Estadística Bayesiana

Modelos Jerárquicos Lineales
Variables Latentes
Mezclas de Distribuciones

Lizbeth Naranjo Albarrán

Facultad de Ciencias, UNAM

April 18, 2022

Capítulo 1

Modelos Jerárquicos Lineales, Variables Latentes, y Mezclas de Distribuciones

1.1 Modelos Jerárquicos Lineales

Existen muchas aplicaciones estadísticas con conjuntos de parámetros que están relacionados entre sí, debido a la estructura particular del problema. Por ejemplo, en un estudio sobre el nivel de educación, se aplica una prueba diagnóstica a los alumnos de una población, si los estudiantes de la escuela i tienen un nivel académico θ_i, entonces sería razonable suponer que estas estimaciones de los parámetros θ_i están relacionados entre sí a través de un parámetro en común ϕ.

Una forma de modelar esta estructura de los datos y/o parámetros consiste en usar una distribución inicial, en la que los parámetros θ_i tendrán una distribución poblacional común, y que a su vez dependen de un hiperparámetro desconocido ϕ. En esta estructura, los datos observados X_i se utilizan para estimar los parámetros. La figura 1.1 muestra la estructura general de un modelo jerárquico simple.

Esta representación gráfica tiene los siguientes elementos. Las k variables observadas X_1, $X_2 \ldots, X_k$, están representadas con nodos rectangulares (las variables observadas usualmente se representan con rectángulos o cuadrados). Éstas dependen de los parámetros desconocidos $\theta_1, \theta_2, \ldots, \theta_k$, respectivamente, representados con nodos ovaladados (las variables latentes o parámetros desconocidos se reprentan con óvalos o circulos). Note que X_i depende de θ_i, para $i = 1, 2, \ldots, k$, y está representada con una flecha direccionada (las depedendencias condicionales se representan con flechas direccionadas). Además, existe un hiperparámetro común desconocido que está representado con un nodo ovalado; para el cual se tiene que θ_i

depende de ϕ, y esta relación de dependencia está representada con una flecha direccionada.

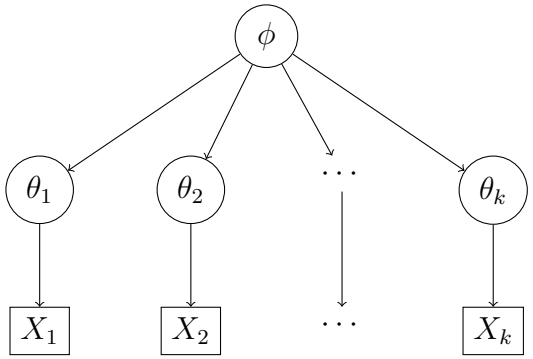


Figura 1.1: Estructura de un modelo jerárquico.

Esta representación gráfica se traduce en un modelo jerárquico con la siguiente estructura de funciones de distribución:

$$\begin{aligned} x_i|\theta_i &\sim F(x_i|\theta_i) \\ \theta_i|\phi &\sim F(\theta_i|\phi) \\ \phi &\sim F(\phi) \end{aligned}$$

equivalentemente, tenemos las funciones de densidad:

$$\begin{aligned} \text{Observaciones} \quad f(\boldsymbol{x}|\boldsymbol{\theta}) &= f(x_1,\ldots,x_k|\theta_1,\ldots,\theta_k) = \prod_{i=1}^{k} f(x_i|\theta_i) \\ \text{Parámetros} \quad f(\boldsymbol{\theta}|\phi) &= \prod_{i=1}^{k} f(\theta_1,\ldots,\theta_k|\phi) \\ \text{Hiperparámetros} \quad f(\phi) & \end{aligned}$$

Esta relación puede interpretarse de la siguiente manera:

- Las observciones $X_1,\ldots,X_k$ provienen de experimentos distintos, pero de alguna manera están relacionados entre sí (ejemplo, prueba diagnóstica sobre el nivel de eduación realizada en k escuelas de una población). Note que X_i depende de θ_i, para $i = 1, 2, \ldots, k$, y se refiere a una dependencia condicional en el sentido de probabilidad, es decir, su función de densidad condicional es $f(x_i|\theta_i)$.

