Universidad Nacional Autónoma de México

Posgrado en Ciencias Matemáticas
Facultad de Ciencias

Métodos computacionales para modelos de mezclas no paramétricos

Tesis

que para obtener el grado académico de
Maestro en Ciencias

presenta
Asael Fabian Martínez Martínez

director de tesis
Dr. Ramsés Humberto Mena Chávez

México, D.F. Diciembre, 2010

A Manuel, Mercedes,
Haziel, Elianai, Eliel
& Matías

Agradecimientos

Agradezco a todos aquellos que de alguna manera me han ayudado a llegar a este punto:

A mis padres, Manuel y Mercedes, por su apoyo incondicional

A Ramsés, por su gran paciencia y por toda su ayuda para realizar este trabajo

A mis maestros del Departamento de Probabilidad y Estadística, por sus enseñanzas

A mis compañeros de la maestría, porque me han permitido aprender mucho de ellos

Por último, pero siempre siendo el primero, gracias Eterno por crear todas las cosas

האף תפר משפטי תרשיעני למען תצדק...

...שמע־נא ואנכי אדבר אשאלך והודיעני
לשמע־אזן שמעתיך ועתה עיני ראתך

איוב מ"ח, מב"ד־ה

Índice general

Introducción

Dentro de la estadística, una de las principales tareas es hacer inferencias acerca de la distribución de probabilidad que mejor modele algún fenómeno a partir de un conjunto de observaciones. Los modelos utilizados deben tener un sustento probabilístico para que se puedan considerar válidos. En la rama de la estadística bayesiana existe un modelo básico que considera a las observaciones como realizaciones de eventos aleatorios que son condicionalmente independientes dada cierta información y con distribución común, la cual no está completamente especificada. Además, dentro de este enfoque, todo aquello que no se conozca con certeza es posible modelarlo a través de distribuciones de probabilidad.

En el enfoque paramétrico de la estadística bayesiana, todo aquello que no se conoce acerca de la distribución se modela a través de un parámetro de dimensión finita. La distribución asignada a este parámetro, junto con la información dada por las observaciones y el teorema de Bayes permiten, entonces, hacer inferencias sobre él, inclusive hacer predicciones acerca del fenómeno bajo estudio.

Un área de interés relativamente joven dentro de la estadística bayesiana es la llamada estadística bayesiana no paramétrica. En este enfoque, esencialmente, en lugar de asignar una distribución de probabilidad a los parámetros de la distribución, es esta última la que se considera desconocida y toma valores en cierto espacio de distribuciones. Para poder hacer inferencias con este tipo de modelos es necesario asignarle una distribución inicial a esta distribución desconocida.

En relación a la elección de la distribución inicial en este enfoque no paramétrico, el proceso Dirichlet se considera la piedra angular, debido a que fue el primer ejemplo concreto y actualmente es la intersección de muchos otros modelos más generales. De hecho, este proceso ha servido como base para el desarrollo de nuevas medidas de probabilidad aleatorias, entre ellas: los procesos con incrementos independientes normalizados, el proceso Poisson–Dirichlet de dos parámetros y la representación *stick–breaking*; todas ellas, distribuciones aleatorias discretas casi seguramente.

Como una extensión a estas medidas de probabilidad aleatorias se han desarrollado modelos que puedan ser aplicados en situaciones en que las observaciones se modelen a través de variables aleatorias continuas. Este tipo de modelos, conocidos como modelos de mezclas, utilizan una distribución inicial que es una mezcla formada por un kernel y una medida de probabilidad aleatoria discreta c.s. De esta manera, se amplía el rango de aplicaciones de los modelos no paramétricos, por ejemplo, la estimación de densidades.

Sin embargo, tratar analíticamente modelos de mezclas resulta prácticamente imposible. La solución más utilizada actualmente es auxiliarse de métodos computacionales, principalmente los conocidos como Monte Carlo vía cadenas de Markov.

Este trabajo, basado en los trabajos de Ishwaran & James (2001), Neal (2000) y Fuentes-García, Mena & Walker (2010), estudia e implementa algunos de estos modelos de mezclas, además de hacer un análisis comparativo entre ellos.

En el primer capítulo se da una introducción a la estadística bayesiana no paramétrica. Existen al menos dos resultados principales que sustentan la inferencia bayesiana: la teoría de decisiones y el concepto de intercambiabilidad junto con el teorema de representación de Bruno de Finetti. En este trabajo se utiliza el segundo resultado debido, principalmente, a que las medidas de probabilidad aleatorias que se estudian están construidas bajo el supuesto de intercambiabilidad y el teorema de representación justifica la existencia de la distribución inicial, necesaria para hacer inferencias.

Una vez que se introducen estos conceptos, se estudian algunas medidas de probabilidad aleatorias. Se comienza con el proceso Dirichlet (Ferguson 1973) y su caracterización bajo el esquema de urnas de Pólya, dado por Blackwell & MacQueen (1973) y generalizado por Pitman (1996), posteriormente se estudian los procesos con incrementos independientes normalizados (Regazzini, Lijoi & Prünster 2003), con especial atención en los procesos Gamma y σ–estable normalizados, se continúa con el proceso Poisson–Dirichlet de dos parámetros (Pitman & Yor 1997) y la representación *stick–breaking* (Ishwaran & James 2001). Por último se estudian las mezclas de medidas de probabilidad aleatorias. Estas proporcionan información importante para realizar análisis de conglomerados. En esta misma parte se estudia una nueva medida de probabilidad aleatoria: el proceso geométrico (Fuentes-García et al. 2010); su estructura es una simplificación de las mezclas de procesos Dirichlet, útil en el ámbito de la estimación de densidades.

En el segundo capítulo se estudian algunos métodos de estimación de densidades, en especial sus algoritmos. Se comienza con una introducción a la simulación estocástica, herramienta indispensable para estos métodos; se presentan dos de los algoritmos MCMC más utilizados en la estadística bayesiana: el Metropolis–Hastings (Metropolis, Rosenbluth, Rosenbluth & Teller 1953) y un caso particular, el *Gibbs sampler* (Geman & Geman 1984). En la segunda sección se presentan los algoritmos de estimación de densidades que posteriormente se implementarán y compararán. El primer algoritmo es la aplicación del esquema de urnas de Pólya a través del *Gibbs sampler*, estudiado, entre otros, por Ishwaran & James (2001); estos mismos autores proponen un nuevo método basado en la representación *stick–breaking* que permite la actualización de parámetros por bloque, el denominado *Gibbs sampler* por bloques. Por otro lado, Neal (2000) propone distintos algoritmos de estimación de densidades; de estos se estudian los Algoritmos 6, 7 y 8, que están diseñados para el caso en que no exista conjugamiento entre la medida base y el kernel de la mezcla. Para terminar el capítulo se continúa el estudio del proceso geométrico, dando su implementación a través del *Gibbs sampler*.

El tercer capítulo está dedicado a la obtención de las distribuciones posteriores y condicionales necesarias para programar cada uno de los métodos antes mencionados. Además, las medidas de probabilidad aleatorias subyacentes a estos métodos, a excepción del proceso geométrico, dependen de uno o dos parámetros, los cuales pueden incluirse en los algoritmos de simulación a través de la asignación de distribuciones iniciales. En la segunda parte de este capítulo se explica cómo incluir estas actualizaciones.

Por último, el cuarto capítulo contiene un análisis comparativo de todos los métodos programados. Se tomaron muestras de distintos modelos de mezclas de distribuciones normales así como un conjunto de datos reales. Para poder comparar los resultados se explican algunas de las estadísticas más usuales dentro del contexto de modelos de mezclas para seleccionar modelos y monitorear su comportamiento. En la parte final, además de mostrar los principales resultados de las simulaciones, se hace una serie de observaciones acerca de los modelos, las cuales permiten analizar su comportamiento.

Debido a la gran cantidad de datos generados por las simulaciones, en el último capítulo se presentan sólo aquellos fuertemente relevantes para las conclusiones; sin embargo, para fines de referencia se incluye un apéndice con todos los resultados obtenidos.

Asimismo, se incluye un segundo apéndice que sirve como guía rápida para utilizar los programas que se anexan a este trabajo.

– I –

Estadística bayesiana no paramétrica

Uno de los propósitos de la estadística es hacer inferencias acerca de la distribución de probabilidad subyacente a cierto fenómeno aleatorio, a partir de un conjunto de observaciones; ya sea que se realice un análisis sobre un fenómeno pasado o predicciones sobre un fenómeno futuro de características similares. Estas inferencias, a su vez, están basadas en un modelo probabilístico que da soporte a las conclusiones, o decisiones, que puedan obtenerse.

Desde la perspectiva bayesiana de la estadística todas las cantidades desconocidas son tratadas como variables aleatorias y se asume, además, que las observaciones provienen de variables aleatorias condicionalmente independientes con distribución común. Esta distribución depende de parámetros desconocidos, por tanto, es necesario asignar una distribución «inicial» a estos parámetros. La justificación de la existencia de una distribución inicial, bajo el supuesto de intercambiabilidad, está dada por el teorema de representación, atribuido a Bruno de Finetti. De esta manera, y haciendo uso del teorema de Bayes, es posible hacer inferencias sobre la distribución «posterior» del parámetro o sobre observaciones futuras, incorporando la información dada por las observaciones.

A pesar del potencial que posee este «modelo bayesiano», su uso fue restringido a formas paramétricas para las distribuciones iniciales, debido, en parte, a la imposibilidad del cálculo de la distribución posterior en casos más generales. Para afrontar esta limitante se han desarrollado modelos «no paramétricos» los cuales utilizan «distribuciones de probabilidad aleatorias» como distribuciones iniciales. Una de las distribuciones aleatorias de mayor trascendencia en la estadística bayesiana no paramétrica es generada por el proceso Dirichlet.

El proceso Dirichlet también permitió el desarrollo de nuevas distribuciones de probabilidad aleatorias, o equivalentemente «medidas de probabilidad aleatorias», algunas de las cuales son extensiones de éste. Entre las distintas clases de medidas de probabilidad aleatorias se tienen: los «procesos con incrementos independientes normalizados», el «proceso Poisson–Dirichlet de dos parámetros» y las medidas de probabilidad aleatorias inducidas mediante la representación *stick–breaking*. Una característica de todas estas medidas de probabilidad aleatorias es que son discretas casi seguramente.

El carácter discreto de estas medidas, sin embargo, puede ser un inconveniente si las observaciones se modelan a través de variables aleatorias continuas. Una manera de solucionarlo es trabajando con distribuciones aleatorias, llamadas «mezclas de distribuciones aleatorias», que son una mezcla de distribuciones, formadas por un kernel absolutamente continuo y una medida de probabilidad aleatoria como las anteriores.

El objetivo de este primer capítulo es presentar los resultados necesarios para el desarrollo de este trabajo sobre la estadística bayesiana no paramétrica. Se comienza con una explicación de las bases de la estadística bayesiana de acuerdo al concepto de intercambiabilidad y el teorema de representación. Este teorema es importante, ya que garantiza la existencia de una distribución inicial.

En la segunda parte se da una introducción a las medidas de probabilidad aleatorias, comenzando con el proceso Dirichlet, introducido por Ferguson (1973) y su caracterización a través del esquema de urnas de Pólya, dada por Blackwell & MacQueen (1973). Asimismo se estudiarán las medidas de probabilidad aleatorias mencionadas anteriormente: procesos con incrementos independientes normalizados (Regazzini et al. 2003), proceso Poisson–Dirichlet de dos parámetros (Pitman & Yor 1997) y la representación *stick–breaking* (Ishwaran & James 2001). Por último, se estudiarán las mezclas de distribuciones aleatorias: una generalización del modelo dado por Lo (1984) para el proceso Dirichlet y el proceso geométrico, una nueva medida de probabilidad aleatoria dada por Fuentes-García et al. (2010).

§ 1 Inferencia bayesiana

Supóngase que se quiere obtener información acerca de algún fenómeno a través de un conjunto de observaciones $x_1, \ldots, x_n$. Para realizar un análisis estadístico, i.e., poder hacer inferencias, se asume que las observaciones son generadas a través de alguna distribución de probabilidad, μ, que pertenece a alguna familia paramétrica de funciones de distribución, indexada por un parámetro θ de dimensión finita. El parámetro θ es desconocido y pertenece al espacio de parámetros $(\Theta, \mathcal{B}(\Theta))$.

En el enfoque bayesiano, toda cantidad desconocida es modelada a través de variables aleatorias, cuyas distribuciones son conocidas, las cuales reflejan el «conocimiento» que se tiene sobre cada una. De esta manera, previo al análisis se tiene cierto conocimiento sobre el parámetro θ que se modela a través de una «distribución inicial», π; posteriormente, para actualizar el conocimiento acerca de θ se hace uso de las observaciones y del teorema de Bayes para obtener la «distribución posterior» de éste

$$\mathbb{P}[\theta \in B \mid X^n \in A] = \frac{\mu_\theta(A)\pi(B)}{\int_\Theta \mu_\theta(A)\mathrm{d}\pi(\theta)},$$

donde $X^n = (X_1, \ldots, X_n)$ es un vector aleatorio con densidad conjunta μ_θ.

En cuanto a la incorporación de la información dada por las observaciones, no es útil suponer que éstas son independientes, ya que se tendría

$$\mathbb{P}[X_{n+1} \in A_{n+1} \mid X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n] = \mathbb{P}[X_{n+1} \in A_{n+1}],$$

por tanto, las observaciones no aportarían información alguna para el análisis, i.e., suponer que las observaciones son independientes no permitiría utilizarlas para actualizar la información acerca del parámetro θ y por consecuencia, no podrían hacerse inferencias.

Se necesita, por tanto, cambiar el supuesto de independencia por alguno que relacione las observaciones, pero sin ser tan restrictivo. La propiedad de «intercambiabilidad» es una de las que permite tener tal relación.

1.1 Intercambiabilidad y teorema de representación

El concepto de intercambiabilidad expresa la relación entre un conjunto de variables aleatorias de una manera simple. Básicamente, este concepto dice que toda la información relevante de las variables está en sus valores, de modo que sus índices no proporcionan información alguna.

Intercambiabilidad. Un conjunto finito $\{X_i\}_{i=1}^n$ de variables aleatorias se dice que son «intercambiables» si, para toda $n \geq 1$, el vector aleatorio $(X_1, \ldots, X_n)$ tiene la misma distribución conjunta que $(X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)})$ para toda permutación σ de $\{1, 2, \ldots, n\}$. Una sucesión infinita $\{X_n\}_{n\geq 1}$ de variables aleatorias es intercambiable si toda subsucesión finita lo es.

Una característica de la intercambiabilidad es que contiene la propiedad de independencia, pero no lo contrario, como se muestra a continuación.

Ejemplo 1. Considérese una sucesión $X_1, X_2, \ldots$ de variables aleatorias, finita o infinita, independientes e idénticamente distribuidas. Claramente $(X_1, \ldots, X_n) \overset{\mathrm{d}}{=} (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ para cualquier permutación σ. Por tanto, toda sucesión de variables aleatorias independientes es intercambiable.

Ejemplo 2. Sea $X_1, \ldots, X_n$ una sucesión intercambiable. Si $n < \infty$, su coeficiente de correlación es $\rho_X \geq -1/(n-1)$, mientras que si $n = \infty$, se tiene $\rho_X \geq 0$ (Aldous 1985). Por lo que intercambiabilidad no implica independencia.

Por otro lado, una caracterización de la intercambiabilidad, utilizada ampliamente dentro de la estadística bayesiana, está dada a por el conocido teorema de representación de Bruno de Finetti.

Teorema de representación. Sea $\{X_n\}_{n\geq 1}$ una sucesión infinita de variables aleatorias con valores en $(\mathbb{X}, \mathscr{X})$. La sucesión $\{X_n\}_{n\geq 1}$ es intercambiable si y sólo si existe una medida de probabilidad aleatoria P en $(\mathbb{X}, \mathscr{X})$ tal que, condicional a que $P = \mu$, $\{X_n\}_{n\geq 1}$ son independientes con distribución μ. Además, si la sucesión es intercambiable, entonces la distribución de P es única y

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(A) \Longrightarrow P(A) \text{ c.s.}$$

cuando $n \to \infty$, para toda $A \in \mathscr{X}$; donde $\Longrightarrow$ denota convergencia débil.

Este teorema se puede reescribir a través de la distribución conjunta de $\{X_n\}_{n\geq 1}$. Para toda $n \geq 1$

$$\mathbb{P}[X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n] = \int_{\mathbb{M}} \prod_{i=1}^{n} \mu(A_i)\Pi(\mathrm{d}\mu), \tag{1}$$

donde $A_i \in \mathscr{X}$, $i \geq 1$, $\mathbb{M}$ es el espacio de medidas de probabilidad en $\mathbb{X}$, Π es la distribución de P y dado $P = \mu$, las variables aleatorias $\{X_n\}_{n\geq 1}$ son independientes con distribución μ.

De esta manera, el teorema de representación garantiza la existencia de la distribución inicial Π.

Un estudio más amplio sobre estos conceptos se puede encontrar, por ejemplo, en Schervish (1995). A continuación solamente se explica cómo estos se aplican en el proceso de inferencia.

1.2 Modelos bayesianos

El concepto de intercambiabilidad junto con el teorema de representación permiten incorporar la información de las observaciones en el proceso de inferencia. El proceso para hacer inferencias se puede resumir como sigue

1. Se tiene un conjunto de observaciones $x = \{x_1, \ldots, x_n\}$ condicionalmente independientes, dado μ_θ, con distribución común μ_θ. Esta distribución comúnmente pertenece a una familia paramétrica indexada por un parámetro finito dimensional θ. Supóngase que μ_θ tiene función de densidad asociada $p(x_i \mid \theta)$. Se tiene entonces

$$p(x \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta),$$

   la cual, vista como función de θ se conoce como «función de verosimilitud».

   Se tiene también una distribución inicial π para θ, con función de densidad $p(\theta)$.

2. Se actualiza la información acerca de θ calculando su «distribución final» a través del teorema de Bayes. En términos de su función de densidad

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int_\Theta p(x \mid t)p(t)\mathrm{d}t},$$

   omitiendo el denominador, i.e., la constante de normalización

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta).$$

3. Una vez obtenida la distribución final se pueden hacer inferencias sobre cualquier función medible de θ, por ejemplo, su valor esperado *a posteriori*

$$\mathbb{E}[\theta \mid x] = \int_\Theta \theta p(\theta \mid x)\mathrm{d}\theta.$$

   También es posible hacer inferencia sobre observaciones futuras, e.g., x_{n+1}, por medio de la llamada «distribución predictiva», cuya función de densidad está dada por

$$p(x_{n+1} \mid x) = \int_\Theta p(x_{n+1} \mid \theta)p(\theta \mid x)\mathrm{d}\theta.$$

El modelo resultante de este tipo de análisis se conoce como «modelo bayesiano paramétrico», el cual generalmente se representa como

$$\begin{aligned} x_i \mid \theta &\overset{\text{iid}}{\sim} \mu_\theta, \quad i = 1, \ldots, n \\ \theta &\sim \pi. \end{aligned} \tag{2}$$

La relación de este modelo con la descomposición (1) es clara cuando la medida de probabilidad aleatoria se restringe a la familia paramétrica en cuestión y se plantea un modelo con población infinita.

Como se ha mencionado, la distribución π refleja el conocimiento previo que se tiene acerca del fenómeno bajo estudio. Además, en estos modelos, la distribución μ_θ comúnmente está restringida a cierta familia de manera que sea lo más fácilmente manejable, por ejemplo, que sea unimodal o forme un par conjugado con π.

Sin embargo, si no se tiene información suficiente sobre el fenómeno bajo estudio o se desea un modelo que refleje lo más posible la información de las observaciones, el parámetro θ se define, entonces, en un espacio infinito dimensional. Esto resulta equivalente a «transferir» la incertidumbre del parámetro a la distribución de los datos; por lo que en lugar de asignar una distribución inicial sobre el espacio de parámetros, Θ, se asigna una distribución inicial sobre el espacio de las distribuciones, $\mathbb{M}$.

Este tipo de problemas dio origen a la estadística bayesiana «no paramétrica». El modelo obteniendo es, entonces,

$$\begin{aligned} x_i \mid \mu &\overset{\text{iid}}{\sim} \mu, \quad i = 1, \ldots, n \\ \mu &\sim \Pi. \end{aligned} \tag{3}$$

La existencia de la distribución Π está garantizada por el teorema de representación.

A pesar de contar con un modelo muy general, su uso se vio frenado debido a la falta de herramientas que ayudaran a manipular las distribuciones posteriores resultantes. No fue hasta que Ferguson (1973) introdujo el proceso Dirichlet que se retomó el interés en la estadística bayesiana no paramétrica; además, los avances computacionales han permitido desarrollar métodos que facilitan aún más su uso. Esto último se discutirá con más detalle en la siguiente sección y en el Capítulo II.

Modelos jerárquicos

Una variante de los modelos (2) y (3) se tiene cuando la distribución inicial depende de parámetros desconocidos. En este caso es necesario asignarles también una distribución inicial, obteniendo así un modelo más complejo, conocido como «modelo jerárquico». Por ejemplo, si al modelo (2) se le asigna una distribución ν a los parámetros de θ se obtiene, esquemáticamente, el modelo jerárquico

$$\begin{aligned} x_i \mid \theta_i &\overset{\text{ind}}{\sim} \mu_{\theta_i}, \quad i = 1, \ldots, n \\ \theta \mid \phi &\overset{\text{iid}}{\sim} \pi_\phi \\ \phi &\sim \nu \end{aligned} \tag{4}$$

donde el parámetro ϕ se conoce como «hiperparámetro». Las θ_i's son ahora una muestra de una distribución poblacional común y las observaciones x_i se pueden ver como muestras de varias subpoblaciones. Una vez realizado el análisis, también es posible hacer inferencias sobre el parámetro ϕ.

Cualquier modelo bayesiano se puede extender a uno jerárquico. Además es posible tener varios «niveles», por ejemplo, el modelo jerárquico anterior consta de tres niveles: observaciones, parámetros e

hiperparámetros; sin embargo, la dificultad al agregar niveles se encuentra al calcular las distribuciones posteriores.

Es importante mencionar que la propiedad de intercambiabilidad no aplica para estos modelos, debido a que las observaciones x_i son muestra de distintas subpoblaciones. Afortunadamente existe la propiedad de «intercambiabilidad parcial», en la cual se tiene una sucesión de variables aleatorias $\{X_n\}_{n\geq 1}$ con una etiqueta adicional, dada por θ_i por ejemplo, y, entonces, una subsucesión de variables aleatorias con la misma etiqueta es intercambiable. Para una mayor referencia véase, por ejemplo, Schervish (1995, Sección 8.1).

§ 2 Medidas de probabilidad aleatorias

Los modelos bayesianos no paramétricos requieren la especificación de una distribución inicial Π; sin embargo, esta distribución está definida en el espacio de distribuciones $(\mathbb{M}, \mathscr{M})$. Recordando, una función de distribución induce una medida de probabilidad, por lo que resulta equivalente hablar de espacios de distribuciones que de espacios de medidas de probabilidad. Por tanto, en esta sección se trabajará con «medidas de probabilidad aleatorias».

Definición 1. Sean $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad, $(\mathbb{X}, \mathscr{X})$ un espacio completo y separable y $(\mathbb{M}, \mathscr{M})$ un espacio de medidas de probabilidad sobre $\mathbb{X}$, equipado con la topología de convergencia débil. Se dice que $P : \mathbb{M} \times \Omega \to \mathbb{X}$ es una «medida de probabilidad aleatoria» si

i. $P(m, \cdot)$ es una variable aleatoria para cada $m \in \mathbb{M}$;

ii. $P(\cdot, \omega)$ es una medida de probabilidad para cada $\omega \in \Omega$.

Además, por convención, se dirá que una sucesión $\{y_i\}_{i=1}^n$ es una «muestra de P» si, dada $P = \mu$, las y_i's son condicionalmente independientes y con distribución común μ.

Uno de los primeros trabajos sobre la construcción de medidas de probabilidad aleatorias fueron las distribuciones *tailfree*, presentadas por Freedman (1963). Sin embargo, el proceso Dirichlet, introducido por Ferguson (1973), se considera la piedra angular de la estadística bayesiana no paramétrica. Este proceso se basa en la distribución Dirichlet.

Definición 2 (Distribución Dirichlet). Sea $\{Z_i\}_{i=1}^n$ una sucesión de variables aleatorias independientes con distribución Gamma con parámetros $(a_i, 1)$, donde $a_i \geq 0$ para toda i y $a_i > 0$ para al menos una i, $i = 1, \ldots, n$. La «distribución Dirichlet» con parámetro $(a_1, \ldots, a_n)$, denotada por $\mathcal{D}ir(a_1, \ldots, a_n)$, es la distribución del vector $(Y_1, \ldots, Y_n)$, donde

$$Y_j = \frac{Z_j}{\sum_{i=1}^n Z_i},$$

para $j = 1, \ldots, n$.

Esta distribución es singular con respecto a la medida de Lebesgue en el espacio n–dimensional, ya que $Y_1 + \cdots + Y_n = 1$. Además, si alguna $a_i = 0$, su correspondiente Y_i es degenerada en cero. Sin embargo, si $a_i > 0$ para toda i, la distribución $(n-1)$–dimensional de $Y_1, \ldots, Y_{n-1}$ es absolutamente continua con densidad

$$f(y_1, \ldots, y_{n-1}; a_1, \ldots, a_n) = \frac{\Gamma(a_1 + \cdots + a_n)}{\Gamma(a_1) \cdots \Gamma(a_n)} \left(\prod_{j=1}^{n-1} y_j^{a_j - 1} \right) \left(1 - \sum_{j=1}^{n-1} y_j \right)^{a_n - 1} \mathbb{1}_{S_{n-1}}(y_1, \ldots, y_{n-1}),$$

donde S_{n-1} es el simplejo $\{(y_1, \ldots, y_{n-1}) : y_j \geq 0, \sum_{j=1}^{n-1} y_j \leq 1\}$.

Utilizando la distribución Dirichlet, Ferguson (1973) define el proceso Dirichlet a partir de sus distribuciones finito dimensionales y demuestra su existencia a través de las condiciones de consistencia de Kolmogorov.

DEFINICIÓN 3 (PROCESO DIRICHLET). Sea α una medida finita, no nula sobre un espacio medible $(\mathbb{X}, \mathscr{X})$. Se dice que P es un «proceso Dirichlet» en $(\mathbb{X}, \mathscr{X})$ con parámetro α, denotado por $\mathcal{D}P(\alpha)$, si, para cada $k \geq 1$ y cada partición medible $\{B_i\}_{i=1}^k$ de $\mathscr{X}$, la distribución de $(P(B_1), \ldots, P(B_k))$ es Dirichlet con parámetro $(\alpha(B_1), \ldots, \alpha(B_k))$.

El proceso Dirichlet ha sido de gran importancia, en parte porque da una forma explícita para las distribuciones finito dimensionales de Π en (1) y tiene la propiedad de conjugamiento, i.e., al utilizarlo como distribución inicial, la distribución posterior también es un proceso Dirichlet con sus parámetros actualizados, lo cual permite desarrollar cálculos analíticamente. Asimismo, se han encontrado diversas formas de caracterizarlo, lo que ha permitido el desarrollo de nuevas medidas de probabilidad aleatorias. A continuación se estudian algunas de ellas: procesos con incrementos independientes normalizados, proceso Poisson–Dirichlet de dos parámetros y las medidas de probabilidad aleatorias inducidas mediante la representación *stick–breaking*. Pero antes se presenta una caracterización del proceso Dirichlet que ha permitido implementar computacionalmente modelos no paramétricos con esta medida de probabilidad aleatoria como distribución inicial.

2.1 ESQUEMA DE URNAS DE PÓLYA

Blackwell & MacQueen (1973) describen la construcción del proceso Dirichlet a través de una generalización del esquema de urnas de Pólya. Esta se resume en la siguiente proposición.

PROPOSICIÓN 1. Una sucesión de variables aleatorias $\{X_n\}_{n \geq 1}$ con valores en $\mathbb{X}$ es una «sucesión de Pólya con parámetro α» si

$$\mathbb{P}[X_1 \in \cdot\,] = \frac{\alpha(\cdot)}{\alpha(\mathbb{X})} \quad \text{y} \quad \mathbb{P}[X_{i+1} \in \cdot \mid X_1, \ldots, X_i] = \frac{\alpha_i(\cdot)}{\alpha_i(\mathbb{X})}, \quad i \geq 1,$$

donde $\alpha_i(\cdot) = \alpha(\cdot) + \sum_{k=1}^{i} \delta_{X_k}(\cdot)$.

Sea $\{X_n\}_{n \geq 1}$ una sucesión de Pólya con parámetro α, entonces

I. $\alpha_i(\cdot)/\alpha_i(\mathbb{X})$ converge c.s., cuando $n \to \infty$, a una medida de probabilidad aleatoria discreta α^*;

II. α^* corresponde al proceso Dirichlet con parámetro α;

III. Dado α^*, las variables aleatorias $X_1, X_2, \ldots$ son independientes con distribución α^*.

Posteriormente, Pitman (1996) generalizó este modelo a través de los «modelos de muestreo de especies».

Modelo de muestreo de especies

Un problema de «muestreo de especies» se puede modelar como sigue. Supóngase que se tiene una muestra aleatoria $X_1, X_2, \ldots$, con valores en $\mathbb{X}$, de una población grande de individuos de varias especies, de manera que X_i representa la especie del i–ésima individuo muestreado. El espacio $\mathbb{X}$ se puede ver como un conjunto arbitrario de etiquetas que identifiquen a las distintas especies. Sea $M_1 = 1$ y

$$M_j = \inf\left\{n : n > M_{j-1}, X_n \notin \{X_1, \ldots, X_{n-1}\}\right\},$$

para $j > 1$, con la convención $\inf\varnothing = \infty$. Defínase $Y_j = X_{M_j}$ si $M_j < \infty$. Entonces, Y_i representa la i–ésima especie. El número N_{jn} de veces que la j–ésima especie Y_j aparece en la muestra $X_1, \ldots, X_n$ se define como

$$N_{jn} = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = Y_j, M_j < \infty\}}, \qquad j \geq 1.$$

Sea $K_n = \max\{j : N_{jn} > 0\}$ el número de especies diferentes en las primeras n observaciones.

Supóngase que X_1 tiene distribución fija H no atómica. Se llamará «regla de predicción» a la regla que especifica la distribución de X_1 y la distribución condicional de X_{n+1} dado $X_1, \ldots, X_n$, para cada $n \geq 1$. Considérese entonces $\{X_n\}_{n\geq 1}$ sujeta a la regla de predicción de la forma

$$\mathbb{P}[X_1 \in \cdot\,] = H(\cdot), \tag{5}$$

$$\mathbb{P}[X_{n+1} \in \cdot \mid X_1, \ldots, X_n, K_n = k] = \sum_{i=1}^{k} p_i(\mathrm{N}_n)\delta_{Y_i}(\cdot) + p_{k+1}(\mathrm{N}_n)H(\cdot), \tag{6}$$

donde $\mathrm{N}_n = \{N_{1n}, N_{2n}, \ldots\}$ es el vector de conteos de las especies observadas en $(X_1, \ldots, X_n)$ con rango $\mathbb{N}^* = \bigcup_{k=1}^{\infty} \mathbb{N}^k$ y $\{p_n\}_{n\geq 1}$ es una sucesión de «funciones de probabilidad predictoras» definidas en $\mathbb{N}^*$. Las funciones $\{p_n\}_{n\geq 1}$ son tales que, dado que después de n observaciones el vector de conteos es $\mathrm{N}_n = \mathrm{n}$, donde $\mathrm{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$, con $\sum_k n_k = n$, la siguiente observación es la j–ésima especie ya observada con probabilidad $p_j(\mathrm{n})$, para $1 \leq j \leq k$, o es una nueva especie con probabilidad $p_{k+1}(\mathrm{n})$, donde $k = k(\mathrm{n})$ es el número de componentes no cero de n. De aquí se observa que cualquier sucesión de funciones $\{p_n\}_{n\geq 1}$ tal que

$$p_j(\mathrm{n}) \geq 0, \quad \text{y} \quad \sum_{j=1}^{k(\mathrm{n})+1} p_j(\mathrm{n}) = 1, \qquad \mathrm{n} \in \mathbb{N}^*, \tag{7}$$

determina la distribución de una sucesión de variables aleatorias $\{X_n\}_{n\geq 1}$ a través de la regla de predicción (6).

Como una primera aproximación al modelo de Blackwell & MacQueen (1973), Proposición 1, se tiene lo siguiente.

