FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

DEPARTAMENTO DE ESTADÍSTICA
MAT2095 - Taller de Iniciación Científica

Estadística Bayesiana No-Paramétrica

Eduardo Alfonso Vásquez Tapia

Profesor guía: Fernando Quintana

Marzo del 2023

Tabla de Contenidos

Prefacio

El presente material, incluyendo la monografía, códigos e ilustraciones, fue confeccionado durante mi Taller de Iniciación Científica (MAT2095), guiado por el profesor Fernando Quintana, durante el segundo semestre del 2022 y verano del 2023. Todo el material se encuentra disponible en el siguiente repositorio de Github.

El objetivo principal de este taller fue implementar el algoritmo SIGN (Ni et al. 2020), que permite aplicar modelos de Mezcla Proceso de Dirichlet (DPM) para bases de datos relativamente grandes. Lo anterior suponía un conocimiento previo de Estadística Bayesiana No Paramétrica, por lo que, en realidad, la mayor parte del trabajo se enfocó en aprender sobre lo anterior, especialmente sobre Procesos de Dirichlet, Mezclas de Procesos de Dirichlet y Modelos de Particiones Aleatorias, incluyendo tanto la teoría como los métodos computacionales disponibles.

En cuanto a la parte computacional, en R existen diferentes paquetes que implementan diferentes modelos Bayesianos no paramétricos, como `DPPackage` (Jara et al. 2011) y `dirichletprocess` (Ross et al. 2020), así como modelos de particiones aleatorias como `salso` (Dahl, Johnson, and Müller 2022) y `ppmSuite` (Page et al. 2022). Por otro lado, en Julia (Bezanson et al. 2017), que recientemente ha aumentado considerablemente el número de usuarios, no existe mucho desarrollo respecto a los modelos anteriores. Considerando lo anterior, todos los códigos fueron implementados en este lenguaje. Como proyecto a futuro, se podría incluso formar una librería con este material.

La monografía incluye una pequeña introducción tanto a la estadística Bayesiana como al enfoque no paramétrico, para así entender cómo se mezclan ambos conceptos. Luego, se presentan los Procesos de Dirichlet (DP), que es probablemente el punto de partida más común en la estadística Bayesiana no paramétrica. Ya entendiendo estos procesos, pasaremos a una extensión que será el enfoque principal de todo este trabajo, que son los Dirichlet Process Mixture Models (DPM), realizando una pequeña revisión histórica de los métodos de simulación. Finalmente, veremos la conexión entre estos modelos con los Modelos de Particiones Aleatorias, para luego aplicarlo en el contexto de clustering.

1 Introducción

El material presentado a continuación se enmarca en el área de la Estadística Bayesiana No-Paramétrica, por lo que, en primer lugar, es una buena idea presentar una breve introducción de ambos conceptos.

1.1 Estadística Bayesiana

1.1.1 Tipos de Incertidumbre

Es común que al presentarnos como estadísticos se nos pregunte acerca de qué es lo que hacemos en nuestro trabajo, ante lo cual solemos responder que nuestro objetivo principal es el de cuantificar la incertidumbre. Explicamos, además, que para lo anterior nos apoyamos sobre la teoría de probabilidades, tomándola como herramienta principal para modelar aquellas incertezas de interés.

Pero, quizás nosotros mismos como estadísticos no hemos reparado acerca de a qué nos referimos exactamente con *incertidumbre*. Esta pregunta es la que nos lleva a las bases mismas de la estadística, así como a entender cómo surgen dos visiones que son diferentes entre sí: la **Estadística frecuentista** (también denominada clásica) y la **Estadística Bayesiana**.

En particular, se distinguen dos tipos de incertidumbres (O'Hagan 2004). Una de ellas la podemos denominar **incerteza ontólogica** (o aleatoria), mientras que la otra toma el nombre de **incerteza epistemológica** [1].

La incerteza ontológica trata acerca de una incerteza que está sujeta a una variabilidad aleatoria innata, que no podemos predecir bajo ninguna cantidad de información. Dentro de los ejemplos de incerteza ontológica se encuentran varios de los ejemplos introductorios a la estadística, como el lanzamiento de un dado o el de ganar la lotería.

