matemática: Una publicación del ICM – ESPOL
2005, Vol. 3, No. 2

MODELOS BAYESIANOS MULTIVARIADOS

Sandoya Fernando[1]

Resumen: *En este artículo se da una introducción corta a la estadística multivariante Bayesiana, sobre todo introduciendo el "pensamiento" bayesiano, y el concepto de "modelo" bayesiano, por medio del análisis bayesiano los parámetros desconocidos de los modelos no son tratados a través de la manera convencional con estimadores puntuales con una confiabilidad asegurada por medio de una prueba de hipótesis, y no se hace una distinción fundamental entre observaciones y parámetros desconocidos, todos los cuales son tratados como variables aleatorias de una consecuencia lógica del análisis condicional bayesiano. Por último se construye un modelo multivariante bayesiano para resolver el problema de la separación de orígenes.*

Palabras claves: Estadística bayesiana, separación de fuentes y análisis bayesiano.

1. INTRODUCCIÓN

La estadística normalmente se enseña desde una perspectiva frecuentista clásica, en la cual la probabilidad tiene su origen en la frecuencia a la larga con la que ocurre cierto evento en un experimento repetido; sin embargo, es muy discutible que un experimento pueda desarrollarse repetidamente (exactamente bajo las mismas condiciones). Por ejemplo, en un partido de fútbol entre dos equipos se podría hablar de la probabilidad de que el equipo A gane el partido, y asociar a esta probabilidad la frecuencia con la que el equipo A ganó al equipo B en el pasado, aunque aquí es dudoso hablar de que el experimento aleatorio (en este caso el partido de fútbol) sea un experimento repetitible, ya que es improbable que un partido se desarrolle exactamente con las mismas condiciones de otro juego.

Contrario a esto, desde otro punto de vista no frecuentista, los modelos estadísticos, e incluso la probabilidad misma, pueden ser a menudo desarrollados desde un punto de vista Bayesiano.

Los siguientes son algunos argumentos que favorecen al análisis bayesiano frente al frecuentista clásico:

- Existe a menudo información a priori sobre los parámetros.
  En muchos modelos se tiene información inicial sobre los parámetros desconocidos.
  Por ejemplo, puede suceder que algunos valores del parámetro no tengan mucho sentido por lo que se refiere a alguna propiedad subyacente que deban cumplir
  (por ejemplo, las elasticidades del precio positivas en una función de demanda). Un análisis Bayesiano hace muy fácil incorporar en un modelo esta información directamente.
- Incertidumbre = Probabilidad.
  Toda persona que ha realizado un análisis estadístico sabe que cualquier conclusión obtenida de su análisis estadístico debe estar siempre ligada a una indicación sobre la incertidumbre de la conclusión. Por ejemplo, un estimador puntual de un parámetro desconocido no es de mucha utilidad sin una indicación acerca de la incertidumbre que está tras esta estimación.

En la estadística clásica frecuentista sólo se puede hablar sobre la incertidumbre en un ambiente de muestreo repetido, y esto es porque según la ley de los grandes números la frecuencia converge a la probabilidad en probabilidad, y obviamente esta convergencia es una tendencia a la cercanía entre ambos valores "a la larga". Otro ejemplo es la construcción de intervalos de confianza, los cuales tienen sentido si el muestreo se realiza repetidamente, así por ejemplo se dirá que el intervalo de confianza para la elasticidad del sueldo es [0.1, 0.18] con un 95% de confianza (antes de obtener la muestra y observar los datos). En cambio, un análisis Bayesiano permitirá conclusiones como "*Dados los datos observados*, creo con 95% de probabilidad que la elasticidad del sueldo está entre 0.1 y 0.18".

- Permite el condicionamiento de los datos.

El análisis Bayesiano condiciona a los datos observados mientras que el análisis clásico promedia sobre todas las posibles estructuras de datos en lo que se conoce como muestreo. Nótese que el condicionamiento es bueno porque las observaciones que se van obteniendo mejoran nuestras expectativas sobre lo desconocido. Por ejemplo si conocemos que en un dia de la temporada de invierno la probabilidad de que

[1] *Sandoya Fernando, M.Sc., Profesor Agregado de la Escuela Superior Politécnica del Litoral (ESPOL);*
(e-mail: fsandoya@espol.edu.ec

llueva es del 60%, este no es un valor muy alto como para que las personas estén impulsadas a sacar el paraguas, pues esta sería nuestra expectativa de que va a llover. Sin embargo, si en un día de invierno al levantarnos oímos en el noticiero el pronóstico del tiempo, el cual nos afirma que se presentarán probables lluvias, probablemente estemos dispuestos a llevar el paraguas ese día, ya que nuestra creencia respecto al evento "va a llover" a aumentado por la nueva información.