PROPOSICIÓN 2. Supóngase que $\{X_n\}_{n\geq 1}$ es una sucesión intercambiable de variables aleatorias sujeta a la regla de predicción de la forma (5) y (6). Sea μ_n la distribución condicional de X_{n+1} dado $X_1, \ldots, X_n$ como en (6). Entonces

I. μ_n converge en la norma de variación total c.s., conforme $n \to \infty$, a la medida de probabilidad aleatoria

$$\mu = \sum_i \tilde{P}_i \delta_{Y_i} + (1 - \sum_i \tilde{P}_i) H, \tag{8}$$

donde $\tilde{P}_i$ es la frecuencia de la i–ésima especie, i.e.,

$$\tilde{P}_i = \lim_{n\to\infty} \frac{\mathrm{N}_{in}}{n} \text{ c.s.}$$

II. Las Y_i's son independientes con distribución H e independientes de $\tilde{P}_i$.

III. $\{X_1, X_2, \ldots\}$ es una muestra de μ.

Es importante mencionar que el número K_∞ de valores distintos en la sucesión infinita $\{X_1, X_2, \ldots\}$ es casi seguramente igual a $\inf\{k : \tilde{P}_1 + \cdots + \tilde{P}_k = 1\}$. En cuanto al inciso (II), significa que condicionalmente dado $\{\tilde{P}_1, \tilde{P}_2, \ldots\}$ con $K_\infty = k$, las Y_i's son independientes con distribución H, para $1 \leq j \leq k+1$.

Por otro lado, esta proposición resulta deficiente en dos aspectos al compararla con la Proposición 1. En primer lugar, en esta proposición se asume que $\{X_n\}_{n\geq 1}$ es intercambiable, lo cual es parte de la conclusión de la primera; además de que no da una forma explícita de la distribución μ. Esto se remedia a continuación.

DEFINICIÓN 4. Se dice que $\{X_n\}_{n\geq 1}$ es una «muestra de especies» si es una sucesión intercambiable sujeta a la regla de predicción de la forma (5) y (6) para una medida H no atómica.

Como una variación de la Proposición 2, $\{X_n\}_{n\geq 1}$ es una muestra de especies si y sólo si $\{X_n\}_{n\geq 1}$ es una muestra de una distribución aleatoria μ de la forma

$$\mu = \sum_i P_i \delta_{Z_i} + (1 - \sum_i P_i) H, \tag{9}$$

para alguna sucesión de variables aleatorias $\{P_n\}_{n\geq 1}$ tal que

$$P_i \geq 0 \quad \text{y} \quad \sum_i P_i \leq 1 \text{ c.s.},$$

y alguna sucesión $Z_1, Z_2, \ldots$ independientes con distribución H e independientes de $\{P_n\}_{n\geq 1}$.

DEFINICIÓN 5. Un «modelo de muestreo de especies» es un modelo formado de una medida de probabilidad aleatoria μ como en (9) y una muestra $\{X_n\}_{n\geq 1}$ de μ.

En esta definición, P_i corresponde a la frecuencia relativa de la i–ésima especie de alguna lista de especies presentes en una población y Z_i es la etiqueta asignada a tal especie.

Función de probabilidad sobre particiones intercambiables

Sea $[n] = \{1, \dots, n\}$ y $|A|$ el número de elementos en A. Si $\{X_n\}_{n\geq 1}$ es intercambiable, entonces, para cada partición de $[n]$ en k subconjuntos no vacíos $A_1, \dots, A_k$, donde los A_i's se asume que están en orden de aparición, i.e., $1 \in A_1$ y, para $2 \leq j \leq k$, el primer elemento de $[n] \setminus (A_1 \cup \cdots \cup A_{j-1})$ pertenece a A_j, se tiene

$$\mathbb{P}\left[\bigcap_{j=1}^{k}\{X_i = Y_j, \text{ para toda } i \in A_j\}\right] = p(|A_1|, \dots, |A_k|), \tag{10}$$

para alguna función simétrica p de k–tuplas de enteros no negativos con suma n. De manera general, cuando n varía, se define $p : \mathbb{N}^* \to [0,1]$. Esta función p determina la distribución de la partición aleatoria de $\mathbb{N}$, cuyas clases son las clases de equivalencia para la relación de equivalencia aleatoria definida por $i \sim j$ si y sólo si $X_i = X_j$. Se dice, entonces, que p es la «función de probabilidad sobre particiones intercambiables» (FPPI) derivada de una sucesión intercambiable $\{X_n\}_{n\geq 1}$.

Identificando $\mathrm{n} = (n_1, \dots, n_n) \in \mathbb{N}^*$ con la sucesión infinita $(n_1, \dots, n_n, 0, 0, \dots)$, tomando $k = k(\mathrm{n})$ como el número de elementos diferentes de cero de n, definiendo $\mathrm{n}^{j+} \in \mathbb{N}^*$, para la sucesión infinita n, incrementando n_j en uno, para $1 \leq j \leq k(n) + 1$, y de (10), una FPPI debe satisfacer

$$p(1) = 1 \quad \text{y} \quad p(\mathrm{n}) = \sum_{j=1}^{k(\mathrm{n})+1} p(\mathrm{n}^{j+}). \tag{11}$$

Conversamente, toda función p simétrica no negativa definida en $\mathbb{N}^*$ y que satisface (11) es la FPPI de alguna sucesión intercambiable $\{X_n\}_{n\geq 1}$.

PROPOSICIÓN 3. Correspondiente a cada pareja (p, H), donde p es una FPPI y H es una medida de probabilidad no atómica, existe una única medida de probabilidad para la muestra de especies $\{X_n\}_{n\geq 1}$ tal que p es la FPPI de $\{X_n\}_{n\geq 1}$ y H es la distribución de X_1.

Como consecuencia de esta proposición y de la Proposición 2, la medida de probabilidad aleatoria μ que gobierna una muestra de especies $\{X_n\}_{n\geq 1}$ se puede recuperar c.s. de $\{X_n\}_{n\geq 1}$. Así, una pareja (p, H) determina las distribuciones finito dimensionales de una medida de probabilidad aleatoria μ tal que una muestra $\{X_n\}_{n\geq 1}$ de μ tiene FPPI p y cada X_n tiene distribución H.

Regla de predicción

Considérese las funciones p_i que determinan la regla de predicción (6) de una muestra de especies $\{X_n\}_{n\geq 1}$. De (10) y del teorema de Bayes, estas funciones se pueden expresar en términos de la FPPI p de $\{X_n\}_{n\geq 1}$ como

$$p_j(\mathrm{n}) = \frac{p(\mathrm{n}^{j+})}{p(\mathrm{n})}, \qquad 1 \leq j \leq k(\mathrm{n}) + 1, \tag{12}$$

siempre que $p(\mathrm{n}) > 0$. Por lo que ahora, de la Proposición 3, la Proposición 2 se mejora de la siguiente manera.

TEOREMA 1. Dada una medida de probabilidad H no atómica y una sucesión de funciones $\{p_n\}_{n\geq 1}$ definidas en $\mathbb{N}^*$ que satisfacen (7), sea $\{X_n\}_{n\geq 1}$ una sucesión gobernada por la regla de predicción (5) y

(6). La sucesión $\{X_n\}_{n\geq 1}$ es intercambiable si y sólo si existe una función p definida en $\mathbb{N}^*$, simétrica, no negativa, tal que (12) se cumple. Entonces $\{X_n\}_{n\geq 1}$ es una muestra de μ como en la Proposición 2 y la FPPI de $\{X_n\}_{n\geq 1}$ es la función p simétrica no negativa que cumple (12) y $p(1) = 1$.

2.2 Procesos con incrementos independientes normalizados

La construcción de medidas de probabilidad aleatorias vía normalización fue presentada por Regazzini et al. (2003). Según los autores, este enfoque resulta interesante, entre otros aspectos, debido a que representa una aproximación natural al problema de definir una medida de probabilidad aleatoria. La idea detrás de estas medidas es ir de distribuciones deterministas a aleatorias, definidas en $\mathbb{R}$, considerando sus incrementos en intervalos disjuntos como variables aleatorias independientes.

Antes de introducir estas medidas es necesario dar una breve exposición de «procesos de Lévy crecientes» o «subordinadores», ya que se basan en ellos para su construcción. (Para un estudio más detallado se puede consultar, por ejemplo, Sato (1999).)

DEFINICIÓN 6 (SUBORDINADOR). Un proceso estocástico $\{\xi_t\}_{t\geq 0}$ definido en un espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$ es un «subordinador» si

I. $\xi_0 = 0$ c.s.

II. Existe $\Omega_0 \in \mathscr{F}$ con $\mathbb{P}[\Omega_0] = 1$, tal que, para todo $\omega \in \Omega_0$, $\xi_t(\omega)$ es continuo por la derecha en $t \geq 0$ y tiene límites por la izquierda en $t > 0$ (trayectorias *càdlàg*)

III. Para cualquier $n \geq 1$ y $0 \leq t_0 < t_1 < \cdots < t_n$, las variables aleatorias $\xi_{t_0}, \xi_{t_1} - \xi_{t_0}, \ldots, \xi_{t_n} - \xi_{t_{n-1}}$ son independientes (incrementos independientes)

IV. La distribución de $\xi_{s+t} - \xi_s$ no depende de s, i.e., $(\xi_{s+t} - \xi_s) \overset{\mathrm{d}}{=} \xi_t$ para toda $s, t \geq 0$ (incrementos estacionarios)

V. $\lim_{h\to 0} \mathbb{P}[|\xi_{t+h} - \xi_t| \geq \varepsilon] = 0$ para toda $t \geq 0$ y toda $\varepsilon > 0$ (continuidad estocástica)

VI. $\xi_t(\omega)$ es creciente c.s. como función de t

Este tipo de procesos tiene una descomposición única, conocida como «descomposición de Lévy–Itó», dada por la terna (G, d, ν). El término G se conoce como «término gaussiano» y define la varianza del componente gaussiano continuo, d es la «deriva» y es la responsable del desarrollo promedio del proceso, y ν es una medida en $\mathbb{R}$ que satisface

$$\nu(\{0\}) = 0 \quad \text{y} \quad \int_{\mathbb{R}} \min(|x|^2, 1)\nu(\mathrm{d}x) < \infty,$$

llamada «medida de Lévy», la cual exhibe la frecuencia y magnitud de los brincos del proceso.

Por otro lado, su exponente característico también está dado por la terna (G, d, ν).

TEOREMA 2 (REPRESENTACIÓN DE LÉVY–KHINTCHINE). Sea $\{\xi_t\}_{t\geq 0}$ un proceso de Lévy en $\mathbb{R}$ con terna (G, d, ν). Entonces

$$\mathbb{E}\left[e^{-iz\xi_t}\right] = e^{t\psi(z)}, \qquad z \in \mathbb{R},$$

donde ψ, llamado «exponente característico», está dado por

$$\psi(z) = idz - \tfrac{1}{2}Gz^2 + \int_{\mathbb{R}} \left(e^{izx} - 1 - izx\mathbb{1}_{|x|\leq 1}\right)\nu(\mathrm{d}x).$$

Un caso específico de subordinadores son aquellos con terna $(0, 0, t\nu)$, los cuales son procesos crecientes de brincos puros. De esta manera, su representación de Lévy–Kintchine se simplifica: su exponente característico está dado por

$$\psi(z) = -\int_{\mathbb{R}^+} (1 - e^{zx})\nu(\mathrm{d}x),$$

también es más conveniente trabajar con la transformada de Laplace, por lo que se tiene

$$\mathbb{E}\left[e^{-z\xi_t}\right] = \exp\left(-t\int_{\mathbb{R}^+} (1 - e^{-zx})\nu(\mathrm{d}x)\right);$$

asimismo, la condición sobre la medida de Lévy se simplifica a

$$\nu(\{0\}) = 0 \quad \text{y} \quad \int_{\mathbb{R}^+} \min(x, 1)\nu(\mathrm{d}x) < \infty.$$

Para la construcción de procesos con incrementos independientes normalizados se trabajará con subordinadores en $\mathbb{R}$ con terna $(0, 0, t\nu)$ y que cumplan con $\nu(\mathbb{R}^+) = \infty$, esto último permitirá que la normalización esté bien definida. Específicamente se trabajará con los siguientes subordinadores.

EJEMPLO 3 (PROCESO GAMMA). Sea $\{\xi_t\}_{t\geq 0}$ un subordinador con brincos Gamma, i.e., para toda $t \in \mathbb{R}^+$, $\xi_t \sim \mathcal{G}a(\xi_t \mid t, \beta)$. La función de densidad de la distribución Gamma con parámetros (t, β) está dada por

$$f(x; t, \beta) = \frac{\beta^t}{\Gamma(t)} x^{t-1} e^{-\beta x}, \qquad x \geq 0.$$

Entonces, su transformada de Laplace es

$$\mathbb{E}\left[e^{z\xi_t}\right] = \left(\frac{\beta + z}{\beta}\right)^{-t} = \exp\left(-t\int_0^\infty (1 - e^{-zx})x^{-1}e^{-\beta x}\mathrm{d}x\right).$$

De aquí se puede observar que

$$\nu(\mathrm{d}x) = x^{-1}e^{-\beta x}\mathrm{d}x,$$

y su exponente característico es

$$\psi(z) = \log\left(\frac{\beta + z}{\beta}\right).$$

Además, se tiene que $\nu(0+) = \infty$ y por tanto $\nu(\mathbb{R}^+) = \infty$.

EJEMPLO 4 (PROCESO σ–ESTABLE). Sea $\{\xi_t\}_{t\geq 0}$ un subordinador con brincos σ–estable, $\sigma \in (0, 1)$. El subordinador con brincos σ–estable se puede definir como

1. $\{\xi_t\}_{t\geq 0}$ es un proceso de Lévy;

II. para cada $a > 0$, $\{\xi_{at}, t \geq 0\} \overset{\mathrm{d}}{=} \{a^{1/\sigma}\xi_t, t \geq 0\}$;

(Kingman 1975). En general, para este proceso no existe una forma analítica de su función de densidad.

La medida de Lévy de una distribución σ–estable es

$$\nu(\mathrm{d}x) = cx^{-(1+\sigma)}\mathrm{d}x, \quad c > 0.$$

De esto se sigue que su transformada de Laplace es

$$\mathbb{E}\left[e^{z\xi_t}\right] = \exp\left(-t\frac{c\Gamma(1-\sigma)z^{\sigma}}{\sigma}\right),$$

y su exponente característico es

$$\psi(z) = \frac{c\Gamma(1-\sigma)z^{\sigma}}{\sigma}.$$

Por último, al igual que el proceso Gamma, se tiene que $\nu(0+) = \infty$ y por tanto $\nu(\mathbb{R}^+) = \infty$.

NORMALIZACIÓN DE SUBORDINADORES

Una vez introducido el concepto de subordinador se pueden definir los procesos con incrementos independientes normalizados.

DEFINICIÓN 7. Sea $\{\xi_t\}_{t\geq 0}$ un subordinador $(0, 0, t\nu)$ con medida de Lévy ν sobre $\mathbb{R}^+$ tal que satisface $\nu(\mathbb{R}^+) = \infty$, sea α una medida finita, no nula, sobre $\mathbb{R}$ con masa total $a := \alpha(\mathbb{R})$ y sea $t = \alpha(-\infty, x]$. Se dice que P es un «proceso con incrementos independientes normalizado» con parámetros (α, ν), denotado como $\mathcal{NP}(\alpha, \nu)$, si

$$P(\cdot) = \frac{\xi_{\alpha(\cdot)}}{\xi_a}.$$

Se puede observar que la operación de normalización ocasiona que el proceso $\mathcal{NP}(\alpha, \nu)$ pierda la propiedad de incrementos independientes. Sin embargo, como se muestra a continuación, algunos momentos son fácilmente deducidos (James, Lijoi & Prünster 2006).

PROPOSICIÓN 4. Sean A, B conjuntos medibles. Si P se distribuye $\mathcal{NP}(\alpha, \nu)$ con masa total a, entonces

I.
$$\mathbb{E}[P(B)] = \frac{\alpha(B)}{a};$$

II.
$$\mathrm{Var}[P(B)] = \frac{\alpha(B)(a - \alpha(B))}{a^2} I_a;$$

III.
$$\mathrm{Cov}[P(A), P(B)] = \frac{a\alpha(A \cap B) - \alpha(A)\alpha(B)}{a^2} I_a;$$

donde $I_a = a \int_{\mathbb{R}^+} ue^{-a\psi(u)} \int_{\mathbb{R}^+} x^2 e^{-ux}\nu(\mathrm{d}x)\mathrm{d}u$.

Es importante mencionar que estas medidas de probabilidad aleatorias, de la misma manera que el proceso Dirichlet, su valor esperado sólo depende de α. Para fines prácticos se hará $H(\cdot) := \alpha(\cdot)/a$.

Otra propiedad importante de estos procesos es su distribución predictiva.

PROPOSICIÓN 5. Sea $X_1, \ldots, X_n$ una muestra de P. Si P se distribuye $\mathcal{NP}(\alpha, v)$ con masa total a, entonces

I.
$$\mathbb{P}[X_2 \in B \mid X_1] = (1 - I_a)H(B) + I_a\delta_{X_1}(B);$$

II.
$$\mathbb{P}[X_{n+1} \in B \mid X_1, \ldots, X_n] = w_n H(B) + \frac{1}{n}\sum_{j=1}^{k} w_{nj}\delta_{X_j^*}(B);$$

donde I_a es como en la proposición anterior, $X_1^*, \ldots, X_k^*$ denotan los k valores distintos de la muestra $X_1, \ldots, X_n$ y w_n y w_{nj} son pesos determinados por la medida de Lévy de P.

Las expresiones para el cálculo de w_n y w_{nj} se pueden consultar en James et al. (2006, Corolario 1). Además, se puede observar que esta distribución predictiva corresponde a la regla de predicción del esquema de urnas de Pólya, donde la FPPI está dada por w_n y w_{nj} (Ecuaciones (5) y (6)).

Para terminar esta sección se dan dos ejemplos de procesos con incrementos independientes normalizados; se dan las expresiones para su valor esperado, varianza, covarianza y distribuciones predictivas.

EJEMPLO 3, CONTINUACIÓN (PROCESO GAMMA NORMALIZADO). Del proceso Gamma se tiene

I.
$$\mathbb{E}[P(B)] = H(B);$$

II.
$$\mathrm{Var}[P(B)] = \frac{H(B)(1 - H(B))}{(1 + a)};$$

III.
$$\mathrm{Cov}[P(B_1), P(B_2)] = \frac{H(B_1 \cap B_2) - H(B_1)H(B_2)}{(1 + a)};$$

IV.
$$\mathbb{P}[X_2 \in B \mid X_1] = \frac{a}{(1 + a)}H(B) + \frac{1}{(1 + a)}\delta_{X_1}(B)$$

y
$$\mathbb{P}[X_{n+1} \in B \mid X_1, \ldots, X_n] = \frac{a}{(n + a)}H(B) + \frac{1}{(n + a)}\sum_{j=1}^{k} n_j\delta_{X_j^*}(B).$$

Este proceso corresponde al proceso Dirichlet y es la definición alternativa dada por Ferguson (1973, Sección 4).

EJEMPLO 4, CONTINUACIÓN (PROCESO σ–ESTABLE NORMALIZADO). Del proceso σ–estable se tiene

I.
$$\mathbb{E}[P(B)] = H(B);$$

II.

$$\text{Var}[P(B)] = H(B)(1 - H(B))(1 - \sigma);$$

III.

$$\text{Cov}[P(B_1), P(B_2)] = (H(B_1 \cap B_2) - H(B_1)H(B_2))(1 - \sigma);$$

IV.

$$\mathbb{P}[X_2 \in B \mid X_1] = \sigma H(B) + (1 - \sigma)\delta_{X_1}(B)$$

y

$$\mathbb{P}[X_{n+1} \in B \mid X_1, \ldots, X_n] = \frac{\sigma k}{n} H(B) + \frac{1}{n} \sum_{j=1}^{k} (n_j - \sigma)\delta_{X_j^*}(B),$$

(James et al. 2006).

2.3 Proceso Poisson–Dirichlet de dos parámetros

La distribución Poisson–Dirichlet de dos parámetros fue presentado por Pitman & Yor (1997). Es una generalización de la distribución Poisson–Dirichlet propuesto por Kingman (1975), el cual surge del estudio de distribuciones asintóticas de frecuencias relativas aleatorias.

Definición 8. Para $0 \leq \sigma < 1$ y $\theta > -\sigma$, sea $V_1, V_2, \ldots$ una sucesión de variables aleatorias independientes tal que $V_k \sim \mathcal{B}e(V \mid 1 - \sigma, \theta + k\sigma)$. Sea

$$U_1 = V_1 \quad \text{y} \quad U_i = V_i \prod_{j=1}^{i-1} (1 - V_j), \quad i \geq 2.$$

Sea $\text{P}_{\sigma,\theta} = (P_1, P_2, \ldots)$ que denota a $(U_1, U_2, \ldots)$ ordenados de manera decreciente. Entonces, la distribución de $\text{P}_{\sigma,\theta}$ se conoce como «Poisson–Dirichlet de dos parámetros», denotada por $PD(\sigma, \theta)$.

Utilizando esta distribución, se define el proceso Poisson-Dirichlet de dos parámetros.

Definición 9. Sea $Z_1, Z_2, \ldots$ una sucesión de variables aleatorias con distribución común H, sea $\text{P}_{\sigma,\theta} \sim PD(\sigma, \theta)$ y sea

$$\tilde{P}(\cdot) = \sum_{i=1}^{\infty} P_i \delta_{Z_i}(\cdot).$$

Entonces, la medida de probabilidad aleatoria $\tilde{P}$ se conoce como «proceso Poisson–Dirichlet de dos parámetros», denotado por $\mathcal{P}\text{D}(\sigma, \theta)$.

Es importante mencionar que las variables aleatorias $P_1, P_2, \ldots$ suman uno c.s.

Por otro lado, una característica de este proceso es que la distribución predictiva para una muestra $\{X_n\}_{n \geq 1}$ se puede caracterizar a través del esquema de urnas de Pólya (Pitman 1996, Ejemplo 16).

PROPOSICIÓN 6. Sea $\{X_i\}_{i=1}^n$ una muestra del proceso $\mathcal{PD}(\sigma,\theta)$ y supóngase que X_1 tiene distribución H no atómica. La distribución predictiva para el proceso $\mathcal{PD}(\sigma,\theta)$ es

$$\mathbb{P}[X_i \in \cdot \mid X_1,\ldots,X_{i-1}] = \frac{\theta+\sigma k}{\theta+i-1}H(\cdot) + \sum_{j=1}^{k}\frac{n_j^*-\sigma}{\theta+i-1}\delta_{X_j^*}(\cdot), \quad i=2,\ldots,n, \tag{13}$$

donde $\{X_1^*,\ldots,X_k^*\}$ es el conjunto de los k valores únicos (o distintos) de $\{X_1,\ldots,X_{i-1}\}$, cada uno con frecuencia n_j^*, $j=1,\ldots,k$.

Las medidas canónicas de este proceso son las siguientes.

EJEMPLO 5. El proceso $\mathcal{PD}(0,\theta)$ corresponde al proceso Dirichlet (Ejemplo 3) con parámetro θ. En este caso, su distribución predictiva está dada por

$$\mathbb{P}[X_n \in \cdot \mid X_1,\ldots,X_{n-1}] = \frac{\theta}{\theta+n-1}H(\cdot) + \frac{1}{\theta+n-1}\sum_{j=1}^{k}\delta_{X_j^*}(\cdot).$$

EJEMPLO 6. El proceso $\mathcal{PD}(\sigma,0)$ corresponde al proceso σ–estable normalizado (Ejemplo 4) y su distribución predictiva está dada por

$$\mathbb{P}[X_n \in \cdot \mid X_1,\ldots,X_{n-1}] = \frac{\sigma k}{n-1}H(\cdot) + \frac{1}{n-1}\sum_{j=1}^{k}(n_j^*-\sigma)\delta_{X_j^*}(\cdot),$$

2.4 REPRESENTACIÓN *STICK–BREAKING*

Otra forma de construir medidas de probabilidad aleatorias discretas c.s. es mediante la representación conocida como «*stick–breaking*» (Ishwaran & James 2001). Una medida de probabilidad aleatoria *stick–breaking* se define como sigue.

DEFINICIÓN 10. Sean $1 \le N \le \infty$, $\mathrm{a} = (a_1, a_2, \ldots)$, $\mathrm{b} = (b_1, b_2, \ldots)$, con $a_i, b_i > 0$, y $\{p_i\}_{i=1}^N$ una sucesión de variables aleatorias, tales que $0 \le p_i \le 1$ y $\sum_{i=1}^N p_i = 1$ c.s. Una medida de probabilidad aleatoria P_N es «*stick–breaking*» con parámetros (a,b), denotada como $\mathcal{SB}_N(\mathrm{a},\mathrm{b})$, si es de la forma

$$P_N(\cdot) = \sum_{i=1}^{N} p_i \delta_{Z_i}(\cdot), \tag{14}$$

donde $Z_i \overset{\text{iid}}{\sim} H$, con H una distribución no atómica y los p_i's son pesos aleatorios, independientes de las Z_i's, tales que

$$p_1 = V_1 \quad \text{y} \quad p_i = V_i\prod_{j=1}^{i-1}(1-V_j), \quad i \ge 2 \tag{15}$$

con $V_i \overset{\text{ind}}{\sim} \mathcal{B}e(V_i \mid a_i, b_i)$.

Esta construcción, de acuerdo a los autores, permite representar varias medidas de probabilidad aleatorias que parecieran no tener relación alguna, entre ellos se encuentra el proceso Dirichlet y el proceso

Poisson–Dirichlet de dos parámetros, vistos anteriormente, así como el proceso Dirichlet multinomial, el proceso Dirichlet finito dimensional y el proceso Beta de dos parámetros, los cuales pueden consultarse en Ishwaran & James (2001) y en las referencias allí incluidas.

El proceso Poisson–Dirichlet con parámetros $\mathcal{P}_D(\sigma, \theta)$ puede representarse por medio de la representación *stick–breaking* haciendo $a_i = 1 - \sigma$ y $b_i = \theta + i\sigma$ para algunas $0 \leq \sigma < 1$ y $\theta > -\sigma$.

Por otro lado, este tipo de medidas de probabilidad aleatorias de dividen de acuerdo al valor de N, si es finito o infinito.

Caso $N < \infty$, distribuciones finito dimensionales

En este caso se tienen $\mathrm{a} = (a_1, \ldots, a_{N-1})$ y $\mathrm{b} = (b_1, \ldots, b_{N-1})$ y es necesario hacer $V_N = 1$ en (15) para que la medida $\mathcal{SB}_N(\mathrm{a}, \mathrm{b})$ esté bien definida. Haciendo esto se garantiza que $\sum_{i=1}^{N} p_i = 1$ c.s., debido a que

$$1 - \sum_{i=1}^{N-1} p_i = (1 - V_1)\cdots(1 - V_{N-1}).$$

Pesos aleatorios Dirichlet generalizados

Considérese los pesos aleatorios $\mathrm{p} = (p_1, \ldots, p_N)$ definidos como en (15). De acuerdo con Ishwaran & James (2001), estos tienen una distribución Dirichlet generalizada. Esta distribución se denotará por $\mathcal{G}_D(\mathrm{a}, \mathrm{b})$, donde $\mathrm{a} = (a_1, \ldots, a_{N-1})$, $\mathrm{b} = (b_1, \ldots, b_{N-1})$ y su densidad está dada por

$$\left(\prod_{k=1}^{N-1} \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)}\right) p_1^{a_1-1}\cdots p_{N-1}^{a_{N-1}-1} p_N^{b_{N-1}-1}(1 - P_1)^{b_1-(a_2+b_2)}\cdots(1 - P_{N-2})^{b_{N-2}-(a_{N-1}+b_{N-1})},$$

donde $P_k = \sum_{i=1}^{k} p_k$. De esto se concluye que todas las medidas de probabilidad aleatorias basadas en pesos aleatorios Dirichlet son medidas $\mathcal{SB}_N(\mathrm{a}, \mathrm{b})$. Más precisamente, sea P una medida *stick–breaking* con pesos aleatorios p, donde

$$\mathrm{p} = (p_1, \ldots, p_N) \sim \mathcal{D}ir(a_1, \ldots, a_N),$$

entonces $\mathrm{p} \sim \mathcal{G}_D(\mathrm{a}, \mathrm{b})$, con $\mathrm{a} = (a_1, \ldots, a_{N-1})$ y $\mathrm{b} = (\sum_{k=2}^{N} a_k, \sum_{k=3}^{N} a_k, \ldots, a_N)$. Por tanto, P es una medida $\mathcal{SB}_N(\mathrm{a}, \mathrm{b})$. Un caso especial de estas medidas es la «distribución Dirichlet finito dimensional», denotada por $\mathcal{DP}_N(\theta H)$, que se obtiene al hacer $a = (\theta/N, \ldots, \theta/N)$ para alguna $\theta > 0$.

Caso $N = \infty$, distribuciones infinito dimensionales

Este tipo de medidas está bien definida si sus pesos suman uno casi seguramente. Para verificar esta condición se puede utilizar el Lema 1 de Ishwaran & James (2001): para los pesos aleatorios de $P_\infty(\mathrm{a}, \mathrm{b})$, $\sum_{i=1}^{\infty} p_i = 1$ c.s. si y sólo si $\sum_{k=1}^{\infty} \mathbb{E}[\log(1 - V_i)] = -\infty$; alternativamente es suficiente probar que $\sum_{k=1}^{\infty} \log(1 - a_i/b_i) = -\infty$.

Estas distribuciones incluyen el proceso Dirichlet, su extensión de dos parámetros, el proceso Poisson–Dirichlet de dos parámetros y el proceso Beta de dos parámetros.

TRUNCAMIENTO CASI SEGURO DE MEDIDAS $S_{B_\infty}(a, b)$

Una clase útil de medidas $S_{B_N}(a, b)$ se obtiene al truncar una medida $S_{B_\infty}(a, b)$ descartando los términos $N+1, N+2, \ldots$ y haciendo $p_N = 1 - p_1 - \cdots - p_{N-1}$, que equivale a hacer $V_N = 1$ en (15).

La determinación del nivel adecuado de truncamiento se puede hacer con base en los momentos de los pesos aleatorios; para ello se tiene la siguiente proposición.

PROPOSICIÓN 7. Sean $\{p_n\}_{n\geq 1}$ los pesos aleatorios de una medida $\mathcal{P}_D(\sigma, \theta)$ dada. Para cada $N \geq 1$ y cada $r \geq 1$, enteros positivos, sean

$$T_N(r, \sigma, \theta) = \left(\sum_{k=N}^{\infty} p_i\right)^r \quad \text{y} \quad U_N(r, \sigma, \theta) = \sum_{k=N}^{\infty} p_i^r.$$

Entonces

$$\mathbb{E}[T_N(r, \sigma, \theta)] = \prod_{k=1}^{N-1} \frac{(\theta + k\sigma)_{r\uparrow}}{(\theta + (k-1)\sigma + 1)_{r\uparrow}}, \qquad N \geq 2,$$

y

$$\mathbb{E}[U_N(r, \sigma, \theta)] = \mathbb{E}[T_N(r, \sigma, \theta)] \frac{(1-\sigma)_{r-1\uparrow}}{(\theta + (N-1)\sigma + 1)_{r-1\uparrow}},$$

donde $(b)_{r\uparrow} := b(b+1)\cdots(b+r-1)$ denota el símbolo de Pochhammer, con la convención $(b)_{0\uparrow} = 1$.

Si se utiliza una medida $S_{B_N}(a, b)$ como distribución inicial en un modelo bayesiano no paramétrico, Ishwaran & James (2001) proponen elegir N de manera que la densidad marginal final sea prácticamente indistinguible de su límite. Supóngase que $X = \{X_i\}_{i=1}^n$ es una muestra del modelo bayesiano jerárquico

$$\begin{aligned} X_i \mid Y_i &\overset{\text{ind}}{\sim} F(X_i \mid Y_i), \quad i = 1, \ldots, n \\ Y_i \mid P &\overset{\text{iid}}{\sim} P \\ P &\overset{\text{iid}}{\sim} S_{B_N}(P \mid a, b); \end{aligned}$$

entonces, la densidad marginal final bajo el truncamiento $P_N = S_{B_N}(a, b)$ es

$$\mu_N = \int \left(\prod_{i=1}^{n} \int f(X_i; Y_i) P(\mathrm{d}Y_i) \right) P_N(\mathrm{d}P),$$

donde f es la función de densidad de F. De esta forma, la densidad μ_N debe estar cercana a su límite μ_∞ bajo la medida $S_{B_\infty}(a, b)$. Para elegir N puede utilizarse la siguiente proposición.

