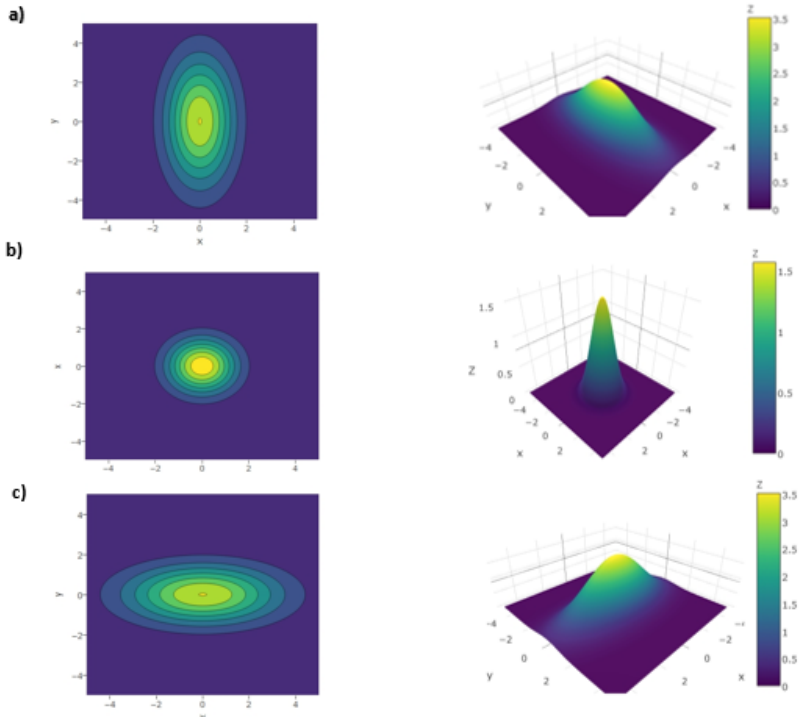

Figura 3.4: Distribución gaussiana bivariada con diferentes varianzas y covarianzas

Análogamente a la distribución gaussiana estándar, si X se distribuye $N_p(x|\mu, \Sigma)$, y para cada entrada X_i del vector aleatorio X consideramos la transformación $Y_i = \frac{X_i - \mu_i}{\sigma}$, entonces el vector aleatorio Y resultante tendrá una distribución $N_p(x|\mathbf{0},\boldsymbol{I})$, donde $\mathbf{0}$ es el vector cero de dimensión $\boldsymbol{p}$, mientras que $\boldsymbol{I}$ es la matriz identidad de dimensión $\boldsymbol{p} \times \boldsymbol{p}$. A esta distribución se le conoce como gaussiana estándar $\boldsymbol{p}$-variada.

Por último, la función generadora de momentos y característica de la distribución normal $\boldsymbol{p}$-variada son las siguiente:

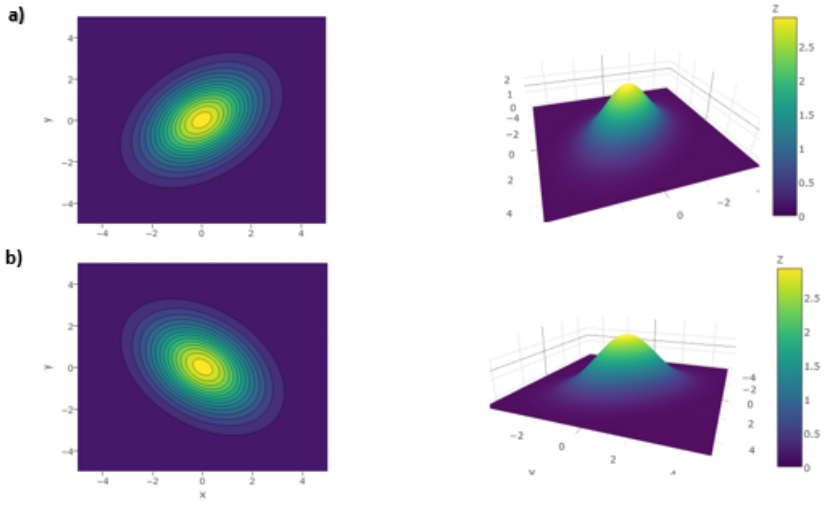

Figura 3.5: Distribución gaussiana bivariada con diferentes valores de covarianza

$$m_x(t) = \exp\{\mu t + \frac{t\Sigma t}{2}\} \tag{3.4}$$

$$\Phi_x(it) = \exp\{\mu it - \frac{t\Sigma t}{2}\} \tag{3.5}$$

3.3. Distribución gamma

Se dice que una variable aleatoria $\boldsymbol{X}$, que toma valores en los números reales positivos, se distribuye gamma con parámetros $\boldsymbol{\lambda}$ y $\boldsymbol{\alpha}$, si su función de densidad de probabilidad es de la siguiente manera:

$$f_x(x|\lambda,\alpha) = \frac{\lambda^\alpha x^{\alpha-1}}{\gamma(\alpha)} \exp\{-\lambda x\} I_{(0,\infty)}(x) \tag{3.6}$$

En la distribución gamma dos parámetros caracterizan la forma de la distribución, λ y γ. El parámetro λ toma valores mayores a cero, mientras que α puede tomar valores mayores o iguales a cero. El parámetro λ también es conocido como parámetro de escala e influye en el tamaño de la densidad respecto al eje y. Por otro lado, el parámetro α influye en la forma de la distribución.

Para referirnos a que X se distribuye gamma con parámetros λ y α, usaremos la notación $X|\lambda,\alpha$ se distribuye $\Gamma(x|\lambda,\alpha)$.

Como se mencionó anteriormente, el parámetro λ es un parámetro de escala, mientras que el parámetro α es un parámetro de forma, lo cual se ilustra a continuación.

En la gráfica 3.6 se muestran cuatro distribuciones gamma con el mismo parámetro forma $\alpha = 1$, pero con distintos parámetros de escala, es decir, con $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$ y $\lambda_4 = 4$, respectivamente.

En la gráfica 3.7 se muestran cuatro distribuciones gamma con al mismo parámetro de escala $\lambda = 1$, pero con distintos parámetros de forma, es decir, con $\alpha_1 = 1$, $\alpha_2 = 2$, $\alpha_3 = 3$ y $\alpha_4 = 4$, respectivamente. Adviertase que conforme α es menor el sesgo de la distribución aumenta a la derecha, mientras que si es menor disminuye a la derecha y aumenta a la izquierda.

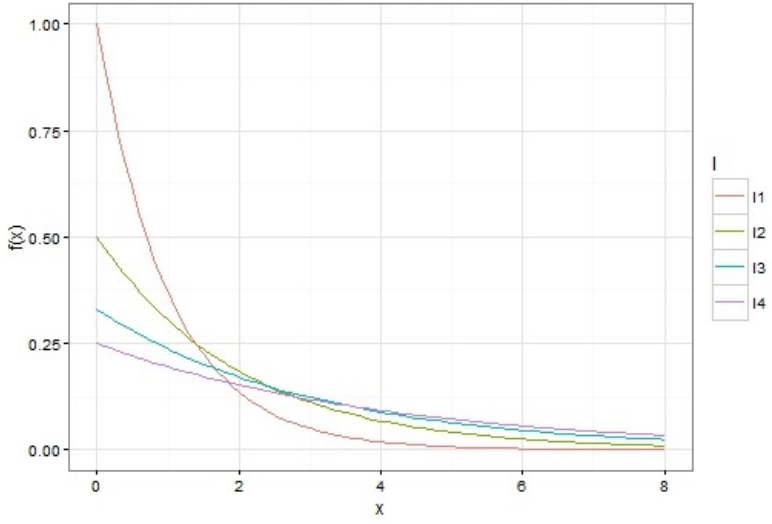

Figura 3.6: Distribución gamma con diferente parámetro $\boldsymbol{\lambda}$

Si la variable aleatoria $\boldsymbol{X}$ se distribuye gamma con parámetros $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, entonces:

$$\begin{aligned} \boldsymbol{E[x]} &= \frac{\boldsymbol{\alpha}}{\boldsymbol{\lambda}}, \\ \boldsymbol{Var[x]} &= \frac{\boldsymbol{\alpha}}{\boldsymbol{\lambda^2}}, \\ \boldsymbol{m_x(t)} &= (\frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda - t}})^{\boldsymbol{\alpha}}, \\ \boldsymbol{\Phi_x(it)} &= (\frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda - it}})^{\boldsymbol{\alpha}}, \end{aligned} \tag{3.7}$$

Por último, es importante mencionar que la distribución gamma es la distribución a priori conjugada de la distribución wishart, la cual se intruduce más adelante.