Por otro lado, la incerteza epistemológica, tal como lo dice el nombre, es una incerteza acerca de lo que sabemos. La diferencia con la anterior es que en este caso sí podemos obtener información para disminuir, e incluso a veces eliminar, la incerteza. Por ejemplo, podemos tener incerteza acerca de la altura del Costanera Center en Santiago, pero podemos fácilmente buscar en internet esta información, eliminando completamente la incerteza [2].

Así, cuando hablamos de cuantificar la incertidumbre con probabilidades, podemos notar que siempre nos hemos estado refiriendo a incertezas ontológicas. En este sentido, las probabilidades se interpretan como la frecuencia de ocurrencia de un evento, considerando un número infinito de repeticiones. Este es el paradigma *frecuentista* de la estadística.

Ahora, ¿por qué no podemos modelar también las incertezas epistemológicas

[1] Es importante mencionar que la ontología es el estudio filosófico del *ser*, mientras que la epistemología es el estudio filosófico del *saber*.

[2] De hecho, el edificio central tiene una altura de 300 metros.

mediante probabilidades?. Esto es precisamente, de manera justificada, lo que propone el paradigma *Bayesiano*. En este caso ya no podemos interpretar las probabilidades como frecuencias, si no que como una *medida racional de incerteza*, lo cual normalmente dependerá de cada persona.

1.1.2 Modelos Bayesianos

Como vimos, el paradigma Bayesiano se basa en la idea de probabilidad subjetiva, donde estas cantidades reflejan el grado de creencia que un individuo tiene con respecto a eventos particulares.

En cuanto al modelamiento, estas creencias son plasmadas en una distribución a priori de los parámetros de interés, $\pi(\theta)$ [3], denominada simplemente **priori** de aquí en adelante. Además, debemos definir la verosimilitud de nuestros datos, $p(\mathbf{y}|\theta)$, que refleja justamente qué tan verosímiles son nuestros datos observados, dado un cierto valor de θ. Finalmente, ambas componentes son utilizadas para definir un modelo conjunto tanto de cantidades observables como no observables, esto es,

$$p(\mathbf{y}, \theta) = p(\mathbf{y}|\theta)\pi(\theta)$$

Luego, a la luz de nueva información, se actualiza nuestra creencia a priori, mediante el teorema de Bayes, obteniendo entonces la distribución a posteriori que llamaremos simplemente **posteriori** en lo que sigue.

$$\begin{aligned} \pi(\theta|\mathbf{y}) &= \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int_\Theta f(\mathbf{y}|\theta)\pi(\theta)d\theta} \\ &\propto f(\mathbf{y}|\theta)\pi(\theta) \end{aligned} \tag{1}$$

Por otro lado, es posible que también tengamos interés en la predicción de valores observables a futuro. Esto se obtiene fácilmente marginalizando la incerteza con respecto a las cantidades no observables, i.e.

$$\begin{aligned} p(y_{n+1}|\mathbf{y}) &= \int p(y_{n+1}, \theta|\mathbf{y})d\theta \\ &= \int p(y_{n+1}|\theta)\pi(\theta|\mathbf{y})d\theta \end{aligned} \tag{2}$$

donde en la segunda ecuación realizamos el supuesto de independencia condicional entre las observaciones, dado los valores del vector de parámetros θ. Ambos resultados, 1 y 2, podemos considerarlos como los productos principales para ser utilizados en la inferencia estadística mediante este paradigma.

[3] θ puede ser multidimensional.

Ejemplo 1: Modelo Normal-Normal

Para aterrizar los conceptos anteriores, consideremos el siguiente modelo:

$$\begin{aligned} y_1, ..., y_n | \theta &\overset{i.i.d.}{\sim} \mathrm{N}(\theta, \sigma^2) \\ \theta &\sim \mathrm{N}(\mu_0, \sigma_0^2) \end{aligned}$$

donde σ^2 es conocido. Utilizando la ecuación (1) es fácil mostrar que

$$\theta | y_1, ..., y_n \sim N(\mu_n, \sigma_n^2)$$

donde

$$\mu_n = \frac{(1/\sigma_0^2)}{1/\sigma_0^2 + n/\sigma^2}\mu_0 + \frac{(n/\sigma^2)}{1/\sigma_0^2 + n/\sigma^2}\bar{y}$$

y

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

Por otro lado, también es posible mostrar que, utilizando la ecuación (2),

$$y_{n+1} | \mathbf{y} \sim \mathrm{N}(\mu_n, \sigma^2 + \sigma_n^2)$$

El ejemplo anterior muestra una de las características principales de la estadística Bayesiana, que es la combinación de la información a priori con la información de los datos en los resultados a posteriori. En particular, notamos que:

- La media a posteriori corresponde a un promedio ponderado de la media a priori y el promedio muestral.
- La varianza a posteriori es un promedio armónico de la varianza a priori y la varianza del promedio muestral.