PROPOSICIÓN 8. Sean $\{p_n\}_{n\geq 1}$ los pesos aleatorios de una medida $S_{B_\infty}(a, b)$ dada. Si $\|\cdot\|_1$ denota la distancia $\mathcal{L}_1$, entonces

$$\|\mu_N - \mu_\infty\|_1 \leq 4 \left\{ 1 - \mathbb{E}\left[\left(\sum_{i=1}^{N-1} p_i \right)^n \right] \right\}, \tag{16}$$

donde n es el tamaño de la muestra. En particular, para el proceso $\mathcal{P}_D(\sigma, \theta)$ se tiene

$$\|\mu_N - \mu_\infty\|_1 \leq 4 \left\{ 1 - \mathbb{E}\left[1 - T_N(1, \sigma, \theta) \right]^n \right\}$$

y para el proceso Dirichlet con parámetro θ

$$\|\mu_N - \mu_\infty\|_1 \sim 4ne^{-(N-1)/\theta}.$$

2.5 Mezcla de medidas de probabilidad aleatorias

Las medidas de probabilidad aleatorias estudiadas anteriormente son discretas casi seguramente. Esto puede ser un inconveniente cuando se trabaja con observaciones modeladas a través de variables aleatorias continuas. En este aspecto, Lo (1984) propuso una generalización de medidas de probabilidad aleatorias mediante el uso de una «mezcla de distribuciones». De esta manera es posible obtener medidas de probabilidad aleatorias continuas.

Supóngase que $\mu(x \mid y)$ es una medida de probabilidad y $\{\pi_i\}_{i\geq 1}$ una sucesión de números positivos, tales que $\sum_{i\geq 1} \pi_i = 1$, entonces la medida de probabilidad

$$F(x) = \sum_{i\geq 1} \pi_i \mu(x \mid y_i),$$

se conoce como «mezcla de medidas de probabilidad». A μ se le conoce como «kernel».

Una generalización esta mezcla se obtiene al suponer que los pesos π_i's y los parámetros y_i's están relacionados de manera que, para alguna variable aleatoria discreta Y, $\mathbb{P}[Y = y_i] = \pi_i$, $i \geq 1$. Además, es posible suponer que la distribución de Y es aleatoria, obteniendo así «mezclas de medidas de probabilidad aleatorias» y si el kernel es absolutamente continuo, la mezcla será, además, continua.

DEFINICIÓN 11. Sea $\{X_i\}_{i=1}^n$ una sucesión de variables aleatorias condicionalmente independientes tales que $(X_i \mid Y_i)$ tienen distribución μ con parámetro Y_i. Si la distribución de Y_i es incierta y modelada como una medida de probabilidad aleatoria discreta casi seguramente, entonces se dice que la sucesión $\{X_i\}_{i=1}^n$ proviene de una «mezcla de medida de probabilidad aleatoria».

En el contexto bayesiano, las mezclas de medidas de probabilidad aleatorias se pueden representar esquemáticamente como un modelo jerárquico

$$\begin{aligned} X_i \mid Y_i &\overset{\text{ind}}{\sim} \mu(X_i \mid Y_i), \qquad i = 1, \ldots, n \\ Y_i \mid P &\overset{\text{iid}}{\sim} P \\ P &\sim \Pi, \end{aligned}$$

donde Π puede sustituirse por alguna de las medidas de probabilidad aleatorias estudiadas anteriormente. Estos modelos se conocen como «modelos semiparamétricos».

Distribución del número de grupos en la mezcla

Una variable de interés al trabajar con mezclas de medidas de probabilidad aleatorias es el número de grupos de la mezcla, o número de componentes. Este número se determina por el número de valores distintos $\{Y_i^*\}_{i=1}^k$ de la sucesión $\{Y_i\}_{i=1}^n$ en la Definición 11 y conocer su distribución permitiría saber la probabilidad de obtener un nuevo valor Y_{k+1}^*.

Sea K_n la variable aleatoria que denota el número de valores distintos en una muestra de tamaño n. Al elegir un proceso normalizado, se induce una distribución sobre K_n, y obtener $\mathbb{P}[K_n = k]$ es de utilidad para la especificación inicial de la medida de probabilidad aleatoria.

EJEMPLO 7. Si $\{Y_i\}_{i=1}^n$ es una muestra del proceso Dirichlet con parámetro α y masa total a, se tiene que

$$\mathbb{P}[Y_{n+1} \notin \{Y_1^*, \ldots, Y_k^*\} \mid Y_1, \ldots, Y_n] = \frac{a}{n+a}.$$

Se tiene, entonces, que la probabilidad de obtener un nuevo valor es positiva. Antoniak (1974) encontró la distribución de K_n, la cual está dada por

$$\mathbb{P}[K_n = k] = \frac{c_{n,k} a^k}{(a)_{n\uparrow}},$$

donde $c_{n,k}$ es el valor absoluto del número de Stirling de primera clase y $(a)_{n\uparrow}$ es el símbolo de Pochhammer. Además, se tiene que

$$\mathbb{E}[K_n] = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}, \tag{17}$$

(Pitman 2002).

EJEMPLO 8. Si $\{Y_i\}_{i=1}^n$ es una muestra del proceso σ–estable normalizado, se tiene que

$$\mathbb{P}[Y_{n+1} \notin \{Y_1^*, \ldots, Y_k^*\} \mid Y_1, \ldots, Y_n] = \frac{k\sigma}{n}.$$

De igual forma, la probabilidad de obtener un nuevo valor es positiva. Lijoi, Mena & Prünster (2007*b*) encuentran la distribución de K_n, la cual está dada por

$$\mathbb{P}[K_n = k] = \frac{1}{\sigma k (n-1)!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_{n\uparrow}.$$

Asimismo se tiene que

$$\mathbb{E}[K_n] = \frac{(\sigma)_{n\uparrow}}{\sigma (n-1)!}, \tag{18}$$

(Pitman 2002).

Distribución del número de especies

De manera similar a las mezclas de procesos con incrementos independientes normalizados, es de interés conocer el número de distintas especies en una muestra $\{X_i\}_{i=1}^n$ en una mezcla del proceso $\mathcal{P}_D(\sigma, \theta)$. Lijoi, Mena & Prünster (2007*a*) estudian este problema. Sea K_n la variable aleatoria que denota el número de especies en una muestra de tamaño n y $\mathrm{N}_{K,n} = (N_1, N_2, \ldots, N_{K_n})$ las frecuencias de las K_n especies. Se tiene entonces, que la distribución conjunta de $(K_n, \mathrm{N}_{K,n})$ está dada por

$$\mathbb{P}[K_n = k, N_{K,n} = n_{k,n}] = \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta+1)_{n-1\uparrow}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1\uparrow}. \tag{19}$$

La distribución del número de especies coincide con

$$\mathbb{P}[K_n = k] = \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{\sigma^k k! (\theta+1)_{n-1\uparrow}} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_{n\uparrow};$$

además, Pitman (2002) obtiene que

$$\mathbb{E}[K_n] = \frac{(\theta+\sigma)_{n\uparrow}}{\sigma(\theta+1)_{n-1\uparrow}} - \frac{\theta}{\sigma}. \tag{20}$$

En la Figura 1, página 26, se muestran las gráficas de la densidad de K_n para cada uno de las tres medidas de probabilidad aleatorias para distintos parámetros de α, σ y (σ, θ). Se puede observar el comportamiento de cada una: el proceso Dirichlet tiene densidades leptocúrticas, a diferencia del σ–estable normalizado que, por lo general, son platicúrticas. Conocer esto es útil en el contexto bayesiano ya que la elección de un proceso u otro refleja la certeza que se tenga sobre el número de grupos. Por otro lado, el proceso Poisson–Dirichlet, al contar con dos parámetros, ofrece mayor flexibilidad al incorporar la información.

2.6 Proceso geométrico

Las mezclas de medidas de probabilidad aleatorias estudiadas en la sección anterior se pueden utilizar en los problemas de estimación de densidades. Otro enfoque que se ha desarrollado para estos problemas consiste en modelos de mezclas basados en un número aleatorio, pero finito, de componentes; en términos de las funciones de densidad

$$f(x \mid N) = \sum_{l=1}^{N} p_l K(x; \theta_l). \tag{21}$$

Basados en estos modelos, Fuentes-García et al. (2010) proponen el siguiente

$$f(x \mid N) = N^{-1} \sum_{l=1}^{N} K(x; \theta_l), \tag{22}$$

donde N es aleatoria, pero con la misma distribución para cada observación, i.e., $\mathbb{P}[N] = q_N$ y se le asigna una distribución inicial a $\{q_N\}$. La elección de $\{q_N\}$ determina los pesos $\{p_l\}$ en (21). Entonces, los autores trabajan con unos pesos que tienen forma similar a una distribución geométrica y obtienen una nueva medida que es sorprendentemente simple y eficiente para realizar estimaciones. A continuación se explica el desarrollo es esta nueva medida de probabilidad aleatoria.

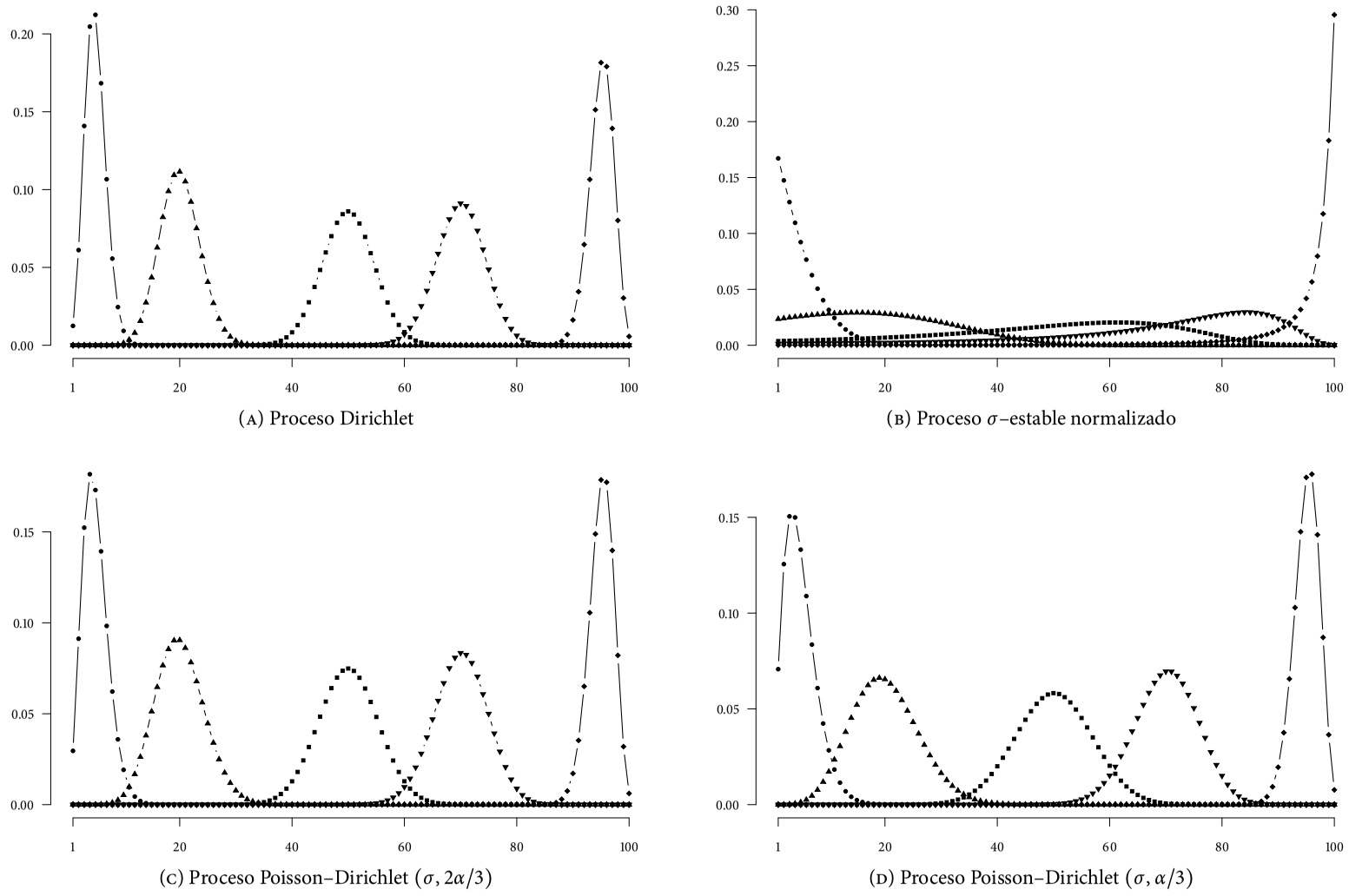
Análisis del modelo

Marginalizando (22) con respecto a N, se obtiene

$$f(x) = \sum_{N=1}^{\infty} \frac{1}{N} \sum_{l=1}^{N} K(x; \theta_l) q_N,$$

que puede reescribirse como

$$f(x) = \sum_{l=1}^{\infty} p_l K(x; \theta_l) = \int K(x; \theta) P(\mathrm{d}\theta),$$



(A) Proceso Dirichlet

(B) Proceso σ–estable normalizado

(C) Proceso Poisson–Dirichlet $(\sigma, 2\alpha/3)$

(D) Proceso Poisson–Dirichlet $(\sigma, \alpha/3)$

FIGURA 1: Densidad de K_n para los distintos procesos con $n = 100$; los puntos se unieron para facilitar su visualización. Las densidades con el mismo tipo de punto corresponden a un mismo valor de $\mathbb{E}[K_n]$: 5 (●), 20 (▲), 50 (■), 70 (▼) y 95 (◆). En el proceso Poisson–Dirichlet, el parámetro α corresponde al utilizado en el proceso Dirichlet.

donde

$$P(\mathrm{d}\theta) = \sum_{l=1}^{\infty} p_l \delta_{\theta_l}(\mathrm{d}\theta)$$

y los pesos $\{p_l\}$ están dados por

$$p_l = \sum_{N=l}^{\infty} q_N / N.$$

Estos pesos suman uno y son decrecientes, lo cual simplifica la complejidad de su distribución.

En cuanto a la distribución de $\{q_N\}$, existen diversas opciones, entre ellas se encuentran las distribuciones Poisson o geométrica; pero antes, es importante estudiar la distribución condicional de N para cada observación, denotada por N_i. Si d_i dada N_i es la indicadora de la componente de la cual proviene x_i, entonces

$$\mathbb{P}[d_i = l \mid N_i] = N_i^{-1} \mathbb{1}_{\{1,\dots,N_l\}}(l);$$

además se tiene que $\mathbb{P}[N_i = l] = q_l$, por lo que

$$\mathbb{P}[N_i = N \mid d_i] \propto q_N / N \mathbb{1}_{\{N \geq d_i\}}.$$

Por tanto, para fines de simulación, es conveniente que $\{q_N/N\}$ sea una sucesión de probabilidades para la cual sea fácil simular versiones truncadas de su distribución. Con este fin, los autores eligieron $q_N \propto N(1-\lambda)^{N-1}$, por lo que $P[N_i \mid d_i]$ es una distribución geométrica truncada. Además, es recomendable asignarle a λ una distribución inicial Beta.

$$P = \sum_{i=1}^{\infty} w_i \delta_{\phi_i}.$$

De ésta se tiene que $w_i = v_i \prod_{l<i}(1 - v_l)$, con $v_i \sim \mathcal{B}e(v_i \mid 1, a)$, por lo que $\mathbb{E}[v_i] = (1+a)^{-1}$, para $a > 0$. Entonces, haciendo $\lambda = (1+a)^{-1}$, $\mathbb{E}[w_l] = \lambda(1-\lambda)^{l-1}$, lo cual corresponde a una distribución geométrica. Además, es común asignarle una distribución inicial a a. Por tanto, este nuevo modelo elimina un nivel de jerarquía en el modelo de mezcla de procesos Dirichlet, sustituyendo $\{v_l\}$ por sus esperanzas.

Entonces, es posible reescribir el proceso Dirichlet ordenando sus pesos de manera decreciente, i.e.,

$$P = \sum_{i=1}^{\infty} \rho_i \delta_{\phi_i}, \tag{23}$$

donde $\rho_1 > \rho_2 > \cdots$ y los ϕ_i's son independientes con alguna distribución común. Por otro lado, utilizando los «pesos geométricos», se tiene la medida de probabilidad aleatoria

$$P_G = \sum_{i=1}^{\infty} w_i(\lambda) \delta_{\theta_i}, \tag{24}$$

donde los θ_i's son independientes e idénticamente distribuidos con la misma distribución que los ϕ_i's, los $w_i(\lambda)$'s son los pesos geométricos y λ tiene alguna distribución.

De esta manera es posible aproximar el proceso Dirichlet (23) con (24) cuando, para alguna sucesión $n_1, n_2, \ldots$, se tiene que

$$|w_1(\lambda) + \cdots + w_{n_1}(\lambda) - \rho_1|, |w_{n_1+1}(\lambda) + \cdots + w_{n_2}(\lambda) - \rho_2|, \ldots,$$

son todas relativamente pequeñas. Igualmente se tiene que

$$\max\{|\theta_1 - \phi_1|, \ldots, |\theta_{n_1} - \phi_1|\}, \max\{|\theta_{n_1+1} - \phi_2|, \ldots, |\theta_{n_2} - \phi_2|\}, \ldots,$$

son todos también relativamente pequeños. Por lo que se puede concluir que la medida de probabilidad aleatoria

$$P_G(\cdot) = \lambda \sum_{i=1}^{\infty} (1-\lambda)^{i-1} \delta_{\theta_i}(\cdot),$$

permite para modelar el proceso Dirichlet, pero con la ventaja de que no requiere de pesos tan elaborados como la representación *stick–breaking* estudiada con anterioridad.

– II –

Métodos para estimación de densidades

Los modelos bayesianos estudiados en el capítulo anterior proveen un mecanismo robusto para hacer inferencias. Sin embargo, en la práctica surgieron varios problemas que limitaron su implementación. Si se quería un modelo que describiera de manera precisa el fenómeno bajo estudio, se corría el riesgo de que el cálculo de las distribuciones posteriores fuera imposible hacerlo analíticamente, además de que la implementación de una solución computacional podría ser muy costosa. Por otro lado, simplificar un modelo puede implicar inferencias erróneas.

Sin embargo, actualmente se cuenta con métodos y tecnologías computacionales potentes y accesibles que han permitido implementar modelos complejos. Dos áreas de estudio de la computación a las que se recurre frecuentemente en la estadística bayesiana son el análisis numérico y la simulación estocástica.

Entre los principales problemas que se encuentran en la estadística bayesiana, que también son objeto de estudio en el análisis numérico, son los problemas de optimización y de integración. Los problemas de optimización surgen, por ejemplo, al encontrar la decisión óptima en un problema de toma de decisiones en ambiente de incertidumbre, mientras que los problemas de integración pueden surgir al momento de calcular las distribuciones posteriores. En este trabajo se estudiará únicamente el problema de integración.

El problema de integración, como se comentó, ha sido ampliamente estudiado dentro del análisis numérico; sin embargo, muchos de los métodos desarrollados son deficientes cuando se trabaja en dimensiones grandes. Uno de los métodos que no sufre de esta limitación es el método Monte Carlo. Este método resuelve el problema de integración a través de una perspectiva probabilística y utiliza métodos de simulación estocástica.

El método Monte Carlo permite resolver problemas relativamente sencillos. Sin embargo, ha ayudado en el desarrollo de métodos más potentes, por ejemplo, aquellos que utilizan cadenas de Markov, los conocidos métodos Monte Carlo vía cadenas de Markov (MCMC); dos ejemplos de estos son: el conocido método Metropolis–Hastings, utilizado en primera instancia en la física, y un caso particular de éste, el *Gibbs sampler*.

Dentro del contexto bayesiano no paramétrico (o semiparamétrico), existen modelos para estimar densidades que utilizan mezclas de distribuciones aleatorias como los estudiados en las últimas secciones del capítulo anterior. Sus respectivos algoritmos hacen uso de los métodos MCMC.

En este capítulo se da una introducción a los métodos MCMC, específicamente al Metropolis–Hastings

y *Gibbs sampler*. Además se estudian los métodos de estimación de densidades para modelos bayesianos no paramétricos, específicamente los métodos desarrollados en Ishwaran & James (2001), Neal (2000) y Fuentes-García et al. (2010).

§ 1 Métodos de simulación

Uno de los objetivos de la simulación estocástica es desarrollar algoritmos que permitan obtener muestras aleatorias de «cualquier» función de distribución. Existen diversos métodos que permiten hacerlo, sin embargo, la mayoría se basan en muestras aleatorias con distribución uniforme en $(0,1)$. Una vez que se tiene la muestra uniforme, por medio de algún método, e.g., transformación, es posible obtener muestras de la distribución deseada. Algunos de los métodos utilizados para este fin son: método de inversión, método de aceptación y rechazo, muestreo por importancia y método de rechazo adaptativo.

Obtener un «buen» generador de números aleatorios, sin embargo, es más complicado. Una de las razones es que se busca un mecanismo o método que genere sucesiones de números que no tengan ningún patrón, pero además es necesario que se pueda generar la misma sucesión más de una vez. Una alternativa dentro del cómputo científico es generar números «pseudo–aleatorios» a través de los «generadores de números pseudo–aleatorios», los cuales son algoritmos que dado un valor inicial, conocido como «semilla», producen una sucesión de números aleatorios $\{u_i\}_{i=1}^n$ que toman valores en $(0,1)$. La sucesión de números aleatorios debe tener las mismas propiedades relevantes de una sucesión de variables aleatorias uniformes en $(0,1)$ independientes. Esto básicamente significa que los números aleatorios generados se deben comportar como si fueran variables aleatorias uniformes independientes en $(0,1)$, lo cual se puede probar aplicando diversas pruebas estadísticas. Existe una serie de pruebas, llamada *diehard*, desarrollada por George Marsaglia[1], que permite saber si un generador se puede considerar «bueno».

Para mayor referencia sobre generadores de números aleatorios y métodos de simulación puede consultarse, por ejemplo, Robert & Casella (2005).

En las siguientes secciones se estudia una manera de resolver el problema de integración, el cual se encuentra frecuentemente en la estadística bayesiana. Por ejemplo, en la Sección 1.1.2 se explicó cómo obtener el valor esperado *a posteriori* del parámetro θ

$$\mathbb{E}[\theta \mid x] = \int_\Theta \theta p(\theta \mid x) \mathrm{d}\theta;$$

en ocasiones, la integral no puede calcularse analíticamente, por lo que se necesita utilizar algún método de integración numérica; sin embargo, en la mayoría de los casos, ni siquiera es posible obtener una forma «útil» de $p(\theta \mid x)$, lo que imposibilita la aplicación de los métodos «convencionales» de integración numérica. Los métodos Monte Carlo y MCMC son una solución, respectivamente, a este tipo de problemas.

[1] El conjunto de pruebas se puede obtener en `http://stat.fsu.edu/pub/diehard/`; también existe una versión en C, llamada *dieharder*, la cual ha sido implementada a su vez en R.

1.1 Método Monte Carlo

Supóngase que se quiere calcular la integral

$$I = \int_A f(x)\mathrm{d}x, \qquad A \subset \mathbb{R}^d \tag{1}$$

con $d \geq 1$, la cual no es posible obtener analíticamente. Una forma de obtenerla es numéricamente. Sin embargo, es importante mencionar que si $A = \mathbb{R}^d$, antes de aplicar cualquier método numérico es necesario verificar que la integral I existe y es finita.

Ejemplo 1. En el caso unidimensional, un método de integración es el «método de punto medio». Si $A = [a, b]$, sea $\{[a_i, a_{i+1}]\}_{i=0}^{n-1}$ una partición de A, entonces

$$\hat{I} = \sum_{i=0}^{n-1} (a_{i+1} - a_i) f\left(\tfrac{a_{i+1}-a_i}{2}\right);$$

comúnmente los elementos de la partición tienen la misma longitud para facilitar aún más el calculo. Sin embargo, la extensión a dimensiones $d \geq 2$ de este y otros métodos de integración se dificulta en gran manera.

Una alternativa para este problema es ver la integral en (1) como un valor esperado. Para esto se necesita encontrar una variable aleatoria Y con soporte A y distribución π, con densidad p, y una función g tales que

$$\mathbb{E}[g(Y)] = \int_A f(y)\mathrm{d}y = \int_A g(y)p(y)\mathrm{d}y = I. \tag{2}$$

Este método se conoce como «método Monte Carlo». Una de sus ventajas es que puede extenderse fácilmente al caso multidimensional.

Utilizando la Ley de los Grandes Números, la integral en (2) se puede estimar por medio una muestra aleatoria $y_1, \ldots, y_n$ de tamaño n con distribución π como

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} g(y_i).$$

De esta forma, se tiene que

$$\mathbb{E}[\hat{I}] = I \quad \text{y} \quad \operatorname{Var}[\hat{I}] = \frac{1}{n} \int_A \left(f(y) - I\right)^2 p(y)\mathrm{d}y,$$

donde esta última se puede estimar por

$$v_n = \frac{1}{n^2} \sum_{i=1}^{n} \left(f(y_i) - \hat{I}\right)^2,$$

la cual sirve para medir la precisión de $\hat{I}$.

Además, es posible utilizar el Teorema Central del Límite para obtener intervalos de probabilidad, ya que

$$\frac{\hat{I} - \mathbb{E}[g(Y)]}{\sqrt{v_n}} \Longrightarrow \mathcal{N}(0,1),$$

cuando $n \to \infty$; donde $\Longrightarrow$ denota convergencia débil.

EJEMPLO 1, CONTINUACIÓN. Retomando el caso unidimensional, $A = [a, b]$, Y se puede tomar con distribución uniforme en $[a, b]$ y $g(y) = (b - a)f(y)$, entonces

$$I = (b - a)\,\mathbb{E}[f(Y)];$$

utilizando el método Monte Carlo, I se estima a través de una muestra aleatoria $y_1, \ldots, y_n$, con $y_i \sim \mathcal{U}(y_i \mid a, b)$, como

$$\hat{I} = \frac{(b - a)}{n} \sum_{i=1}^{n} f(y_i).$$

1.2 Monte Carlo vía cadenas de Markov

Una extensión al método Monte Carlo que se utiliza cuando no es posible simular directamente de π, son los métodos conocidos como «Monte Carlo vía cadenas de Markov» (MCMC). La idea de estos métodos es construir una cadena de Markov con distribución estacionaria π.

Para comprender el funcionamiento de los métodos MCMC es necesario tener presentes algunos conceptos sobre cadenas de Markov.

En la siguiente definición y teoremas se presentan de manera resumida algunos de ellos; para mayor referencia puede consultarse, por ejemplo, Robert & Casella (2005).

DEFINICIÓN 1. Sea $\xi = \{\xi_t\}_{t \geq 0}$ una sucesión de variables aleatorias definidas en algún espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$ con valores en $(\mathbb{X}, \mathscr{X})$ y sea $E \in \mathscr{X}$ su conjunto de estados.

I. Un «kernel de transición» es una función K definida en $(\mathbb{X}, \mathscr{X})$ tal que $K(\cdot, B$ es medible para todo $B \subset \mathscr{X}$ y $K(x, \cdot)$ es una medida de probabilidad para toda $x \in \mathbb{X}$. Similarmente, si se denota $K^{(1)}(x, B) = K(x, B)$, el «kernel de n transiciones» está dado por

$$K^{(n)}(x, B) = \int K^{(n-1)}(y, B) K(x, \mathrm{d}y).$$

II. Dado un kernel de transición K, para $seq[n]B \subset E$, se dice que ξ es una «cadena de Markov» si

$$\mathbb{P}[\xi_n \in B_n \mid \xi_0 \in B_0, \ldots, \xi_{n-1} \in B_{n-1}] = \mathbb{P}[\xi_n \in B_n \mid \xi_{n-1} \in B_{n-1}].$$

Si además, la distribución de $(\xi_{t_1}, \xi_{t_2}, \ldots, \xi_{t_k})$ dado $\xi_{t_0} = x_{t_0}$ es la misma que la distribución de $(\xi_{t_1 - t_0}, \xi_{t_2 - t_0}, \ldots, \xi_{t_k - t_0})$ dado $\xi_{t_0} = x_{t_0}$, para toda k y todos $t_0 \leq t_1 \leq \cdots \leq t_k$, se dice que la cadena es «homogénea en el tiempo».

III. Una cadena de Markov es «ϕ–irreducible» para alguna medida ϕ definida en $\mathscr{X}$ si para todo $B \in \mathscr{X}$ tal que $\phi(B) > 0$ y para todo $x \in \mathscr{X}$

$$\mathbb{P}[\tau_B < \infty \mid \xi_0 = x] > 0,$$

donde $\tau_B = \inf\{n \geq 1 : \xi_n \in B\}$ es el «tiempo de primera llegada» a B.

IV. Un conjunto C es «pequeño» si existe $m \in \mathbb{N}$ y una medida no nula ν_m tal que

$$K^{(m)}(x, B) \geq \nu_m(B),$$

para todo $x \in C$ y todo $B \in \mathscr{X}$. Una cadena ξ phi–irreducible tiene un «ciclo de tamaño d» si existe un conjunto pequeño C, un entero M y una distribución ν_M, tal que

$$d = \text{mcd}\{m \geq 1 : \exists \delta_m > 0 \text{ tal que } C \text{ es pequeño para } \nu_m \geq \delta_m \nu_M\}.$$

El «periodo» de ξ es el mayor entero d que satisface lo anterior y ξ es aperiódica si $d = 1$.

V. Un conjunto B es «Harris recurrente» si

$$\mathbb{P}\left[\sum_{n=1}^{\infty} \mathbb{1}_B(\xi_n) = \infty \,\middle|\, \xi_0 = x\right] = 1.$$

para todo $x \in B$. La cadena ξ es «Harris recurrente» si existe una medida ϕ tal que ξ es ϕ–irreducible y para cada B con $\phi(B) > 0$, B es Harris recurrente.

VI. Se dice que π es una «distribución estacionaria» de la cadena si cumple con

$$\pi(B) = \int K(x, B)\pi(\mathrm{d}x),$$

para todo $B \in \mathscr{X}$.

VII. Se dice que una cadena de Markov es «reversible», si para alguna función f se cumple

$$K(x, y)f(x) = K(y, x)f(y).$$

Si además f es función de densidad, se tiene que

$$\int K(y, B)f(y)\mathrm{d}y = \pi(B),$$

esto es

$$\pi(B) = \int K(y, B)\pi(\mathrm{d}y),$$

es decir, π es la distribución estacionaria de la cadena.

Los siguientes teoremas son las versiones equivalentes a la Ley de los Grandes Números y al Teorema Central de Límite, respectivamente, para cadenas de Markov.

TEOREMA 1 (TEOREMA ERGÓDICO). Sea $\{\xi_t\}_{t\geq 0}$ una cadena de Markov Harris recurrente con distribución estacionaria π y sea g una función medible tal que $\mathbb{E}_\pi[|g|] < \infty$, entonces

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(\xi_i) = \int g(x)\pi(\mathrm{d}x), \quad \text{c.s.}$$

TEOREMA 2. Sea $\{\xi_t\}_{t\geq 0}$ una cadena de Markov Harris recurrente y reversible con distribución estacionaria π y g como en el teorema anterior, entonces

$$\frac{1}{n}\sum_{i=1}^{n}\big[g(\xi_i) - \mathbb{E}_\pi[g]\big] \Longrightarrow \mathcal{N}(0,\gamma^2),$$

donde $0 < \gamma^2 = \mathbb{E}_\pi\left[\bar{g}^2(\xi_0)\right] + 2\sum_{i=1}^{\infty}\mathbb{E}_\pi\left[\bar{g}(\xi_0)\bar{g}(\xi_i)\right] < \infty$, con $\bar{g} = g - \mathbb{E}_\pi[g]$ y $\Longrightarrow$ denota convergencia débil.

Utilizando lo anterior, si se desea utilizar algún método MCMC se debe construir una cadena de Markov con kernel de transición K que converja a π, sea aperiódica, Harris recurrente y reversible. Con estas propiedades se garantiza que la cadena pueda alcanzar cualquier conjunto sin importar el estado donde se inicie, además de que se garantiza la existencia y unicidad de la distribución estacionaria. Asimismo, será posible estimar la integral en (1) como en el método Monte Carlo, al generar observaciones del kernel de transición, y calcular intervalos de probabilidad, gracias a los Teoremas 1 y 2, respectivamente.

Sin embargo, al implementar estos métodos, no siempre es posible verificar la convergencia de la cadena debido a la complejidad del modelo. Así que es necesario realizar algunas pruebas para saber si la cadena alcanzó su estado estacionario, de lo contrario, los estimadores que se obtengan no serán correctos. Existen varias sugerencias para solucionar este problema. Una de ellas, la más utilizada, es dejar correr la cadena cierto número de iteraciones, durante las cuales la cadena llega a su estado estacionario; después de este periodo, que se conoce como «periodo de calentamiento», se espera que las observaciones obtenidas provengan de la distribución de interés y por tanto son las utilizadas para las estimaciones. También se puede verficar la convergencia corriendo varias veces la cadena, cada una con distinto valor inicial; si hubo convergencia, el efecto del valor inicial debe desaparecer y por tanto, las estimaciones deben ser las mismas en todas las corridas.