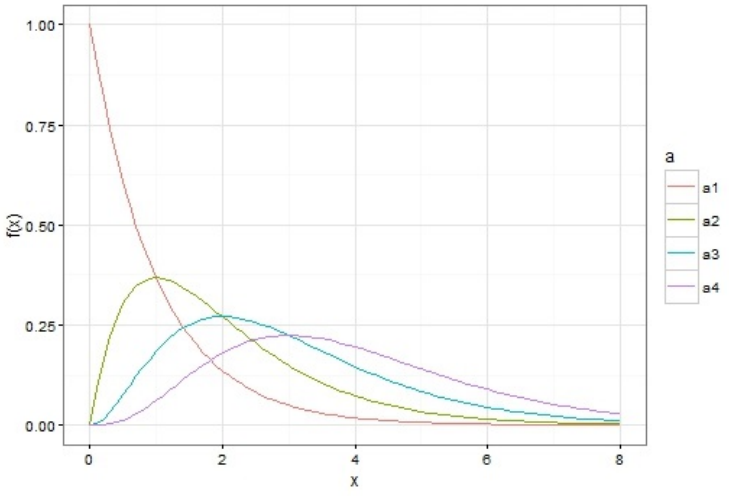

Figura 3.7: Distribución gamma con diferente parámetro $\boldsymbol{\alpha}$

3.4. Distribución gamma inversa

La distribución gamma inversa resulta de aplicar la transformación $\boldsymbol{X = \frac{1}{y}}$ a la variable aleatoria $\boldsymbol{Y}$, donde $\boldsymbol{Y}$ se distribuye gamma con parámetros $\boldsymbol{\lambda}$ y $\boldsymbol{\alpha}$. Dicho lo anterior se tiene la siguiente definición.

Se dice que una variable aleatoria $\boldsymbol{X}$, que toma valores en los números reales positivos, se distribuye gamma inversa con parámetros $\boldsymbol{\lambda}$ y $\boldsymbol{\alpha}$, si su función de densidad de probabilidad es de la siguiente manera:

$$\boldsymbol{f_x(x|\lambda,\alpha) = \frac{\lambda^{\alpha} x^{1-\alpha}}{\gamma(\alpha)} \exp\{-\frac{\lambda}{x}\} I_{(0,\infty)}(x).} \tag{3.8}$$

La distribución gamma inversa hereda sus dos parámetros de la distribución gamma, por lo que tanto $\boldsymbol{\lambda}$ como $\boldsymbol{\alpha}$ tienen las mismas restricciones que en la distribución gamma, y también caracterizan la forma de la distribución. Al igual que en la distribución gamma, el parámetro $\boldsymbol{\lambda}$ es conocido como parámetro de escala e influye en el tamaño de la densidad respecto al

eje $\boldsymbol{y}$. Mientras que el parámetro $\boldsymbol{\alpha}$ influye en la forma de la distribución.

Para referirnos a que $\boldsymbol{X}$ se distribuye gamma inversa con parámetros $\boldsymbol{\lambda}$ y $\boldsymbol{\alpha}$, usaremos la notación $\boldsymbol{X|\lambda,\alpha}$ se distribuye $\boldsymbol{\Gamma Inv(x|\lambda,\alpha)}$.

Como se mencionó anteriormente, el parámetro $\boldsymbol{\lambda}$ es un parámetro de escala, mientras que el parámetro $\boldsymbol{\alpha}$ es un parámetro de forma, lo cual se ilustra a continuación.

En la gráfica 3.8 se muestran cuatro distribuciones gamma inversa con el mismo parámetro de forma $\boldsymbol{\alpha = 1}$, pero con distintos parámetros de escala, es decir, con $\boldsymbol{\lambda_1 = 1}$, $\boldsymbol{\lambda_2 = 2}$, $\boldsymbol{\lambda_3 = 3}$ y $\boldsymbol{\lambda_4 = 4}$, respectivamente.

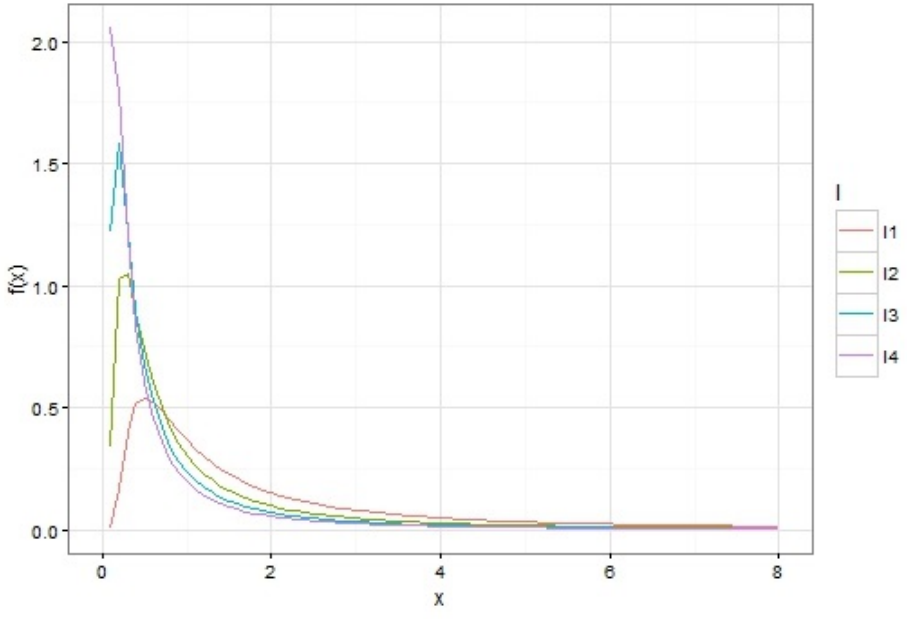

Figura 3.8: Distribución gamma inversa con diferente parámetro $\boldsymbol{\lambda}$

En la gráfica 3.9 se muestran cuatro distribuciones gamma inversa con al mismo parámetro de escala $\boldsymbol{\lambda = 1}$, pero con distintos parámetros de forma, es decir, con $\boldsymbol{\alpha_1 = 1}$, $\boldsymbol{\alpha_2 = 2}$, $\boldsymbol{\alpha_3 = 3}$ y $\boldsymbol{\alpha_4 = 4}$, respectivamente.

Adviertase que conforme $\boldsymbol{\alpha}$ es menor la curtosis de la distribución aumenta, mientras que si es mayor disminuye.