- Los parámetros $\theta_1,\ldots,\theta_k$ se suponen *intercambiables* (ejemplo, θ_i podría representar el nivel de eduación de cada escuela). Los parámetros θ_i depende de ϕ, y esta dependencia se traduce en una función de densidad condicional $f(\theta_i|\phi)$.

- El parámetro ϕ describe alguba característica relevante de la población (ejemplo, ϕ podría representar el nivel de eduación global de la población bajo estudio). Será necesario incluir una función de densidad para el hiperparámetro ϕ, es decir, $f(\phi)$.

Este tipo de estructurar jerárquicas pueden ser muy útiles al analizar modelos con muchos parámetros. Se debe tener cuidado en elegir la estructura jerárquica, que tenga un sentido relacional con respecto a los datos. También será necesario tener cuidado al NO sobre-parametrizar el modelo porque se tendría un problema de sobre-ajuste, así como NO poner pocos parámetros porque el modelo no podrá ajustar adecuadamente los datos. Esta estructura se puede generalizar para construir modelos jerárquicos con más niveles en la estructura jerárquica.

El el análisis para la estimación de los parámetros será de interés obtener la estimación de todos los parámetros, $\theta_1, \ldots, \theta_k$ y ϕ. La distribución inicial apropiada es:

$$f(\boldsymbol{\theta}, \phi) \quad = \quad f(\boldsymbol{\theta}|\phi) f(\phi)$$

La distribución final correspondiente es:

$$f(\boldsymbol{\theta}, \phi|\boldsymbol{x}) \quad \propto \quad f(\boldsymbol{x}|\boldsymbol{\theta}, \phi) f(\boldsymbol{\theta}, \phi) = f(\boldsymbol{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\phi) f(\phi)$$

Note que la distribucón de las observaciones sólo depende de $\boldsymbol{\theta}$, el hiperparámetro ϕ afecta a $\boldsymbol{x}$ sólo a través de θ, por tanto, $f(\boldsymbol{x}|\boldsymbol{\theta}, \phi) = f(\boldsymbol{x}|\boldsymbol{\theta})$, dicho de otra forma, $\boldsymbol{x}$ y ϕ son condicionalmente independientes dato ϕ.

Ejemplo: ver Ejercicio 12.3 de Koop et al. (2007) y Notas de Gutiérrez-Peña (1998).
`Código R:` Bayes10_1JerarquicoRats.R

1.2 Estructuras Aumentadas con Variables Latentes

Un modelo de clases latentes generalmente trae consigo un conjunto de variables observadas llamadas variables manifiestas, y un conjunto de variables aleatorias no observadas o no observables llamadas variables latentes. Los modelos más usados de este tipo son los modelos latentes condicionalmente independientes, que establecen que todas las variables manifiestas son condicionalmente independientes dadas las variables latentes.

Albert y Chib (1993) desarrollaron métodos Bayesianos para modelar datos de respuesta categórica usando la idea de aumento de datos combinada con técnicas de Monte Carlo vía cadenas de Markov. Por ejemplo, el modelo de regresión probit para datos binarios se supone que tiene una estructura de regresión normal sobre datos continuos latentes. Estos autores generalizan esta idea para modelos de respuesta multinomial, incluyendo el caso donde las categorías multinomiales están ordenadas. En este último caso, los modelos enlazan las probabilidades acumulativas de las respuestas con una estructura de regresión lineal.

Este enfoque tiene un número de ventajas, especialmente en el sistema multinomial, donde puede ser difícil evaluar la función de verosimilitud. Para muestras pequeñas, esta aproximación Bayesiana generalmente se desempeñará mejor que los métodos de máxima verosimilitud tradicionales, los cuales se basan en resultados asintóticos. Además, uno puede elaborar el modelo probit usando mezclas apropiadas de la distribución normal para modelar los datos latentes.