- Teoría de distribuciones exactas
  La teoría de distribución de estimadores de parámetros, en la estadística clásica, requiere de aproximaciones asintóticas, dadas fundamentalmente por la ley de los grandes números y el teorema del límite central. Estas aproximaciones a veces son buenas, y otras veces son malas. En cambio la distribución Bayesiana siempre es exacta y nunca requiere el uso de aproximaciones asintóticas.
- Coherencia y Racionalidad.
  Se ha demostrado [1] que cualquier análisis estadístico que no es Bayesiano debe violar algún axioma del "sentido común" básico de conducta. Esto se relaciona al hecho que un análisis Bayesiano está directamente basado en la teoría de utilidad axiomática. Esta teoría establece que para fundamentar el uso de la esperanza matemática de la utilidad como criterio de elección en el futuro aleatorio y determinar una función de utilidad que permita ordenar las eventualidades, se deben cumplir los siguientes axiomas:
  - De preferencia
  - De transitividad
  - De independencia estricta
  - De unicidad
  - De ordenación
  - De no saciedad

  De manera formal se puede decir que los tres primeros axiomas mencionados permiten identificar los distintos tipos de preferencias de los consumidores, por ejemplo se necesita el axioma de transitividad porque si las preferencias no fueran transitivas se podría encontrar un conjunto de alternativas tal que ninguna de ellas fuese mejor y por lo tanto no se podría decidir por una de ellas. Los tres últimos axiomas en cambio indican que el orden de las preferencias puede ser representado por una función de utilidad.

  En definitiva, el cumplimiento de estos axiomas se reduce a la satisfacción de dos hipótesis: i) las decisiones se toman de manera completamente racional y ii) estas decisiones se las elige de un gran número de posibilidades.
- El análisis Bayesiano es óptimo desde una perspectiva clásica.
  Se ha mostrado en numerosos trabajos [1] que siempre que se encuentre una regla de decisión óptima desde una perspectiva clásica (óptimo con respecto a algún principio aceptable) esto corresponde a alguna regla de decisión de Bayes.
- Una ventaja operacional: "¡Siempre se conoce qué hacer!"

  Los investigadores a menudo se enfrentan con problemas del tipo "¿Cómo estimo los parámetros de este modelo estadístico de una manera buena?". En la teoría estadística se determinan las características que deberían tener estos estimadores para considerarlos de "buena calidad", en general se requiere que sean consistentes, suficiente, eficientes, insesgados y robustos. Y aunque tempranamente se desarrollaron métodos para obtener estimadores que satisfagan la mayoría de estos criterios, tales como el método de los momentos (Pearson, 1891) o el de máxima verosimilitud (Fisher, 1912), a veces no es muy claro como obtener estos estimadores. En el análisis Bayesiano esto siempre se hace de la misma manera y normalmente con una respuesta buena.
- Desarrollo computacional.
  En el pasado era a menudo muy difícil llevar a cabo de manera práctica un análisis Bayesiano debido a la necesidad de integración analítica, o de optimizar alguna función. Con la introducción de la computación y el desarrollo de los métodos numéricos de integración: Newton Cotes, Cuadratura Gaussiana, métodos de Montecarlo, etc. ahora es posible estimar modelos con muchos parámetros. Así mismo se han desarrollado métodos numéricos para optimizar fácilmente funciones complejas.
- Permite la incertidumbre del parámetro
  Muchas veces es razonable, e incluso necesario, pensar en el verdadero valor del parámetro θ de una población como una realización de una variable aleatoria Θ con una distribución desconocida. Esta distribución no siempre corresponde a un experimento físicamente realizable, sino más bien es una medida de la "creencia" del experimentador con respecto al verdadero valor de θ antes de que se observe cualquier dato. En este sentido la inferencia estadística resultante se vuelve subjetiva.

2. LA PROBABILIDAD DESDE UN PUNTO DE VISTA FRECUENTISTA Y SUBJETIVO

Normalmente se introducen los espacios de probabilidad en la forma de los axiomas de Kolmogorov publicados en (1933). En forma resumida podemos plantear la definición así: Un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ consiste de un espacio muestral Ω, un conjunto de eventos $\mathcal{F}$ (subconjuntos de Ω) y una medida de probabilidad P con las propiedades:

i) $P(A) \geq 0$, para todo $A \in \mathcal{F}$

ii) $P(\Omega) = 1$

iii) Para una colección disjunta $\{A_j \in F\}$,

$$P\left(\bigcup_j A_j\right) = \sum_j P(A_j)$$

La interpretación clásica del número $P(A)$ es que este es el límite de la frecuencia relativa, es decir es la frecuencia relativa con que A ocurre en un *experimento aleatorio repetido* cuando el número de ensayos va al infinito. Esta idea de concebir a la probabilidad parece originaria de Laplace quien a inicios del siglo XIX, definió a la probabilidad como el número de eventos exitosos sobre los intentos observados. De esta manera se puede simplemente repetir el experimento u observar el fenómeno un largo tiempo para determinar la probabilidad de recurrencia. Esta forma de ver a la probabilidad es muy útil, pero el problema es que frecuentemente no es posible obtener un gran número de resultados desde exactamente un experimento realizado bajo las mismas condiciones.