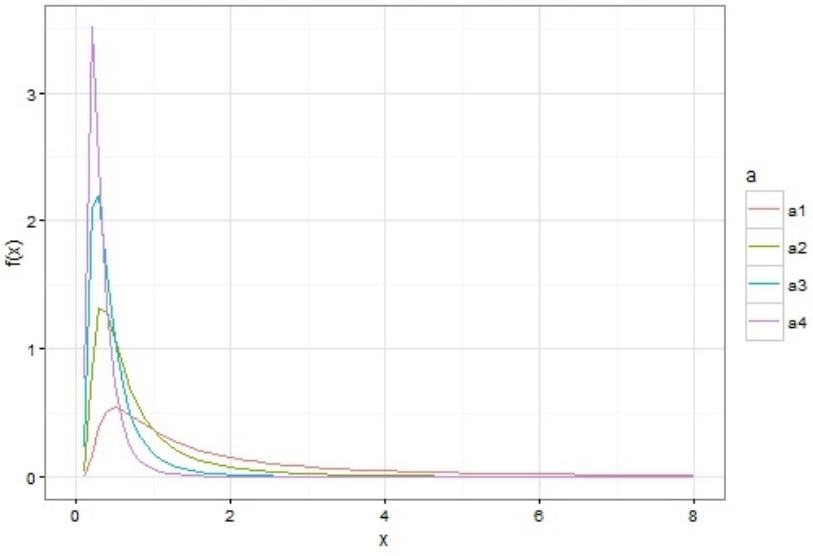

Figura 3.9: Distribución gamma inversa con diferente parámetro $\boldsymbol{\alpha}$

Si la variable aleatoria $\boldsymbol{X}$ se distribuye gamma inversa con parámetros $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, entonces:

$$\begin{aligned} \boldsymbol{E[x]} &= \frac{\boldsymbol{\alpha}}{\boldsymbol{\alpha-1}}, \\ \boldsymbol{Var[x]} &= \frac{\boldsymbol{\lambda^2}}{\boldsymbol{(\alpha-1)^2(\alpha-1)}}, \\ \boldsymbol{\Phi_x(it)} &= (\frac{\boldsymbol{2(-\lambda it)^{\alpha/2}}}{\boldsymbol{\gamma(\alpha)}})\boldsymbol{\kappa}(\sqrt{\boldsymbol{-4\lambda it}}), \end{aligned} \tag{3.9}$$

Por último, es importante mencionar que la distribución Gamma Inversa es la distribución apriori conjugada de la distribución Wishart Inversa, la cual se intruduce más adelante.

3.5. Distribución Wishart

La distribución whishart es utilizada como distribución de la matriz de varianza-covarianza de vectores aleatorios normales de dimensión $\boldsymbol{p}$, y se deriva de la siguiente manera:

Supongamos que tenemos $\boldsymbol{n}$ vectores aleatorios $\boldsymbol{X}$ de dimensión $\boldsymbol{p}$ que se distribuyen $\boldsymbol{N_p(X|0,\Sigma)}$, luego un estimador de la matriz de varianza-covarianza es $\boldsymbol{S = \Sigma_{i=1}^{n} x_i x_i / n}$, que es una matriz positiva y simétrica de dimensión $\boldsymbol{p \times p}$, por lo que para los distintos valores que tomen los vectores aleatorios $\boldsymbol{X}$, $\boldsymbol{S}$ también tomará un valor diferente, por lo que es natural preguntarse por la distribución de $\boldsymbol{S}$, con lo cual se llega a la siguiente caracterización.

Se dice que una matriz $\boldsymbol{S}$ de dimensión $\boldsymbol{pxp}$, simétrica y positiva definida, se distribuye wishart con $\boldsymbol{n}$ grados de libertad, si su densidad es de la siguiente forma:

$$\boldsymbol{f_S(S|\Sigma,n,p) = c\frac{|S|^{(n-p-1)/2}}{|\Sigma|^{n/2}}\exp(-\frac{1}{2}tr(\Sigma^{-1}S)),}$$

donde $\boldsymbol{c = \left(2^{np/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\gamma(\frac{n+1-i}{2})\right)^{-1}}$, $\boldsymbol{\Sigma}$ es una matriz simétrica y positiva definida de dimensión $\boldsymbol{pxp}$, y se le conoce como matriz de escala, $\boldsymbol{n}$ es elnúmero de vectores disponibles, y es mayor a la dimensión de los vectores, es decir, mayor que $\boldsymbol{p}$. En este caso usaremos la notación $\boldsymbol{S}$ se distribuye $\boldsymbol{W(S|\Sigma,n,p)}$.

Algunas características numéricas de la distribución Whisart son las siguientes:

$$\boldsymbol{E[S] = p\Sigma,}$$

La distribución marginal del $\boldsymbol{i}$-ésimo componente de la distribución wishart se distribuye $\boldsymbol{\Gamma}(\frac{\boldsymbol{1}}{\boldsymbol{2}},\frac{\boldsymbol{1}}{\boldsymbol{2}})$.

3.6. Distribución Wishart inversa

Se dice que una matriz $\boldsymbol{G}$ de dimensión $\boldsymbol{p}\times\boldsymbol{p}$, simétrica y positiva definida, se distribuye wishart inversa con $\boldsymbol{n}$ grados de libertad, si su función de densidad es de la siguiente forma:

$$\boldsymbol{f_G(G)} = \boldsymbol{c}\frac{|\boldsymbol{K}|^{(n-p-1)/2}}{|\boldsymbol{G}|^{n/2}}\exp(-\frac{\boldsymbol{1}}{\boldsymbol{2}}\boldsymbol{tr}(\boldsymbol{G^{-1}K})),$$

donde $\boldsymbol{c} = \left(\boldsymbol{2^{(n-p-1)p/2}\pi^{p(p-1)/4}}\prod_{\boldsymbol{i=1}}^{\boldsymbol{p}}\boldsymbol{\gamma}(\frac{\boldsymbol{n-p-i}}{\boldsymbol{2}})\right)^{\boldsymbol{-1}}$, $\boldsymbol{K}$ es una matriz simétrica y definida positiva, y se le conoce como matriz de escala. En este caso usaremos la notación $\boldsymbol{G}$ se distribuye $\boldsymbol{W^{-1}(G|K,n,p)}$.

Una propiedad importante que liga a la distribución wishart y wishart inversa, es la siguiente.

Si una matriz aleatoria $\boldsymbol{\Sigma}$ de dimensión $\boldsymbol{p}\times\boldsymbol{p}$ se distribuye whisart con matriz de escala $\boldsymbol{S}$, y con $\boldsymbol{n}$ grados de libertad, entonces su matriz inversa, $\boldsymbol{\Sigma^{-1}}$, se distribuye whisart inversa con matriz de escala $\boldsymbol{S^{-1}}$, y con $\boldsymbol{n+p+1}$ grados de libertad.

3.7. Distribución gaussiana inversa generalizada

Se dice que la variable aleatori $\boldsymbol{X}$, que toma valores en los números reales positivos, tiene una distribución gaussiana inversa generalizada, denotada por $\boldsymbol{GIG(x|\lambda,\xi,\Psi)}$, si su densidad es de la siguiente forma:

$$f(x) = \frac{\xi^{-\lambda}\sqrt{\xi\Psi}^{\lambda}x^{\lambda-1}\exp\frac{-1}{2}(\xi x^{-1}+\Psi x)}{2\kappa_{\lambda}(\sqrt{\xi\Psi})}I_{(0,\infty)}(x),$$

donde $\kappa_{\lambda(.)}$ es una función modificada de Bessel de tercer tipo, y si $\lambda < 0$, entonces $\xi > 0$, $\Psi \geq 0$; si $\lambda = 0$, entonces $\xi > 0$, $\Psi > 0$, si $\lambda > 0$, entonces $\xi \geq 0$, $\Psi > 0$.

En la distribución gaussiana inversa generalizada tres parámetros caracterizan la forma de la distribución. Tanto el parámetro ξ como el parámetro Ψ influyen en la escala de la distribución, mientras que el parámetro λ en la forma. El punto anterior se ilustra a continuación.

En la gráfica 3.10 se muestran cuatro distribuciones GIG, todas con los mismos parámetros $\lambda = 1$ y $\xi = 1$, pero con diferentes parámetros $\psi_1 = 0{,}5$, $\psi_2 = 1$, $\psi_3 = 2$ y $\psi_4 = 4$, respectivamente. Es importante notar que entre menor es el valor de ψ, mayor es la curtosis de la distribución.

En la gráfica 3.11 se observan cuatro distribuciones GIG, todas con paámetros $\lambda = 1$, $\psi = 1$, pero con diferentes parámetros $\xi_1 = 0{,}5$, $\xi_2 = 1$, $\xi_3 = 2$ y $\xi_4 = 4$, respectivamente. Es importante notar que entre mayor es el valor del parámetro ξ, mayor es el sesgo a la derecha de la distribución.