Además, en el ejemplo anterior obtuvimos un resultado bastante conveniente. Al considerar una función de verosimilitud Normal, así como una priori Normal para θ, obtuvimos que la posteriori sigue siendo una distribución Normal. Este tipo de modelos se denominan **conjugados** y serán importantes en los siguientes capítulos. Presentamos la definición formal a continuación.

Definición 1: Priori conjugada

Decimos que una clase $\mathcal{P}$ de distribuciones a priori es **conjugada** para la verosimilitud $f(\mathbf{y}|\theta)$ si

$$\pi(\theta) \in \mathcal{P} \implies \pi(\theta|\mathbf{y}) \in \mathcal{P}$$

esto es, la distribución a posteriori de θ sigue teniendo la misma distribución que la priori.

Los modelos conjugados son convenientes ya que obtenemos resultados analíticamente tractables, así como generalmente intuitivos. Ahora, nada nos debe restringir a ocupar modelos conjugados. En particular, vemos que, en un principio, tanto para la verosimilitud como para la priori podemos ocupar cualquier distribución de probabilidad, que deberán ser elegidas de acorde al problema.

Lo anterior provocó uno de los cuellos de botella más importantes de la estadística Bayesiana, que frenó su amplio uso en la práctica. Principalmente, el problema radicaba en los cálculos computacionales necesarios para realizar la inferencia a posteriori.

- Métodos computacionales
  - Rejection Sampling
  - Importance Sampling
  - MCMC: Gibbs sampler, Slice sampling
  - MCMC: Metropolis y Metropolis-Hastings
  - MCMC: Hamiltonian Monte Carlo and No-U-Turn Sampler
  - MCMC: Sequential Monte Carlo
- Lenguajes probabilísticos
  - Stan
  - Turing
  - PyMC3

1.2 Estadística No-Paramétrica

Normalmente, en estadística asumimos

$$y_1, ..., y_n | G \overset{i.i.d.}{\sim} G$$

Suponemos que la densidad de G, g, pertenece a

$$\mathcal{G} = \{g_\theta \colon \theta \in \Theta \subset \mathbb{R}^p\}$$

- Ejemplo
- Figura

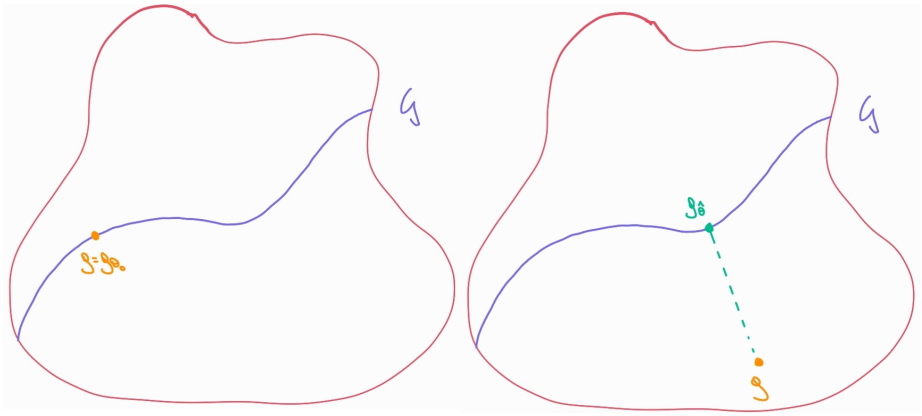

Figura 1: Necesidad de métodos flexibles

- Nos gustaría ir un poco más allá: estimación de densidades (figura) y regresión

1.3 Estadística Bayesiana No-Paramétrica

2 Procesos de Dirichlet

2.1 Definición

Consideremos en primer lugar el problema de estimación de densidades ...

Considerando lo anterior, debemos entonces definir medidas de probabilidad sobre medidas de probabilidad. Una de estas opciones es la del Proceso de Dirichlet, que definimos a continuación.