Otro aspecto importante de mencionar es que, por la construcción del método, las observaciones no son independientes. Pero si se desea una muestra que se comporte como si las observaciones fueran independientes, existen dos propuestas para obtenerlo. La primera de ellas consiste en que una vez que la cadena convergió, las observaciones se tomen cada cierto número de iteraciones, con lo cual se espera que disminuya su correlación. La otra opción consiste en correr varias cadenas simultáneamente de manera independiente y una vez que han convergido se toma una muestra de cada una, de manera que las muestras en conjunto sean del tamaño deseado. (Debido a que la no correlación no implica independencia, se mencionó que las observaciones se comportarían «como si fueran» independientes.)

Junto con estas pruebas, es recomendable hacer algunas gráficas para monitorear el comportamiento de la cadena. Por ejemplo, si se grafican los promedios ergódicos y la cadena alcanzó su límite, estos deben aparecer como una línea recta; y si se desea ver la correlación de la muestra se puede hacer un correlograma.

Dentro de la estadística bayesiana, los métodos MCMC son muy utilizados, debido a que la distribución final, en la mayoría de los casos, es difícil de obtener analíticamente o no es posible simular direc-

tamente de ella. Existen dos métodos que son los más utilizados en estos casos: el primero, propuesto por Metropolis et al. (1953) para la física mecánica y generalizado por Hastings (1970), se conoce como Metropolis–Hastings, y el segundo, conocido como *Gibbs sampler*, es un caso particular de Metropolis–Hastings propuesto por Geman & Geman (1984). Ambos métodos se explican a continuación.

Metropolis–Hastings

Supóngase que se quiere simular de una distribución con densidad p, definida excepto por una constante, y supóngase que se tiene una densidad $q(y \mid x)$. Defínase la siguiente cadena de Markov, cuya transición de x_n a x_{n+1} se define como

I. generar $y \sim q(y \mid x_n)$;

II. definir

$$x_{n+1} = \begin{cases} y & \text{con probabilidad } \alpha(x_n, y) \\ x_n & \text{con probabilidad } 1 - \alpha(x_n, y), \end{cases}$$

donde

$$\alpha(x, y) = \min\left\{1, \frac{p(y)q(x \mid y)}{p(x)q(y \mid x)}\right\},$$

con la convención de que $\alpha(x, y) = 0$ si $p(y) = p(x) = 0$.

El kernel de transición de la cadena está dado de la siguiente manera. Estando en x_n sólo se puede pasar a x_{n+1} de dos formas

I. $x_{n+1} = y$; lo que significa que se genera y y se acepta, por lo que

$$K(x_n, y) = q(y \mid x_n)\alpha(x_n, y);$$

II. $x_{n+1} = x_n$; en este caso se rechaza todo y generado, por tanto

$$K(x_n, x_{n+1}) = \left[1 - \int q(y \mid x_n)\alpha(x_n, y)\mathrm{d}y\right] \mathbb{1}_{\{x_n\}}(x_{n+1}).$$

Entonces

$$K(x, y) = q(y \mid x)\alpha(x, y) + \left[1 - \int q(t \mid x)\alpha(x, t)\mathrm{d}t\right] \mathbb{1}_{\{x\}}(y).$$

Para probar que $K(x, y)$ es reversible, sólo es necesario trabajar con $q(y \mid x)\alpha(x, y)$. De acuerdo a la definición de $\alpha(x, y)$, se tienen dos casos

I. si $\alpha(y, x) = 1$,

$$\alpha(x, y) = \frac{p(y)q(x \mid y)}{p(x)q(y \mid x)},$$

por lo que $\alpha(x, y)p(x)q(y \mid x) = p(y)q(x \mid y)$. Entonces

$$\alpha(x, y)q(y \mid x)p(x) = \alpha(y, x)q(x \mid y)p(y);$$

II. si $\alpha(x, y) = 1$, de forma análoga

$$\alpha(y, x) = \frac{p(x)q(y \mid x)}{p(y)q(x \mid y)},$$

por lo que $\alpha(y, x)p(y)q(x \mid y) = p(x)q(y \mid x)$. Entonces

$$\alpha(y, x)q(x \mid y)p(y) = \alpha(x, y)q(y \mid x)p(x).$$

En ambos casos se tiene que $\alpha(x, y)q(y \mid x)$ es reversible respecto a p.

Por otro lado, se necesitan condiciones sobre q para asegurar la convergencia a p. Una condición suficiente para que la cadena sea aperiódica es que

$$\mathbb{P}\big[q(y_n \mid x_n)p(x_n) \leq q(x_n \mid y_n)p(y_n)\big] < 1.$$

Esto implica que los eventos $[X_{n-1} = X_n]$ tienen probabilidad positiva, por lo que la cadena tiene probabilidad positiva de permanecer en X_n.

La condición de ϕ–irreductibilidad se cumple si se asegura que $0 < q(x \mid y) < \infty$ para toda pareja (x, y), ya que esto significa que cualquier conjunto puede ser alcanzado en un paso. Esto implica que la cadena es Harris recurrente. Un estudio más detallado de estas condiciones se puede encontrar en Tierney (1994).

Por último, es necesario elegir una $q(x \mid y)$ adecuada. Esta elección varía de acuerdo al problema específico a tratar, aunque existen tres propuestas que son las más comunes

- Independencia. Se supone que $q(y \mid x) = q(y)$, lo que implica que la nueva variable es generada independientemente de x.
- Caminata aleatoria. La idea aquí es tomar en cuenta el valor que se generó en la etapa n para la generación del valor en la etapa $n + 1$.
- Simetría. Esta es la propuesta original de Metropolis et al. (1953) y consiste en suponer que $q(y|x) = q(x \mid y)$.

Para terminar con esta sección, se presenta el algoritmo del método

1. Inicializar $x^{(0)}$ y hacer $k = 0$
2. Generar $y \sim q(y \mid x^{(k)})$
3. Definir
$$\alpha(x^{(k)}, y) = \min\left\{1, \frac{p(y)q(x^{(k)} \mid y)}{p(x^{(k)})q(y \mid x^{(k)})}\right\}$$
4. Generar $u \sim \mathcal{U}(u \mid 0, 1)$
5. Si $\alpha(x^{(k)}, y) \leq u$, hacer $x^{(k+1)} = y$, de lo contrario hacer $x^{(k+1)} = x^{(k)}$
6. Hacer $k = k + 1$ y regresar a 2 hasta obtener convergencia

Gibbs sampler

Un caso particular del método Metropolis–Hastings que se utiliza constantemente en estadística bayesiana es el *Gibbs sampler*.

DEFINICIÓN 2. Sea $x \in \mathbb{R}^d$, $d > 1$, con función de densidad conjunta $p(x)$. Si x_{-i} denota al vector x eliminando la i–ésima componente, se define la «densidad condicional completa» de x_i dado x_{-i} como

$$p(x_i \mid x_{-i}) = \frac{p(x_i, x_{-i})}{p(x_{-i})} = \frac{p(x)}{\int p(x)\mathrm{d}x_i}.$$

Supóngase que las densidades condicionales completas son todas conocidas y sus distribuciones correspondientes son fáciles de simular. Haciendo $q(x_i \mid x) = p(x_i \mid x_{-i})$ en el método Metropolis–Hastings se tiene

$$\alpha(x_i, x) = \min\left\{1, \frac{p(x_{-i})p(x_i \mid x_{-i})}{p(x_i)p(x_{-i} \mid x_i)}\right\} = 1,$$

por lo que los valores candidatos se eligen con probabilidad uno.

El kernel de transición es función de las densidades condicionales completas $\{p(x_i \mid x_{-i})\}_{i=1}^d$ y está dado por

$$K(x, y) = p(y_1 \mid x_2, \ldots, x_d)p(y_2 \mid y_1, x_3, \ldots, x_d)\cdots p(y_d \mid y_1, \ldots, x_{d-1}), \qquad x, y \in \mathbb{R}^d.$$

Entonces

$$\int K(x, B)p(x)\mathrm{d}x = \int_B p(x)\mathrm{d}x, \qquad x \in \mathbb{R}^d,$$

es decir, p es la distribución estacionaria de la cadena.

Si el integrando $K(x, y)$ es positivo, se puede demostrar que la cadena es ϕ–irreducible, aperiódica y Harris recurrente (Tierney 1994).

Por último, se tiene que el algoritmo de este método es el siguiente

1. Inicializar $x_1^{(0)}, \ldots, x_d^{(0)}$ y $k = 1$
2. Obtener $x_1^{(k)}, \ldots, x_d^{(k)}$ a partir de $x^{(k-1)}$ generando

$$\begin{aligned}
x_1^{(k)} &\sim p(x_1 \mid x_2^{(k-1)}, x_3^{(k-1)}, \ldots, x_d^{(k-1)}) \\
x_2^{(k)} &\sim p(x_2 \mid x_1^{(k)}, x_3^{(k-1)}, \ldots, x_d^{(k-1)}) \\
x_3^{(k)} &\sim p(x_3 \mid x_1^{(k)}, x_2^{(k)}, \ldots, x_d^{(k-1)}) \\
&\vdots \\
x_{d-1}^{(k)} &\sim p(x_{d-1} \mid x_1^{(k)}, x_2^{(k)}, \ldots, x_d^{(k-1)}) \\
x_d^{(k)} &\sim p(x_d \mid x_1^{(k)}, x_2^{(k)}, \ldots, x_{d-1}^{(k)})
\end{aligned}$$

3. Hacer $k = k + 1$ y repetir 2 hasta obtener convergencia

Una generalización de este método consiste en introducir variables latentes. Supóngase que existe una densidad $f(x, y)$ tal que al marginalizar sobre y se obtiene $p(x)$ y que f es más fácil de simular que p. Entonces se puede aplicar el *Gibbs sampler* utilizando la densidad f en lugar de p. Las observaciones de y pueden no tener ningún significado para el modelo original, sin embargo, mejoran el funcionamiento del método.

§ 2 Métodos de estimación de densidades para modelos no paramétricos

En las siguientes secciones se estudian los métodos de estimación de densidades para modelos bayesianos no paramétricos desarrollados por Ishwaran & James (2001), Neal (2000) y Fuentes-García et al. (2010).

2.1 *Gibbs sampler* para urnas de Pólya

Supóngase que se tiene una muestra $X = \{X_1, \ldots, X_n\}$ obtenida del siguiente modelo semiparamétrico

$$\begin{aligned} X_i \mid Y_i, \phi &\overset{\text{ind}}{\sim} \mu(X_i \mid Y_i, \phi), \qquad i = 1, \ldots, n \\ Y_i \mid P &\overset{\text{iid}}{\sim} P \\ \phi &\sim \nu(\phi) \\ P &\sim \Pi. \end{aligned} \tag{3}$$

donde Π es una distribución aleatoria. Ishwaran & James (2001) suponen que Π corresponde al proceso Poisson–Dirichlet de dos paramétros, pero se puede extender a cualquier medida de probabilidad aleatoria discreta casi seguramente que tenga una regla de predicción explícita. Por tanto, utilizando las medidas de probabilidad aleatorias del Capítulo I, este método se puede utilizar con distribuciones que se puedan caracterizar a través del esquema de urnas de Pólya.

Integrando sobre P se obtiene

$$\begin{aligned} X_i \mid Y_i, \phi &\overset{\text{ind}}{\sim} \mu(X_i \mid Y_i, \phi), \qquad i = 1, \ldots, n \\ Y_1, \ldots, Y_n &\sim \pi(Y_1, \ldots, Y_n) \\ \phi &\sim \nu(\phi). \end{aligned} \tag{4}$$

El método propuesto por Ishwaran & James (2001, Sección 4) se basa en el *Gibbs sampler* y utiliza el hecho de que las Y_i's son intercambiables y, por tanto, sólo es necesario obtener la distribución condicional completa para una de ellas.

Sea Y_{-i} el vector $Y = \{Y_1, \ldots, Y_n\}$ eliminando la i–ésima entrada. Recordando, la distribución predicitva para $Y = \{Y_i\}_{i=1}^n$ una muestra del proceso $\mathcal{P}\text{D}(\sigma, \theta)$ está dada por

$$\mathbb{P}[Y_i \in \cdot \mid Y_{-i}] = \frac{\theta + \sigma m}{\theta + n - 1} H(\cdot) + \sum_{j=1}^{m} \frac{n_j^* - \sigma}{\theta + n - 1} \delta_{Y_j^*}(\cdot),$$

donde $\{Y_1^*, \ldots, Y_m^*\}$ es el conjunto de valores únicos de Y_{-i}, cada uno con frecuencia n_j^* para $j = 1, \ldots, m$ y H es la distribución, no atómica, de Y_1 (Teorema 1.6).

Entonces, para obtener observaciones de la distribución posterior $\pi(Y, \phi \mid X)$ se necesita simular valores de $(Y_i \mid Y_{-i}, \phi, X)$, $i = 1, \ldots, n$, y de $(\phi \mid Y, X)$ iterativamente. En particular, cada iteración del *Gibbs sampler* genera las siguientes observaciones

I. $(Y_i \mid Y_{-i}, \phi, X)$, $i = 1, \ldots, n$. La distribución condicional está definida por

$$\mathbb{P}[Y_i \in \cdot \mid Y_{-i}, \phi, X] = q_0 \, \mathbb{P}[Y_i \in \cdot \mid \phi, X_i] + \sum_{j=1}^{m} q_j \delta_{Y_j^*}(\cdot),$$

donde

$$q_0 \propto \frac{\theta + \sigma m}{\theta + n - 1} \int_{\mathbb{Y}} f(X_i; y, \phi) H(\mathrm{d}y) \quad \text{y} \quad q_j \propto \frac{n_j^* - \sigma}{\theta + n - 1} f(X_i; Y_j^*, \phi), \tag{5}$$

y f es la densidad correspondiente a μ. Además, los pesos q_i, $i = 0, \ldots, m$, se deben normalizar para que sumen uno.

II. $(\phi \mid Y, X)$. Aplicando el teorema de Bayes, se obtiene la densidad

$$g(\phi \mid Y, X) \propto \nu(\mathrm{d}\phi) \prod_{i=1}^{n} f(X_i; Y_i, \phi).$$

Aceleración de la mezcla

Un inconveniente en este método se tiene cuando los valores q_j se vuelven más grandes que q_0. Cuando esto ocurre, la simulación puede quedarse estancada en los valores únicos $Y_1^*, \ldots, Y_m^*$ de Y y pueden pasar muchas iteraciones antes de que un nuevo valor Y' sea generado. Una propuesta para solucionar esto consiste en remuestrar los valores únicos $Y_1^*, \ldots, Y_m^*$ al final de cada iteración.

En este «paso de aceleración», Y se reexpresa en términos de sus valores únicos Y^* y un vector de membrecías $C = \{C_1, \ldots, C_n\}$, el cual se define como $C_i = j$ si y sólo si $Y_i = Y_j^*$. De esta manera, al final de cada iteración, después de generar Y se calcula C y entonces se actualizan los valores únicos Y_j^* dado C. Esto equivale a agregar el siguiente paso

III. Obtener una muestra de las distribuciones posteriores para $(Y_j^* \mid C, \phi, X)$ para $j = 1, \ldots, m$. En particular, para cada j, la densidad posterior requerida es

$$h(Y_j^* \mid C, \phi, X) \propto H(\mathrm{d}Y_j^*) \prod_{\{i : C_i = j\}} f(X_i; Y_j^*, \phi).$$

Después, actualizar Y utilizando Y^* y C.

Cabe mencionar que este método para urnas de Pólya es equivalente a los Algoritmos 1 al 3 de Neal (2000) desarrollados para el proceso Dirichlet.

LIMITACIONES DEL MÉTODO

A pesar de la versatilidad del método, éste tiene algunas limitaciones. En primer lugar, para aplicar el *Gibbs sampler* se necesitan las distribuciones condicionales completas y la actualización se realiza coordenada por coordenada, lo cual produce coeficientes q_j muy grandes y como consecuencia la cadena de Markov del modelo converge lentamente, aún cuando se aplique el paso de aceleración.

Otra limitación se tiene cuando H y μ no son conjugadas, lo cual dificulta el cálculo de q_0 y de las distribuciones necesarias para el método de aceleración.

Por último, marginalizar sobre P permite hacer inferencias de la distribución final de P sólo a través de los valores finales de Y.

2.2 *GIBBS SAMPLER* POR BLOQUES

Una manera de tratar las limitaciones del método anterior, de acuerdo a Ishwaran & James (2001), es utilizar el *Gibbs sampler* por bloques. Este método se puede utilizar para modelos no paramétricos y semiparamétricos como (3) tomando Π como una distribución *stick–breaking* $\mathcal{SB}_N(\mathrm{a}, \mathrm{b})$ truncada en N. De esta manera, el modelo quedará totalmente determinado por un número finito de componentes y como consecuencia, será posible actualizar bloques de parámetros en cada iteración a través de distribuciones multivariadas.

El modelo (3) queda redefinido, entonces, como

$$\begin{aligned} X_i \mid Z, K, \phi &\overset{\text{ind}}{\sim} \mu(X_i \mid Z_{K_i}, \phi), \qquad i = 1, \ldots, n \\ K_i \mid p &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} p_k \delta_k(\cdot) \\ Z_k &\overset{\text{iid}}{\sim} H(Z_k), \qquad k = 1, \ldots, N \\ p &\sim \mathcal{GD}(p \mid a, b) \\ \phi &\sim \nu(\phi), \end{aligned} \tag{6}$$

donde $K = (K_1, \ldots, K_n)$ y $p = (p_1, \ldots, p_N)$.

De este nuevo modelo es posible obtener observaciones directamente de la distribución posterior $\Pi(\cdot \mid X)$. El método obtiene iterativamente observaciones de las distribuciones condicionales

$$\begin{aligned} &(Z \mid K, \phi, X) \\ &(K \mid Z, p, \phi, X) \\ &(p \mid K) \\ &(\phi \mid Z, K, X) \end{aligned}$$

Haciendo esto, eventualmente (cuando la cadena alcance su límite) se obtendrán observaciones de la

distribución posterior de $(Z, K, p, \phi \mid X)$ y como cada (Z, K, p, ϕ) define una distribución

$$P(\cdot) = \sum_{k=1}^{N} p_k \delta_{Z_k}(\cdot),$$

también se obtendrán observaciones de la distribución posterior $\Pi(\cdot \mid X)$. La distribución posterior $\pi(Y \mid X)$ puede calcularse haciendo $Y_i = Z_{K_i}$.

Si $K^* = \{K_1^*, \ldots, K_m^*\}$ denota al conjunto de los m valores únicos de K, entonces, cada iteración del *Gibbs sampler* por bloques obtiene observaciones en el siguiente orden

I. $(Z \mid K, \phi, X)$. Obtener $(Z_{K_j^*} \mid K, \phi, X)$ de la densidad

$$h(Z_{K_j^*} \mid K, \phi, X) \propto H(\mathrm{d}Z_{K_j^*}) \prod_{\{i:K_i = K_j^*\}} f(X_i; Z_{K_j^*}, \phi), \qquad j = 1, \ldots, m$$

y simular $Z_k \overset{\text{iid}}{\sim} H$ para $k \in K \setminus K^*$.

II. $(K \mid Z, p, \phi, X)$. Obtener observaciones de la siguiente manera

$$K_i \mid Z, p, \phi, X \overset{\text{ind}}{\sim} \sum_{k=1}^{N} p_{k,i} \delta_k(\cdot), \qquad i = 1, \ldots, n,$$

donde

$$(p_{1,i}, \ldots, p_{N,i}) \propto (p_1 f(X_i; Z_1, \phi), \ldots, p_N f(X_i; Z_N, \phi)).$$

III. $(p \mid K)$. De la defínción de p (Ecuación I.15) se obtiene

$$p_1 = V_1 \quad \text{y} \quad p_k = V_k \prod_{i=1}^{k-1} (1 - V_i), \qquad k = 2, \ldots, N-1,$$

donde

$$V_k \overset{\text{ind}}{\sim} \mathcal{B}e\Big(V_k \Big| a_k + M_k, b_k + \textstyle\sum_{i=k+1}^{N} M_i\Big), \qquad k = 1, \ldots, N-1$$

y M_k es el número de K_i iguales a k, y

$$p_N = 1 - \sum_{k=1}^{N-1} p_k.$$

IV. $(\phi \mid Z, K, X)$. Como en el método anterior

$$g(\phi \mid Z, K, X) \propto \nu(\mathrm{d}\phi) \prod_{i=1}^{n} f(X_i; Z_{K_i}, \phi).$$

Gracias a la posibilidad de actualizar las variables Z, K y p por bloque, el método posee buenas propiedades de mezcla. Además, el mejoramiento de este método con respecto al anterior se logró al introducir variables latentes, permitiendo así introducir la distribución inicial Π.

Considérese, por ejemplo, la medida $\mathcal{DP}_N(\alpha H)$ (página 21). Utilizando el *Gibbs sampler* para urnas de Pólya, el ajuste en (3) involucra obtener observaciones de (Y, ϕ) de la densidad proporcional a

$$\prod_{i=1}^{n} f(X_i; Y_i, \phi) \prod_{i=1}^{n} \left[\frac{\alpha(1-m/N)}{\alpha+i-1} H(\mathrm{d}Y_i) + \sum_{j=1}^{m} \frac{n^* + \alpha N}{\alpha+i-1} \delta_{Y_j^*}(\mathrm{d}Y_i) \right] \nu(\mathrm{d}\phi)$$
$$= \prod_{i=1}^{n} f(X_i; Y_i, \phi) \left[\iint \prod_{i=1}^{n} P(\mathrm{d}Y_i) \Pi_{\alpha\xi_N}(\mathrm{d}P) H^N(\mathrm{d}Z) \right] \nu(\mathrm{d}\phi),$$

donde $\Pi_{\alpha\xi_N}$ corresponde al proceso Dirichlet con medida finita $\alpha\xi_N(Z, \cdot)$ para un valor fijo Z.

Si ahora se incluyen la variable latente Z y la medida P, se obtienen observaciones (Y, ϕ, Z, P) de la densidad proporcional a

$$\prod_{i=1}^{n} f(X_i; Y_i, \phi) \prod_{i=1}^{n} P(\mathrm{d}Y_i) \Pi_{\alpha\xi_N}(\mathrm{d}P) H^N(\mathrm{d}Z) \nu(\mathrm{d}\phi)$$
$$= \prod_{i=1}^{n} f(X_i; Y_i, \phi) \prod_{i=1}^{n} \left[\sum_{k=1}^{N} p_k \delta_{Z_k}(\mathrm{d}Y_i) \right] \pi(\mathrm{d}p) H^N(\mathrm{d}Z) \nu(\mathrm{d}\phi),$$

donde $\pi(\mathrm{d}p)$ es la densidad de la distribución Dirichlet con parámetro $(\alpha/N, \ldots, \alpha/N)$.

Debido a que $\mathcal{DP}_N(\alpha H)$ es una buena aproximación al proceso Dirichlet $\mathcal{DP}(\alpha H)$, el *Gibbs sampler* por bloques es, por tanto, también una buena aproximación al *Gibbs sampler* para urnas de Pólya.

Por otro lado, este método se puede aplicar aún en el caso no conjugado, ya que sólo se debe modificar la manera de obtener la distribución condicional para Z, utilizando Metropolis–Hastings, por ejemplo. Además, la distribución condicional para Z tiene básicamente el mismo esquema que el paso de aceleración, por lo que no es necesario agregar ningún paso adicional.

2.3 Metropolis–Hastings para distribuciones no conjugadas

Una forma de tratar el caso donde las distribuciones del kernel y de la medida base no son conjugadas fue propuesta por Neal (2000). Para ello utiliza el modelo semiparamétrico

$$\begin{aligned} X_i \mid Y_i &\overset{\text{ind}}{\sim} \mu(X_i \mid Y_i), \qquad i = 1, \ldots, n \\ Y_i \mid P &\overset{\text{iid}}{\sim} P \\ P &\sim \mathcal{DP}(\alpha H) \end{aligned} \tag{7}$$

y su equivalente, un modelo de mezcla finita con N componentes

$$\begin{aligned} X_i \mid Z_i, K_i &\overset{\text{ind}}{\sim} \mu(X_i \mid Z_{K_i}), \qquad i = 1, \ldots, n \\ K_i \mid p &\overset{\text{iid}}{\sim} \text{Discreta}(K_i \mid p) \\ Z_{K_i} &\overset{\text{iid}}{\sim} H(Z_{K_i}) \\ p &\sim \mathcal{D}ir(p \mid \alpha/N, \ldots, \alpha/N). \end{aligned} \tag{8}$$

Estos modelos se pueden generalizar sustituyendo el proceso Dirichlet $\mathcal{D}P(\alpha H)$ por el proceso $\mathcal{P}D(\sigma, \theta)$ en (7). A lo largo de las siguientes secciones se trabajará con el modelo general.

En su Algoritmo 5, Neal (2000) propone actualizar K_i a través del método Metropolis–Hastings utilizando su distribución inicial como distribución propuesta. De esta manera, integrando sobre p, se actualiza cada K_i y el Z_{K_i} correspondiente. En la actualización de cada K_i, la distribución final es proporcional a la verosimilitud de la observación X_i multiplicada por la distribución condicional para $(K_i \mid K_{-i})$, por lo que la probabilidad de aceptación queda definida como

$$\alpha(K^*, K_i) = \min\left\{1, \frac{f(X_i; Z_{K^*})}{f(X_i; Z_{K_i})}\right\}, \tag{9}$$

Entonces, este método obtiene observaciones de la siguiente manera

I. $(K \mid X, Z)$. Para $i = 1, \ldots, n$, actualizar K_i r veces: Obtener un candidato k de la distribución inicial para K_i definida por

$$\mathbb{P}[K_i = K_j^* \text{ para alguna } j \mid K_{-i}] = \frac{n_j^* - \sigma}{n - 1 + \theta} \quad \text{y} \quad \mathbb{P}[K_i \neq K_j^* \text{ para toda } j \mid K_{-i}] = \frac{\theta + \sigma m}{n - 1 + \theta},$$

donde n_j^* es la frecuencia de K_j^* y m es el número de valores únicos.
Si $k \notin \{K_1, \ldots, K_n\}$ simular Z_k de H. Calcular la probabilidad de aceptación (9) y hacer $K_i = k$ con esa probabilidad, de lo contrario dejar K_i sin cambio.

II. $(Z \mid X, K)$. Para cada $k \in \{K_1^*, \ldots, K_m^*\}$ actualizar $Z_k \mid X_i$.

La actualización de $(Z \mid X, K)$ puede hacerse utilizando *Gibbs sampler*, por ejemplo.

Una característica de este método es que al considerar cambiar una K_i, existe mayor probabilidad de ser asignada a un componente de mezcla con más observaciones que a uno con menos, a diferencia del *Gibbs sampler* para urnas de Pólya de la Sección 2.1, el cual considera todos los componentes existentes.

2.4 Metropolis–Hastings y *Gibbs sampler*

Una desventaja del método anterior es que la creación de nuevos componentes de mezcla puede ser ineficiente. Por tal razón, Neal (2000), en su Algoritmo 7, soluciona esta deficiencia modificando la distribución propuesta para actualizar los K_i's.

Si $K_i = K_j$ para alguna $j \neq i$, se puede proponer cambiar K_i a un nuevo componente, con su correspondiente Z_k simulada de H. Por otro lado, para permitir que un componente desaparezca, la distribución propuesta del K_i único (i.e., $K_i \neq K_j$ para toda $j \neq i$) se puede limitar a aquellos componentes asociados con otras observaciones, con probabilidades proporcionales a n_j^*. Se puede observar que, cuando el valor actual K_i no es único, la probabilidad de proponer un nuevo componente es un factor de $(n - 1 + \theta)/(\theta + \sigma m)$, mientras que si K_i es único, la probabilidad de proponer algún componente existente es un factor de $(n-1+\theta)/(n-1-\sigma m)$. Por tanto, la probabilidad de aceptar un valor propuesto debe ajustarse con el cociente de estos factores.

Estos cambios son suficientes para producir una cadena de Markov ergódica, ya que existe probabilidad positiva de que, en una iteración, cada K_i esté asociada a componentes diferentes. Sin embargo, puede resultar ineficiente debido a que una K_i puede pasar de un componente existente a otro sólo después de haber permanecido algunas iteraciones en un componente donde la K_i es única. Una forma de evitar este estancamiento es combinar las actualizaciones del Metropolis–Hastings con actualizaciones parciales utilizando *Gibbs sampler*; estas últimas se aplicarán sólo a aquellas K_i's que no sean únicas, permitiendo que cambien sólo a componentes que tengan al menos una K_j, $i \neq j$, asociada.

Con estos cambios, se obtienen observaciones de la siguiente manera

I. $(K \mid X, Z)$. Para $i = 1, \ldots, n$: Si K_i no es única, sea k un nuevo componente con su correspondiente Z_k simulada de H. Hacer $K_i = k$ con probabilidad

$$a(k, K_i) = \min\left\{1, \frac{\theta + \sigma m}{n - 1 - \sigma m} \frac{f(X_i; Z_k)}{f(X_i; Z_{K_i})}\right\},$$

o dejar K_i sin cambio.

De lo contrario, K_i es única. Entonces, elegir k de K_{-i} haciendo $k = K_j$ con probabilidad $(n_j^* - \sigma)/(n - 1 + \theta)$. Hacer $K_i = k$ con probabilidad

$$\alpha(k, K_i) = \min\left\{1, \frac{n - 1 - \sigma m}{\theta + \sigma m} \frac{f(X_i; Z_k)}{f(X_i; Z_{K_i})}\right\},$$

o dejar K_i sin cambio.

II. Para $i = 1, \ldots, n$: Si K_i es única no hacer nada. De lo contrario, elegir un nuevo valor para K_i de $\{K_1^*, \ldots, K_m^*\}$ con probabilidad

$$\mathbb{P}[K_i = K_j^* \mid K_{-i}, X_i, Z] \propto (n_j^* - \sigma) f(X_i; Z_{K_j^*})$$

III. $(Z \mid X, K)$. Para cada $k \in \{K_1^*, \ldots, K_m^*\}$ actualizar $Z_k \mid X_i$.

2.5 *Gibbs sampler* con parámetros auxiliares

Otra manera de tratar el modelo semiparamétrico (7), o equivalentemente (8), en el caso no conjugado es utilizando *Gibbs sampler* junto con una serie de variables latentes. Utilizando este enfoque, es posible actualizar cada K_i sin necesidad de integrar sobre H, una de las dificultades del método de la Sección 2.1, ya que al actualizar los K_i se introducen las variables latentes, las cuales representan posibles valores de los parámetros de los componentes Z_{K_i} que no están asociados con ninguna otra observación.

La distribución condicional inicial de $(K_i \mid K_{-i})$ queda determinada, entonces, en términos de l componentes auxiliares y sus parámetros asociados, donde los K_i toman valores en $\{1, \ldots, m\}$, con m el número de valores únicos en K_{-i}. Por tanto, la probabilidad de que K_i sea igual a una $K_j \in \{1, \ldots, m\}$ es $(n_j - \sigma)/(\theta + n - 1)$, mientras que la probabilidad de que K_i tenga otro valor es $(\theta + \sigma m)/(\theta + n - 1)$, el cual se puede dividir equitativamente entre las l variables latentes introducidas.

Con estos cambios, se obtienen observaciones de la siguiente manera

I. $(K \mid X, Z)$. Para $i = 1, \ldots, n$: Sea $h = m + l$, donde m son los valores únicos en K_{-i}, y supóngase que K_j toma valores en $\{1, \ldots, m\}$. Si K_i no es única, simular valores de H de manera independiente para Z_k tal que $m < k \leq h$. Si K_i es único, sea $K_i = m + 1$ y simular valores de H de manera independiente para Z_k tal que $m + 1 < k \leq h$.
Simular un nuevo valor para K_i de $\{1, \ldots, h\}$ con probabilidad

$$\mathbb{P}[K_i = K_j \mid K_{-i}, X_i, Z_1, \ldots, Z_h] \propto \begin{cases} (n_j^* - \sigma) f(X_i; Z_{K_j}) & \text{para } 1 \leq K_j \leq m \\ ((\theta + \sigma m)/l)\, f(X_i; Z_{K_j}) & \text{para } m + 1 \leq K_j \leq h. \end{cases}$$

II. $(Z \mid X, K)$. Para cada $k \in \{K_1^*, \ldots, K_m^*\}$ actualizar $Z_k \mid X_i$.

2.6 *Gibbs sampler* para el proceso geométrico

El modelo propuesto por Fuentes-García et al. (2010), que utiliza como medida de probabilidad aleatoria el proceso geométrico (Sección 1.2.6), puede implementarse a través del *Gibbs sampler*. Este modelo puede expresar como sigue

$$\begin{aligned} X_i \mid d_i, N_i &\overset{\text{ind}}{\sim} \mu(X_i \mid Z_{d_i}), \qquad i = 1, \ldots, n \\ d_i \mid N_i &\overset{\text{ind}}{\sim} \mathcal{U}(d_i \mid 1, N_i) \\ \mathbb{P}[N_i = N] &= N\lambda^2(1-\lambda)^{N-1} \\ \lambda &\sim \mathcal{B}e(\lambda \mid a, b) \end{aligned} \tag{10}$$

donde las Z_l's son independientes con distribución común H y a y b se asumen conocidos.