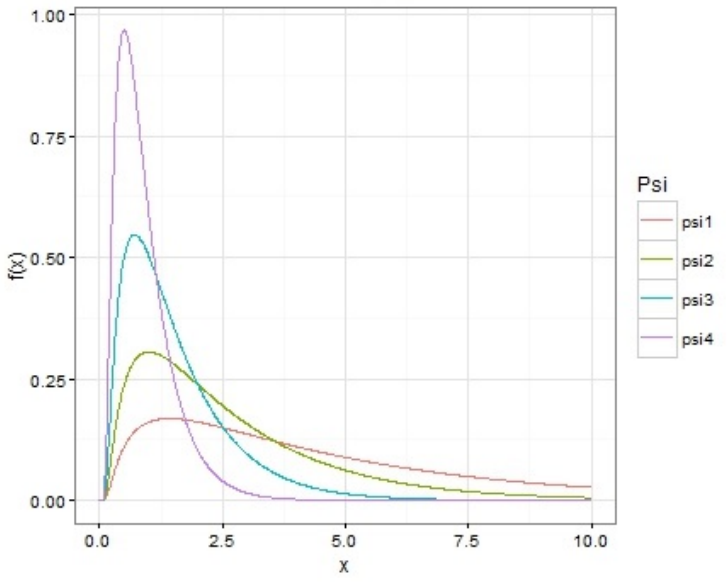

Figura 3.10: Distribución GIG con diferentes valores del parámetro $\boldsymbol{\psi}$.

En la gráfica 3.12 se observan cuatro distribuciones GIG, todas con parámetros $\boldsymbol{\psi = 1}$, $\boldsymbol{\xi = 1}$, pero con diferentes parámetros $\boldsymbol{\lambda_1 = -0{,}5}$, $\boldsymbol{\lambda_2 = 0}$, $\boldsymbol{\lambda_3 = {,}5}$ y $\boldsymbol{\lambda_4 = 1}$, respectivamente. En este caso el parámetro $\boldsymbol{\lambda}$ también influye en la curtosis de la distribución, ero sobre todo influye en la cola de esta. También el parámetro $\boldsymbol{\lambda}$ influye en la failia paramétrica, pues si la variable aleatoria $\boldsymbol{X}$ se distribuye GIG, con parámetro $\boldsymbol{\lambda = 0}$ entonces se obtiene la distribución hiperbólica, mientras que si $\boldsymbol{\lambda = -0{,}5}$ se obtiene la distribución Gaussiana Inversa de la cual se hablará posteriormente.

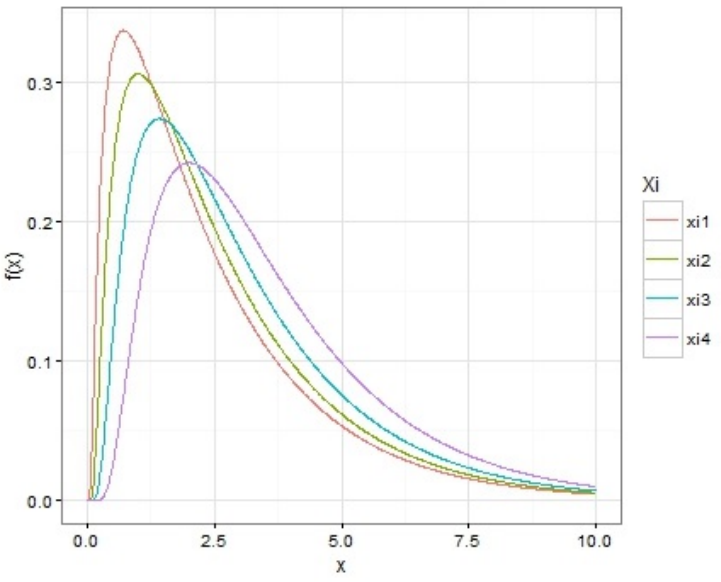

Figura 3.11: Distribución GIG con diferentes valores del parámetro $\boldsymbol{\xi}$

Ahora veamos algunos resultados de la distribución GIG. Si $\boldsymbol{X}$ se distribuye $\boldsymbol{\mathcal{N}(x|\lambda,\xi,\Phi)}$, entonces su función generadora de momentos es:

$$\begin{aligned}\boldsymbol{\Phi(it)} &= \int_{-\infty}^{\infty} \boldsymbol{\exp(itx)} \frac{\boldsymbol{\xi^{-\lambda}\sqrt{\xi\Psi}^{\lambda} x^{\lambda-1} \exp\frac{-1}{2}(\xi x^{-1}+\Psi x)}}{\boldsymbol{2\kappa_{\lambda}(\sqrt{\xi\Psi})}} \boldsymbol{dx} \\ &= \boldsymbol{c}\int_{-\infty}^{\infty} \boldsymbol{\xi^{-\lambda}} \frac{\boldsymbol{\sqrt{\xi(2it+\Psi)}^{\lambda}}}{\boldsymbol{2\kappa_{\lambda}(\sqrt{\xi(2it+\Psi)})}} \boldsymbol{\exp\frac{-1}{2}(\xi x^{-1}+(2it+\Psi)x)dx},\end{aligned}$$

con $\boldsymbol{c} = \frac{\sqrt{\xi\Psi}^{\lambda}}{2\kappa_{\lambda}(\sqrt{\xi\Psi})} \frac{2\kappa_{\lambda}(\sqrt{\xi(2it+\Psi)})}{\sqrt{\xi(2it+\Psi)}^{\lambda}}$. La última integral vale uno por ser una densidad $\boldsymbol{\mathcal{N}(\lambda,\xi,2it+\psi)}$ integrada sobre su soporte, por lo que:

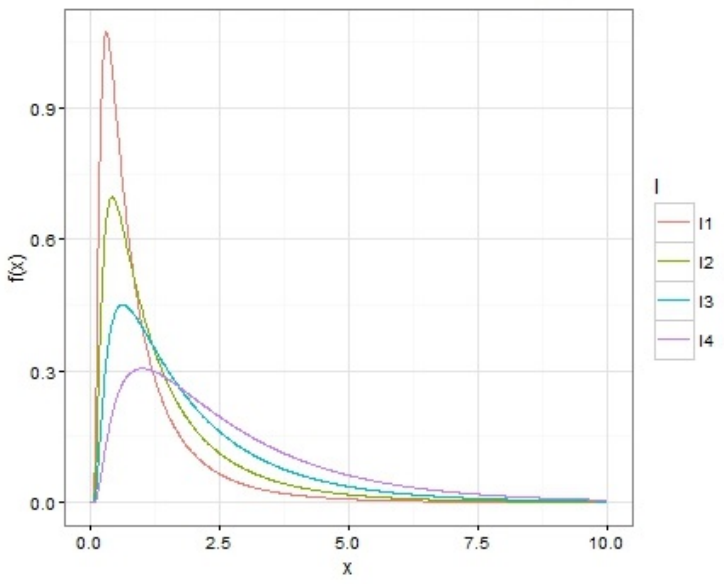

Figura 3.12: Distribución GIG con diferentes valores de $\boldsymbol{\lambda}$

$$\boldsymbol{\Phi(it) = \frac{\sqrt{\xi\Psi}^{\lambda}}{\kappa_{\lambda}(\sqrt{\xi\Psi})} \frac{\kappa_{\lambda}(\sqrt{\xi(2it+\Psi)})}{\sqrt{\xi(2it+\Psi)}^{\lambda}}.}$$