Ejemplo 1.1 (Regresión Binaria Probit)**.** Sea $y_1, \ldots, y_n$ una muestra de n v.a. independientes Bernoulli(p_i), $i = 1, \ldots, n$, donde p_i se relaciona a un conjunto de k variables explicativas $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$ a través de la relación

$$p_i \;=\; \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$$

donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ es el vector de coeficientes de regresión y $\Phi(\cdot)$ es la función de densidad de una Normal(0,1). Note que la función de verosimilitud es:

$$\begin{aligned} f(y_1, \ldots, y_n|\boldsymbol{\beta}) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})^{y_i}(1-\Phi(\boldsymbol{x}_i'\boldsymbol{\beta}))^{1-y_i} \\ &= \prod_{i=1}^{n} \left\{ \int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta}} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w_i^2} \mathrm{d}w_i \right\}^{y_i} \times \left\{ 1 - \int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta}} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w_i^2} \mathrm{d}w_i \right\}^{1-y_i} \end{aligned}$$

Para calcular las estimaciones se requiere usar un algoritmo de Metropolis-Hastings.

Para el modelo probit se puede utilizar una representación aumentada con variables latentes de la siguiente manera. Sean $Z_1, \ldots, Z_n$ donde Z_i son v.a. latentes latentes (no observadas) independientes con distribución normal con media igual al predictor lineal $\boldsymbol{x}_i'\boldsymbol{\beta}$ y varianza 1,

es decir $Z_i \sim \text{Normal}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$, donde

$$Y_i \;=\; \begin{cases} 1 & \text{si } Z_i > 0 \\ 0 & \text{si } Z_i \leq 0 \end{cases}$$

Note que esta representación es válida ya que la probabilidad p_i sigue teniendo la misma forma, es decir,

$$\begin{aligned} p_i &= \mathbb{P}(Y_i = 1) \;=\; \mathbb{P}(Z_i > 0) \\ &= \mathbb{P}(Z_i - \boldsymbol{x}_i'\boldsymbol{\beta} > -\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= 1 - \mathbb{P}(Z_i - \boldsymbol{x}_i'\boldsymbol{\beta} \leq -\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= 1 - \Phi(-\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}) \end{aligned}$$

esto es válido por las probiedades de la distribución normal, de simetría, $\Phi(w) = 1 - \Phi(-w)$, y de estandarización,

$$\begin{aligned} Z_i &\sim \text{Normal}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1) \\ Z_i - \boldsymbol{x}_i'\boldsymbol{\beta} &\sim \text{Normal}(0, 1) \end{aligned}$$

Note que la función de verosimilitud ahora es una *función de verosimilitud aumentada*:

$$\begin{aligned} f(y_1, \ldots, y_n, z_1, \ldots, z_n | \boldsymbol{\beta}) &= \prod_{i=1}^{n} f(y_i|z_i) f(z_i|\boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} \left\{ \mathbb{I}[y_i = 1]\mathbb{I}[z_i > 0] + \mathbb{I}[y_i = 0]\mathbb{I}[z_i \leq 0] \right\} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2} \end{aligned}$$

Esto permite aplicar el algoritmo de Gibbs. ■

Ejemplo 1.2 (Regresión Ordinal)**.** Albert y Chib (1993) modelan las categorías ordenadas de la siguiente manera. Suponga que $Y_1, \ldots, Y_N$ son variables aleatorias observadas, donde Y_i toma valores en alguna de las J categorías ordenadas, $1, \ldots, J$. Sea $p_{ij} = p(Y_i = j)$ y definimos las probabilidades acumulativas como $\eta_{ij} = \sum_{k=1}^{j} p_{ik}$, $j = 1, \ldots, J-1$. Un modelo de regresión para los $\{p_{ij}\}$ esta dado por $\eta_{ij} = \Phi(\gamma_j - \mathbf{x}_i'\beta)$, $i = 1, \ldots, N,\ j = 1, \ldots, J-1$. Uno puede motivar este modelo suponiendo que existe una variable aleatoria continua latente Z_i que se distribuye $Normal(\mathbf{x}_i'\beta, 1)$, y que observamos Y_i, donde $Y_i = j$ si $\gamma_{j-1} < Z_i \leq \gamma_j$ (definimos $\gamma_0 = -\infty$ y $\gamma_J = \infty$). Este problema es un problema de regresión normal donde las variables respuesta están en forma de datos agrupados.