De esta manera, cada iteración del *Gibbs sampler* obtiene observaciones de la siguiente manera

I. $(Z \mid X, d)$. La distribución condicional completa para cada Z_j está dada por

$$h(Z_j \mid X, d) \propto H(\mathrm{d}Z_j) \prod_{d_i = j} f(X_i; Z_j),$$

donde f es la función de densidad de μ.

II. $(d \mid X, Z, N)$ La distribución condicional completa para cada d_i está dada por

$$\mathbb{P}[d_i = l \mid X, Z, N] \propto f(X_i; Z_l) \mathbb{1}_{\{1, \ldots, N_i\}}(l).$$

III. $(N \mid d)$. En la Sección 1.2.6 se obtiene la distribución condicional para N_i. De acuerdo a la elección de $\{q_N / N\}$, se tiene que la distribución condicional completa para N_i está dada por

$$\mathbb{P}[N_i = N \mid d_i] = (1-\lambda)^{N-1} \mathbb{1}_{\{N \geq d_i\}},$$

la cual corresponde a una distribución geométrica truncada.

IV. $(\lambda \mid N_i)$. Finalmente, la distribución condicional para λ está dada por

$$\pi(\lambda \mid N_i) \propto \lambda^{a-1}(1-\lambda)^{b-1} \prod_{i=1}^{n} \lambda^2 (1-\lambda)^{N_i - 1},$$

para a y b conocidos, la cual corresponde a una distribución Beta con parámetros $a + 2n$ y $b - n + \sum_{i=1}^{n} N_i$.

– III –

Programación de métodos para modelos no paramétricos

Una vez estudiados algunos de los métodos de estimación de densidades para modelos no paramétricos, es conveniente desarrollarlos para casos particulares y hacer un análisis de su comportamiento.

Como ya se estudió, para desarrollar un caso particular, es necesario especificar una distribución como medida base en cada proceso y así como un kernel. Por tanto, en este capítulo se fijan estas distribuciones y se obtienen las distribuciones condicionales necesarias.

Con esto en mente, en este capítulo se desarrollarán los casos particulares de los métodos de simulación que se implementarán; esencialmente son los descritos en el capítulo anterior (Sección II.2): *Gibbs sampler* para urnas de Pólya, *Gibbs sampler* por bloques, Metropolis–Hastings para distribuciones no conjugadas, Metropolis–Hastings y *Gibbs sampler*, *Gibbs sampler* con parámetros auxiliares y *Gibbs sampler* para el proceso geométrico. Todos estos modelos se implementarán para el caso en donde la medida base es la distribución Normal–Gamma y el kernel corresponde a la distribución Normal.

Adicionalmente, se estudiará el caso en donde los parámetros de cada proceso son incluidos en los esquemas de simulación, asignándoles una distribución inicial. Específicamente se hará para los procesos Dirichlet, σ–estable normalizado y Poisson–Dirichlet de dos parámetros. Como se verá, la inclusión de estos parámetros en los esquemas de simulación, a excepción del proceso Dirichlet, trae consigo una serie de dificultades desde el punto de vista computacional debido, principalmente, a los grandes números que se obtienen de la proporcional a la distribución condicional correspondiente. Por tanto, se harán algunos comentarios para explicar cómo solucionarlas.

§ 1 Cálculo de distribuciones condicionales

Se comienza obteniendo las distribuciones condicionales para cada uno de los métodos estudiados en el capítulo anterior. Para ello será útil el siguiente lema.

Lema 1. Sea $x = \{x_1, \ldots, x_n\}$ una muestra tal que $x_i \overset{\text{iid}}{\sim} \mathcal{N}(x_i \mid \mu, 1/\tau)$, para $i = 1, \ldots, n$, con $\mu \in \mathbb{R}$ y $\tau \in \mathbb{R}^+$. Bajo la distribución inicial $\mathcal{N}(\mu \mid \omega, t/\tau)\mathcal{G}a(\tau \mid a, b)$, para ω, t, a y b conocidos, su distribución posterior está dada por

$$\mathcal{N}\left(\mu \,\middle|\, \tfrac{\omega+tn\bar{x}}{tn+1}, \tfrac{t}{\tau(tn+1)}\right) \mathcal{G}a\left(\tau \,\middle|\, a + \tfrac{n}{2}, b + \tfrac{S}{2} + \tfrac{n(\omega-\bar{x})^2}{2(tn+1)}\right),$$

donde $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ y $S = \sum_{i=1}^n (x_i - \bar{x})^2$.

Demostración
Bajo los supuestos, se tiene que la función de verosimilitud $p(x \mid \mu, \tau)$ está dada por

$$p(x \mid \mu, \tau) \propto \prod_{i=1}^{n} \tau^{1/2} \exp\Big(-\tfrac{1}{2}\tau(x_i - \mu)^2\Big) = \tau^{n/2} \exp\Big(-\tfrac{\tau}{2} \sum_{i=1}^{n} (x_i - \mu)^2\Big).$$

La suma $\sum_{i=1}^{n}(x_i - \mu)^2$ se simplifica a $\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\mu - \bar{x})^2$, donde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Por tanto, se tiene

$$\begin{aligned} p(x \mid \mu, \tau) &\propto \tau^{n/2} \exp\Big(-\tfrac{\tau n}{2}(\mu - \bar{x})^2\Big) \exp\Big(-\tfrac{\tau}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2\Big) \\ &\propto \mathcal{N}\Big(\mu \,\Big|\, \bar{x}, 1/(\tau n)\Big) \mathcal{G}a\Big(\tau \,\Big|\, (n+1)/2, S/2\Big), \end{aligned}$$

donde $S = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Por otro lado, se sabe que $p(\mu, \tau \mid x) \propto p(x \mid \mu, \tau) p(\mu, \tau)$, entonces

$$\begin{aligned} p(\mu, \tau \mid x) &\propto \mathcal{N}\Big(\mu \,\Big|\, \bar{x}, 1/(\tau n)\Big)\, \mathcal{G}a\left(\tau \,\middle|\, (n+1)/2, S/2\right) \mathcal{N}\Big(\mu \,\Big|\, \omega, t/\tau\Big)\, \mathcal{G}a\left(\tau \,\middle|\, a, b\right) \\ &\propto \tau^{a + \frac{n+1}{2} - 1} \exp\Big(-\tfrac{\tau n}{2}(\mu - \bar{x})^2 - \tfrac{\tau}{2t}(\mu - \omega)^2 - \tau(S/2 + b)\Big) \\ &\propto \tau^{a + \frac{n+1}{2} - 1} \exp\left(-\frac{\tau}{2t}\left\{(tn+1)\left(\mu - \frac{\omega + tn\bar{x}}{tn+1}\right)^2 + \frac{tn(\omega - \bar{x})^2}{tn+1}\right\} - \tau(S/2 + b)\right) \\ &\propto \mathcal{N}\left(\mu \,\middle|\, \tfrac{\omega + tn\bar{x}}{tn+1}, \tfrac{t}{\tau(tn+1)}\right) \mathcal{G}a\left(\tau \,\middle|\, a + \tfrac{n}{2}, b + \tfrac{S}{2} + \tfrac{n(\omega - \bar{x})^2}{2(tn+1)}\right). \end{aligned}$$

□

1.1 *Gibbs sampler* para urnas de Pólya

De la Sección II.2.1, para una muestra $x = \{x_1, \ldots, x_n\}$, se obtiene el siguiente modelo

$$\begin{aligned} x_i \mid \mu_i, \tau_i &\overset{\text{ind}}{\sim} \mathcal{N}(x_i \mid \mu_i, \tau_i^{-1}), \qquad i = 1, \ldots, n \\ \mu_i, \tau_i \mid P &\overset{\text{iid}}{\sim} P \\ P &\sim \mathcal{P}\mathcal{D}(\sigma, \theta), \end{aligned} \tag{1a}$$

con medida base

$$H(\mu, \tau) = \mathcal{N}(\mu \mid \omega, t/\tau)\mathcal{G}a(\tau \mid a, b), \tag{1b}$$

donde ω, t, a y b se asumen conocidos.

El *Gibbs sampler* requiere de la distribución condicional completa para los parámetros μ_i, τ_i y de los pesos $\{q_j\}_{j=0}^{m}$, los cuales se obtienen a continuación.

PROPOSICIÓN 1. Bajo el Modelo (1)

1. La distribución condicional completa $(\mu_i, \tau_i \mid \mu_{-i}, \tau_{-i}, x)$, para $i = 1, \ldots, n$, está dada por

$$\mathcal{N}\left(\mu_i \,\middle|\, \tfrac{tx_i + \omega}{t+1}, \tfrac{t}{\tau(t+1)}\right) \mathcal{G}a\left(\tau \,\middle|\, a + \tfrac{1}{2}, b + \tfrac{(x_i - \omega)^2}{2(t+1)}\right); \tag{2}$$

II. Los pesos $\{q_j\}_{j=0}^m$ están dados por

$$q_0 \propto \frac{\theta + \sigma m}{\theta + n - 1} St\left(x; 2a, \omega, b(t+1)/a\right), \tag{3a}$$

$$q_j \propto \frac{n_j^* - \sigma}{\theta + n - 1} \mathcal{N}\left(x_i; \mu_j^*, \tau_j^{*\,-1}\right), \quad j = 1, \ldots, m, \tag{3b}$$

para $i = 1, \ldots, n$, donde $St(x; \alpha, \mu, \sigma^2)$ corresponde a la densidad Student con α grados de libertad, media μ y varianza σ^2.

Demostración

I. Utilizando el Lema 1 con $n = 1$ se obtiene (2).

II. Para q_0, de (II.5) se tiene que

$$q_0 \propto \frac{\theta + \sigma m}{\theta + n - 1} \int_{\mathbb{R}\times\mathbb{R}^+} \mathcal{N}(x; \mu, \tau^{-1}) \mathcal{N}(\mu; \omega, t/\tau) \mathcal{G}a(\tau; a, b) \mathrm{d}\mu \mathrm{d}\tau,$$

la doble integral se calcula como en el Lema 1 con $n = 1$, pero ahora es necesario conservar todos los términos constantes. Así, se obtiene

$$\begin{aligned}
&\mathcal{N}\left(x; \mu, \tau^{-1}\right) \mathcal{N}(\mu; \omega, t/\tau) \mathcal{G}a(\tau; a, b) = \\
&\qquad = \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(\mu - x)^2\right) \frac{\tau^{1/2}}{\sqrt{2\pi t}} \exp\left(-\frac{\tau}{2t}(\mu - \omega)^2\right) \frac{b^a \tau^{a-1}}{\Gamma(a)} \exp\left(-b\tau\right) \\
&\qquad = \frac{(b\tau)^a}{2\pi t^{1/2} \Gamma(a)} \exp\left(-\frac{\tau}{2}(\mu - x)^2 - \frac{\tau}{2t}(\mu - \omega)^2 - b\tau\right) \\
&\qquad = \frac{b^a \sqrt{2\pi} t^{1/2} \Gamma(a + 1/2)}{2\pi t^{1/2} \Gamma(a)(t+1)^{1/2} \Gamma(a)} \left(\frac{(x - \omega)^2}{2(t+1)} + b\right)^{-(a+1/2)} \times \\
&\qquad\quad \times \mathcal{N}\left(\mu; \tfrac{\omega + tx}{t+1}, \tfrac{t}{\tau(t+1)}\right) \mathcal{G}a\left(\tau; a + 1/2, b + \tfrac{(\omega - x)^2}{2(t+1)}\right).
\end{aligned}$$

Integrando la última igualdad, se tiene

$$\begin{aligned}
&\int_{\mathbb{R}\times\mathbb{R}^+} \mathcal{N}\left(x; \mu, \tau^{-1}\right) \mathcal{N}(\mu; \omega, t/\tau) \mathcal{G}a(\tau; a, b) \mathrm{d}\mu \mathrm{d}\tau = \\
&\qquad\qquad = \frac{\Gamma(a + 1/2)}{\Gamma(a)\sqrt{2a\pi}} \left(\frac{a}{b(t+1)}\right)^{1/2} \left(1 + \frac{a(x - \omega)^2}{2ab(t+1)}\right)^{-(a+1/2)},
\end{aligned}$$

lo cual corresponde a la densidad Student con $2a$ grados de libertad, media ω y varianza $b(t+1)/a$. De esta forma se obtiene q_0 como en (3a).

Para q_j, $j = 1, \ldots, m$, de (II.5) se obtiene directamente (3b), donde (μ_j^*, τ_j^*) es el j–ésimo elemento único de $\{(\mu_i, \tau_i)\}_{i=1}^n$, $j = 1, \ldots, m$, el cual tiene frecuencia n_j^*. □

Si se quiere utilizar el método de aceleración se debe obtener, además, la distribución condicional $(\mu_j^*, \tau_j^* \mid C, x)$ para $j = 1, \ldots, m$, donde C es un vector de membrecías $\{C_1, \ldots, C_n\}$, definido como $C_i = j$ si y sólo si $(\mu_i, \tau_i) = (\mu_j^*, \tau_j^*)$.

PROPOSICIÓN 2. Las distribución condicional $(\mu_j^*, \tau_j^* \mid C, x)$ para $j = 1, \ldots, m$, bajo el Modelo (1), está dada por

$$\mathcal{N}\left(\mu_j^* \,\middle|\, \tfrac{tn_j^*\bar{x}_j^*+\omega}{tn_j^*+1}, \tfrac{t}{\tau_j^*(tn_j^*+1)}\right) \mathcal{G}a\left(\tau_j^* \,\middle|\, \theta + \tfrac{n_j^*}{2}, \beta + \tfrac{S_j}{2} + \tfrac{n_j^*(\bar{x}_j^*-\omega)^2}{2(tn_j^*+1)}\right),$$

donde

$$I_j = \{i : C_i = C_j^*\}, \quad \bar{x}_j^* = \frac{1}{n_j^*}\sum_{k\in I_j} x_k \quad \text{y} \quad S_j = \sum_{k\in I_j}(x_k - \bar{x}_j^*)^2.$$

Demostración
Se obtiene directamente del Lema 1, ya que los elementos únicos $C^* = \{C_1^*, \ldots, C_m^*\}$ de C forman una partición de $\{x_i\}_{i=1}^n$ y de $\{(\mu_i, \tau_i)\}_{i=1}^n$, donde cada elemento de esta tiene tamaño n_j^*. □

1.2 *Gibbs sampler* por bloques

El modelo de la Sección II.2.2, para una muestra $x = \{x_1, \ldots, x_n\}$, queda definido como sigue

$$\begin{aligned} x_i \mid \mu_{K_i}, \tau_{K_i}, K_i &\overset{\text{ind}}{\sim} \mathcal{N}(x_i \mid \mu_{K_i}, \tau_{K_i}^{-1}), \qquad i = 1, \ldots, n \\ K_i \mid p &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} p_k \delta_k(\cdot) \\ \mu_k, \tau_k &\overset{\text{iid}}{\sim} \mathcal{N}(\mu_k \mid \omega, t/\tau_k)\mathcal{G}a(\tau_k \mid a, b), \quad k = 1, \ldots, N \\ p &\sim \mathcal{G}D(p \mid c, d), \end{aligned} \tag{4}$$

donde $p = (p_1, \ldots, p_N)$ y ω, t, a, b, c y d se asumen conocidos.

En este caso, sólo es necesario obtener la distribución condicional para $(\mu_{K_i}, \tau_{K_i} \mid K_i, x)$, ya que las demás se obtienen directamente de la Sección II.2.2.

PROPOSICIÓN 3. Bajo el modelo (4), la distribución condicional $(\mu_{K_i}, \tau_{K_i} \mid K_i, x)$, para $i = 1, \ldots, n$, está dada por

$$\mathcal{N}\left(\mu_{K_i} \,\middle|\, \tfrac{tn_j^*\bar{x}_j^*+\omega}{tn_j^*+1}, \tfrac{t}{\tau_{K_i}(tn_j^*+1)}\right) \mathcal{G}a\left(\tau_{K_i} \,\middle|\, a + \tfrac{n_j^*}{2}, b + \tfrac{S_j}{2} + \tfrac{n_j^*(\bar{x}_j^*-\omega)^2}{2(tn_j^*+1)}\right),$$

donde

$$I_j = \{i : K_i = K_j^*\}, \quad \bar{x}_j^* = \frac{1}{n_j^*}\sum_{k\in I_j} x_k \quad \text{y} \quad S_j = \sum_{k\in I_j}(x_k - \bar{x}_j^*)^2.$$

Demostración
Se puede ver que K_j corresponde a C_j, $j = 1, \ldots, m$, en la Proposición 2, por tanto, el resultado se sigue directamente de ésta. □

1.3 MÉTODOS PARA DISTRIBUCIONES NO CONJUGADAS

En los métodos de las Secciones II.2.3, II.2.4 y II.2.5 se trabaja con el siguiente modelo, para una muestra $x = \{x_1, \ldots, x_n\}$

$$\begin{aligned} x_i \mid \mu_i, \tau_i, K_i &\overset{\text{ind}}{\sim} \mu(X_i \mid \mu_{K_i}, \tau_{K_i}^{-1}), \qquad i = 1, \ldots, n \\ K_i \mid p &\overset{\text{iid}}{\sim} \text{Discreta}(K_i \mid p) \\ \mu_{K_i}, \tau_{K_i} &\overset{\text{iid}}{\sim} \mathcal{N}(\mu_{K_i} \mid \omega, t/\tau_{K_i}) \mathcal{G}a(\tau_{K_i} \mid a, b) \\ p &\sim \mathcal{G}\text{D}(p \mid c, d), \end{aligned} \tag{5}$$

donde $p = (p_1, \ldots, p_N)$ y ω, t, a, b, c y d se asumen conocidos.

En este modelo, sólo se necesita obtener la distribución condicional $(\mu_{K_i}, \tau_{K_i} \mid K_i, x)$. Sin embargo, se puede ver que corresponde a la Proposición 3.

1.4 *GIBBS SAMPLER* PARA EL PROCESO GEOMÉTRICO

El modelo de la Sección II.2.6, para una muestra $x = \{x_1, \ldots, x_n\}$, queda definido de la siguiente manera

$$\begin{aligned} x_i \mid \mu_i, \tau_i, d_i &\overset{\text{ind}}{\sim} \mu(x_i \mid \mu_{d_i}, \tau_{d_i}^{-1}), \qquad i = 1, \ldots, n \\ d_i \mid N_i &\overset{\text{ind}}{\sim} \mathcal{U}(d_i \mid 1, N_i) \\ \mu_{d_i}, \tau_{d_i} &\overset{\text{iid}}{\sim} \mathcal{N}(\mu_{d_i} \mid \omega, t/\tau_{d_i}) \mathcal{G}a(\tau_{d_i} \mid a, b) \\ \mathbb{P}[N_i = N] &= q_N \end{aligned} \tag{6}$$

donde ω, t, a y b se asumen conocidos.

Como se explicó en la Sección I.2.6, existen varias opciones para la distribución de $\{q_N\}$, pero se tomará la estudiada en Fuentes-García et al. (2010): $q_N = N\lambda^2(1-\lambda)^{N-1}$. Además, es recomendable asignar una distribución a λ para incluirla en el *Gibbs sampler*; una opción es asignarle una distribución Beta con parámetros conocidos α y β.

Para este modelo se necesita obtener las distribuciones condicionales $(\mu, \tau | x, d)$, $(N | d, \lambda)$ y $(\lambda | N, x)$. La distribución condicional $(d \mid x, \mu, \tau, N)$ se obtiene directamente de la Sección II.2.6.

PROPOSICIÓN 4. Bajo el modelo (6)

1. La distribución condicional $(\mu_{d_i}, \tau_{d_i} \mid \mu_{-d_i}, \tau_{-d_i}, x, d)$, para $i = 1, \ldots, n$, está dada por

$$\mathcal{N}\left(\mu_{d_i} \,\middle|\, \tfrac{tn_j\bar{x}_j+\omega}{tn_j+1}, \tfrac{t}{\tau_{d_i}(tn_j+1)}\right) \mathcal{G}a\left(\tau_{d_i} \,\middle|\, a + \tfrac{n_j}{2}, b + \tfrac{S_j}{2} + \tfrac{n_j(\bar{x}_j-\omega)^2}{2(tn_j+1)}\right),$$

donde

$$n_j = \sum_{d_i=j} 1, \quad \bar{x}_j = \frac{1}{n_j} \sum_{d_i=j} x_i \quad \text{y} \quad S_j = \sum_{d_i=j} (x_i - \bar{x}_j)^2;$$

II. La distribución condicional $(N_i \mid N_{-i}, d, \lambda)$, para $q_N = N\lambda^2(1-\lambda)^{N-1}$, está dada por

$$\mathbb{P}[N_i = N \mid d_i, \lambda] = \lambda(1-\lambda)^{N-1}\mathbb{1}_{\{N \geq d_i\}},$$

la cual corresponde a una distribución Geométrica truncada en d_i y con parámetro λ.

III. La distribución condicional completa $(\lambda \mid N, x)$ está dada por

$$\mathcal{B}e\Big(\lambda \,\Big|\, \alpha + 2n, \beta - n + \textstyle\sum_{i=1}^{n} N_i\Big),$$

para α y β conocidos.

Demostración

I. Este resultado se sigue de la Proposición 2, observando que d_i y C_i se definen de la misma manera.

II. En la Sección 1.2.6 se obtuvo que

$$\mathbb{P}[N_i = N \mid d_i] \propto q_N/N\mathbb{1}_{\{N \geq d_i\}}.$$

Haciendo $q_N = N\lambda^2(1-\lambda)^{N-1}$, se sigue el resultado.

III. La distribución condicional final para q está dada por

$$\pi(q \mid \dots) \propto \prod_{i=1}^{n} q_{N_i}\pi(q);$$

para la elección de $\{q_N\}$ y tomando la distribución inicial $\pi(q) = \mathcal{B}e(\lambda|\alpha, \beta)$, para α y β conocidos, esta distribución se convierte en

$$\begin{aligned}\pi(\lambda \mid N, x) &\propto \lambda^{\alpha-1}(1-\lambda)^{\beta-1}\prod_{i=1}^{n}\lambda^2(1-\lambda)^{N_i-1} \\ &\propto \mathcal{B}e\Big(\lambda \,\Big|\, \alpha + 2n, \beta - n + \textstyle\sum_{i=1}^{n} N_i\Big).\end{aligned}$$

□

§ 2 Asignación de distribuciones iniciales adicionales

Para todos los métodos es posible asignar distribuciones iniciales adicionales, por ejemplo a los parámetros de la medida base o a los parámetros de las medidas de probabilidad aleatorias. En este trabajo únicamente se tomará el segundo caso, por tanto, en esta sección se asignarán distribuciones iniciales a los parámetros α, σ y (σ, θ) de los procesos Dirichlet, σ–estable normalizado y Poisson–Dirichlet de dos parámetros, respectivamente, para así incluirlos en los algoritmos de simulación de los distintos métodos.

2.1 Proceso Dirichlet

De la Sección 1.2.5, Ejemplo 1.7, para el proceso Dirichlet con parámetro $\alpha > 0$, la distribución de K_n está dada por

$$\mathbb{P}[K_n = k \mid \alpha, n] = \frac{c_{n,k}\alpha^k}{(\alpha)_{n\uparrow}}. \tag{7}$$

West (1992) estudia cómo incluir α en los métodos de simulación. Para ello, se asume que los datos X son inicialmente condicionalmente independientes de α dados todos los demás parámetros Y, obteniendo así

$$p(\alpha \mid k, Y, X) \propto p(\alpha \mid k) \propto p(\alpha)p(k \mid \alpha),$$

con $p(k \mid \alpha)$ dada por (7). Una de sus propuestas para la distribución inicial de α es una distribución Gamma con parámetros a, b conocidos. La obtención de la distribución posterior de α se resume en la siguiente proposición.

Proposición 5. La distribución posterior de α, con función de verosimilitud dada por (7) y distribución inicial Gamma con parámetros a, b conocidos, es

$$p(\alpha \mid \eta, k) = \pi_\eta \mathcal{G}a(\alpha \mid a + k, b - \log\eta) + (1 - \pi_\eta)\mathcal{G}a(\alpha \mid a + k - 1, b - \log\eta), \tag{8}$$

donde $\eta \sim \mathcal{B}e(\eta \mid \alpha + 1, n)$ y

$$\frac{\pi_\eta}{1 - \pi_\eta} = \frac{a + k - 1}{n(b - \log\eta)},$$

con n el tamaño de la muestra.

Demostración

Para $\alpha > 0$, el recíproco del símbolo de Pochhammer en (7) se puede reescribir como

$$\frac{1}{(a)_{n\uparrow}} = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} = \frac{(\alpha + n)\beta(\alpha + 1, n)}{\alpha\Gamma(n)},$$

donde $\beta(\cdot, \cdot)$ es la función beta. Entonces, para $k = 1, \ldots, n$

$$\begin{aligned} p(\alpha \mid k) &\propto p(\alpha)\alpha^{k-1}(\alpha + n)\beta(\alpha + 1, n) \\ &\propto p(\alpha)\alpha^{k-1}(\alpha + n)\int_0^1 \eta^\alpha(1 - \eta)^{n-1}\mathrm{d}\eta. \end{aligned}$$

Esto implica que $p(\alpha \mid k)$ es la distribución marginal de una distribución conjunta para α y una variable η continua, con $0 < \eta < 1$, tal que

$$p(\alpha, \eta \mid k) \propto p(\alpha)\alpha^{k-1}(\alpha + n)\eta^\alpha(1 - \eta)^{n-1}.$$

Las distribuciones condicionales $p(\alpha \mid \eta, k)$ y $p(\eta \mid \alpha, b)$, quedan determinadas, por tanto, como sigue

I. $p(\alpha \mid \eta, k)$. Bajo la distribución inicial $p(\alpha) = \mathcal{G}a(\alpha \mid a, b)$ se tiene

$$\begin{aligned} p(\alpha \mid k) &\propto \alpha^{a+k-2}(\alpha + n)\exp(-\alpha(b - \log\eta)) \\ &\propto \alpha^{a+k-1}\exp(-\alpha(b - \log\eta)) + n\alpha^{a+k-2}\exp(-\alpha(b - \log\eta)), \end{aligned}$$

lo que se reduce a

$$p(\alpha \mid k) \propto c_1 \mathcal{G}a(\alpha \mid a + k, b - \log\eta) + c_2 \mathcal{G}a(\alpha \mid a + k - 1, b - \log\eta),$$

donde

$$c_1 = \frac{\Gamma(a+k)}{(b - \log\eta)^{a+k}} \quad \text{y} \quad c_2 = \frac{n\Gamma(a+k-1)}{(b - \log\eta)^{a+k-1}}.$$

De estas constantes se obtiene

$$\frac{c_1}{c_2} = \frac{a + k - 1}{n(b - \log\eta)}.$$

II. $p(\eta \mid \alpha, k)$. En este caso se tiene que

$$p(\eta \mid \alpha, k) \propto \eta^{\alpha}(1 - \eta)^{n-1},$$

con $0 < \eta < 1$, por lo que se tiene que $\eta \sim \mathcal{B}e(\eta \mid \alpha + 1, n)$. □

INCLUSIÓN EN LOS ALGORITMOS DE SIMULACIÓN

El resultado anterior permite ver cómo actualizar α en cada iteración. Después de realizar las actualizaciones para los parámetros Y (que corresponden a μ, τ para el caso implementado) y para las variables indicadoras (C o K), y asumiendo que Y es de tamaño n y contiene k valores únicos, hacer lo siguiente

I. Simular $\eta \sim \mathcal{B}e(\eta \mid \alpha + 1, n)$ utilizando el valor más reciente de α,

II. Simular el nuevo valor para α a través de (8) utilizando el η simulado y el valor más reciente de k.

2.2 PROCESO σ–ESTABLE NORMALIZADO

El proceso σ–estable normalizado, como se mencionó en el Ejemplo I.6, corresponde a la medida canónica del proceso Poisson–Dirichlet $(\sigma, 0)$. Por tanto, se puede tomar (I.19) como función de verosimilitud para σ haciendo $\theta = 0$, obteniendo así

$$\mathbb{P}[K_n = k, N_{K,n} = n_{k,n} \mid \sigma] = \frac{\Gamma(k)\sigma^{k-1}}{\Gamma(n)} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1\uparrow},$$

donde k es el número de grupos en una muestra de tamaño n y $n_{k,n} = (n_1, \ldots, n_k)$ son las frecuencias de cada grupo.

Como $\sigma \in (0, 1)$, se le puede asignar una distribución inicial Beta con parámetros a, b conocidos. De esta manera se obtiene la distribución posterior

$$p(\sigma \mid k, n_1, \ldots, n_k, a, b) \propto \sigma^{k+a-2}(1 - \sigma)^{b-1} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1\uparrow}, \tag{9}$$

Se puede observar que (9) no corresponde a ninguna función de densidad «fácil» de simular. Además, el producto de símbolos de Pochhammer por lo general arroja números muy grandes y, desde el punto de vista computacional, puede ocasionar desbordamiento al trabajar con tipos de datos estándar.

Por tanto, para simular de ella se utilizará el método Metropolis de rechazo adaptativo (ARMS), el cual es una generalización del método ARS para densidades que no son log–cóncavas (Gilks, Best & Tan 1995). La función que se utilizará es, entonces, la log–densidad (proporcional) condicional para σ, que está dada por

$$(k+a-2)\log(\sigma)+(b-1)\log(1-\sigma)-k\log\Gamma(1-\sigma)+\sum_{j=1}^{k}\log\Gamma(n_j-\sigma). \tag{10}$$

Al utilizar (10), los números arrojados por el log–producto de símbolos de Pochhammer resultan considerablemente menores, solventando así la limitación computacional antes mencionada.

Inclusión en los algoritmos de simulación

Para incluir la actualización de σ simplemente se simulará un número aleatorio de (9) a través de (10) utilizando el método ARMS[1] con los valores más recientes de k y $n_{n,k}$ al final de cada iteración.

2.3 Proceso Poisson–Dirichlet de dos parámetros

De manera análoga al número de grupos, en el proceso Poisson–Dirichlet de dos parámetros se puede modelar el número de especies a través de la variable aleatoria K_n, como en la Sección 1.2.5, la cual tiene distribución conjunta

$$\mathbb{P}[K_n=k, N_{K,n}=n_{k,n}]=\frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1\uparrow}}\prod_{j=1}^{k}(1-\sigma)_{n_j-1\uparrow}$$

Si se asignan distribuciones iniciales independientes a σ y θ, se tienen, respectivamente, las distribuciones posteriores

$$p(\sigma\mid\theta,k,n_1,\ldots,n_k)\propto p(\sigma)\sigma^{k-1}\left(\tfrac{\theta}{\sigma}+1\right)_{k-1\uparrow}\prod_{j=1}^{k}(1-\sigma)_{n_j-1\uparrow}$$

y

$$p(\theta\mid\sigma,n,k)\propto p(\theta)\frac{\left(\frac{\theta}{\sigma}+1\right)_{k-1\uparrow}}{(\theta+1)_{n-1\uparrow}},$$

donde $n_1,\ldots,n_k$ denota la frecuencia de los k valores distintos $\{Y_i^*\}_{i=1}^{k}$ de una muestra $\{Y_i\}_{i=1}^{n}$ de tamaño n.

Debido al soporte de σ y θ, una opción para sus distribuciones iniciales son Beta con parámetros a, b y Gamma con parámetros c, d, respectivamente, con a, b, c y d conocidos. De esta forma, se obtiene

$$p(\sigma\mid\theta,k,n_1,\ldots,n_k,a,b)\propto\sigma^{k+a-2}(1-\sigma)^{b-1}\left(\tfrac{\theta}{\sigma}+1\right)_{k-1\uparrow}\prod_{j=1}^{k}(1-\sigma)_{n_j-1\uparrow}, \tag{11}$$

[1] En `http://www.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/web_page/Welcome.html` es posible descargar una implementación de este método.

y

$$p(\theta \mid \sigma, n, k, c, d) \propto \frac{\left(\frac{\theta}{\sigma}+1\right)_{k-1\uparrow}}{(\theta+1)_{n-1\uparrow}} \theta^{c-1} e^{-d\theta}. \tag{12}$$

Inclusión en los algoritmos de simulación

De la misma manera que en los procesos anteriores, la actualización de estos parámetros se puede hacer después de actualizar los parámetros del método.