Si $\boldsymbol{X}$ se distribuye $\boldsymbol{\mathcal{N}(x|\lambda,\xi,\Phi)}$, entonces su $\boldsymbol{r}$-ésimo momento es:

$$\boldsymbol{E[X^r] = \int_{-\infty}^{\infty} x^r \frac{\xi^{-\lambda}\sqrt{\xi\Psi}^{\lambda} x^{\lambda-1} \exp\frac{-1}{2}(\xi x^{-1} + \Psi x)}{2\kappa_{\lambda}(\sqrt{\xi\Psi})} dx}$$

$$\boldsymbol{= \frac{x^r 2\kappa_{\lambda+r}(\sqrt{\xi\Psi})}{\sqrt{\xi\Psi}^r 2\kappa_{\lambda}(\sqrt{\xi\Psi})} \int_{-\infty}^{\infty} \frac{\xi^{-\lambda-r}\sqrt{\xi\Psi}^{\lambda+r} x^{\lambda+r-1} \exp\frac{-1}{2}(\xi x^{-1}+\Psi x)}{2\kappa_{\lambda+r}(\sqrt{\xi\Psi})} dx}$$

Por lo tanto:

$$\boldsymbol{E[X^r] = (\frac{\xi}{\Psi})^{\frac{r}{2}} \frac{\kappa_{\lambda+r}(\sqrt{\xi\Psi})}{\sqrt{\kappa_\lambda(\sqrt{\xi\Psi})}}}$$

3.8. Distribución gaussiana inversa

Se dice que la variable aleatoria $\boldsymbol{X}$ tiene una distribución gaussiana inversa si su función de densidad es de la siguiente forma:

$$\boldsymbol{f_x(X|\lambda,\psi) = \sqrt{\frac{\lambda}{2\pi x^3}}\exp(-\frac{\lambda(x-\Psi)^2}{2\Psi^2 x}),}$$

para $\boldsymbol{\lambda}$, $\boldsymbol{\psi}$ y $\boldsymbol{X}$ mayores a cero. Como notación, si X se distribuye gaussiana inversa con parámetros $\boldsymbol{\lambda}$, $\boldsymbol{\Psi}$, diremos que $\boldsymbol{X|\lambda,\psi}$ se distribuye $\boldsymbol{GI(X|\lambda,\Psi)}$.

Ahora veamos como es la función característica de esta distribución. Si $\boldsymbol{X}$ se distribuye $\boldsymbol{GI(x|\lambda,\Psi)}$, entonces su función característica es:

$$\begin{aligned}\boldsymbol{\Phi_x(it)} &= \boldsymbol{\int_0^\infty \exp^{itx}\sqrt{\frac{\lambda}{2\pi x^3}}\exp^{-\frac{\lambda(x-\Psi)^2}{2\Psi^2 x}}dx} \\ &= \boldsymbol{\int_0^\infty \sqrt{\frac{\lambda}{2\pi x^3}}\exp(-\frac{\lambda}{2\Psi^2}(x-2\Psi-(itx2\Psi^2/\lambda)+\Psi^2 x))dx}\end{aligned}$$

Trabajando únicamente con el exponente de la expresión anterior, y facto-

rizando el término $(\mathbf{1} - \frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})$ tenemo que:

$$
\begin{aligned}
&-\frac{\boldsymbol{\lambda}}{\boldsymbol{2\Psi^2}}(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})(x-\boldsymbol{2\Psi}/(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})+\boldsymbol{\Psi^2}/(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})\\
&=-\frac{\boldsymbol{\lambda}}{\boldsymbol{2\Psi^2}}(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})(x-\boldsymbol{2\Psi}/(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})+\boldsymbol{\Psi^2}/x(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})\\
&+\boldsymbol{2\Psi}/\sqrt{1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}}}-\boldsymbol{2\Psi}/\sqrt{1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}}})\\
&=\frac{\boldsymbol{\lambda}}{\boldsymbol{\Psi}}-\frac{\boldsymbol{\lambda}}{\boldsymbol{\Psi}}\sqrt{(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})}-\frac{\boldsymbol{\lambda}}{\boldsymbol{2x\Psi^2}}(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})(x-\boldsymbol{\Psi}/\sqrt{1-\boldsymbol{2it\Psi^2}/\boldsymbol{\lambda}})^2
\end{aligned}
$$

Por lo que la función característica queda de la siguiente forma:

$$
\begin{aligned}
\boldsymbol{\Phi_x(it)} &= (\exp^{\frac{\boldsymbol{\lambda}}{\boldsymbol{\Psi}}-\frac{\boldsymbol{\lambda}}{\boldsymbol{\Psi}}\sqrt{(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})}})\\
&\quad(\int_0^\infty \sqrt{\frac{\boldsymbol{\lambda}}{\boldsymbol{2\pi x^3}}}\exp^{\frac{\boldsymbol{\lambda}}{\boldsymbol{2x\Psi^2}}(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})(x-\boldsymbol{\Psi}/\sqrt{1-\boldsymbol{2it\Psi^2}/\boldsymbol{\lambda}})^2}\,\boldsymbol{dx})\\
&= \exp(\frac{\boldsymbol{\lambda}}{\boldsymbol{\Psi}}-\frac{\boldsymbol{\lambda}}{\boldsymbol{\Psi}}\sqrt{(1-\frac{\boldsymbol{it2\Psi^2}}{\boldsymbol{\lambda}})})
\end{aligned}
$$

El resultado anterior se sigue de que la integral previa es una distribución Gaussiana Inversa con parámetros $\boldsymbol{\lambda}$, $\boldsymbol{\Psi}/\sqrt{\mathbf{1}-\boldsymbol{(2it\Psi^2)/\lambda}}$ integrada sobre su soporte, por lo cual vale $\mathbf{1}$, y por lo tanto:

$$\boldsymbol{\Phi_x(it)} = \exp^{\frac{\boldsymbol{\lambda}}{\boldsymbol{\Psi}}-\frac{\boldsymbol{\lambda}}{\boldsymbol{\Psi}}\sqrt{(1-\frac{\boldsymbol{2it\Psi^2}}{\boldsymbol{\lambda}})}} \tag{3.2}$$

Con el resultado anterior se puede probar que si $\boldsymbol{X}$ se distribuye $\boldsymbol{GI(x|\lambda,\Psi)}$,

entonces:

$$\boldsymbol{m_x(t) = \exp^{\frac{\lambda}{\Psi} - \frac{\lambda}{\Psi}\sqrt{(1-\frac{2t\Psi^2}{\lambda})}}}, \qquad \boldsymbol{E[X] = \Psi},$$
$$\boldsymbol{Var[X] = \frac{\Psi^3}{3}}, \qquad \boldsymbol{Sesgo[X] = \frac{2\Psi^4}{\lambda} + \frac{3\Psi^5}{\lambda^2}}.$$

Entoces, al igual que la distibución gaussiana, la distribución gaussiana inversa tiene un parámetro de localización, es decir, $\boldsymbol{\Psi}$. Es importante notar como afecta el parámetro $\boldsymbol{\lambda}$ a la dispersión de la variable aleatpria $\boldsymbol{X}$, y además, esta distribución siempre tiene un sesgo positivo ya que $\boldsymbol{\Psi}$ y $\boldsymbol{\lambda}$ son mayores a cero. Por lo que para valores muy pequeños de $\boldsymbol{\lambda}$ o muy grandes de $\boldsymbol{\Psi}$ la distribución tendrá más peso en la cola.

Bibliografía

Sheldon Ross. *A First Course in Probability.* Pearson/Prentice Hall, New York, 1976.

Apéndice A

Estimación por máxima verosimilitud

El método de máxima verosimilitud tiene un enfoque frecuentista, en el que los datos son generados por una función de densidad de probabilidad $\boldsymbol{f_x}$, con la cual es posible construir una función de verosimilitud, la cual tiene como argumento los parámetros a estimar, y a su vez depende de los datos ya observados. Por lo regular la función de verosimilitud se construye de acuerdo al evento, concerniente a los datos de interés, por lo que muchas veces es la misma función de densidad, pues se considera el evento en que la variable aleatoria $\boldsymbol{X_i}$ tome el valor $\boldsymbol{x_i}$. De manera intuitiva, la función de verosimilitud representa la probabilidad de haber observado la variable aleatoria $\boldsymbol{X}$ bajo el modelo dado, la cual se pretende maximizar con respecto al parámetro de interés para así obtener qué tan probable es que el modelo dado haya generado los datos.