Definición 1: (Ferguson, 1973)

Sea $\alpha > 0$ y G_0 una medida de probabilidad definida sobre S. Un **Proceso de Dirichlet (DP)** de parámetros (α, G_0), denotado por $\text{DP}(\alpha, G_0)$, es una medida de probabilidad aleatoria G definida en S que asigna probabilidad $G(B)$ a todo conjunto medible B tal que, para toda partición medible finita $\{B_1, ..., B_k\}$ de S, la distribución conjunta del vector $(G(B_1), ..., G(B_k))$ es Dirichlet con parámetros

$$(\alpha G_0(B_1), ..., \alpha G_0(B_k))$$

Los parámetros G_0 y α se denominan la **medida de centralización** y la **precisión**, respectivamente. También se suele denominar αG_0 como la **medida base**.

Ferguson muestra que G existe para todo G_0. Además, señala algunas de las propiedades estadísticas de los DP, tales como:

Propiedades 1: (Propiedades Proceso de Dirichlet)

Sea $G \sim \text{DP}(\alpha, G_0)$, B, B_1 y B_2 conjuntos medibles, con $B_1 \cap B_2 = \emptyset$. Luego,

- El soporte de G coincide con el de G_0. Esto es,

$$G_0(B) = 0 \implies P(G(B) = 0) = 1$$

y

$$G_0(B) > 0 \implies P(G(B) > 0) > 1$$

- $\text{E}(G(B)) = G_0(B)$
- $\text{Var}(G(B)) = \frac{G_0(B)(1-G_0(B)}{1+\alpha}$
- $\text{Cov}(G(B_1), G(B_2)) = \frac{-G_0(B_1)G_0(B_2)}{1+\alpha}$

Las propiedades anteriores nos muestran la razón por la que G_0 se denomina la medida de centralización, así como el por qué α se denomina el parámetro de precisión. La covarianza muestra que la covarianza entre dos conjuntos cualesquiera es siempre negativa (acá hay extensiones, mencionar que están fuera del alcance de este trabajo)

> **Nota:** *Algunos se preguntarán, como yo lo hice la primera vez que aprendí sobre esto, el por qué se denomina un* ***proceso****. La razón es bastante sencilla, y es que el DP es un proceso estocástico que, en vez de estar indexado por índices comunes como el tiempo o coordenadas geográficas, está indexado por todos los conjuntos medibles, esto es, para cada conjunto medible B tenemos la variable aleatoria $G(B)$.*

Una propiedad muy importante que muestra Ferguson, que será central en el transcurso de esta monografía, es que G es casi-seguramente discreta. Este resultado nos dice que G se puede escribir como una suma ponderada de masas puntuales, también denominados átomos, esto es,

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{m_h}(\cdot)$$

donde $\sum_{h=1}^{\infty} w_h = 1$ y $\delta_x(\cdot)$ denota la medida de Dirac en x. En la Figura 2 se presenta gráficamente un Proceso de Dirichlet. A la izquierda se ilustran los átomos con puntos morados, donde los largos indican la masa que aporta cada uno. A la derecha se muestra cómo se calcularía la probabilidad para un cierto conjunto medible B, que es simplemente tomar la suma de las masas de los átomos que se encuentran dentro de este conjunto.

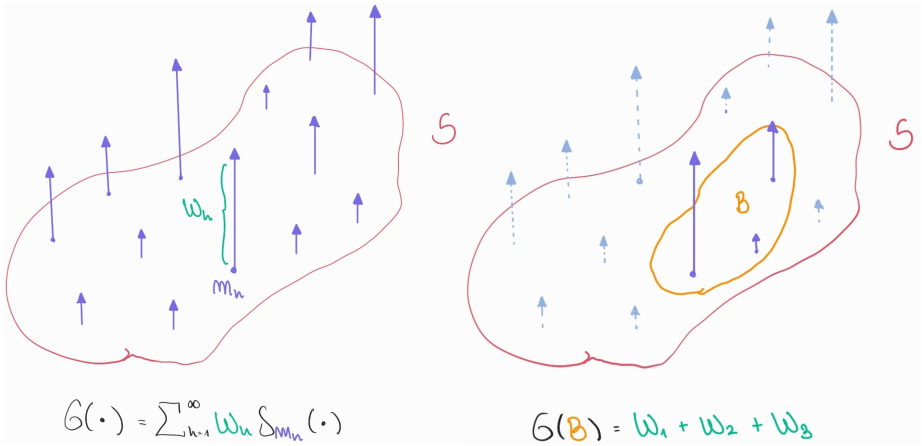

Figura 2: Naturaleza discreta del Proceso de Dirichlet

Por último, Ferguson tambien demuestra que un DP es conjugada para una muestra i.i.d. de esta distribución, donde se considera un promedio ponderado entre la medida de centralización G_0 y la función de distribución empírica de los datos.