A partir de aquí surgen dos preguntas importantes:

- ¿Se debe basar la teoría de probabilidad exactamente en estos axiomas?,
  La respuesta a esta pregunta es NO, de hecho muchos investigadores han criticado estos axiomas como arbitrarios, aunque hay que reconocer que en la práctica funcionan bien para describir a la probabilidad.
- ¿Se pueden plantear principios más profundos que parezcan menos arbitrarios?,
  La respuesta a esta pregunta es Sí, lo cual nos lleva a a una interpretación alternativa del número $P(A)$, al que se le denomina probabilidad subjetiva y que está íntimamente asociado con la frase "grado de creencia".

Notemos que en una gran parte de nuestras vidas, nuestros cerebros están comprometidos en el razonamiento de lo que consideramos creíble. Así según este enfoque, la teoría de probabilidad no es una teoría sobre límites de frecuencias relativas en experimentos aleatorios, sino más bien una formalización del proceso de razonamiento creíble y la interpretación de la probabilidad como:
$P(A)$= "el grado de creencia en el evento A".
Esta definición subjetiva de probabilidad puede usarse para formalizar la idea de aprender en un ambiente incierto. Supongamos que el grado de creencia en A es $P(A)$. Entonces aprendo que la proposición B es verdadera. Si creo que hay alguna conexión entre A y B. Yo he aprendido entonces alguna cosa sobre A. En particular, las leyes de probabilidad (o, según la teoría anterior, las leyes de razonamiento creíble) me dicen que:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B \mid A)\mathrm{P}(A)}{\mathrm{P}(B)}$$

La cual, es la conocida regla de Bayes. Si no hay ninguna conexión lógica entre A y B entonces $\mathrm{P}(B \mid A) = \mathrm{P}(B)$ y en este caso $\mathrm{P}(A \mid B) = \mathrm{P}(A)$ y no se ha aprendido nada nuevo observando B. Por otro lado, si $\mathrm{P}(B \mid A) \neq \mathrm{P}(B)$ entonces B contiene información sobre A y por consiguiente debo actualizar mis creencias sobre A.

3. EL ENFOQUE BAYESIANO DE LA ESTADÍSTICA

El enfoque Bayesiano de la estadística está basado en aplicar las leyes de probabilidad a la inferencia estadística. Para ver lo que esto implica, reemplacemos en lo anterior A y B por
A = el vector de parámetros no observados, θ,
B = el vector de los datos observados, y.
Reemplazando probabilidades con funciones de distribuciones de probabilidad se obtiene:

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)} \qquad (1)$$

Donde:

- $p(\theta)$ se denomina la distribución *a priori* de θ.
- $p(y \mid \theta)$ es la distribución muestral de los datos dado θ, denominada también la verosimilitud de la muestra
- $p(\theta \mid y)$ es la distribución a posteriori (o simplemente posterior) de θ, y

- $p(y) = \int p(y|\theta)p(\theta)\,d\theta$, según la fórmula de la probabilidad total es la distribución marginal de los datos.

La notación usual es estadística bayesiana para la formula de bayes (1) es:

$$p(\theta|y) \alpha\ p(y|\theta)\ p(\theta)$$

Así que la estadística Bayesiana es nada más que un *modelo formal de aprendizaje en un ambiente de incertidumbre aplicado a la inferencia estadística*. Lo a priori expresa mis creencias sobre θ *antes* de observar los datos; mientras que la distribución $p(\theta|y)$ expresa mis creencias actualizadas sobre θ *después* de observar los datos.

A partir de esto, llevar a cabo un análisis Bayesiano es ilusoriamente simple y *siempre* se procede de la siguiente manera:

- Formular la distribución muestral $p(y|\theta)$ y la distribución a priori $p(\theta)$.
- Calcular la distribución posterior $p(\theta|y)$ según Bayes

¡Eso es todo! Toda la información acerca de θ está contenida ahora en la distribución posterior. Por ejemplo la probabilidad que $\theta \in A$ es:

$$P(\theta \in A|y) = \int_A p(\theta|y)d\theta$$

Ya que la distribución posterior es la representación completa de sus creencias sobre θ, a veces es conveniente informar una sola estimación, por ejemplo el valor más probable de θ, este se denomina entonces el estimador bayesiano de θ.

4. DISTRIBUCIONES ESCALARES, VECTORIALES Y MATRICIALES:

De acuerdo con Rowe [2], usaremos las siguientes notaciones para las distribuciones de las variables aleatorias escalares, vectoriales y matriciales:

a) Distribuciones Escalares

Binomial:

$$x|\xi \sim Bin(n,\xi)$$

Beta:

$$\xi|\alpha,\beta \sim B(\alpha,\beta)$$

Es usada como la distribución a priori de la probabilidad de éxito ξ de una binomial.