En términos generales los estimadores máximo verosimiles se construyen de la siguiente forma:

1. Construir la función de verosimilitud, la cual llamaremos $\boldsymbol{LIK(\Theta|X)}$

2. Plantear el problema de maximización:

$$\boldsymbol{max_{\Theta}(LIK(\Theta|X))}$$

3. La solucion $\boldsymbol{\Theta*}$ del problema de maximización es el estimador buscado.

Aunque el algoritmo para encontrar estimadores máximo verosimiles parece sencillo, no siempre lo es, pues la dificultad del proceso de maximización aumenta conforme más variables se tienen, lo cual puede llevar a procesos computacionales poco eficientes. Estas dificultades incentiva el uso del algoritmo EM.

Apéndice B

Algoritmo EM

El algoritmo Expectation Maximization (EM) fue formalizado por Dempster et al (1977); este algoritmo tiene un enfoque frecuentista, y tiene como objetivo encontrar los estimadores máximo verosímiles de funciones de densidad con datos no observados, lo cual es ideal para manejar distribuciones tipo mezcla, pues la variable de mezcla toma el lugar de los datos no observados.

Para implementar el algoritmo EM se requiere un conjunto de datos u observaciones $\boldsymbol{x_i}$, de las cuales se conoce su función paramétrica de densidad; el parámetro de dicha distribución, $\boldsymbol{\Theta}$, es desconocido y es lo que se pretende estimar; también se requiere la distribución, parametrizada por $\boldsymbol{\Theta}$, de los datos no observados y finalmente una suposición inicial, $\boldsymbol{\Theta^0}$, sobre el parámetro $\boldsymbol{\Theta}$.

De manera intuitiva, el algorimo EM se desarrolla de la siguiente forma:llamemos a la función de densidad asociada a $\boldsymbol{x_i}$ como $\boldsymbol{f(x_i|\Theta)}$, después supongamos que existen datos no observados $\boldsymbol{z}$, y una función determinista $\boldsymbol{H(.)}$ que tiene como dominio los datos no observados y como imagen los datos observados. Esto es, que para cada dato no observado se cumple que $\boldsymbol{H(Z_s) = x_i}$, para $\boldsymbol{Z_s}$ un subconjunto de $\boldsymbol{Z}$. Luego nos interesaría

encontrar el valor de $\boldsymbol{\Theta}$ que maximiza la probabilidad de haber obtenido los datos no observados dado $\boldsymbol{\Theta}$, pero justamente no conocemos los datos no observados $\boldsymbol{X}$, por lo que la probabilidad anterior la ponderamos por la probabilidad de haber obtenido los datos no observados dados los datos sí observados y la suposición inicial del parámetro de interés, luego nos interesa maximizar $\boldsymbol{f(Z|\Theta)f(Z|X,\Theta^0)}$, lo cual no nos libra del problema de los datos no observados. Entonces, nos fijamos en un promedio de todas las posibles posibilidades de $\boldsymbol{Z}$, por lo que nuestra función a maximizar con respecto a $\boldsymbol{\Theta}$ se convierte en:

$$\boldsymbol{E_Z[f(Z|\Theta)] = \int_Z f(Z|\Theta)f(Z|X,\Theta^0)dz}$$

Por último, si se define a la esperanza descrita previamente como una función de $\boldsymbol{\Theta}$, es decir, que $\boldsymbol{Q(\Theta|\Theta^0) = E_Z[f(Z|\Theta)]}$, entonces el problema planteado se resume en dado un valor $\boldsymbol{\Theta^0}$, maximizar con respecto a $\boldsymbol{\Theta}$ la función $\boldsymbol{Q(\Theta|\Theta^0)}$. De esta forma se crea un proceso iterativo, donde el valor $\boldsymbol{\Theta^*}$ que maximiza la función $\boldsymbol{Q(.)}$ se convierte ahora en $\boldsymbol{\Theta^0}$, y el procedimiento se repite de nuevo; de esta forma se asegura que la esperanza converge a la verosimilitud buscada, y a su vez, $\boldsymbol{\Theta^*}$ converge al estimador máximo verosímil. (citar dónde viene la demostración de esto)

En términos generales el estimador máximo verosimil a través del algoritmo EM se construye de la siguiente manera:

1. Dar un valor inicial $\boldsymbol{\Theta^0}$.

2. Obtener las funciones de densidad $\boldsymbol{f(Z|\Theta)}$, $\boldsymbol{f(Z|X,\Theta^0)}$.

3. Calcular $\boldsymbol{Q(\Theta|\Theta^0) = E_Z[f(Z|\Theta)]}$, algunas veces resulta más sencillo calcular $\boldsymbol{E_Z[log(f(Z|\Theta))]}$.

4. Maximizar $\boldsymbol{Q(\Theta|\Theta^0)}$.

5. Una vez obtenido $\boldsymbol{\Theta}^*$ sustituir por $\boldsymbol{\Theta}^{\mathbf{0}}$, y repetir los pasos **3**, **4** y **5**, hasta que el algoritmo converja.

Apéndice C

Slice sampler

Los métodos de simulación slice sampler consisten en generar una cadena de Markov que converja a la distribución que se desea muestrear. La idea intuitiva de estos métodos consiste en suponer que el vector aleatorio bivariado $(\boldsymbol{X}, \boldsymbol{Y})$ se distribuye uniforme en la región que está por debajo de la función de densidad, de la cual obtendríamos una muestra $(\boldsymbol{X_0}, \boldsymbol{Y_0})$, y de está solo nos quedariamos con $\boldsymbol{X_0}$, siendo esta nuestra variable de interés.

Ahora, para generar una muestra uniforme en dicha región, se utiliza un algoritmo recursivo con el fin de tener una cadena de markov que converja a la muestra deseada.

En términos ilustrativos el procedimiento es como sigue:

1) Supongamos que tenemos el kernel o una densidad de una variable aleatoria univarada $\boldsymbol{X}$, de la cual nos interesaría obtener una muestra aleatoria, y que además la gráfica de esta densidad es de la siguiente manera, por ejemlo:

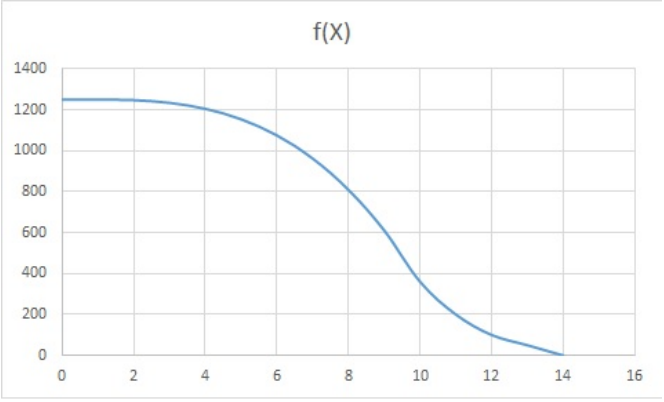

Figura C.1

2) Luego definimos el vector aleatorio bivariado $(\boldsymbol{X}, \boldsymbol{Y})$, el cual suponemos que se distribuye uniforme en la región que está por debajo de la gráfica anterior, y a esta región la denotamos como $\boldsymbol{U}$, por lo que $\boldsymbol{U}$ sería el conjunto de parejas $(\boldsymbol{x}, \boldsymbol{y})$ con la propiedad de que $\boldsymbol{x}$ es menor a $\boldsymbol{y}$, siendo $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$.
Entonces la región de la cual nos interesa obtener una muestra se ve de la siguiente forma:

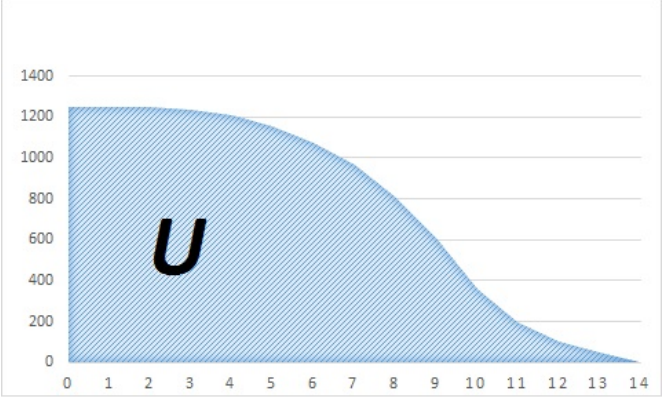

Figura C.2

3) Ahora, proponemos un valor inicial $\boldsymbol{x_0}$, y por el anexo (algo), si un vector aleatorio bivariado se distribuye uniforme en alguna región, entonces para cada valor que tome la variable $\boldsymbol{X}$, la distribución marginal de $\boldsymbol{Y}$ dado $\boldsymbol{X = x_0}$ se distribuye uniforme dentro del segmento de recta $\boldsymbol{L(y) = ((y, x_0)|y \in R)}$, donde $\boldsymbol{L(y)}$ está dentro de $\boldsymbol{U}$. De esta manera, podríamos generar una realización de $\boldsymbol{Y}$ dado $\boldsymbol{x_0}$, así obtendríamos algún valor $\boldsymbol{y_0}$, después fijamos éste valor $\boldsymbol{y_0}$, y tendríamos ahora que $\boldsymbol{X}$ dado $\boldsymbol{y_0}$ se distribuye uniforme en $\boldsymbol{U \cap L(x) = ((x, y_0)|x \in R)}$, así obtendríamos una realización de $\boldsymbol{X}$ dado $\boldsymbol{y_0}$ de donde obtenríamos nuevamente un valor $\boldsymbol{x_0}$. Así repetiriamos el proceso, pues de esta manera tenemos una cadena de Markov que converge a la realización de haber obtenido una muestra uniforme sobre $\boldsymbol{U}$, donde sólo nos interesa el valor de $\boldsymbol{X}$.
En la siguiente gráfica se ilustra el punto **3)**, por simplicidad se supuso que el primer valor que tomó la variable $\boldsymbol{X}$ fue 0, luego $\boldsymbol{Y}$ dado $\boldsymbol{x}$ igual a cero se distribuye uniforme en el intervalo $\boldsymbol{(0, 1200)}$, de donde ahora obtenemos una realización de $\boldsymbol{Y}$ dado $\boldsymbol{x}$ igual a cero, donde Y tomó ahora el valor de **500**, con esta nueva información actualizamos el valor de $\boldsymbol{Y}$, por lo que ahora obtenemos una realización de $\boldsymbol{X}$ dado $\boldsymbol{Y}$ igual a 500, y así seguimos actualizando los valores de $\boldsymbol{X}$ y $\boldsymbol{Y}$ hasta obtener la muestra deseada.

C.1. Simulación de distribuciones tipo mezcla normal en esperanza varianza

Para generar distribuciones tipo mezcla normal en esperanza varianza basta con seguir los siguieentes pasos:

1. Proponer valores para el vector de medias $\boldsymbol{\mu}$, el vector de sesgo $\boldsymbol{\beta}$, y la matriz de varianza-covarianza $\boldsymbol{\Sigma}$.

2. Simular una realización de una variable aleatoria con soporte en los reales positivos, esta será nuestra variable de mezcla $\boldsymbol{u}$.

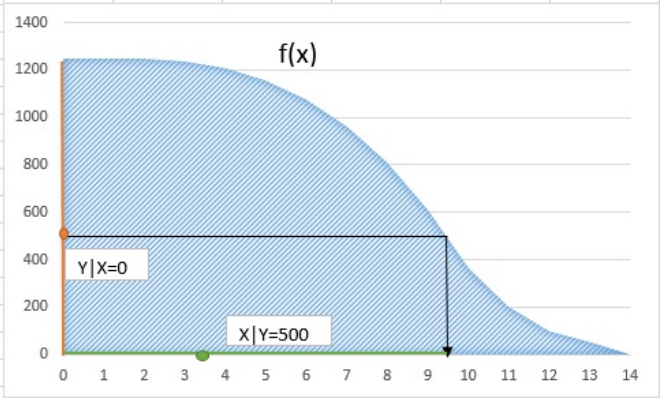

Figura C.3

3. Realizar una simulación de $\boldsymbol{X}$, donde $\boldsymbol{X}$ se distribuye normal con vector de medias $\boldsymbol{\mu + u\beta}$ y matriz de varianza-covarianza $\boldsymbol{\Sigma}$.

4. Realizar los pasos **2**) y **3**) hasta obtener la muestra deseada.

C.2. Distribución tipo mezla normal p variada

Se dice que el vector aleatorio $\boldsymbol{X \in R^p}$ con densidad f(x) puede ser expresado como una distribución de mezcla normal $\boldsymbol{N(X|\mu, \beta, \Sigma)}$ con variable de mezcla $\boldsymbol{u}$ con densidad $\boldsymbol{f(u)}$ y soporte en $\boldsymbol{R^+}$, si:

$$\boldsymbol{f(X|\mu, \beta, \Sigma) = \int_0^\infty N_p(X|\mu, u\beta, u\Sigma) f_u(u) du}$$

Donde, $\boldsymbol{N_p(X|\mu, u\beta, u\Sigma)}$ es una distribución normal p-variada con vector de medias $\boldsymbol{\mu + u\beta}$, y matriz de varianza-covarianza $\boldsymbol{u\Sigma}$

C.3. Covarianza de un vector aleatorio p variado

Se define la covarianza de un vector aleatorio p variado como:

$$\boldsymbol{COV}(\boldsymbol{X}) = \boldsymbol{E}((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})')$$

Siempre y cuando el vector de medias (del vector aleatorio $\boldsymbol{X}$) $\boldsymbol{\mu}$ exista.

C.4. Covarianza de una distribución tipo mezcla

Sea $\boldsymbol{X}$ dado $\boldsymbol{u}$ una distribución tipo mezcla de dimención p, y $\boldsymbol{u}$ una variable de mezcla con soporte en $\boldsymbol{R}_+$, entonces la covarianza de $\boldsymbol{X}$ se puede calcular como:

$$\boldsymbol{COV}(\boldsymbol{X}) = \boldsymbol{E}_{\boldsymbol{u}}(\boldsymbol{COV}(\boldsymbol{X}|\boldsymbol{u})) - \boldsymbol{COV}_{\boldsymbol{u}}(\boldsymbol{E}(\boldsymbol{X}|\boldsymbol{u}))$$

Demostración: De la definición de $\boldsymbol{COV(X)}$ tenemos que:

$$\begin{aligned}
\boldsymbol{COV(X)} &= \boldsymbol{E_x((X-\mu)(X-\mu)')} \\
&= \boldsymbol{E_x(XX' - 2X\mu + \mu\mu')} \\
&= \boldsymbol{E_x(XX') - 2E(X)\mu + \mu\mu'} \\
&= \boldsymbol{E_x(XX') - E_x(X)E_x(X)'} \\
&= \boldsymbol{E_x(XX') - E(X)_xE_x(X)' + E(X)_xE(X)'_x -} \\
&\quad \boldsymbol{2E(X)_xE(X)'_x + E(X)_xE(X)_x} \\
&= \boldsymbol{E_u(E_x(XX'|u) - E_x(X|u)E_x(X|u)') +} \\
&\quad \boldsymbol{E_u(E_x(X|u)E_x(X|u)') - 2E_u(E_x(X|u)E(X)')} \\
&\quad \boldsymbol{+E_u(E(X|u))E_u(E(X|U))} \\
&= \boldsymbol{E_u(E_x(XX'|u) - 2E_x(X|u)E_x(x|u)' +} \\
&\quad \boldsymbol{E_x(X|u)E_x(X|u)') + E_u(E_(X|u)E_x(X|u)'} \\
&\quad \boldsymbol{-2E_x(X|u)E_u(E_x(X|u)) +}
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{E_uE_x(X|u)E_uE_x(X|u)')} \\
= \; &\boldsymbol{E_u(E_x(XX' - 2XE_x(X|u)} \\
&\boldsymbol{+E_x(X|u)E_x(X|U)) +} \\
&\boldsymbol{E_u((E_x(X|u) - E_u(E_x(X|u)))} \\
&\boldsymbol{(E_x(X|u) - E_u(E_x(X|u)))')} \\
= \; &\boldsymbol{E_u((E_x(X) - E_x(X|u))(E_x(X) - E_x(X|u))') +} \\
&\boldsymbol{COV_u(E_x(X|u))} \\
= \; &\boldsymbol{E_u(COV_x(X|u)) + COV_u(E_x(X|u)).}
\end{aligned}$$