En el modelo, el vector de regresión β y las cotas $\gamma_1, \ldots, \gamma_{J-1}$ son desconocidas. Para asegurar que los parámetros son identificables, es necesario imponer una restricción sobre

las cotas; sin pérdida de generalidad, tomamos $\gamma_1 = 0$. La densidad final conjunta de β y $\gamma = (\gamma_2, \ldots, \gamma_{J-1})$ está dada por

$$\pi(\beta, \gamma|\mathbf{y}) = C\pi(\beta, \gamma) \prod_{i=1}^{N} \sum_{j=1}^{J} 1_{(y_i=j)} \left[\Phi(\gamma_j - \mathbf{x}_i'\beta) - \Phi(\gamma_{j-1} - \mathbf{x}_i'\beta)\right],$$

donde $\pi(\beta, \gamma)$ es la distribución inicial. Encontramos la moda de la distribución final de (β, γ) usando el algoritmo de Newton-Raphson y se obtienen las desviaciones estándar aproximadas de las distribución final de (β, γ) usando la segunda derivada del logaritmo de las distribuciones finales evaluada en la moda.

Podemos generalizar el algoritmo de Gibbs para esta situación. Introducimos las variables aleatorias latentes no observadas $Z_1, \ldots, Z_n$ definidas previamente y simuladas de la distribución final conjunta de $(\beta, \gamma, \mathbf{Z})$. Si asignamos una distribución inicial para (β, γ), entonces esta densidad final conjunta está dada por

$$\Pi(\beta, \gamma, \mathbf{Z}|\mathbf{y}) = C \prod_{i=1}^{N} \left[\sqrt{\frac{1}{2\pi}} \exp\left(-\frac{(Z_i - \mathbf{x}_i'\beta)^2}{2} \right) \left\{ \sum_{j=1}^{J} 1_{(Y_i=j)} 1_{(\gamma_{j-1} < Z_i < \gamma_j)} \right\} \right].$$

La distribución final de β condicional a $\mathbf{y}$ y $\mathbf{Z}$ está dada por la distribución normal multivariada

$$Normal_k(\hat{\beta}_Z, (\mathbf{X}'\mathbf{X})^{-1}). \tag{1.1}$$

Las distribuciones finales condicionales completas de $Z_1, \ldots, Z_N$ son independientes con

$$\begin{gathered} Z_i|\beta, \gamma, y_i = j \text{ que se distribuye } Normal(\mathbf{x}_i'\beta, 1) \\ \text{truncada en el lado izquierdo (derecho) por } \gamma_{j-1}(\gamma_j). \end{gathered} \tag{1.2}$$