Normal:

$$x|\mu,\theta^2 \sim N(\mu,\theta^2)$$

$$p(x|\mu,\theta^2) = (2\pi\theta^2)^{-1/2} e^{-\frac{(x-\mu)^2}{2\theta^2}}$$

Gamma:

$$\xi|\alpha,\beta \sim G(\alpha\beta)$$

$$p(\xi|\alpha,\beta) = \frac{\xi^{\alpha-1}e^{-\xi/\beta}}{\xi(\alpha)\beta^{\alpha}},\quad \xi\in\mathbb{R}^+,\alpha\in\mathbb{R}^+,\beta\in\mathbb{R}^+$$

b) Distribuciones Vectoriales

Un vector de observaciones *p*-variado x es una colección de *p* observaciones escalares, $x_1, \ldots, x_p$.

Normal Multivariada

La distribución normal *p*-variada x es usada para describir simultáneamente una colección de *p* variables aleatorias continuas real valuadas.

Una variable aleatoria que sigue una distribución normal multivariada con vector de medias μ y matriz de covarianzas Σ es representado con:

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

Donde (μ,Σ) parametriza la distribución la cual está dada por:

$$p(x|\mu,\Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} e^{-1/2(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

$$\text{Con } x\in\mathbb{R}^p,\quad \mu\in\mathbb{R}^p,\quad \Sigma>0$$

Proposición.:

$$E(x|\mu,\Sigma) = \mu;$$

$$\text{Moda}\,(x|\mu,\Sigma) = \mu;$$

$$\text{Var}(x|\mu,\Sigma) = \Sigma$$

La distribución normal *p*-variada es la distribución, a la cuál otras con momentos primero y segundo tienden en promedio de acuerdo al teorema del límite central.

c) Distribuciones Matriciales

Normal Matricial

La distribución normal matricial *n x p* puede ser obtenida como un caso especial de la distribución normal multivariada *np*-variada cuando la matriz de covarianzas es separable. Denotamos una distribución normal multivariada *np*-dimensional con media *np*

dimensional μ y matriz de covarianzas *np* x *np* Ω por:

$$p(x|\mu,\Omega) = (2\pi)^{-\frac{np}{2}} |\Omega|^{-1/2} e^{-1/2(x-\mu)^T \Omega^{-1}(x-\mu)}$$

Una matriz separable es aquella de la forma $\Omega = \Phi \otimes \Sigma$ donde $\otimes$ es el producto de kronecker el cual multiplica cada entrada de su primera matriz argumento por toda la segunda matriz argumento.

Así, el producto de kronecker entre Φ y Σ que son matrices *n* y *p* dimensionales respectivamente, es:

$$\Phi \otimes \Sigma = \begin{pmatrix} \Phi_{11}\Sigma & \cdots & \Phi_{1n}\Sigma \\ \vdots & \ddots & \vdots \\ \Phi_{n1}\Sigma & \cdots & \Phi_{nn}\Sigma \end{pmatrix}$$

Sustituyendo la matriz de covarianzas separable se tiene:

$$p(x|\mu,\Sigma,\Phi) = (2\pi)^{-\frac{np}{2}} |\Phi \otimes \Sigma|^{-1/2} e^{-1/2(x-\mu)^T (\Phi \otimes \Sigma)^{-1}(x-\mu)}$$

Por propiedades de matrices:

$$|\Phi \otimes \Sigma|^{-1/2} = |\Phi|^{-p/2} |\Sigma|^{-n/2}$$

$$(x-\mu)^T (\Phi \otimes \Sigma)^{-1}(x-\mu) = \text{tr}\,\Phi^{-1}(X-M)\Sigma^{-1}(X-M)^T$$

Donde $x = (X^T) = (x_1^T, \ldots, x_n^T)^T$

$X^T = (x_1, \ldots, x_n)$

μ = vec (M') = $(\mu'_1, \ldots, \mu'_n)$; vec (.) amontona las columnas de esta matriz argumento de izquierda a derecha en un sólo vector tr (.) suma los elementos de la diagonal de una matriz cuadrada

Así:

$$p(X|M,\Sigma,\Phi) = (2\pi)^{-\frac{np}{2}} |\Phi|^{-p/2} |\Sigma|^{-n/2} e^{-1/2\,\text{tr}\,\Phi^{-1}(X-\mu)\Sigma^{-1}(X-M)'}$$

Una variable aleatoria que sigue una distribución matricial normal se representa con:

$$X|M,\Sigma,\Phi \sim N(M, \Phi \otimes \Sigma)$$

Donde (M,Σ,Φ) parametrizan la distribución anterior con:

$$X \in \mathbb{R}^{n \times p}, \quad M \in \mathbb{R}^{n \times p}, \quad \Sigma, \Phi > 0$$

Las matrices Σ y Φ son comúnmente denominadas matrices de covarianzas dentro y entre.