C.5. Kernel de una distribución normal p variada

El kernel un vector aleatorio es la parte de la función de densidad que únicamente depende de dicho vector, y a su vez nos permite identificar de que familia proviene la distribución. Por ejemplo, en el caso de la distribución normal p variada

$$\begin{aligned} \boldsymbol{f(X|\mu,\Sigma)} &= \frac{\mathbf{1}}{\boldsymbol{(2\Pi)^{p/2}|\Sigma|^{1/2}}} \\ & \quad \boldsymbol{exp(-\frac{1}{2}(x'\Sigma^{-1}x - x'\Sigma^{-1}\mu - \mu'\Sigma^{-1}x + \mu'\Sigma^{-1}\mu))} \end{aligned}$$

Tenemos que el Kernel correspondiente es:

$$\boldsymbol{exp(-\frac{1}{2}(x'\Sigma^{-1}x - 2x'\Sigma^{-1}\mu))}$$

donde $\boldsymbol{\mu'\Sigma^{-1}x = x'\Sigma^{-1}\mu}$, ya que es una forma cuadrática.

C.6. Kernel de una distribución Wishart

Análogamente al caso normal p variado, si nos concentramos en la parte de la densidad Wishart, con matriz de escala $\boldsymbol{S}$ y con n grados de libertad, que únicamente depende de $\boldsymbol{\Sigma}$, tenemos que el correspondiente kernel es:

$$\frac{\boldsymbol{\exp(-\frac{1}{2}tr(\Sigma^{-1}S))}}{\boldsymbol{|\Sigma|^{\frac{n}{2}}}}$$

C.7. Kernel del producto de n distribuciones normales p variadas con mismo vector de medias y misma matriz de varianza covarianza

Supongamos que $\boldsymbol{X_i}$ se distribuye normal p variada con vector de media $\boldsymbol{\mu}$ y matriz de varianza covarianza $\boldsymbol{\Sigma}$, entonces la función de interés es de la siguiente manera:

$$\prod_{i=1}^{n} \boldsymbol{N(X_i|\mu,\Sigma)} = \prod_{i=1}^{n} \frac{1}{\boldsymbol{(2\Pi)^{p/2}|\Sigma|^{1/2}}} \boldsymbol{exp}(-\frac{1}{2}\boldsymbol{(X_i-\mu)'\Sigma^{-1}(X_i-\mu)})$$

De aquí, trabajando únicamente con el exponente de la función exponencial tenemos que:

$$\prod_{i=1}^{n} \boldsymbol{exp}(-\frac{1}{2}\boldsymbol{(X_i-\mu)'\Sigma^{-1}(X_i-\mu)}) =$$
$$\boldsymbol{exp}(-\frac{1}{2}\sum_{i=1}^{n}\boldsymbol{(X_i-\mu)'\Sigma^{-1}(X_i-\mu)}) =$$
$$\boldsymbol{exp}(-\frac{1}{2}\sum_{i=1}^{n}\boldsymbol{(X_i'\Sigma^{-1}X_i - 2X_i'\Sigma^{-1}\mu + \mu'\Sigma^{-1}\mu)}) =$$
$$\boldsymbol{exp}(-\frac{1}{2}\boldsymbol{(X'\Sigma^{-1}X - 2X'\Sigma^{-1}\mu + n\mu\Sigma^{-1}\mu)})$$

Donde la jésima coordenada del vector X es $\sum_{i=1}^{n} \boldsymbol{X_{i,j}}$ Por último,como sólo nos interesan los términos donde aparece $\boldsymbol{X_i}$, llegamos a que el kernel es:

$$\boldsymbol{exp}(-\frac{1}{2}\boldsymbol{(X'\Sigma^{-1}X - 2X'\Sigma^{-1}\mu)})$$

C.8. Kernel del vector de medias de una distribución normal p variada multiplicada por la distribución del vector de medias

En este caso tenemos que $\boldsymbol{X}$ se ditribuye normal con vector de medias $\boldsymbol{\mu}$ y matriz de varianza covarianza $\boldsymbol{\Sigma}$, mientras que $\boldsymbol{\mu}$ se distribuye normal con vector de medias $\boldsymbol{\mu_0}$ y matriz de varianza covarianza $\boldsymbol{\Sigma_0}$, y ahora nos interesa conocer el Kernel correspondiente a $\boldsymbol{\mu}$, entonces tendríamos que el producto de las funciones de densidad es:

$$\boldsymbol{f_X(X|\mu,\Sigma)f_\mu(\mu|\mu_0,\Sigma_0)} =$$

$$\frac{\boldsymbol{exp(-1/2(X-\mu)'\Sigma^{-1}(X-\mu))}}{\boldsymbol{(2\Pi)^{p/2}|\Sigma|^{1/2}}}\frac{\boldsymbol{exp(-1/2(\mu-\mu_0)'\Sigma_0^{-1}(\mu-\mu_0))}}{\boldsymbol{(2\Pi)^{p/2}|\Sigma_0|^{1/2}}}$$

Luego de cada función de densidad tomamos lo que dependa de $\boldsymbol{\mu}$, para así obtener sus respectivos kerneles según el anexo 4,lo cual implica que:

$$\begin{aligned}
&\boldsymbol{Kernel} = \\
&\boldsymbol{exp(-\frac{1}{2}(\mu'\Sigma^{-1}\mu - 2x'\Sigma^{-1}\mu))exp(-\frac{1}{2}(\mu'\Sigma_0^{-1}\mu - 2\mu_0'\Sigma_0^{-1}\mu)) =} \\
&\boldsymbol{exp(-\frac{1}{2}(\mu'\Sigma^{-1}\mu - 2x'\Sigma^{-1}\mu + \mu'\Sigma_0^{-1}\mu - 2\mu_0'\Sigma_0^{-1}\mu)) =} \\
&\boldsymbol{exp(-\frac{1}{2}(\mu'(\Sigma^{-1} + \Sigma_0^{-1})\mu - 2(X'\Sigma^{-1} + \mu_0'\Sigma_0^{-1})\mu) =} \\
&\boldsymbol{exp(-\frac{1}{2}(\mu'(\Sigma^{-1} + \Sigma_0^{-1})\mu - 2(X'\Sigma^{-1} + \mu_0'\Sigma_0^{-1})(\Sigma^{-1} + \Sigma_0^{-1})^{-1}} \\
&\boldsymbol{(\Sigma^{-1} + \Sigma_0^{-1})\mu))}
\end{aligned}$$

De aquí se tiene que $\boldsymbol{\mu}$ se distribuye normal p variada con matriz de varianza

covarianza $\boldsymbol{\Sigma + \Sigma_0}$, y vector de medias $\boldsymbol{(X'\Sigma^{-1} + \mu_0'\Sigma_0^{-1})(\Sigma^{-1} + \Sigma_0^{-1})^{-1}}$

C.9. Probabilidad condicional

La probabilidad condicional se define de la siguiente manera **?**:
Si $\boldsymbol{P(F) > 0)}$, entonces:

$$\boldsymbol{P(E\ F) = \frac{P(E \bigwedge F)}{P(F)}}$$