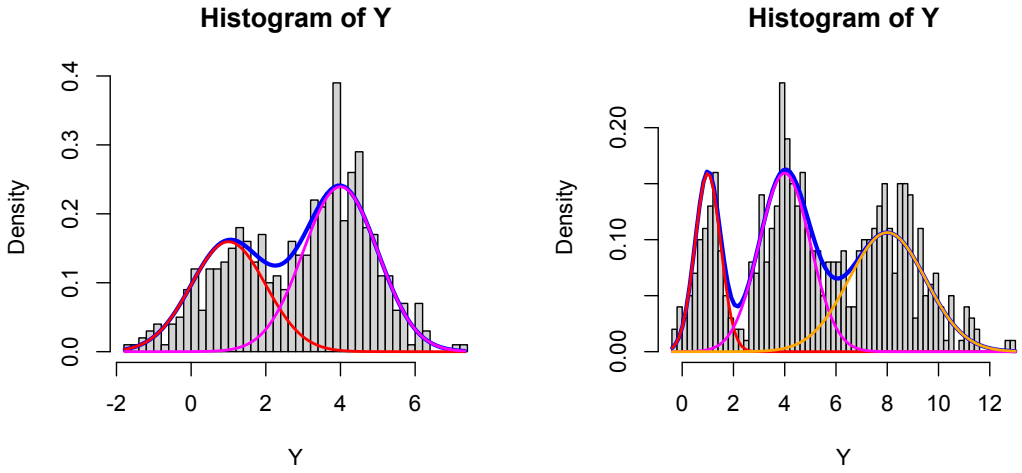
Finalmente, la densidad condicional completa de γ_j dada $\mathbf{Z}$, $\mathbf{y}$, β, y $\{\gamma_k, k \neq j\}$ está dada (hasta una constante de proporcionalidad) por

$$\prod_{i=1}^{N} \left[1_{(Y_i=j)} 1_{(\gamma_{j-1} < Z_i < \gamma_j)} + 1_{(Y_i=j+1)} 1_{(\gamma_j < Z_i < \gamma_{j+1})} \right]. \tag{1.3}$$

Esta distribución condicional puede verse como una distribución condicional uniforme sobre el intervalo $[\max\{\max\{Z_i : Y_i = j\}, \gamma_{j-1}\}, \min\{\min\{Z_i : Y_i = j+1\}, \gamma_{j+1}\}]$. Para implementar el muestreo de Gibbs, comenzamos con (β, γ) inicializada en el valor del estimador máximo verosímil y simulamos de las distribuciones de las ecuaciones (1.3), (1.2) y (1.1), en ese orden. ■

1.3 Mezclas Discreta de Distribuciones Normales

En capítulos previos hemos considerado datos provenientes de funciones de densidad unimodales, sin embargo existen datos que requieren ser modelados con distribuciones cuyas funciones de densidad tengan más de una moda. En algunas escenarios una mezcla discreta de subpoblaciones puede reflejar mejor la densidad de los datos.



El análisis más común para este tipo de datos es suponer que se conoce a priori el número de clases o modas, estimar los posibles modelos considerando posibles alternativas paras las categorizaciones, y comparar las estimaciones usando algún criterio como el criterio de informaciín de Akaike (AIC) o el criterio de información Bayesiano (BIC).

Trabajaremos con el caso más sencillo que es una mezcla de 2 distribuciones normales. Este caso se puede generalizar de manera sencilla al caso de una mezcla de k distribuciones normales.

Sea X una v.a. que pertenece a una distribución $f_X(x)$ que es una mezcla de 2 distribuciones normales:

$$f_X(x) \quad = \quad p_1 f_1(x) + p_2 f_2(x)$$

donde p_1 y p_2 son parámetros tales que $0 < p_1, p_2 < 1$ y $p_1 + p_2 = 1$, y tal que $f_1(x)$ y $f_2(x)$ son funciones de densidad de distribuciones Normales, $Normal(\mu_1, \sigma_1^2)$ y $Normal(\mu_2, \sigma_2^2)$, respectivamente, donde los parámetros para las medias μ_1 y μ_2 y las varianzas σ_1^2 y σ_2^2 son desconocidos.

Ejemplo: `Código R:` Bayes10_2Mixture2normal.R

Para resolver el problema, lo usual es introducir variables aleatorias latentes. Sean $X_1, \ldots, X_n$ las observaciones que supone siguen una fdp $f_X(x)$ que es una mezcla de 2 distribuciones normales:

$$f_X(x) \;=\; p_1 f_1(x) + p_2 f_2(x)$$

Se introducen v.a. latentes $Z_1, \ldots, Z_n$ tal que,

$$Z_i = \begin{cases} 1 & \text{si } x_i \sim f_1(x|\mu_1, \sigma_1^2) \\ 2 & \text{si } x_i \sim f_2(x|\mu_2, \sigma_2^2) \end{cases}$$

esto significa que se contruye una v.a. latente categórica, con dos categorías,

$$Z_i \sim \text{Categorica}(p, 1-p)$$

La función de verosimilitud es:

$$f(x_1, \ldots, x_n|p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \;=\; \prod_{i=1}^{n} f_X(x|\boldsymbol{\theta})$$