OBSERVACIÓN: De acuerdo a la teoría estadística se cumple que:

$$E(X|M,\Sigma,\Phi) = M;$$

$$\text{Moda}(X|M,\Sigma,\Phi) = M;$$

$$\text{Var}(\text{vec}(X')|M,\Sigma,\Phi) = \Phi \otimes \Sigma$$

Si X sigue una distribución normal matricial, las distribuciones condicional y marginal de cualquier subconjunto fila o columna son distribuciones normales multivariadas.

También se puede notar que la media de la *i*-ésima fila de X, x_i' es la correspondiente *i*-ésima fila de M, μ_i'; y la covarianza de la *i*-ésima fila de X es $\Phi_{ii}\Sigma$, donde Φ_{ii} es el elemento en la *i*-ésima fila e *i*-ésima columna de Φ. La covarianza entre la *i*-ésima columna y la *i*-ésima fila de X es Φ_{ii},Σ, donde Φ_{ii}, es el elemento en la *i*-ésima fila e *i*-ésima columna de Φ.

Similarmente, la media de la *j*-ésima columna de X es la *j*-ésima columna de M y la covarianza entre la *j*-ésima y *j*-ésima columnas de X es $\theta_{ii}\Phi$.

Simplemente, ponemos,

$$X = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix} = (X_1, \ldots, X_p)$$

$$M = \begin{pmatrix} \mu'_1 \\ \vdots \\ \mu'_n \end{pmatrix} = (M_1, \ldots, M_p)$$

Si Φ_{ii}, denota el *ii*' elemento de Φ y θ_{ii}, el *jj*' elemento de Σ, entonces,

$$\text{Var}(x_i|\mu_i,\Phi_{ii},\Sigma) = \Phi_{ii}\Sigma$$

$$\text{Cov}(x_i, x_{i'}|\mu_i,\mu_{i'},\Phi_{ii'},\Sigma) = \Phi_{ii'}\Sigma$$

$$\text{Var}(x_j|M_j,\theta_{jj},\Phi) = \theta_{jj}\Phi$$

$$\text{Cov}(x_j, x_{j'}|M_{j'},\theta_{jj'},\Phi) = \theta_{jj'}\Phi$$

También se pueden obtener generalizaciones multivariadas de las distribuciones t, Wishart, Gamma, binomial, beta, etc.

5. ASIGNACIÓN DE DISTRIBUCIONES A PRIORI:

Para el análisis bayesiano se puede especificar cualquier distribución a priori, pero la elección debe guiarse por el rango de los valores de los parámetros. Por ejemplo, la probabilidad de éxito en un experimento binomial tiene un rango (0,1) de posibles valores, la media de una distribución normal tiene un rango (-∞,+∞), y la varianza de una distribución normal (0, +∞).
En general existen tres tipos de distribuciones a priori:

1. Vagas (no informativas o difusas)
2. Conjugadas
3. Conjugadas generalizadas

Aunque puede ser usada cualquier distribución definida únicamente en el rango de los parámetros, la elección de distribuciones a priori conjugadas tienen propiedades de actualización naturales que pueden simplificar la estimación, y por esto cuantifican de mejor manera nuestra información a priori disponible.

5.1 A PRIORI VAGAS:

La distribución a priori vaga puede ser usada tanto si el parámetro es acotado (rango finito de valores) o no acotado (rango infinito de valores). Si una distribución vaga a priori es usada sobre un parámetro θ que tiene un rango finito de valores sobre el intervalo (a,b), entonces la distribución a priori es una distribución uniforme sobre (a,b) indicando que todos los valores en este rango son a priori igualmente verosímiles, es decir:

$$p(\theta)=\begin{cases}\dfrac{1}{b-a} & \text{si } a<\theta<b\\ 0 & \text{si no}\end{cases}$$

$$p(\theta)\propto(\text{una constante})$$

En el caso en el cual el parámetro es no acotado, la situación es un poco diferente. Consideremos $\theta=\mu$ la media de una distribución normal, si deseamos ubicar una distribución a priori uniforme, tenemos:

$$p(\mu)=\begin{cases}\dfrac{1}{2a} & \text{si } -a<\theta<a\\ 0 & \text{si no}\end{cases}$$

Donde $a\to\infty$, y se escribe:

$$p(\mu)\propto(\text{una constante})$$

De igual manera, si el parámetro $\theta=\sigma^2$ es la varianza de una distribución normal, tomamos el $\log(\sigma^2)$ que consideramos uniforme en $(-\infty,\infty)$, y tenemos:

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

Vector variadas:

Una distribución a priori para una media vector variada de una distribución normal multivariada es la misma que para la distribución escalar normal:

$$p(\mu)\propto(\text{una constante}),$$

Donde $\mu=(\mu_1, \mu_2, \ldots, \mu_n)$.