Proposición 1: (Ferguson, 1973)

Sea $y_1, ..., y_n | G \overset{i.i.d}{\sim} G$ y $G \sim \text{DP}(\alpha, G_0)$. Luego,

$$G|y_1, ..., y_n \sim \text{DP}\left(\alpha + n, \frac{\alpha G_0 + n\hat{f}_n}{\alpha + n}\right)$$

donde $\hat{f}_n$ es la distribución empírica obtenida a partir de los datos, i.e.

$$\hat{f}_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}(\cdot)$$

Ahora, todo lo anterior aún no nos dice mucho acerca de como trabajar con esta distribución, ya que de momento solo sabemos que existen tales procesos, así como algunas propiedades. En la práctica, se ha trabajado principalmente de dos formas. La primera es marginalizando la medida de probabilidad aleatoria G, esto es, trabajar directamente con

$$p(y_1, ..., y_n) = \int p(y_1, ..., y_n|G) d\pi(G)$$

La otra forma es considerar la construcción de un DP mediante una representación basada en cortar una varilla de largo unitario de manera sucesiva e indefinida, denominada *Stick-Breaking*.

2.2 Construcción mediante Urnas de Pólya

Una de las formas de poder trabajar con un Proceso de Dirichlet es, irónicamente, no trabajar con él. En probabilidades esto lo logramos marginalizando con respecto a la medida que no es de interés.

Considerando una muestra aleatoria $y_1, ..., y_n | G \sim G$, Blackwell y MacQueen (Blackwell and MacQueen 1973) formulan una representación de la densidad marginal $p(y_1, ..., y_n)$ mediante una representación por Urnas de Pólya. En particular, se tiene que

$$p(y_1, ..., y_n) = p(y_1) \prod_{i=2}^{n} p(y_i|y_1, ..., y_{i-1})$$

y lo que muestran es que

$$p(y_i|y_1,...,y_{i-1}) = \frac{1}{M+i-1}\sum_{h=1}^{i-1}\delta_{y_h}(y_i) + \frac{M}{M+i-1}G_0(y_i)$$

Hay dos cosas bastante importantes:

- Dada la intercambiabilidad, las condicionales completas toman la misma forma
- La predictiva toma la misma forma para $i = n+1$
- Lo anterior se puede simplificar considerando solo los valores iguales

Ejemplo 1: Urnas de Pólya

Para ilustrar, consideramos el ejemplo de obtener datos de un Proceso de Dirichlet con medida de centralización Gamma(6, 4) y precisión $\alpha = 1, 10, 50, 100, 1000, 10000$.

2.3 Construcción Stick-Breaking

La forma de trabajar directamente con un Proceso de Dirichlet vino dada por una construcción stick-breaking indefinida.

Teorema 1: (Sethuraman, 1994)

Sea $w_h = \upsilon\prod_{l<h}(1-\upsilon_l)$ con $\upsilon_h \overset{i.i.d.}{\sim} \text{Beta}(1,\alpha)$ y $m_h \overset{i.i.d.}{\sim} G_0$, donde (υ_h) y (m_h) son independientes entre sí. Luego,

$$G(\cdot) = \sum_{h=1}^{\infty} w_h\delta_{m_h}(\cdot)$$

define un Proceso de Dirichlet de parámetros α y G_0.

En la Figura 4 se pequeña una pequeña ilustración del proceso stick-breaking para obtener un Proceso de Dirichlet.

Ahora, lo anterior sigue teniendo un pequeño problema, y es que claramente no podemos repetir el proceso una cantidad infinita de veces para obtener las secuencias infinitas de pesos y localizaciones. Para arreglar esto, se propone simplemente truncar la representación hasta un valor H fijo, considerando $\upsilon_H = 1$, u obtener los pesos w_h hasta cubrir un cierto número fijo, cercano a 1, de probabilidad.