En el caso del parámetro σ, se puede actualizar de manera análoga al proceso σ–estable normalizado, a través del método ARMS, utilizando la log–densidad (proporcional) condicional

$$(k+a-2)\log(\sigma) + (b-1)\log(1-\sigma) + \log\Gamma(\theta/\sigma + k) - \log\Gamma(\theta/\sigma+1) - k\log\Gamma(1-\sigma) + \sum_{j=1}^{k} \log\Gamma(n_j - \sigma).$$

En cuanto al parámetro θ, se tiene lo siguiente. El símbolo de Pochhammer se puede reescribir como

$$(x)_{n\uparrow} = \sum_{i=1}^{n} c_{n,i} x^i,$$

donde $c_{n,i}$ es el valor absoluto de los números de Stirling de primer orden con parámetros n, i; además, se tiene que $(x+1)_{n-1\uparrow} = (x)_{n\uparrow}/x$. Entonces (12) se puede reescribir como

$$p(\theta \mid \sigma, n, k, c, d) \propto \left(\tfrac{\theta}{\sigma}\right)_{k\uparrow} \frac{\theta^{c-1} e^{-d\theta}}{(\theta)_{n\uparrow}} \propto \sum_{i=1}^{k} \frac{c_{k,i}}{\sigma^i} \frac{\theta^{c+i-1} e^{-d\theta}}{(\theta)_{n\uparrow}}.$$

El término $\theta^{c+i-1}e^{-d\theta}/(\theta)_{n\uparrow}$ se puede reescribir como en la Proposición (5), obteniendo

$$\frac{\theta^{c+i-1}e^{-d\theta}}{(\theta)_{n\uparrow}} = \frac{(\theta+n)\theta^{c+i-2}e^{-d\theta}}{\Gamma(n)} \int_0^1 \eta^{\theta}(1-\eta)^{n-1} \mathrm{d}\eta,$$

y de aquí se tiene la distribución condicional conjunta

$$p(\theta, \eta \mid \sigma, n, k, c, d) \propto \sum_{i=1}^{k} \frac{c_{k,i}}{\sigma^i} (\theta+n)\theta^{c+i-2} e^{-d\theta} \eta^{\theta} (1-\eta)^{n-1}.$$

La distribución condicional completa para η, como en la Proposición 5, corresponde a la distribución Beta con parámetros $\theta+1, n$. Mientras que, para la distribución condicional completa para θ, se tiene

$$\begin{aligned} p(\theta \mid \eta, \sigma, n, k, c, d) &\propto \sum_{i=1}^{k} \frac{c_{k,i}}{\sigma^i} (\theta+n)\theta^{c+i-2} e^{\theta(d-\log\eta)} \\ &\propto \sum_{i=1}^{k} w_i \big(\pi \mathcal{G}a(\theta \mid c+i, d-\log\eta) + (1-\pi)\mathcal{G}a(\theta \mid c+i-1, d-\log\eta)\big), \end{aligned} \tag{13}$$

donde

$$\pi = \frac{c+i-1}{n(d-\log\eta) + c + i - 1}$$

y

$$w_i = \frac{c_{k,i}(c+i-1+n(d-\log\eta))\Gamma(c+i-1)}{\big(\sigma(d-\log\eta)\big)^i}, \qquad i = 1,\ldots,k,$$

con los w_i's normalizados para que sumen uno.

Por tanto, la actualización del parámetro θ se puede realizar de la siguiente manera

I. Simular $\eta \sim \mathcal{B}e(\eta \mid \theta+1, n)$ utilizando el valor más reciente de θ,

II. Elegir la i–ésima componente de (13) con probabilidad proporcional a w_i, $i = 1,\ldots,k$, utilizando el η simulado y los valore más recientes de k y σ,

III. Simular de la mezcla de distribuciones Gamma con el η simulado y el valor de i seleccionado.

– IV –

Análisis comparativo

Una vez implementados los distintos métodos de estimación de densidades estudiados en los capítulos anteriores, es posible hacer un análisis comparativo entre ellos. Para este análisis se utilizaron muestras provenientes de distintos modelos de mezclas y se toma también un conjunto de datos reales.

Como se explicó en el Capítulo II, monitorear la convergencia de un método MCMC no es tarea fácil, debido principalmente a su complejidad. Sin embargo, para métodos de estimación de densidades existen ciertas estadísticas que ayudan a esto.

Este capítulo comienza con una descripción de los modelos de mezclas y de los datos reales y continúa con una explicación de algunas de las estadísticas más usuales que se utilizan para monitorear la convergencia. Además, todos los métodos requieren de un conjunto de parámetros, por lo que en la tercera sección se especifican todos ellos. Por último, se presentan los resultados de las simulaciones.

§ 1 Especificación de modelos y datos

Para probar los métodos de simulación estudiados con anterioridad se utilizan muestras de distintos modelos de mezclas, así como un conjunto de datos reales.

Los modelos que se utilizan son algunos de los mencionados en Kalli, Griffin & Walker (2009), los cuales consisten en una mezcla de distribuciones normales. Específicamente, se trabajó con una muestra de cada uno de los siguientes modelos de mezclas

I. modelo «separado»

$$0.1\,\mathcal{N}(x;-7,1) + 0.4\,\mathcal{N}(x;0,1) + 0.3\,\mathcal{N}(x;4,0.5^2) + 0.2\,\mathcal{N}(x;8,1.5^2).$$

II. modelo «platicúrtico»

$$0.2\,\mathcal{N}(x;-4,1) + 0.2\,\mathcal{N}(x;-2,1) + 0.2\,\mathcal{N}(x;0,1) + 0.2\,\mathcal{N}(x;2,1) + 0.2\,\mathcal{N}(x;4,1),$$

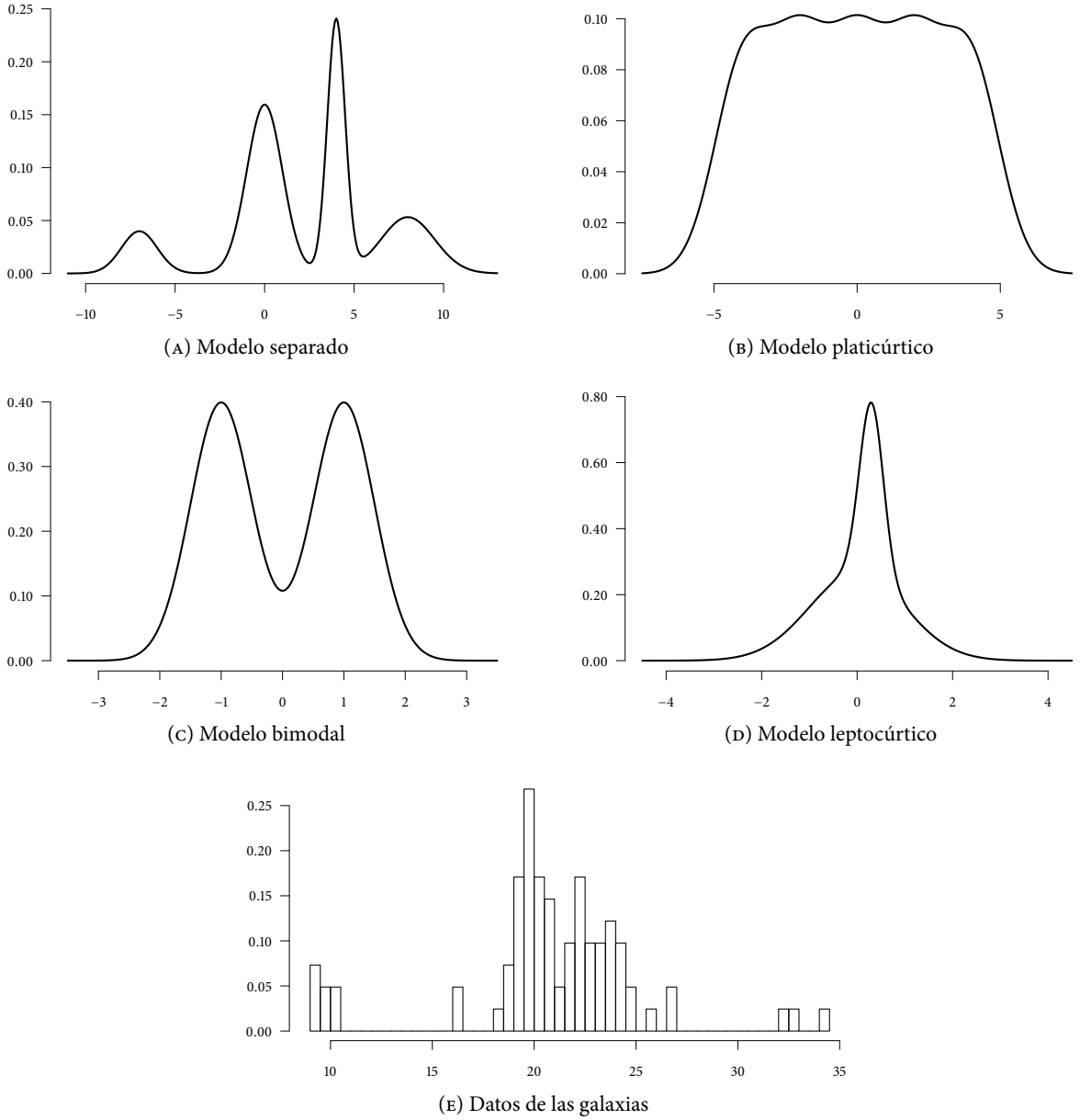
III. modelo «bimodal»

$$0.5\,\mathcal{N}(x;-1,0.5^2) + 0.5\,\mathcal{N}(x;1,0.5^2),$$

IV. modelo «leptocúrtico»

$$0.67\,\mathcal{N}(x;0,1) + 0.33\,\mathcal{N}(x;0.3,0.25^2),$$

Para cada uno de los modelos se tomó una muestra de tamaño 100. El conjunto de datos reales que se utilizó fue los conocidos datos de las galaxias, que consta de las velocidades de 82 galaxias que se alejan de la nuestra. En la Figura 1 se muestran las densidades de los modelos y el histograma del conjunto de datos.



(A) Modelo separado

(B) Modelo platicúrtico

(C) Modelo bimodal

(D) Modelo leptocúrtico

(E) Datos de las galaxias

FIGURA 1: Modelos de prueba e histograma de los datos reales

§ 2 Medidas de ajuste

Como se comentó en el Capítulo II, una vez que se implementa algún método de simulación es necesario verificar que converja; sin embargo, debido a su complejidad, sólo se pueden realizar algunas pruebas para saber si se alcanzó su estado estacionario.

En el caso de los modelos de estimación de densidades, como los estudiados anteriormente, una de las estadísticas más utilizadas es la devianza, y por otro lado, en la mayoría de estos modelos, una variable de interés es el número de grupos. Por tanto, es importante verificar que ambas variables converjan. De esta manera se podrá monitorear el desempeño de cada modelo (Kalli et al. 2009). En cuanto a la eficiencia de los métodos, Kalli et al. (2009) estiman el «tiempo de autocorrelación integrado». A continuación se explican con detalle cada una de estas estadísticas.

2.1 Devianza

Para una muestra $x = \{x_1, \ldots, x_n\}$, la calidad del ajuste de un punto estimado $\hat{h}(x_i)$ de su densidad se puede resumir definiendo su «devianza» asociada

$$D = -2 \sum_{i=1}^{n} \log \hat{h}(x_i). \tag{1}$$

Desde la perspectiva bayesiana, la densidad estimada corresponde a la distribución predictiva para la siguiente observación. En particular, utilizando la densidad predictiva incondicional para una nueva observación z dada la muestra x, que está dada por

$$p(z \mid x) = \mathbb{E}\left[\textstyle\sum_j w_j K(z; \theta_j) \,\middle|\, x\right],$$

se tiene que la devianza está dada por

$$D = -2 \sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} w_j K(x_i; \theta_j) \right). \tag{2}$$

En algunos de los modelos estudiados, las proporciones $\{w_1, \ldots, w_m\}$ están dadas por $w_j = n_j^*/n$, donde m es el número de grupos, cada uno con frecuencia n_j^* y n es el tamaño de la muestra.

2.2 Número de grupos

El número de grupos en cada método, a excepción del *Gibbs sampler* para el proceso geométrico, está determinado por la variable aleatoria K_n. Durante cada iteración, se obtiene un vector $k^{(l)} = \{k_1^{(l)}, \ldots, k_n^{(l)}\}$ que asocia a cada observación x_i con la componente $k_i^{(l)}$. Además se puede obtener el vector $k^{*(l)} = \{k_1^{*(l)}, \ldots, k_n^{*(l)}\}$, el cual contiene los valores únicos de $k^{(l)}$, por lo que al final de cada iteración se obtiene $K_n^{(l)} = |k^{*(l)}|$. Por tanto, se busca para cada modelo que $K_n^{(l)}$ converja.

2.3 Tiempo de autocorrelación integrado

Además de monitorear la convergencia de la devianza y del número de grupos, la eficiencia de cada método se obtiene estimando el «tiempo de autocorrelación integrado», τ, para ambas variables. El tiempo de autocorrelación integrado se define como

$$\tau = \frac{1}{2} + \sum_{l=1}^{\infty} \rho_l$$

donde ρ_l es la autocorrelación con retraso l.

El tiempo de autocorrelación integrado es de interés debido a que controla el error estadístico en las estimaciones Monte Carlo de una función f. Supóngase que se tiene el estimador de Monte Carlo $\hat{f}$, su varianza es

$$\operatorname{Var}\left[\hat{f}\right] \approx \frac{2\tau}{M} V,$$

donde V es la varianza marginal de f y M es el número de iteraciones (efectivas). Como se observa, esta varianza es 2τ veces mayor que la varianza cuando las observaciones son independientes. Por tanto, una corrida de M iteraciones únicamente contiene $M/2\tau$ observaciones «efectivamente independientes», así que el método con el menor τ será más eficiente.

En cuanto a la estimación de τ, la mayor dificultad está en estimar las autocorrelaciones ρ_l. Kalli et al. (2009) utilizan el estimador

$$\hat{\tau} = \frac{1}{2} + \sum_{l=1}^{C-1} \hat{\rho}_l,$$

donde $\hat{\rho}_l$ es la autocorrelación muestral con retraso l y $C = \min\left\{l : |\hat{\rho}_l| < 2/\sqrt{M}\right\}$. (Se pueden consultar las referencias en Kalli et al. (2009) para mayor detalle.)

2.4 Densidad estimada

Cada uno de los métodos permiten obtener una estimación de la densidad de la que proviene la muestra. En la mayoría de los métodos, la densidad estimada está dada por

$$\hat{f}(x) = \frac{1}{M} \sum_{t=1}^{M} \sum_{j=1}^{K_n^{(t)}} w_j^{(t)} K(x; \theta_j^{(t)}), \tag{3}$$

donde $w_j^{(t)}$ es la proporción de las $K_n^{(t)}$ componentes, cada una con parámetros $\theta_j^{(t)}$ para el kernel K, en la iteración t. Para el caso del *Gibbs sampler* para el proceso geométrico, la densidad estimada está dada por

$$\hat{f}(x) = \frac{1}{Mn} \sum_{t=1}^{M} \sum_{i=1}^{n} \frac{1}{N_i^{(t)}} \sum_{j=1}^{N_i^{(t)}} K(x; \theta_j^{(t)}). \tag{4}$$

Además, es posible calcular la devianza de $\hat{f}$, que se obtiene haciendo $\hat{h} = \hat{f}$ en (1). La devianza calculada de esta manera permite hacer comparaciones entre modelos, eligiendo aquel con menor devianza, a diferencia de la calculada en (2), que permite monitorear la convergencia de la simulación.

§ 3 Configuración de parámetros

Los métodos de estimación estudiados en capítulos anteriores se compararon utilizando los modelos de mezclas de la Sección 1. El análisis se realizó de acuerdo a las siguientes configuraciones

Iteraciones. Cada método se corrió 10000 iteraciones tras un periodo de calentamiento de 5000.

Medida base. Los parámetros de la medida base fueron los mismos en todos los casos; estos se tomaron de acuerdo a lo explicado en Richardson & Green (1997): $\omega = M$, $t = R^2$, $a = 2$ y $b = 0.02R^2$, donde R es el rango de los datos y $M = \min\{x_i\} + R/2$.

Medidas de probabilidad aleatorias. Todos los métodos, a excepción del *Gibbs sampler* para el proceso geométrico, utilizan una medida de probabilidad aleatoria generada a través de alguno de los siguientes procesos: Dirichlet, σ–estable normalizado o Poisson–Dirichlet de dos parámetros; cualquiera de estos requiere de uno o dos parámetros, por tanto, se probaron todos los procesos en cada método de la siguiente manera

1. Fijando los parámetros de manera que $\mathbb{E}[K_n] = k$ *a priori*, para distintas k, dependiendo del modelo de mezcla;
2. Asignando distintas distribuciones iniciales a sus parámetros, incluyéndolos así en los esquemas de simulación (Sección III.2).

La esperanza *a priori* para K_n se obtuvo de las ecuaciones (I.17), (I.18) o (I.20), dependiendo del proceso. La asignación de distribuciones iniciales se hizo de la siguiente manera: para el parámetro θ se tomó una distribución Gamma y para σ una distribución Beta. Los parámetros de cada distribución se eligieron de manera que $\mathbb{E}[\theta] = \theta_k$, $\mathrm{Var}[\theta] = 2$, $\mathbb{E}[\sigma] = \sigma_k$ y $\mathrm{Var}[\sigma] = 0.005$, donde θ_k y σ_k fueron los valores obtenidos al resolver $\mathbb{E}[K_n] = k$. Asimismo, se tomaron distribuciones no informativas; para el caso de θ: $a = b = 0.00001$, y para σ: $a = b = 1$.

En el caso del proceso geométrico, la medida de probabilidad aleatoria se genera a través del parámetro λ, por lo que se le asignaron distintos parámetros a su distribución inicial. La elección de los parámetros se hizo de manera que $\mathbb{E}[\lambda] \in \{0.1, 0.5, 0.9\}$ y $\mathrm{Var}[\lambda] = 0.005$. Se tomó también una distribución no informativa: $a = b = 1$, así como los parámetros utilizados en Fuentes-García et al. (2010): $a = b = 0.5$.

Parámetros adicionales. Algunos métodos de estimación requieren de un parámetro adicional

1. El *Gibbs sampler* por bloques (Sección II.2.2) trabaja con un truncamiento de una medida *stick–breaking* utilizando únicamente sus primeros N términos. Esta N se toma de manera que se satisfaga

$$4\left\{1 - \sum_{k=0}^{n} (-1)^k \binom{n}{k} \prod_{i=1}^{N-1} \frac{(\theta + i\sigma)_{k\uparrow}}{(\theta + 1 + (i-1)\sigma)_{k\uparrow}}\right\} \leq \varepsilon, \tag{5}$$

donde ε es un error dado.

Como se explicó en la Sección I.2.4, no siempre es posible obtener una aproximación cuando $\sigma > 0$. Por tanto, se hicieron simulaciones obteniendo N para $\varepsilon \in \{10^{-6}, 10^{-10}\}$ utilizando los θ_k y σ_k obtenidos anteriormente; se tomó como máximo $N = 150$. Además se hizo una simulación fijando $N = 30$.

2. El método Metropolis–Hastings para distribuciones no conjugadas (Sección II.2.3) requiere de un parámetro r que indica el número de actualizaciones de K_i. En todos los modelos se hicieron simulaciones tomando los valores $r \in \{1, 5, 15\}$.

3. En el *Gibbs sampler* con parámetros auxiliares (Sección II.2.5), la probabilidad de crear una nueva componente se divide equiprobablemente en l. Se hicieron simulaciones tomando los valores $l \in \{1, 2, 10\}$.

En la Tabla A.1 se muestran los parámetros de la medida base para cada modelo, así como los valores de k utilizados en las simulaciones. Además, en las Tablas A.2 y A.3 se muestran los parámetros que se utilizaron para los datos simulados y los datos reales, respectivamente. Los hiperparámetros del *Gibbs sampler* para el proceso geométrico no dependen del tamaño de la muestra, por lo que se pueden utilizar los mismos para todos los modelos de mezclas; estos se muestran en la Tabla A.4[1].

§ 4 Resultados

Los resultados de las simulaciones se pueden consultar en el Apéndice A.2. Se muestran allí debido a la gran cantidad de gráficas y tablas resultantes. A continuación se presentan, para cada modelo de datos y para el conjunto de datos de las galaxias, las simulaciones seleccionadas de acuerdo a los siguientes criterios

1. Para cada proceso con el mismo $\mathbb{E}[K_n]$ *a priori*, se eligió el de menor devianza para la densidad estimada, calculada de acuerdo a (3); de igual manera para las simulaciones que actualizaron los parámetros de los procesos.

2. Con las simulaciones seleccionadas en el punto anterior, se seleccionaron aquellas con menor $\hat{\tau}_D$ y, por otro lado, aquellas con menor $\hat{\tau}_{K_n}$; obteniendo así, dos simulaciones por método.

3. Únicamente para cada una de las simulaciones seleccionadas, se muestra la densidad estimada y la distribución posterior de K_n.

4. En el caso de que en la simulación seleccionada se hayan actualizado los parámetros del proceso, se muestran sus distribuciones posteriores.

[1] Debido a la dimensión de algunas de las tablas, todas se colocaron en el Apéndice A; solamente se muestran aquellas que sean importantes para el desarrollo del capítulo.

Lo anterior no aplica al *Gibbs sampler* para el proceso geométrico, debido a que su estructura es distinta a los demás métodos. Por lo que para este método se muestran los resultados de todas sus configuraciones en el mismo apéndice, y en este capítulo únicamente se muestran sus densidades estimadas y las distribuciones posteriores de λ.

Conforme a lo que se explicó en la sección anterior, para el proceso Poisson–Dirichlet los métodos se corrieron con dos pares de parámetros: $(\sigma_1, \alpha/3)$ y $(\sigma_2, 2\alpha/3)$, donde α es el parámetro del proceso Dirichlet. Para facilitar la explicación en el resto del capítulo, se denotará como $\mathcal{PD}_\sigma$ al proceso Poisson–Dirichlet utilizando el primer par de parámetros y como $\mathcal{PD}_\theta$ utilizando el segundo par.

4.1 Modelo separado

Las simulaciones seleccionadas para el modelo de datos separado con menor $\hat{\tau}_D$ son

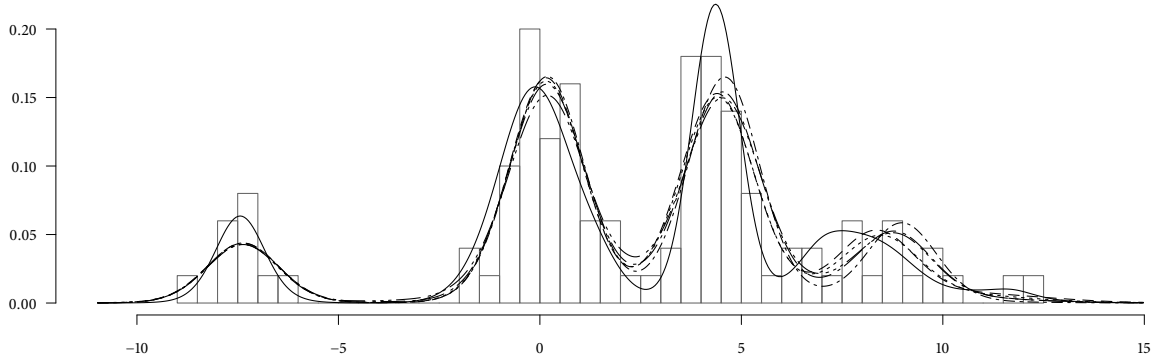
1. *Gibbs sampler* para urnas de Pólya; proceso Dirichlet y actualización de α con distribución inicial no informativa
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; $\mathbb{E}[K_n] = 2$, proceso $\mathcal{PD}_\sigma$ y actualización de (σ, θ)
3. Metropolis–Hastings y *Gibbs sampler*; proceso σ–estable normalizado y actualización de σ con distribución inicial no informativa
4. *Gibbs sampler* por bloques; $N = 48$, $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
5. Metropolis–Hastings para distribuciones no conjugadas; $r = 15$, $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
6. *Gibbs sampler* con 2 parámetros auxiliares; proceso $\mathcal{PD}_\sigma$ y actualización de (σ, θ) con distribuciones iniciales no informativas

Mientras que las de menor $\hat{\tau}_{K_n}$ son

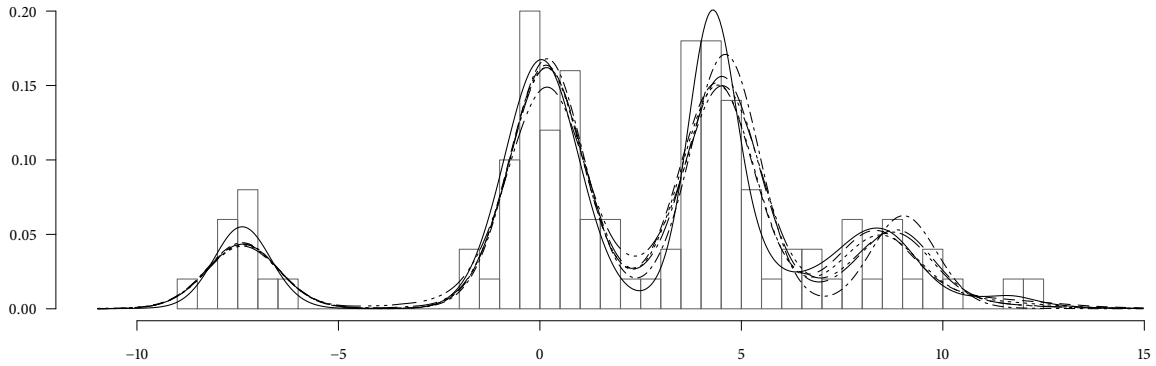
1. *Gibbs sampler* para urnas de Pólya; $\mathbb{E}[K_n] = 8$ y proceso σ–estable normalizado
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; $\mathbb{E}[K_n] = 8$ y proceso $\mathcal{PD}_\theta$
3. Metropolis–Hastings y *Gibbs sampler*; $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
4. *Gibbs sampler* por bloques; $N = 48$, $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
5. Metropolis–Hastings para distribuciones no conjugadas; $r = 15$, $\mathbb{E}[K_n] = 2$ y proceso σ–estable normalizado
6. *Gibbs sampler* con un parámetro auxiliar; $\mathbb{E}[K_n] = 4$ y proceso $\mathcal{PD}_\sigma$

La devianza estimada para cada simulación seleccionada se muestra en la siguiente tabla

	Devianza con menor	
Método	$\hat{\tau}_D$	$\hat{\tau}_{K_n}$
G. s. para urnas de Pólya	509.090	507.903
G. s. para urnas de Pólya con paso de aceleración	514.650	514.779
M. H. y *Gibbs sampler*	519.644	520.475
Gibbs sampler por bloques	519.482	519.482
M. H. para distr. no conjugadas	523.198	527.014
G. s. con parámetros auxiliares	518.670	518.550

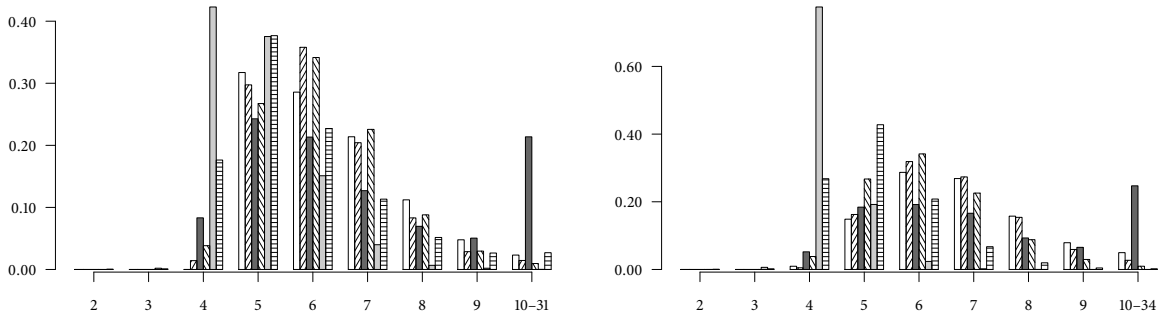


(A) Métodos con menor $\hat{\tau}_D$



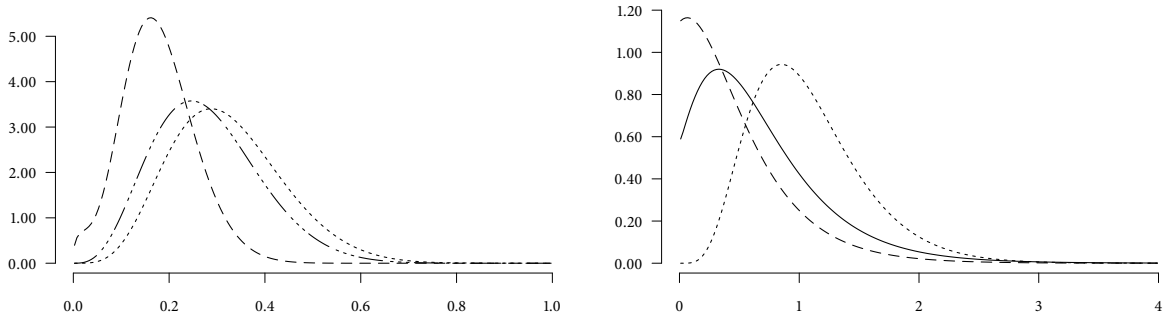
(B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 2: Densidades estimadas, modelo separado, para los distintos métodos: —— *G. s.* para urnas de Pólya, – – – *G. s.* para urnas de Pólya con paso de aceleración, — ··· M. H. y *Gibbs sampler* , ····· *Gibbs sampler* por bloques, – · – ·· M. H. para distr. no conjugadas, — · – *G. s.* con parámetros auxiliares.



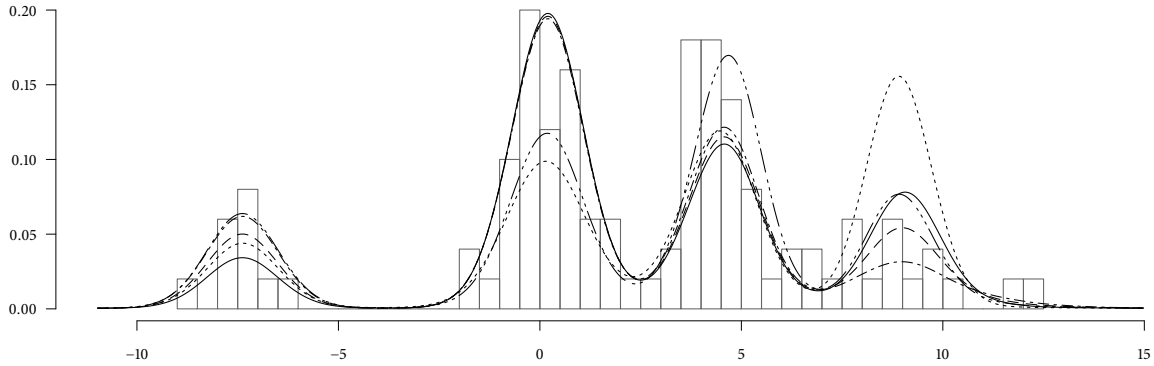
(A) Métodos con menor $\hat{\tau}_D$ (B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 3: Distribuciones posteriores de K_n, modelo separado, para los distintos métodos: ▭ *G. s.* para urnas de Pólya, ▨ *G. s.* para urnas de Pólya con paso de aceleración, ■ M. H. y *Gibbs sampler*, ▧ *Gibbs sampler* por bloques, ▭ M. H. para distr. no conjugadas, ▤ *G. s.* con parámetros auxiliares.

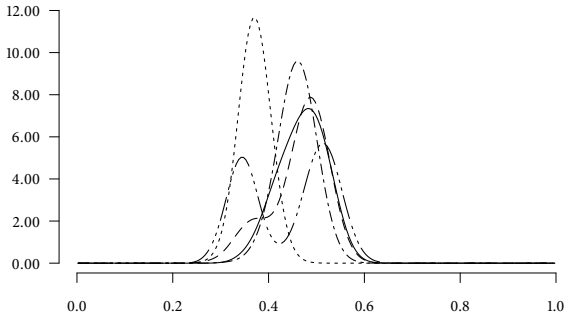


(A) Parámetro σ (B) Parámetro θ

FIGURA 4: Distribuciones posteriores de los parámetros de los procesos, modelo separado, para los métodos con menor $\hat{\tau}_D$: —— *G. s.* para urnas de Pólya, – – – *G. s.* para urnas de Pólya con paso de aceleración, — ··· M. H. y *Gibbs sampler* y ····· *G. s.* con parámetros auxiliares.