Matriz variadas:
Una distribución a priori para una media matriz variada de una distribución normal matriz variada es:

$$p(M)\propto(\text{una constante})$$

Donde la matriz $M=(\mu_1, \mu_2, \ldots, \mu_n)'$. Las filas de M son los vectores individuales μ.

5.2 ASIGNACIÓN DE A PRIORI CONJUGADAS

De acuerdo a Rowe [2] las distribuciones a priori conjugadas son más informativas, y se obtienen de forma natural a partir de la estadística clásica. Parten de la siguiente idea: si un conjunto de datos es dividido en dos partes, entonces un análisis que toma la primera parte como una información inicial (a priori) para la segunda parte es equivalente a un análisis que toma las dos partes simultáneamente. La distribución a priori conjugada para un parámetro es de gran utilidad y es obtenida a través de la verosimilitud, intercambiando los roles de la variable aleatoria y el parámetro, y "enriqueciendo" la distribución hasta que no dependa del conjunto de datos. Esta distribución tiene la propiedad que cuando se la combina con la verosimilitud, el resultado posterior es de la misma familia de distribuciones.
Escalar variadas:
Beta: El número de caras x cuando una moneda es lanzada n_0 veces, sigue una distribución Binomial:

$$p(x/\zeta)\propto\zeta^x(1-\zeta)^{n_0-x}$$

Ahora implementemos el procedimiento conjugado para obtener la distribución a priori para $\theta=\zeta$. Primero, intercambiando los roles de x y ζ:

$$p(\zeta/x)\propto\zeta^x(1-\zeta)^{n_0-x}$$

Y ahora la "enriquecemos" de tal manera que no dependa de los datos, para obtener:

$$p(\zeta) \propto \zeta^{\alpha-1}(1-\zeta)^{\beta-1}$$

Que es la distribución Beta. Este procedimiento conjugado implica que una buena elección es usar la distribución Beta para cuantificar la información a priori disponible respecto a la probabilidad de éxito en un experimento binomial. Las cantidades α y β se denominan hiperparámetros (parámetros de la distribución a priori), los cuales deben ser estimados (avaluados).

Como se mencionó previamente, el uso de la distribución a priori conjugada tiene la ventaja que la distribución posterior resultante es de la misma familia.

En la aplicación anterior:

A priori: $p(\zeta) \propto \zeta^{\alpha-1}(1-\zeta)^{\beta-1}$

Verosimilitud: $p(x/\zeta) \propto \zeta^{x}(1-\zeta)^{n_0-x}$

Posterior: $p(\zeta/x) \propto p(\zeta)p(x/\zeta)$

$$\propto \zeta^{(\alpha+x)-1}(1-\zeta)^{(\beta+n_0-x)-1}$$

Que también es de la familia Beta.

Normal: Si $x/\mu,\sigma^2 \sim N(\mu,\sigma^2)$ con σ^2 conocido o desconocido. La verosimilitud es:

$$p(x/\mu,\sigma^2) \propto (\sigma^2)^{-1/2} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Que es llamada el "kernel" de una distribución normal.

Si intercambiamos los roles de x y μ, obtenemos:

$$p(\mu) \propto (\sigma^2)^{-1/2} e^{\frac{(\mu-x)^2}{2\sigma^2}}$$

Esto implica que se puede seleccionar la distribución para μ desde la familia normal.

Seleccionamos como distribución a priori:

$$p(\mu/\sigma^2) \propto (\sigma^2)^{-1/2} e^{\frac{(\mu-\mu_0)^2}{2\sigma^2}}$$

Donde hemos "enriquecido" la distribución a priori con el uso de μ_0 (que no depende de los datos). La cantidad μ_0 es un hiperparámetro a ser estimado. Especificando las cantidades escalares μ_0 y σ^2, la distribución normal a priori queda completamente especificada. De igual manera se pueden describir las distribuciones a priori conjugadas vector variadas y matriz variadas.

5.3. ESTIMACIÓN DE LOS HIPERPARÁMETROS.

Los hiperparámetros de las distribuciones a priori necesitan ser estimados para que estas distribuciones queden completamente identificadas. Existen dos maneras en que los parámetros pueden ser estimados: o bien de una manera puramente subjetiva que expresa el conocimiento y la creencia de un experto, o por uso de datos desde experimentos similares previos.

6. MODELO PARA LA SEPARACIÓN DE ORÍGENES Y DESCOMPOSICIÓN DE SEÑALES MEZCLADAS

Hay dos posibles enfoques para resolver el problema de separación de orígenes. El primero impone restricciones en el modelo y en la verosimilitud tales como la independencia de los orígenes (fuentes); mientras que el segundo incorpora el conocimiento disponible acerca de los parámetros del modelo. El enfoque estadístico bayesiano no solo permite estimar las fuentes y los coeficientes mezclados, sino también otro tipo de inferencias.