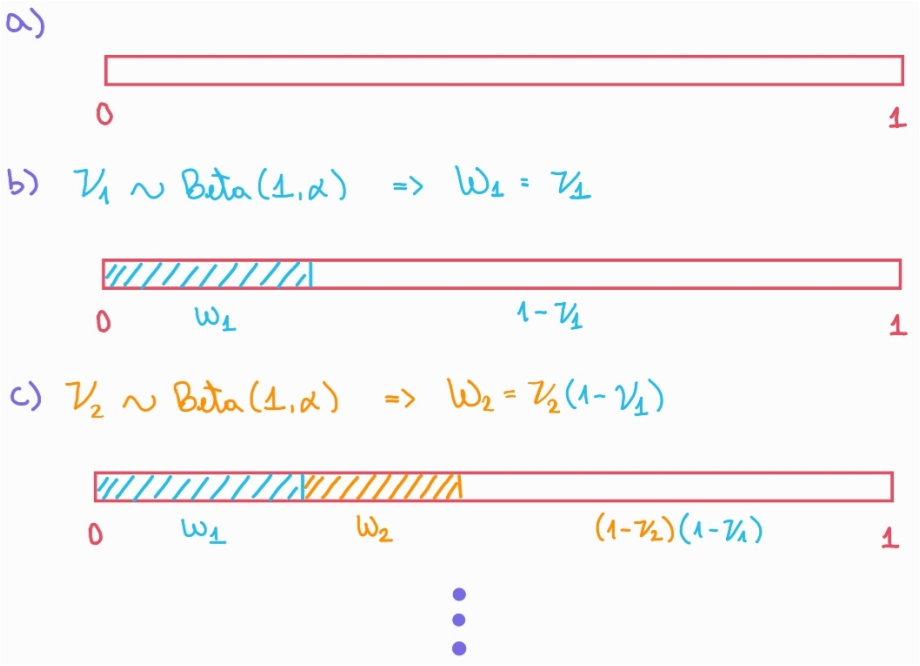

Figura 3: Ilustación del proceso de Stick-Breaking

Ejemplo 2: Stick-Breaking

> **Nota:** *En este punto del trabajo fue donde decidí cambiarme de R a Julia, ya que las simulaciones anteriores tomaban demasiado tiempo. Utilizando Julia obtuve una mejoría en rapidez de casi 100 veces.*

2.4 Algunos resultados asintóticos

Resultados de Antoniak, Korwar & Hollander.

3 Dirichlet Process Mixture Models

3.1 Definición

Podemos notar de la sección anterior que obtener distribuciones discretas con probabilidad 1 puede no ser adecuado para diferentes problemas. Considerando lo anterior, nos interesa extender los procesos de Dirichlet para formular modelos adecuados a estos casos. Utilizaremos la notación de (Neal 2000)

Una opción es usar estas medidas de probabilidad aleatoria como la *mixing distribution* en un modelo de mezcla, esto es,

$$f_G(y) = \int f_\theta(y) dG(\theta)$$

Equivalentemente:

$$\begin{aligned} y_i|\theta_i &\overset{ind.}{\sim} F(\theta_i) \\ \theta_i|G &\overset{i.i.d.}{\sim} G \\ G &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

Resultado de Antoniak

De manera más general podemos incluir inferencia sobre α e hiperparámetros de la medida de centro G_0. Procedemos entonces con el siguiente modelo:

$$\begin{aligned} y_i|\theta_i &\overset{ind.}{\sim} F(\theta_i) \\ \theta_i|G &\overset{i.i.d.}{\sim} G \\ G &\sim \mathrm{DP}(\alpha, G_\eta) \\ (\alpha, \eta) &\sim \pi \end{aligned} \tag{3}$$

3.2 Simulación a posteriori

- Muestreo de Gibbs (Geman and Geman 1984)
- Posteriori condicional completa de θ.

$$\theta_i|\theta_{-i}, y_i \sim \sum_{j \neq i} q_{i,j}\delta(\theta_j) + r_i H_i$$

donde

$$q_{i,j} = bF(y_i, \theta_j)$$
$$r_i = b\alpha \int F(y_i, \theta) dG_0(\theta)$$

donde b es tal que $\sum_{j \neq i} q_{i,j} + r_i = 1$ y H_i es la posteriori que se obtiene al considerar un modelo con G_0 como priori y una única observación de la verosimilitud $F(y_i, \theta_i)$.