(A) Densidades estimadas



(B) Distribuciones posteriores de λ

FIGURA 5: Resultados del *Gibbs sampler* para el proceso geométrico, modelo separado, para los distintos hiperparámetros de λ: —— $(1.0, 1.0)$, – – – $(0.5, 0.5)$, — ··· $(15.3, 1.7)$, ····· $(24.5, 24.5)$, –·–·· $(1.7, 15.3)$.

4.2 Modelo platicúrtico

Las simulaciones seleccionadas para el modelo de datos platicúrtico con menor $\hat{\tau}_D$ son

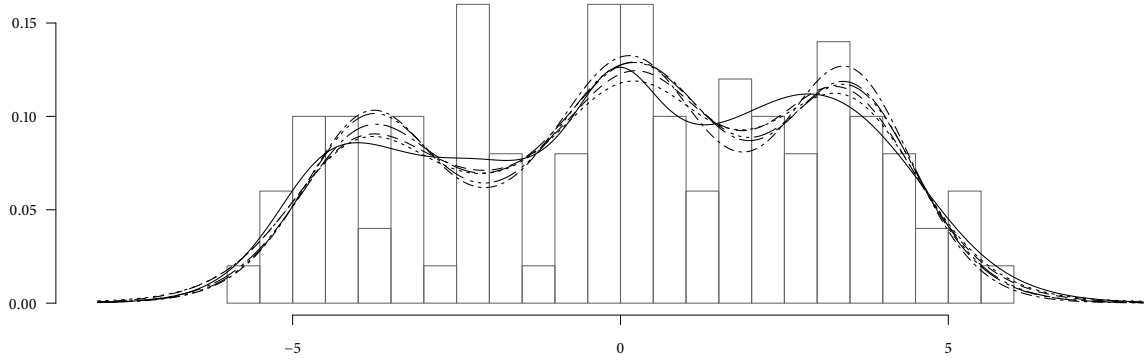
1. *Gibbs sampler* para urnas de Pólya; $\mathbb{E}[K_n] = 10$ y proceso $\mathcal{P}_{D\theta}$
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; $\mathbb{E}[K_n] = 10$ y proceso Dirichlet
3. Metropolis–Hastings y *Gibbs sampler*; $\mathbb{E}[K_n] = 2$ y proceso Dirichlet
4. *Gibbs sampler* por bloques; $N = 30$, $\mathbb{E}[K_n] = 10$ y proceso Dirichlet
5. Metropolis–Hastings para distr. no conjugadas; $r = 15$, $\mathbb{E}[K_n] = 2$ y proceso σ–estable normalizado
6. *Gibbs sampler* con 10 parámetros auxiliares; $\mathbb{E}[K_n] = 2$ y proceso σ–estable normalizado

Mientras que las de menor $\hat{\tau}_{K_n}$ son

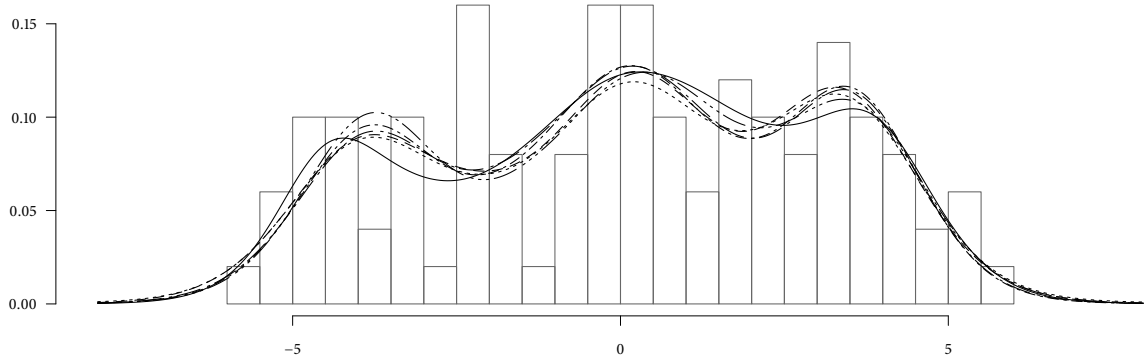
1. *Gibbs sampler* para urnas de Pólya; $\mathbb{E}[K_n] = 2$, proceso Dirichlet y actualización de α
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; $\mathbb{E}[K_n] = 10$ y proceso Dirichlet
3. Metropolis–Hastings y *Gibbs sampler*; $\mathbb{E}[K_n] = 10$ y proceso Dirichlet
4. *Gibbs sampler* por bloques; $N = 30$, $\mathbb{E}[K_n] = 10$ y proceso Dirichlet
5. Metropolis–Hastings para distribuciones no conjugadas; $r = 15$, $\mathbb{E}[K_n] = 10$ y proceso Dirichlet
6. *Gibbs sampler* con un parámetro auxiliar; $\mathbb{E}[K_n] = 10$ y proceso Dirichlet

La devianza estimada para cada simulación seleccionada se muestra en la siguiente tabla

	Devianza con menor	
Método	$\hat{\tau}_D$	$\hat{\tau}_{K_n}$
G. s. para urnas de Pólya	488.941	490.787
G. s. para urnas de Pólya con paso de aceleración	491.886	491.886
M. H. y *Gibbs sampler*	492.647	492.795
Gibbs sampler por bloques	495.253	495.253
M. H. para distr. no conjugadas	493.460	492.436
G. s. con parámetros auxiliares	492.335	492.421

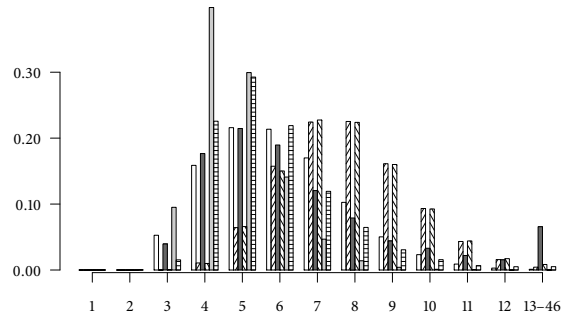


(A) Métodos con menor $\hat{\tau}_D$

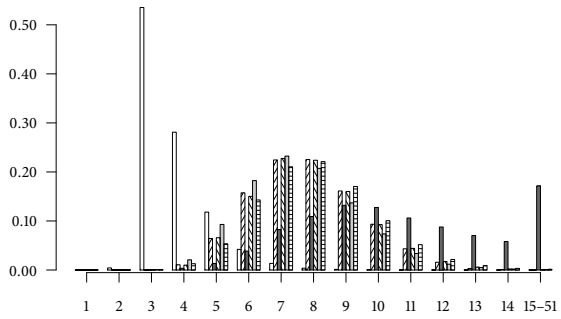


(B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 6: Densidades estimadas, modelo platicúrtico, para los distintos métodos: —— *G. s.* para urnas de Pólya, – – – *G. s.* para urnas de Pólya con paso de aceleración, — ··· M. H. y *Gibbs sampler* , ····· *Gibbs sampler* por bloques, –·–·· M. H. para distr. no conjugadas, — ·– *G. s.* con parámetros auxiliares.

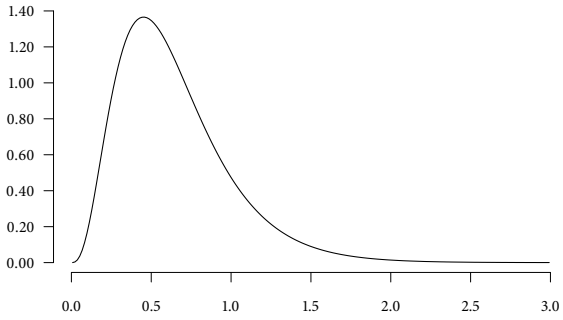


(A) Métodos con menor $\hat{\tau}_D$



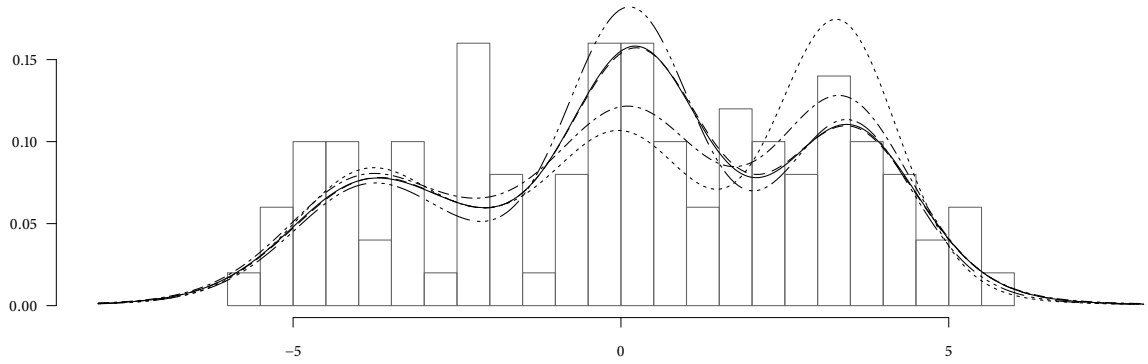
(B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 7: Distribuciones posteriores de K_n, modelo platicúrtico, para los distintos métodos: ▭ *G. s.* para urnas de Pólya, ▨ *G. s.* para urnas de Pólya con paso de aceleración, ■ M. H. y *Gibbs sampler*, ▩ *Gibbs sampler* por bloques, ▭ M. H. para distr. no conjugadas, ▤ *G. s.* con parámetros auxiliares.

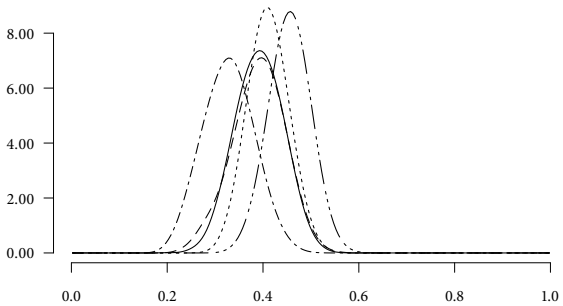


(A) Parámetro θ

FIGURA 8: Distribuciones posteriores de los parámetros de los procesos, modelo platicúrtico, para el método con menor $\hat{\tau}_{K_n}$: —— *G. s.* para urnas de Pólya.



(A) Densidades estimadas



(B) Distribuciones posteriores de λ

FIGURA 9: Resultados del *Gibbs sampler* para el proceso geométrico, modelo platicúrtico, para los distintos hiperparámetros de λ: —— $(1.0, 1.0)$, – – – $(0.5, 0.5)$, — ··· $(15.3, 1.7)$, ····· $(24.5, 24.5)$, – ·– ·· $(1.7, 15.3)$.

4.3 Modelo bimodal

Las simulaciones seleccionadas para el modelo de datos bimodal con menor $\hat{\tau}_D$ son

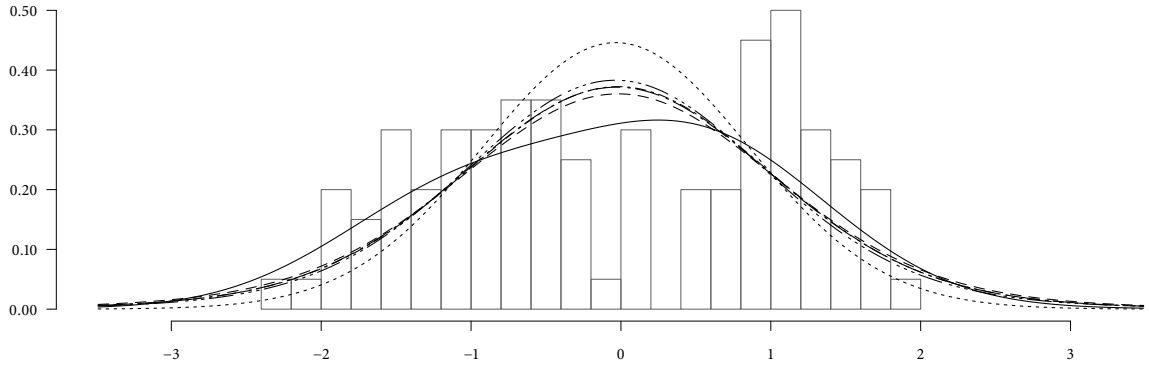
1. *Gibbs sampler* para urnas de Pólya; $\mathbb{E}[K_n] = 8$, proceso $\mathcal{P}_{D_\sigma}$ y actualización de (σ, θ)
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
3. Metropolis–Hastings y *Gibbs sampler*; $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
4. *Gibbs sampler* por bloques; $N = 43$, proceso Dirichlet y actualización de α con distribución inicial no informativa
5. Metropolis–Hastings para distribuciones no conjugadas; $r = 5$, $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
6. *Gibbs sampler* con 2 parámetros auxiliares; $\mathbb{E}[K_n] = 8$ y proceso Dirichlet

Mientras que las de menor $\hat{\tau}_{K_n}$ son

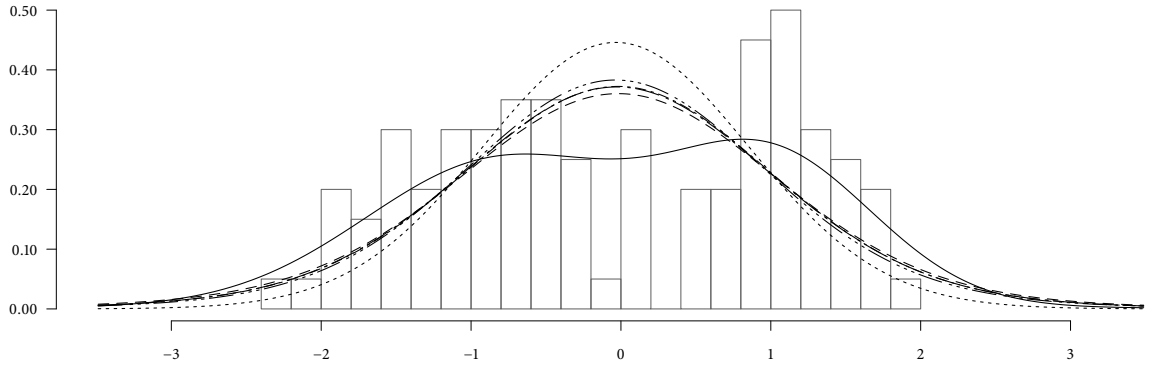
1. *Gibbs sampler* para urnas de Pólya; $\mathbb{E}[K_n] = 8$ y proceso $\mathcal{P}_{D_\theta}$
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
3. Metropolis–Hastings y *Gibbs sampler*; $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
4. *Gibbs sampler* por bloques; $N = 43$, proceso Dirichlet y actualización de α con distribución inicial no informativa
5. Metropolis–Hastings para distribuciones no conjugadas; $r = 5$, $\mathbb{E}[K_n] = 8$ y proceso Dirichlet
6. *Gibbs sampler* con 2 parámetros auxiliares; $\mathbb{E}[K_n] = 8$ y proceso Dirichlet

La devianza estimada para cada simulación seleccionada se muestra en la siguiente tabla

	Devianza con menor	
Método	$\hat{\tau}_D$	$\hat{\tau}_{K_n}$
G. s. para urnas de Pólya	303.400	297.342
G. s. para urnas de Pólya con paso de aceleración	316.345	316.345
M. H. y *Gibbs sampler*	317.239	317.239
Gibbs sampler por bloques	323.772	323.772
M. H. para distr. no conjugadas	315.991	315.991
G. s. con parámetros auxiliares	315.954	315.954

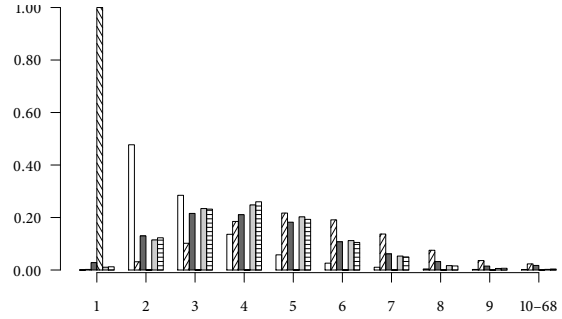


(A) Métodos con menor $\hat{\tau}_D$

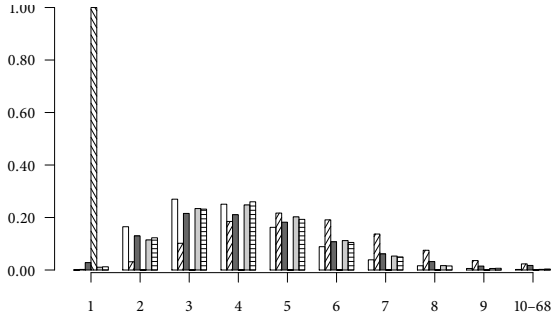


(B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 10: Densidades estimadas, modelo bimodal, para los distintos métodos: —— *G. s.* para urnas de Pólya, – – – *G. s.* para urnas de Pólya con paso de aceleración, — ··· M. H. y *Gibbs sampler* , ····· *Gibbs sampler* por bloques, –·–·– M. H. para distr. no conjugadas, — ·– *G. s.* con parámetros auxiliares.

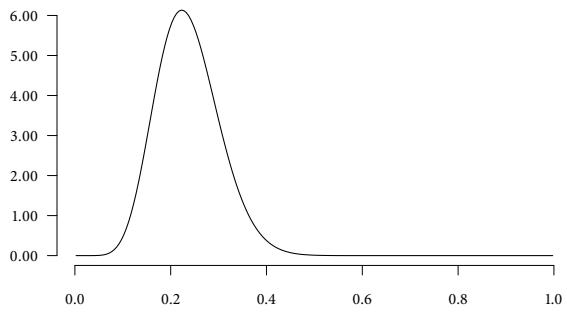


(A) Métodos con menor $\hat{\tau}_D$

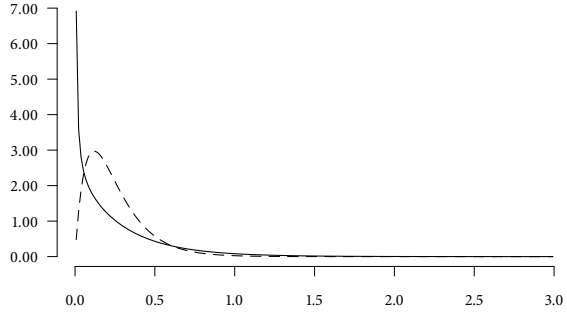


(B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 11: Distribuciones posteriores de K_n, modelo bimodal, para los distintos métodos: ▭ *G. s.* para urnas de Pólya, ▨ *G. s.* para urnas de Pólya con paso de aceleración, ■ M. H. y *Gibbs sampler*, ▧ *Gibbs sampler* por bloques, ▭ M. H. para distr. no conjugadas, ▤ *G. s.* con parámetros auxiliares.

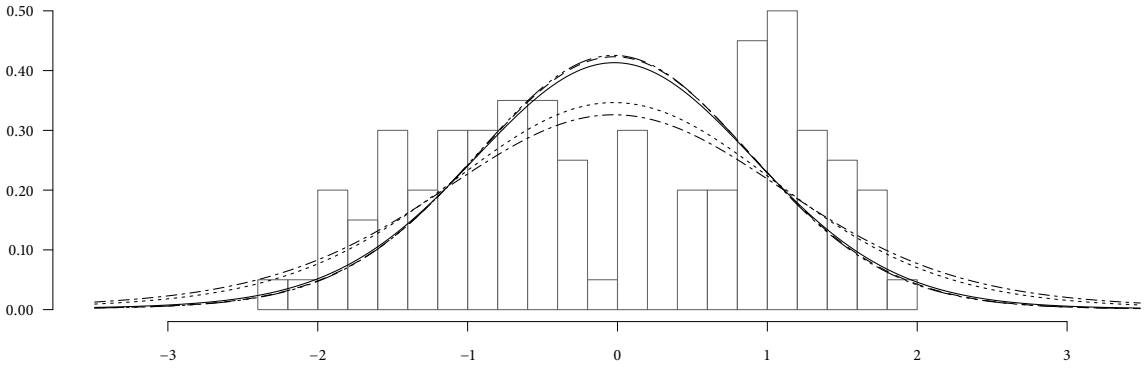


(A) Parámetro σ

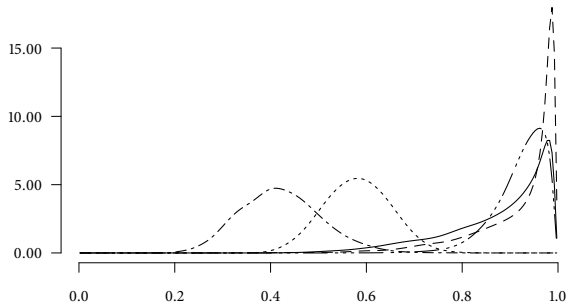


(B) Parámetro θ

FIGURA 12: Distribuciones posteriores de los parámetros de los procesos, modelo bimodal, para el método con menor $\hat{\tau}_D$ y $\hat{\tau}_{K_n}$: – – – *Gibbs sampler* por bloques; y menor $\hat{\tau}_{K_n}$: —— *G. s.* para urnas de Pólya.



(A) Densidades estimadas



(B) Distribuciones posteriores de λ

FIGURA 13: Resultados del *Gibbs sampler* para el proceso geométrico, modelo bimodal, para los distintos hiperparámetros de λ: —— $(1.0, 1.0)$, – – – $(0.5, 0.5)$, — ··· $(15.3, 1.7)$, ····· $(24.5, 24.5)$, – · – · $(1.7, 15.3)$.

4.4 Modelo leptocúrtico

Las simulaciones seleccionadas para el modelo de datos leptocúrtico con menor $\hat{\tau}_D$ son

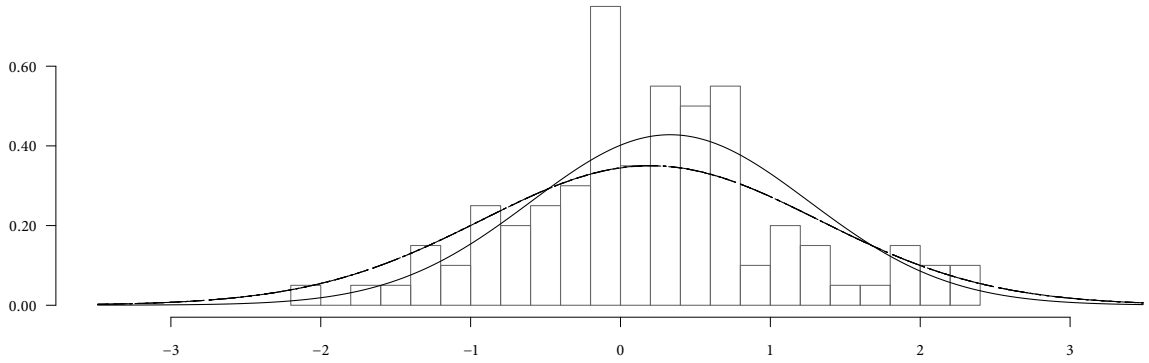
1. *Gibbs sampler* para urnas de Pólya; proceso $\mathcal{P}_{D_\sigma}$ y actualización de (σ, θ) con distribuciones iniciales no informativas
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; $\mathbb{E}[K_n] = 2$ y proceso $\mathcal{P}_{D_\sigma}$
3. Metropolis–Hastings y *Gibbs sampler*; $\mathbb{E}[K_n] = 2$ y proceso σ–estable normalizado
4. *Gibbs sampler* por bloques; $N = 30$, $\mathbb{E}[K_n] = 2$ y proceso σ–estable normalizado
5. Metropolis–Hastings para distribuciones no conjugadas; $r = 15$, $\mathbb{E}[K_n] = 2$ y proceso $\mathcal{P}_{D_\sigma}$
6. *Gibbs sampler* con 10 parámetros auxiliares; $\mathbb{E}[K_n] = 2$ y proceso σ–estable normalizado

Mientras que las de menor $\hat{\tau}_{K_n}$ son

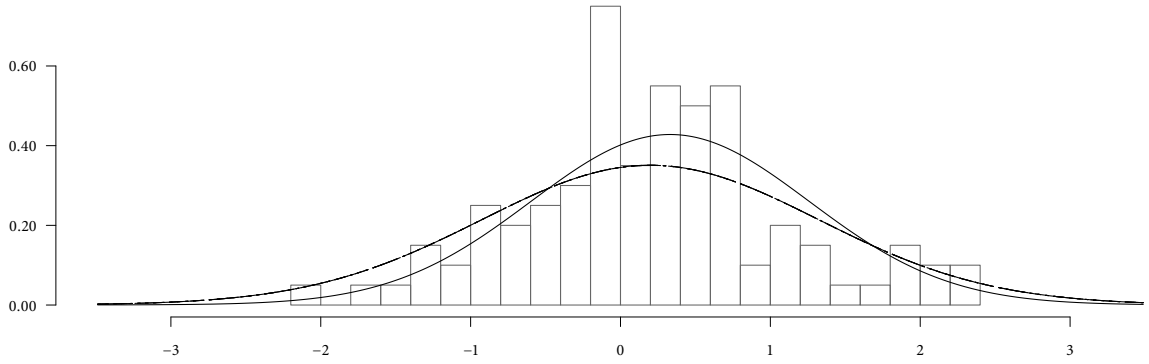
1. *Gibbs sampler* para urnas de Pólya; proceso $\mathcal{P}_{D_\sigma}$ y actualización de (σ, θ) con distribuciones iniciales no informativas
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; proceso Dirichlet y actualización de α con distribución inicial no informativa
3. Metropolis–Hastings y *Gibbs sampler*; proceso Dirichlet y actualización de α con distribución inicial no informativa
4. *Gibbs sampler* por bloques; $N = 30$, $\mathbb{E}[K_n] = 2$ y proceso σ–estable normalizado
5. Metropolis–Hastings para distribuciones no conjugadas; $r = 5$, proceso Dirichlet y actualización de α con distribución inicial no informativa
6. *Gibbs sampler* con 2 parámetros auxiliares; proceso Dirichlet y actualización de α con distribución inicial no informativa

La devianza estimada para cada simulación seleccionada se muestra en la siguiente tabla

	Devianza con menor	
Método	$\hat{\tau}_D$	$\hat{\tau}_{K_n}$
G. s. para urnas de Pólya	262.690	262.690
G. s. para urnas de Pólya con paso de aceleración	271.049	270.795
M. H. y *Gibbs sampler*	270.944	270.780
Gibbs sampler por bloques	270.817	270.817
M. H. para distr. no conjugadas	270.905	270.716
G. s. con parámetros auxiliares	270.887	270.756

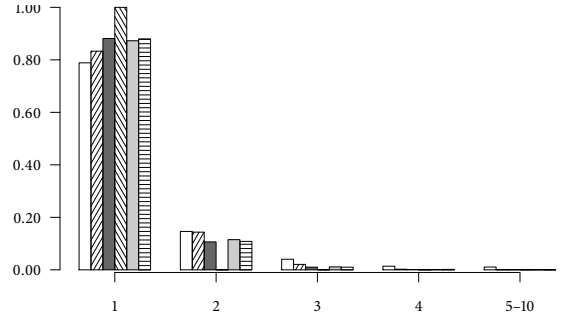


(A) Métodos con menor $\hat{\tau}_D$

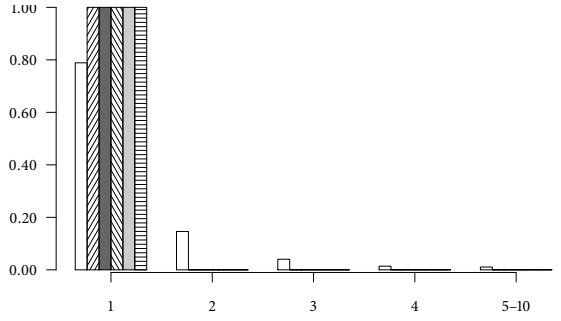


(B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 14: Densidades estimadas, modelo leptocúrtico, para los distintos métodos: —— *G. s.* para urnas de Pólya, – – – *G. s.* para urnas de Pólya con paso de aceleración, — ··· M. H. y *Gibbs sampler* , ····· *Gibbs sampler* por bloques, –·–·· M. H. para distr. no conjugadas, — ·– *G. s.* con parámetros auxiliares.

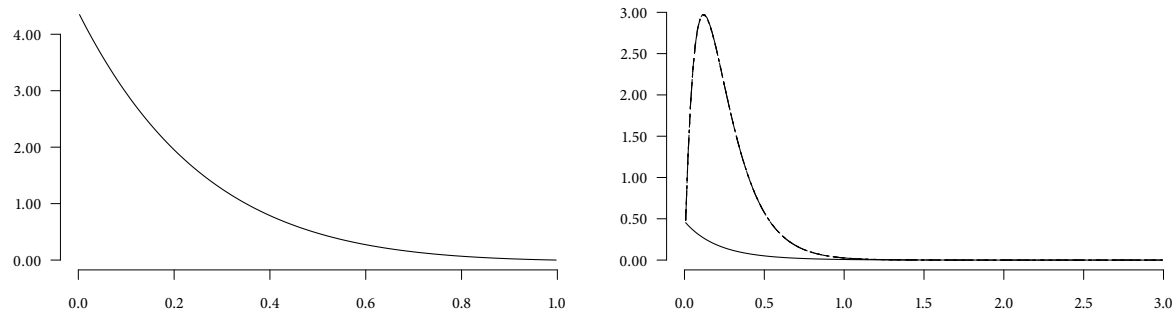


(A) Métodos con menor $\hat{\tau}_D$



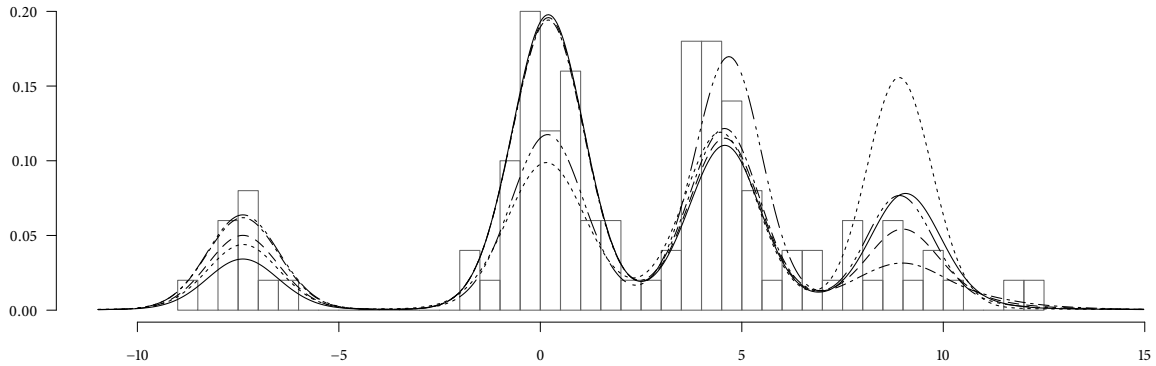
(B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 15: Distribuciones posteriores de K_n, modelo leptocúrtico, para los distintos métodos: ▭ *G. s.* para urnas de Pólya, ▨ *G. s.* para urnas de Pólya con paso de aceleración, ▬ M. H. y *Gibbs sampler*, ▧ *Gibbs sampler* por bloques, ▭ M. H. para distr. no conjugadas, ▤ *G. s.* con parámetros auxiliares.

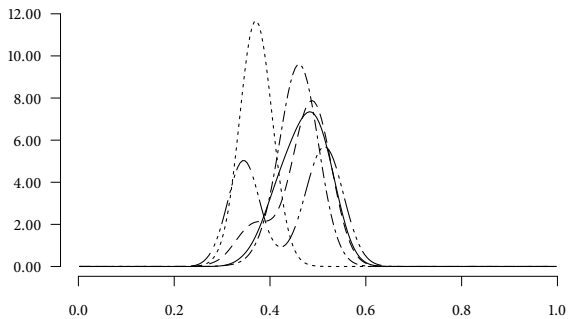


(A) Parámetro σ (B) Parámetro θ

FIGURA 16: Distribuciones posteriores de los parámetros de los procesos, modelo leptocúrtico, para el método con menor $\hat{\tau}_D$ y $\hat{\tau}_{K_n}$: —— *G. s.* para urnas de Pólya; y menor $\hat{\tau}_{K_n}$: – – – *G. s.* para urnas de Pólya con paso de aceleración, — ··· M. H. y *Gibbs sampler*, ····· M. H. para distr. no conjugadas y –·–·· *G. s.* con parámetros auxiliares.



(A) Densidades estimadas



(B) Distribuciones posteriores de λ

FIGURA 17: Resultados del *Gibbs sampler* para el proceso geométrico, modelo leptocúrtico, para los distintos hiperparámetros de λ: —— $(1.0, 1.0)$, – – – $(0.5, 0.5)$, — ··· $(15.3, 1.7)$, ····· $(24.5, 24.5)$, –·–·· $(1.7, 15.3)$.