Existen diversos problemas de varias disciplinas tales como Acústica, genética, portafolios de inversión, radares y vigilancia, márgenes de resonancia magnética (FMRI), MEG y EEG (magneto encefalograma y electro encefalograma), etc., que pueden ser vistos como un problema de separación de fuentes u orígenes. Cualquier problema en el cual una señal está formada como una combinación de señales elementales es un problema de separación de orígenes.

6.1 INTRODUCCIÓN: LA CONVERSACION GRABADA

El modelo de separación de orígenes puede ser fácilmente explicado de la siguiente manera: En una fiesta hay algunas personas manteniendo conversaciones, al mismo tiempo que algunos micrófonos registran las conversaciones de estas personas parlantes denominadas las fuentes (u orígenes) subyacentes (o fundamentales).

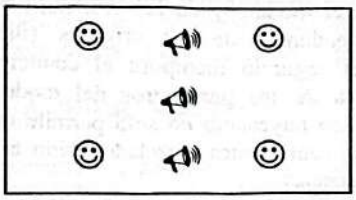
En la fiesta típicamente hay varios grupos pequeños de oradores manteniendo conversaciones. Y en cada grupo, típicamente solo una persona habla al mismo tiempo. Por ejemplo en un grupo de 2, una de las personas habla y la otra escucha, después habla esta última y así por el estilo. Los oradores están obviamente correlacionados negativamente, así en el modelo bayesiano los oradores se permiten estar

correlacionados y no restringidos a ser sólo independientes, como es usual en el modelo clásico.

Consideremos que en la fiesta hay p micrófonos que registran, graban u observan m personas u oradores en n instantes de tiempo. Esta notación es consistente con la estadística multivariada tradicional. Las conversaciones grabadas consisten de mezclas de conversaciones reales no observadas. Los micrófonos no están ubicados en las bocas de los oradores sino que graban las conversaciones mezcladas. El problema es desmezclar o recuperar las conversaciones originales desde las conversaciones mezcladas grabadas.

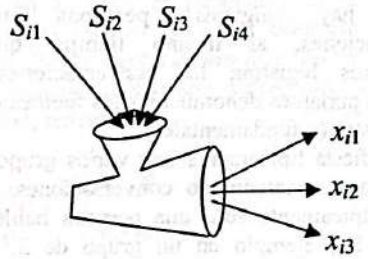
Consideremos el siguiente ejemplo: hay una fiesta con $m = 4$ oradores y $p = 3$ micrófonos como se muestra en la figura 1

FIG. 1
Modelos Bayesianos Multivariados
El proceso de registro de las conversaciones mezcladas



En el instante i; $i = 1, \ldots, n$ la conversación emitida por el orador 1 es S_{i1}, por el orador 2 es S_{i2}, para el orador 3 es S_{i3} y orador 4 es S_{i4}. Luego, la conversación grabada en el micrófono 1 es x_{i1}, en el micrófono 2 es x_{i2}, en el micrófono 3 es x_{i3}. Hay una función desconocida f ilustrada en la FIGURA 2, llamada la función mezcladora, la cual toma las señales de origen emitidas y las mezcla para producir las señales mezcladas observadas.

FIG. 2
Modelos Bayesianos Multivariados
La función mezcladora



6.2 EL MODELO DE SEPARACIÓN DE FUENTES:

Los p micrófonos graban mezclas de los m oradores en cada uno de los n instantes de tiempo. Así, lo que es emitido por los m oradores en el instante i son m valores distintos, representados matricialmente como:

$$S_i = \begin{pmatrix} S_{i1} \\ S_{i2} \\ \vdots \\ S_{im} \end{pmatrix}$$

Y lo que es grabado en el instante i por los p micrófonos son p valores distintos, representados matricialmente como:

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Así, la meta es separar la señal p-dimensional observada en la señal original no observada y fundamental m-dimensional.

El proceso que mezcla las conversaciones de los oradores es instantáneo, constante e independiente en el tiempo. El número de oradores es conocido (aunque puede construirse un modelo más complejo relajando esta suposición). El modelo de separación de orígenes para todo instante i es:

$$\left(x_i \middle| \, S_i\right)_{p\times 1} = f\left(S_i\right)_{p\times 1} + \varepsilon_{i\ p\times 1}$$

Donde $\mathrm{f}\left(S_i\right)$ es una función que mezcla las señales originales y ε_i es el error aleatorio.