La función de verosimilitud aumentada es:

$$\begin{aligned} & f(y_1, \ldots, y_n, z_1, \ldots, z_n|p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \\ &= \prod_{i=1}^{n} \Big\{ \mathbb{I}[z_i = 1] p f_1(x_i|\mu_1, \sigma_1^2) + \mathbb{I}[z_i = 2](1-p) f_2(x_i|\mu_2, \sigma_2^2) \Big\} \end{aligned}$$

Para la estimación de los parámetros es conveniente considerar las siguientes restricciones:

- Usar alguna crestricción en los parámetros de las medias, lo usual es $\mu_1 < \mu_2$. De lo contrario se pueden tener 'saltos' entre las modas, que llevaría a una falta de identificabilidad de los parámetros.
- Usar distribuciones informativas para las prior en los parámetros de las varianzas. De lo contrario, algunas veces las estimaciones de las varianzas son no identificables.

1.4 Mezclas de Escalas y Continuas de Distribuciones

Existen otro tipo de mezclas llamadas mezclas de escalas usando estructuras de variables latentes. Las mezclas de escalas más comunes son para las distribuciones normales y las uniformes.

Ejemplo 1.3 (Distribución t Student como Mezcla de Escala de Normales)**.** La distribución t de Student puede representarse como una mezcla de escalas de distribuciones Normales. Sea

$$\begin{aligned} Y|\beta,\sigma^2,\lambda &\sim \text{Normal}(x\beta,\lambda\sigma^2) \\ \lambda|\nu &\sim \text{InversaGamma}\left(\tfrac{\nu}{2},\tfrac{2}{\nu}\right) \end{aligned}$$

La función de densidad de $Y|\beta,\sigma^2$ es una t Student:

$$f(y|\beta,\sigma^2) = \int_0^\infty f(y|\beta,\sigma^2,\lambda)f(\lambda)\mathrm{d}\lambda$$

Ver por ejemplo Ejericio 15.1 de Koop et al. (2007). ■

Ejemplo 1.4 (Distribución Binomial Negativa como Mezcla Continua de Poisson)**.** Sea X una v.a. (de conteos) con distribución binomial negativa, BinomialNegativa(α,p) con $x=0,1,2,\ldots$, $\alpha>0$, $0<p<1$, si su fdp es

$$f(x|\alpha,\theta) = \frac{\Gamma(\alpha+x)}{x!\gamma(\alpha)}\theta^\alpha(1-\theta)^x$$

Considere que

$$\begin{aligned} X &\sim \text{Poisson}(\lambda) \\ \lambda &\sim \text{Gamma}(\alpha,\beta) \\ \beta &= \frac{\theta}{1-\theta} \end{aligned}$$

`Código R:` Bayes10_3DistNegBinom.R ■

Bibliografía

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*(422), 669–679.

Box, G. E. P. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Massachusetts: Addison-Wesley.

Congdon, P. (2006). *Bayesian Statistical Modelling* (Second ed.). Chichester: John Wiley & Sons.

Congdon, P. D. (2010). *Applied Bayesian Hierarchical Methods*. Boca Raton, Florida: Chapman & Hall/CRC.

Dey, D. K., S. K. Ghosh, and B. K. Mallick (2000). *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.

Gutiérrez-Peña, E. (1998). *Análisis Bayesiano de Modelos Jerárquicos Lineales*. México: IIMAS, UNAM.

Koop, G., D. J. Poirier, and J. L. Tobias (2007). *Bayesian Econometric Methods*. Econometric Exercises 7. Cambridge University Press.