- Resultado α: depende solo del número de valores únicos en θ.
- Resultado η: proporcional a su priori multiplicada por $\prod_c G_0(\theta_i)$.

3.2.1 Caso conjugado

De la ecuación anterior hay dos problemas:

- Simular de H_i
- Calcular la integral $\int F(y_i, \theta) dG_0(\theta)$

Así, en primera instancia se consideraron modelos conjugados.

3.2.1.1 Escobar & West (1995)

Ejemplo 1: (Datos de Galaxias)

Modelo propuesto por Escobar y West en el paper del 1995

To finalize the article, the authors present a final extension of the previous algorithm that now includes learning about the precision parameter α.

This final model is represented as,

$$\begin{aligned} Y_i|\pi_i &\overset{ind.}{\sim} \mathrm{N}(\mu_i, V_i), \quad i = 1, ..., n \\ \pi_1, ..., \pi_n &\overset{i.i.d.}{\sim} G \\ G &\sim DP(\alpha, G_0) \\ G_0 &= N - \Gamma^{-1}(m, 1/\tau, s/2, S/2) \\ \tau &\sim \Gamma^{-1}(w/2, W/2) \\ m &\sim N(0, A), \quad A \to \infty \\ \alpha &\sim \Gamma(a, b) \end{aligned}$$

It follows that,

* The full conditional of α is given by,

$$\alpha|\pi, m, \tau, D_n \equiv \alpha|\eta, k \sim \pi_\eta \Gamma(a+k, b-\log\eta) + (1-\pi_\eta)\Gamma(a+k-1, b-\log\eta)$$

where $\pi_\eta/(1-\pi_\eta) = (a+k-1)/[n(b-\log\eta)]$. Here we introduced an auxiliary variable η that satisfies

$$\eta|\alpha, k \sim \text{Beta}(\alpha+1, n)$$

* The full conditional of m is given by,

$$m|\tau, \pi, \alpha, D_n \equiv m|\pi, \tau \sim N\Big[x\bar{V}\sum(V_j^*)^{-1}\mu_j^*\,;\, x\tau\bar{V}\Big]$$

where $x = A/(A+\tau\bar{V})$ and $\bar{V}^{-1} = \sum(V_j^*)^{-1}$.

* The full conditional of τ is given by,

$$\tau|\pi, m, \alpha, D_n \equiv \tau|\pi, m \sim \Gamma^{-1}((w+k)/2, (W+K)/2)$$

where $K = \sum_{j=1}^{k}(\mu_j^* - m)^2/V_j^*$.

* The full conditionals of π are given by,

$$\pi_i|\pi^{(i)}, m, \tau, \alpha, D_n \sim q_0 G_i(\pi_i) + \sum_{j\neq i} q_j \delta_{\pi_j}(\pi_i)$$

where $G_i(\pi_i) \equiv N - \Gamma^{-1}(x_i, 1/X; (1+s)/2, S_i/2)$ and

$$\begin{aligned} q_0 &\propto \alpha c(s)[1+(y_i-m)^2/(sM)]^{-(1+s)/2}/M^{1/2} && \propto \alpha \cdot t_s(m, \sqrt{M}) \\ q_j &\propto \exp[-(y_i-\mu_j)^2/(2V_j)](2V_j)^{-1/2} && \propto N(\mu_j, V_j) \end{aligned}$$

$$\sum_{j=0, j\neq i}^{n} q_j = 1$$

with

- $x_i = (m+\tau y_i)/(1+\tau)$
- $X = \tau/(1+\tau)$
- $S_i = S + (y_i - m)^2/(1+\tau)$
- $M = (1+\tau)S/s$

- $c(s) = \Gamma((1+s)/2)\Gamma(s/2)^{-1}s^{-1/2}$

Thus, the algorithm proceeds as follows:

1. Sample initial values in the following way: * Sample $\alpha \sim \Gamma(a, b)$ * Sample $\tau \sim \Gamma^{-1}(w/2, W/2)$ * Sample $m \sim N(0, 1)$ * Sample π given m, τ 2. For $t = 1, ..., N$: * Sample $\eta_{(t)}|\alpha_{(t-1)}, k_{(t-1)}$ * Sample $\alpha_{(t)}|\eta_{(t)}, k_{(t-1)}$ * Sample $m_{(t)}|\pi_{(t-1)}, \tau_{(t-1)}$ * Sample $\tau_{(t)}|\pi_{(t-1)}, m_{(t)}$ * Sample $\pi_{i,(t)}|\pi^{(i)}_{(t-1)}, m_{(t)}, \tau_{(t)}, \alpha_{(t)}, D_n$ from its full conditional. Be aware that $\pi^{(i)}_{(t-1)}$ may contain already updated values of the form $\pi_{1,(t)}, ..., \pi_{i-1,(t)}, \pi_{i+1,(t-1)}, ..., \pi_{n,(t-1)}$.

3.2.1.2 Problema de "sticky-clusters"

Problema: sticky clusters

Bush & MacEachern, separación de valores.

Figura de cómo recuperar valores utilizando las indicadoras de clusters

3.2.2 Caso no-conjugado

- No-gaps de MacEachern & Muller
- Algoritmo 8 de Neal

3.2.3 Otras opciones

Inferencia variacional, DP ϵ-finito.

4 Modelos de Particiones Aleatorias y Clustering

5 Aplicación: Modelo CAPM

- El modelo de valorización de activos financieros (CAPM) fue desarrollado en los años 60 de forma independiente por Jack Treynor, William Sharpe, John Linter y Jan Mossin.
- El modelo propone

$$\mathrm{E}(R) = r_f + \beta(\mathrm{E}(R_m) - r_f)$$

donde R es el retorno del activo, r_f es la tasa de retorno libre de riesgo, R_m es el retorno del mercado y β es el riesgo sistemático del activo bajo estudio.

- Normalmente se considera el modelo de regresión

$$Y_j \equiv r_j - r_{fj} = \alpha + \beta(r_{mj} - r_{fj}) + \varepsilon_j, \quad j = 1, ..., n$$

6 Referencias

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2017. "Julia: A Fresh Approach to Numerical Computing." *SIAM Review* 59 (1): 65–98. https://doi.org/10.1137/141000671.

Blackwell, David, and James B. MacQueen. 1973. "Ferguson Distributions Via Polya Urn Schemes." *The Annals of Statistics* 1 (2): 353–55. https://doi.org/10.1214/aos/1176342372.

Dahl, David B., Devin J. Johnson, and Peter Müller. 2022. "Search Algorithms and Loss Functions for Bayesian Clustering." *Journal of Computational and Graphical Statistics* 31 (4): 1189–1201. https://doi.org/10.1080/10618600.2022.2069779.

Geman, Stuart, and Donald Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6 (6): 721–41. https://doi.org/10.1109/TPAMI.1984.4767596.

Jara, Alejandro, Timothy Hanson, Fernando A. Quintana, Peter Müller, and Gary L. Rosner. 2011. "DPpackage: Bayesian Semi- and Nonparametric Modeling in R." *Journal of Statistical Software* 40 (April): 1–30. https://doi.org/10.18637/jss.v040.i05.

Neal, Radford M. 2000. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics* 9 (2): 249–65. https://doi.org/10.2307/1390653.

Ni, Yang, Peter Müller, Maurice Diesendruck, Sinead Williamson, Yitan Zhu, and Yuan Ji. 2020. "Scalable Bayesian Nonparametric Clustering and Classification." *Journal of Computational and Graphical Statistics* 29 (1): 53–65. https://doi.org/10.1080/10618600.2019.1624366.

O'Hagan, Tony. 2004. "Dicing with the Unknown." *Significance* 1 (3): 132–33. https://doi.org/10.1111/j.1740-9713.2004.00050.x.

Page, Garritt L., Jose J. Quinlan, S. McKay Curtis, and Radford M. Neal. 2022. "ppmSuite: A Collection of Models That Employ a Product Partition Distribution as a Prior on Partitions." https://CRAN.R-project.org/package=ppmSuite.

Ross, Gordon J., Dean Markwick, Kees Mulder, and Giovanni Sighinolfi. 2020. "Dirichletprocess: Build Dirichlet Process Objects for Bayesian Modelling." https://CRAN.R-project.org/package=dirichletprocess.