Usando la serie de expansión de Taylor alrededor de un vector c, la función f con las condiciones de suavidad apropiadas puede ser escrita como:

$$f\left(S_i\right) = f\left(c\right) + f'\left(c\right)\left(S_i - c\right) + \ldots$$

Y considerando los primeros dos términos (como en el modelo de regresión):

$$\begin{aligned} f\left(S_i\right) &\approx f\left(\mathrm{c}\right) + f'\left(c\right)\left(S_i - c\right) \\ &\approx \left[f\left(c\right) - f'\left(c\right)c\right] + f'\left(c\right) S_i \\ &\approx \mu + \Lambda\, S_i \end{aligned}$$

Donde $f(c)$ y Λ son matices p x m. Este es llamado el modelo lineal de síntesis como implica la figura 3, la señal original emitida desde cada una de las bocas de los oradores se multiplica por un coeficiente de mezcla que determina la fuerza de su contribución, pero también hay en general un ruido de fondo en cada micrófono, y por tanto un error aleatorio es grabado en el proceso de mezclado. Más formalmente, el modelo adoptado es:

$$\left(x_i \middle| \mu, \Lambda, S_i\right)_{px1} = \mu_{px1} + \Lambda_{pxm} + S_{i\,mx1} + \varepsilon_{ipx1}$$

Donde $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$ es un vector no observado de la media poblacional, $\Lambda = \begin{pmatrix} \lambda_1' \\ \vdots \\ \lambda_p' \end{pmatrix}$ es una matriz p x m de coeficientes de mezclado no observados, Si es el i –ésimo vector m×1 de orígenes no observados y

$\varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{ip} \end{pmatrix}$ es el vector de errores o de ruido del i –ésimo vector de señal observado.

La señal mezclada observada x_{ij} es el j –ésimo elemento del i –ésimo vector de señales observadas mezcladas, que puede ser pensada como la conversación grabada en el instante i; $i = 1, \ldots, n$ para el micrófono j; $j = 1, \ldots, p$. La señal observada x_{ij} es una mezcla de las conversaciones no observadas originales S_i con error, en los instantes $i = 1, \ldots, n$.

La señal original no observada S_{ik} es el k –ésimo elemento del vector original no observado S_i, el cual puede ser pensado como la señal no observada original correspondiente a la conversación de orador k; $k = 1, \ldots, m$ en el instante $i = 1, \ldots, n$.

El modelo describe el proceso de mezcla escribiendo la señal observada x_{ij} como la suma de una parte media general (de fondo) μ, mas una combinación lineal de los componentes de la señal original no observadas S_{ik} y el error de la observación ε_{ij}:

$$\left(x_{ij} \middle| \mu_j, \lambda_j, S_i\right) = \mu_j + \sum_{k=1}^{m} \lambda_{jk} S_{ik} + \varepsilon_{ij}$$

$$= \mu_j + \lambda_j' S_i + \varepsilon_{ij}$$

En pocas palabras, la conversación observada por un micrófono dado consiste de un sonido de fondo medio general en el micrófono más contribuciones desde cada uno de los oradores y el error aleatorio. La contribución de un orador a la conversación grabada depende del coeficiente respectivo. El problema es descomponer los orígenes no observados y obtener información mirando el proceso de mezcla para determinar los parámetros restantes del modelo.

Por ejemplo, suponiendo que en el proceso de mezclado se conocen los valores dados en la TABLA 1, se puede calcular fácilmente el vector de valores observados x_i.

TABLA 1
Modelos Bayesianos Multivariados
Valores observados

μ	Λ				S_i	ε_i
1	5	5	3	1	2	3
2	3	5	5	3	4	4
3	1	3	5	5	6	5
					8	

CONCLUSIONES

- El análisis bayesiano resulta más robusto que el análisis clásico, en el método que no requiere muchos supuestos sobre la distribución de las poblaciones observadas, este hecho es también útil en el análisis multivariado, pues es conocido que en el análisis multivariado clásico se parte de muchos supuestos, tales como la normalidad de los datos o de los residuos, lo cual muchas veces no se cumple.

- Actualmente este tipo de análisis se pueden realizar eficientemente gracias a la gran capacidad informática que permite velocidad en los cálculos, esto es muy importante en cualquier análisis bayesiano porque el procesamiento de distribuciones a priori y a posteriori involucra gran cantidad de operaciones.

- El análisis bayesiano tiene la ventaja, frente a otros tipos de análisis, en que permite fácilmente la incorporación de nueva información, lo cual actualiza la distribución a posteriori.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1] **KARSTEN T.**, (2002), ***"Introduction to bayesian econometrics and decision theory"*** Chapman & Hall 1era edition New York.

[2] **ROWE D.**, (2003), ***"Multivariate bayesian statistics"***, Chapman & Hall 1era edition (New York).

[3] **MENDENHOLL.**, (1998), ***"Estadística Matemática con aplicaciones"***, Iberoamericana, Segunda Edición, (México)

[4] **GILL J.**, (2002), "***Bayesian Métodos"***, Chapmam, 1era edition, (New York).