4.5 Datos de las galaxias

Las simulaciones seleccionadas para el conjunto de datos de las galaxias con menor $\hat{\tau}_D$ son

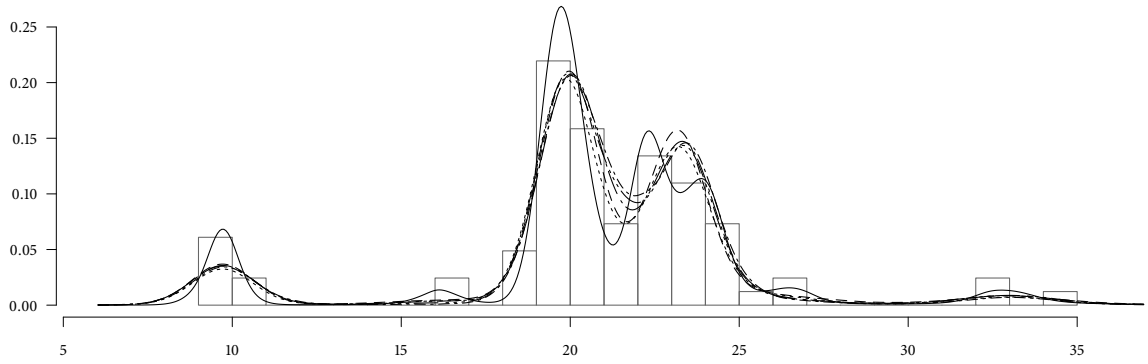
1. *Gibbs sampler* para urnas de Pólya; $\mathbb{E}[K_n] = 12$, proceso σ–estable normalizado y actualización de σ
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; $\mathbb{E}[K_n] = 3$ y proceso Dirichlet
3. Metropolis–Hastings y *Gibbs sampler*; $\mathbb{E}[K_n] = 3$ y proceso σ–estable normalizado
4. *Gibbs sampler* por bloques; $N = 30$, $\mathbb{E}[K_n] = 12$, proceso Dirichlet y actualización de α
5. Metropolis–Hastings para distribuciones no conjugadas; $r = 15$, $\mathbb{E}[K_n] = 12$ y proceso Dirichlet
6. *Gibbs sampler* con 10 parámetros auxiliares; $\mathbb{E}[K_n] = 6$ y proceso Dirichlet

Mientras que las de menor $\hat{\tau}_{K_n}$ son

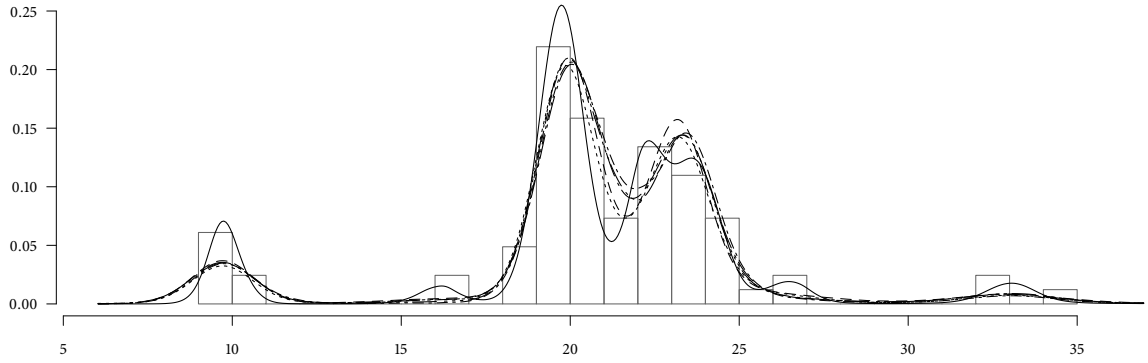
1. *Gibbs sampler* para urnas de Pólya; proceso Dirichlet y actualización de α con distribución inicial no informativa
2. *Gibbs sampler* para urnas de Pólya con paso de aceleración; $\mathbb{E}[K_n] = 3$ y proceso Dirichlet
3. Metropolis–Hastings y *Gibbs sampler*; $\mathbb{E}[K_n] = 6$, proceso Dirichlet y actualización de α
4. *Gibbs sampler* por bloques; $N = 30$, $\mathbb{E}[K_n] = 12$, proceso Dirichlet y actualización de α
5. Metropolis–Hastings para distribuciones no conjugadas; $r = 15$, $\mathbb{E}[K_n] = 12$ y proceso Dirichlet
6. *Gibbs sampler* con 10 parámetros auxiliares; $\mathbb{E}[K_n] = 12$ y proceso $\mathcal{PD}_\sigma$

La devianza estimada para cada simulación seleccionada se muestra en la siguiente tabla

	Devianza con menor	
Método	$\hat{\tau}_D$	$\hat{\tau}_{K_n}$
G. s. para urnas de Pólya	388.213	387.137
G. s. para urnas de Pólya con paso de aceleración	402.118	402.118
M. H. y *Gibbs sampler*	405.002	405.250
Gibbs sampler por bloques	410.614	410.614
M. H. para distr. no conjugadas	408.199	408.199
G. s. con parámetros auxiliares	403.635	404.053

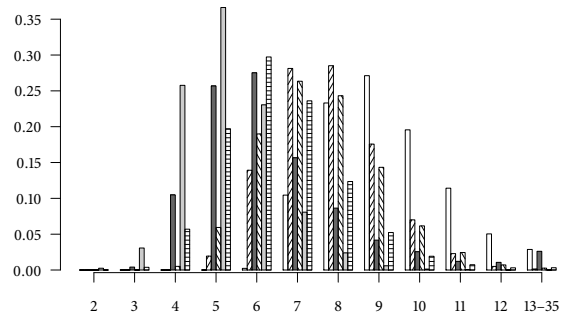


(A) Métodos con menor $\hat{\tau}_D$

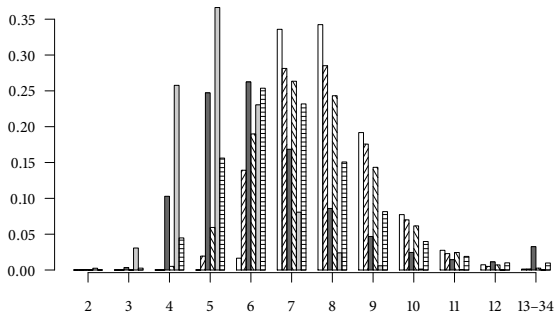


(B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 18: Densidades estimadas, datos de las galaxias, para los distintos métodos: —— *G. s.* para urnas de Pólya, – – – *G. s.* para urnas de Pólya con paso de aceleración, — ··· M. H. y *Gibbs sampler* , ····· *Gibbs sampler* por bloques, – · – · · M. H. para distr. no conjugadas, — · – *G. s.* con parámetros auxiliares.

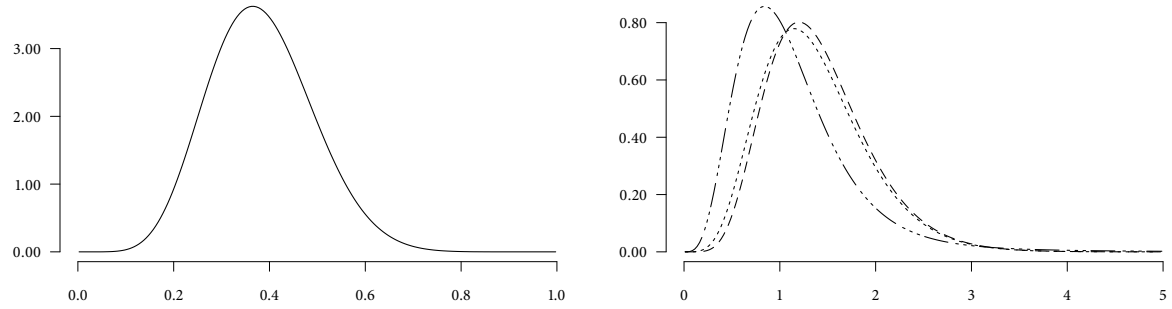


(A) Métodos con menor $\hat{\tau}_D$



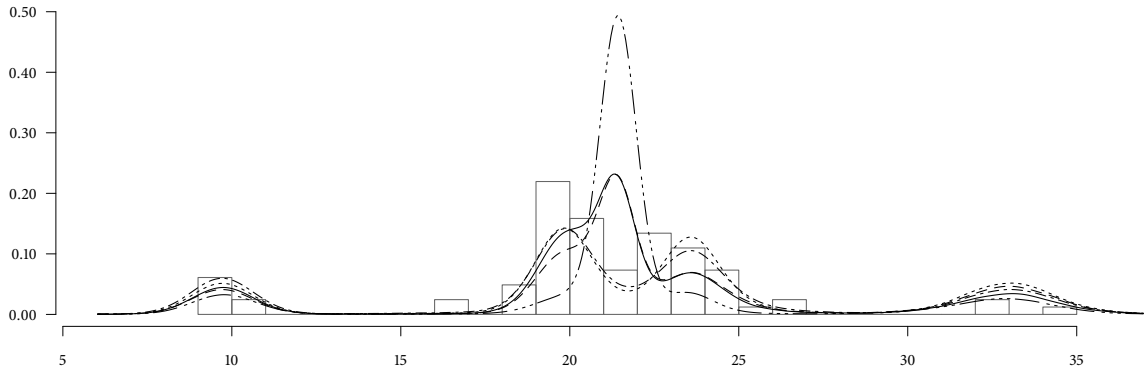
(B) Métodos con menor $\hat{\tau}_{K_n}$

FIGURA 19: Distribuciones posteriores de K_n, datos de las galaxias, para los distintos métodos: ▭ *G. s.* para urnas de Pólya, ▨ *G. s.* para urnas de Pólya con paso de aceleración, ■ M. H. y *Gibbs sampler*, ▧ *Gibbs sampler* por bloques, ▭ M. H. para distr. no conjugadas, ▤ *G. s.* con parámetros auxiliares.

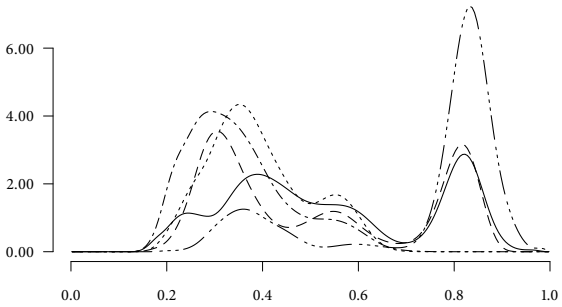


(A) Parámetro σ (B) Parámetro θ

FIGURA 20: Distribuciones posteriores de los parámetros de los procesos, datos de las galaxias, para los métodos con menor $\hat{\tau}_D$: —— *G. s.* para urnas de Pólya; con menor $\hat{\tau}_{K_n}$: – – – *G. s.* para urnas de Pólya y — ··· M. H. y *Gibbs sampler*; y con ambos menores: – – – *Gibbs sampler* por bloques.



(A) Densidades estimadas



(B) Distribuciones posteriores de λ

FIGURA 21: Resultados del *Gibbs sampler* para el proceso geométrico, datos de las galaxias, para los distintos hiperparámetros de λ: —— $(1.0, 1.0)$, – – – $(0.5, 0.5)$, — ··· $(15.3, 1.7)$, ····· $(24.5, 24.5)$, – · – · $(1.7, 15.3)$.

4.6 Comportamiento de los métodos

Una vez presentados los resultados de las simulaciones hechas para todos los métodos, se pueden hacer algunas observaciones que permitan hacer una comparación entre cada uno de ellos. En primer lugar se harán comentarios sobre los primeros seis métodos estudiados, de esta manera se puede estudiar el comportamiento de cada método; posteriormente se harán algunas observaciones sobre los conjuntos de datos.

Por último se revisarán los resultados del *Gibbs sampler* para el proceso geométrico. Como se ha explicado, la estructura de este método es distinta, por lo que no es posible compararlo directamente con los demás.

Gibbs sampler para urnas de Pólya

Los resultados del *Gibbs sampler* para urnas de Pólya no permiten distinguir el efecto de la medida de probabilidad aleatoria o sus parámetros. Sin embargo, algo que se puede observar es que los tiempos de autocorrelación integrado son los más altos.

En cuanto a la distribución posterior de K_n, el efecto de los valores iniciales para α, σ y (σ, θ) se pierde, obteniendo, por lo general, la misma moda para cada simulación.

También se puede observar que en las densidades estimadas para el modelo bimodal, una de las simulaciones seleccionada genera la densidad bimodal y, en general, las modas tienen mayor densidad. Esto se puede atribuir a que, como se comentó en la Sección II.2.1, el método puede quedarse estancado en un mismo conjunto de parámetros (μ, τ).

Paso de aceleración

Al incluir el paso de aceleración se observa que los procesos Dirichlet y $\mathcal{P}_{D_\theta}$ mejoran la devianza estimada al compararlos con el otro par de procesos. Asimismo, los tiempos de autocorrelación integrado disminuyen notablemente.

En cuanto a las densidades estimadas, el paso de aceleración las modifica, debido a la actualización de (μ, τ) que se hace para cada grupo.

Metropolis–Hastings y *Gibbs sampler*

Los resultados de este método no permiten distinguir el efecto de la medida de probabilidad aleatoria o sus parámetros. Los tiempos de autocorrelación integrado, tampoco difieren mucho de los demás métodos. Sin embargo, una observación importante es que la distribución posterior de K_n, en general tiene una forma platicúrtica y con un soporte, por lo general, mayor que para el resto de los métodos.

Gibbs sampler por bloques

Por lo que se puede observar en las devianzas estimadas, este método funciona mejor cuando se trabaja con el proceso Dirichlet o con el proceso $\mathcal{P}_{D_\theta}$. Además se puede observar que cuando se utiliza el

proceso Dirichlet de manera que α es tal que $E[K_n]$ *a priori* corresponde al «verdadero» número de grupos, los resultados son mejores. También se observa que los tiempos de autocorrelación integrados son comparables con el resto de los métodos, sólo resultando un poco mayores en $\hat{\tau}_{K_n}$.

En cuanto al truncamiento de la medida *stick–breaking*, parece no tener mucho efecto en general, ya que en muchos de los resultados, la devianza estimada, la distribución posterior de K_n o la densidad estimada para las configuraciones que varían sólo en N son muy parecidos.

Sobre el número de grupos, la información inicial dada en los parámetros de las medidas de probabilidad aleatorias, parece influir en la distribución posterior, lo cual desaparece al asignarles distribuciones iniciales. Por otro lado, esta información inicial sí afecta, en general, a la devianza estimada, ya que conforme los valores de los parámetros aumentan $\mathbb{E}[K_n]$ *a priori*, la devianza estimada también lo hace.

Por último, las densidades estimadas son prácticamente iguales, sin importar la configuración de los parámetros y son parecidas a las de otros métodos.

Metropolis–Hastings para distribuciones no conjugadas

La principal observación sobre este método es el efecto del número de actualizaciones de las K_i's, mientras más se hicieron, la devianza estimada fue menor; además esta diferencia fue más notoria si los datos formaban grupos separados (por ejemplo, los modelos separado y bimodal, incluso se observa en los datos de las galaxias). El efecto de las actualizaciones se refleja también en la densidad estimada.

En cuanto al tiempo de autocorrelación integrado en ambos casos es, por lo general, parecido a los demás métodos. Además, se puede observar que la devianza estimada, en general, disminuye ligeramente conforme los valores de los parámetros de la medida de probabilidad aleatoria incrementan $\mathbb{E}[K_n]$ *a priori*; sin embargo, esto no tiene efecto en la distribución posterior de K_n.

Gibbs sampler con parámetros auxiliares

Para el método de *Gibbs sampler* con parámetros auxiliares se observa que el número de parámetros auxiliares afecta a la devianza estimada, ya que cuando se corrieron las simulaciones con 10 parámetros, las devianzas fueron mayores.

También se observa un comportamiento parecido al método Metropolis–Hastings para distribuciones no conjugadas, pero con respecto a la medida de probabilidad aleatoria, cuando se tienen grupos separados, el proceso Dirichlet y el proceso $\mathcal{PD}_\theta$ son, en general, mejores.

Modelos de datos

Además del comportamiento de cada método, es posible observar el comportamiento en general de todos ellos para un mismo conjunto de datos.

Para cada modelo de datos, como se explicó al principio de este capítulo, se tomó una muestra de tamaño 100, por lo que es posible calcular la devianza poblacional para cada muestra y compararla con la devianza estimada. Las devianzas poblacionales son

Modelo	*Devianza*
Separado	521.7788
Platicúrtico	492.4947
Bimodal	281.9806
Leptocúrtico	263.5306

Comparando las devianzas con las estimaciones, se puede observar que para los primeros dos modelos, en general, los métodos tienen una devianza estimada muy parecida a la poblacional; sólo en los últimos dos modelos, las devianzas estimadas son mayores.

Una posible explicación para este comportamiento sería la separación entre grupos. Los métodos son más eficientes cuando existen grupos separados y en casos como el modelo leptocúrtico, que los grupos están más cercanos, es difícil ajustarlos. La distribución posterior de K_n puede ayudar también a observar esto, la mayoría de las simulaciones para el modelo leptocúrtico obtienen una distribución posterior con moda en $k = 1$ con probabilidad cercana a uno.

También es importante mencionar es el efecto de asignar distribuciones iniciales a los parámetros de los procesos de las medidas de probabilidad aleatorias; observando las devianzas estimadas, se tiene que en pocos casos se mejora al actualizarlos.

Gibbs sampler para el proceso geométrico

El *Gibbs sampler* para el proceso geométrico, como se ha comentado en capítulos anteriores, es un modelo nuevo y tiene una estructura diferente a los demás métodos estudiados. Como se explicó en el Capítulo 1, este modelo se puede pensar como un modelo de mezcla de procesos Dirichlet para el cual se sustituyen sus pesos por unos ordenados, lo que se logra tomando sus esperanzas. El resultado es un modelo de «mezcla de mezcla de distribuciones» con pesos aleatorios y ordenados. Además, debido a su novedad, aún no se ha estudiado si existe alguna variable que determine el número de grupos.

Considerando esto, se pueden analizar los resultados de las simulaciones. En general se observa, para todos los modelos de datos, que las devianzas estimadas se encuentran por arriba de los demás resultados; además, a pesar de que las densidades muestran las modas localizadas en las mismas posiciones con respecto a los demás métodos, los distintos hiperparámetros modifican su densidad. En cuanto al cambio tan marcado en las densidades, se puede justificar al observar la distribución posterior de λ, ya que éste determina el peso que se le da a cada par de parámetros (μ, τ).

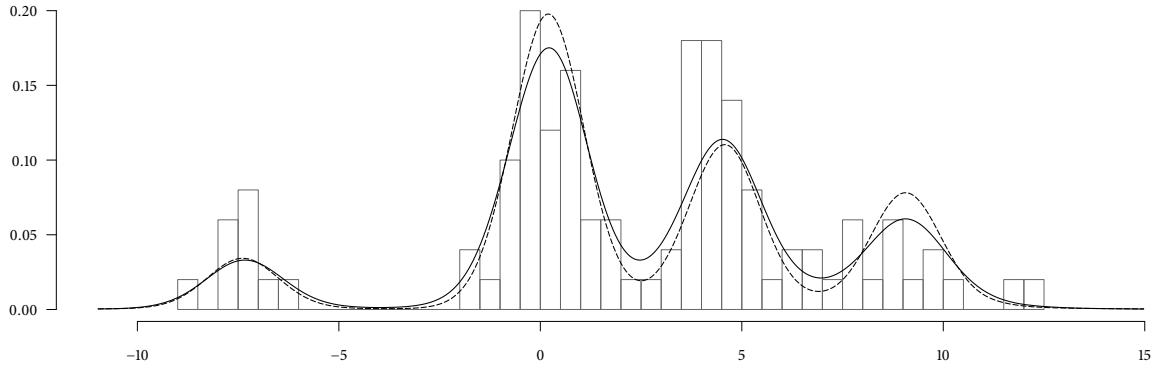
En general se podría pensar que las simulaciones no han convergido aún. Sin embargo, es importante analizar la estructura del modelo, ya que al ser más simple, cada parámetro puede tener más efecto en los resultados. Comparándolo con los demás métodos, el proceso geométrico, además de encontrar los parámetros de cada «grupo», los debe encontrar «en el orden correcto», i.e., el primer parámetro debe corresponder al «grupo» con mayor densidad, el segundo parámetro al segundo «grupo» con mayor densidad y así sucesivamente, debido al orden decreciente de sus pesos y a que sólo existe un parámetro que

los determina.

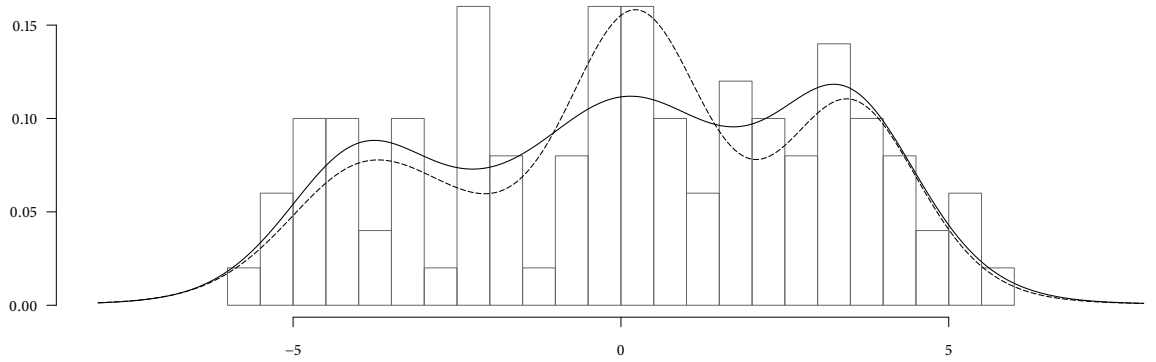
Entre las opciones que existen para mejorar los resultados de este método, se puede pensar en correr el método durante más iteraciones, asignar distribuciones iniciales a los parámetros de la distribución de λ o modificar los parámetros de la medida base. Con respecto al último punto, Richardson & Green (1997) han estudiado su influencia para modelos de mezcla de procesos Dirichlet y hacen énfasis en que modificar estos parámetros puede afectar drásticamente las inferencias. Por tanto, se hicieron nuevamente corridas para este modelo modificando los parámetros t y b de la medida base, haciendo $t = R$ y $b = 0.2R$ y tomando una distribución inicial no informativa para λ, i.e., una distribución Beta con parámetros $(1, 1)$. Los resultados se muestran en la Tabla 1 y en la Figura 22.

Modelo	*Devianza*	$\hat{\tau}_D$
Separado	525.677	2.024
	535.119	7.364
Platicúrtico	493.351	2.669
	497.712	3.104
Bimodal	317.585	9.701
	320.083	4.428
Leptocúrtico	272.200	1.375
	272.769	1.330
Galaxias	408.015	7.064
	440.114	148.590

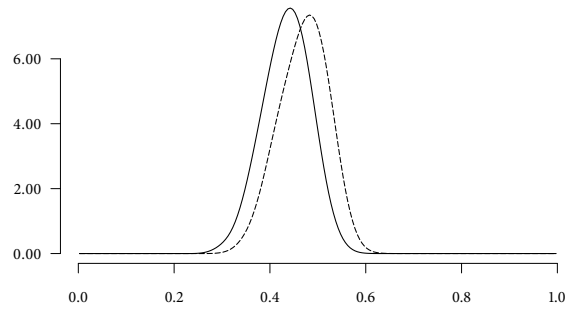
TABLA 1: Resultados de las nuevas simulaciones para el proceso geométrico; el primer renglón corresponde a la simulación con la medida base modificada, el segundo corresponde a la primera medida base.



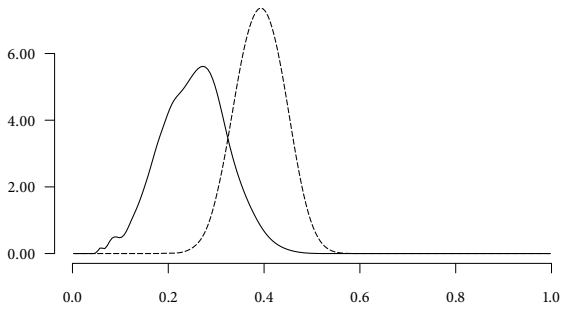
(A) Densidad estimada, modelo separado



(B) Densidad estimada, modelo platicúrtico

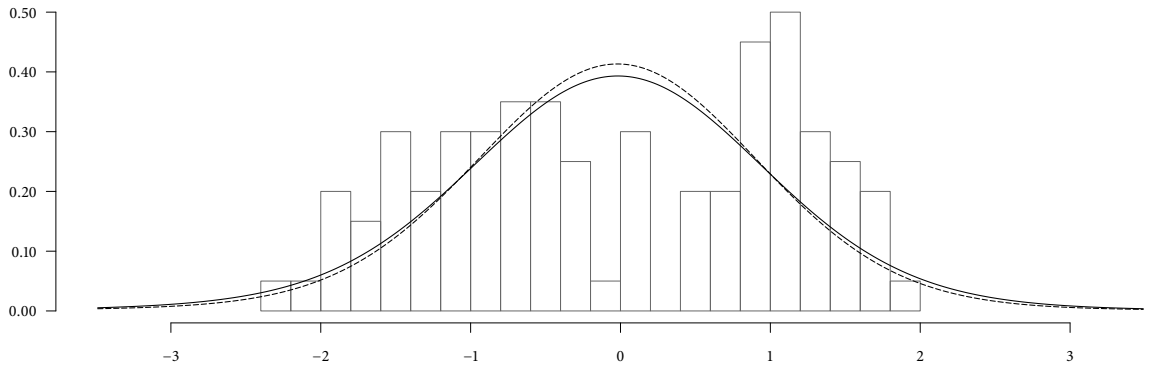


(C) Distribución posterior, modelo separado

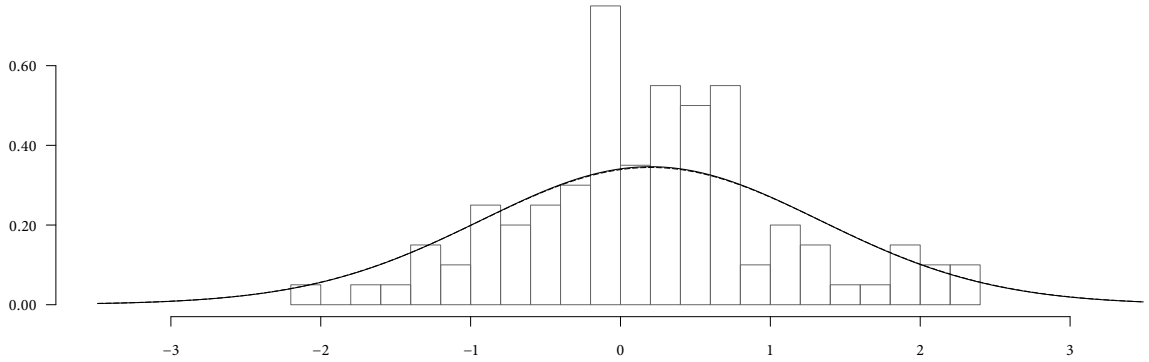


(D) Distribución posterior, modelo platicúrtico

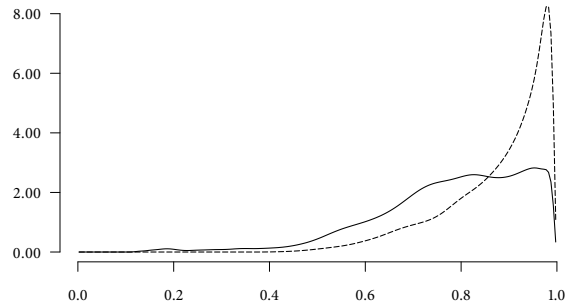
FIGURA 22: Densidad estimada y distribución posterior de λ para cada conjunto de datos; la línea —— corresponde a la medida base modificada, mientras que la línea ------- corresponde a la primera medida base.



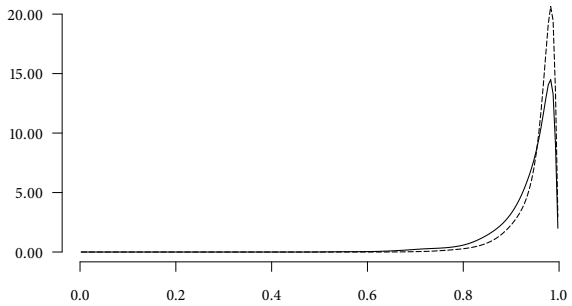
(E) Densidad estimada, modelo bimodal



(F) Densidad estimada, modelo leptocúrtico

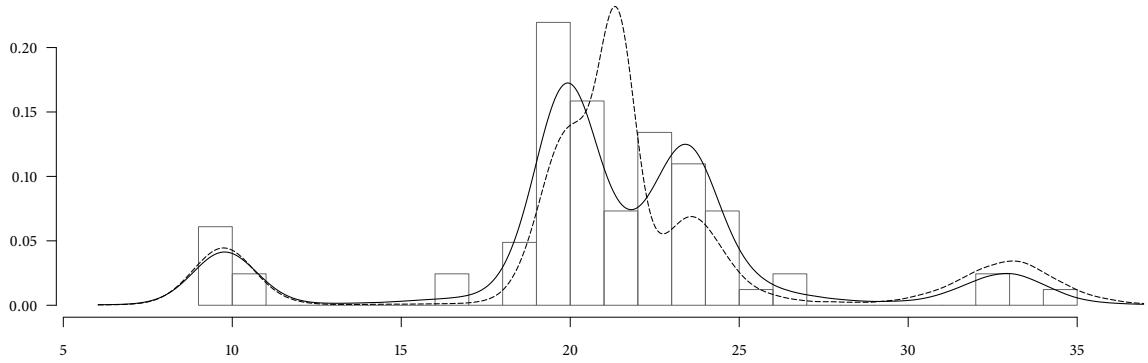


(G) Distribución posterior, modelo bimodal

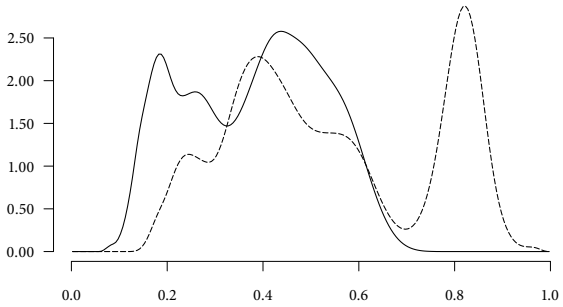


(H) Distribución posterior, modelo leptocúrtico

FIGURA 22: Densidad estimada y distribución posterior de λ para cada conjunto de datos, continuación; la línea —— corresponde a la medida base modificada, mientras que la línea ------ corresponde a la primera medida base.



(1) Densidad estimada, datos de las galaxias



(1) Distribución posterior, datos de las galaxias

FIGURA 22: Densidad estimada y distribución posterior de λ para cada conjunto de datos, continuación; la línea —— corresponde a la medida base modificada, mientras que la línea ------- corresponde a la primera medida base.

Como se observa, los resultados obtenidos al modificar la medida base, en general, cambian considerablemente, excepto por el modelo leptocúrtico. La devianza estimada y $\hat{\tau}_D$ son menores. En cuanto a la densidad estimada, los cambios más notables se tienen para el modelo platicúrtico y para el conjunto de datos de las galaxias, lo cual también se puede observar en la distribución posterior de λ. Por tanto, se puede decir que la medida base para el proceso geométrico tiene fuerte impacto en las inferencias.

Conclusiones

Los modelos bayesianos no paramétricos ofrecen una mayor flexibilidad al no hacer demasiados supuestos sobre la distribución de los datos; sin embargo, su uso se vio frenado debido a que se trabaja sobre un espacio demasiado grande, el espacio de distribuciones, y obtener una distribución posterior no era tarea fácil. Actualmente, gracias al uso de medidas de probabilidad aleatorias y a los métodos computacionales MCMC, el uso de estos modelos ha ido en aumento.

En cuanto a las medidas de probabilidad aleatorias, en este trabajo se estudiaron algunas extensiones del proceso Dirichlet, proceso que se considera fundamental para la estadística bayesiana no paramétrica: procesos con incrementos independientes normalizados, proceso Poisson–Dirichlet de dos parámetros y representación *stick–breaking*. Los primeros permiten incorporar información a través de su medida de Lévy, como el proceso Gamma normalizado (que corresponde al proceso Dirichlet) y el proceso σ–estable normalizado. Por otro lado, el proceso Poisson–Dirichlet de dos parámetros surge del estudio de un problema distinto, pero sus medidas canónicas coinciden con los antes mencionados; de esta manera se puede combinar la información disponible a través de sus dos parámetros. Por último, la representación *stick–breaking* es una forma más general (y posiblemente más simple) de crear medidas de probabilidad aleatorias.

Una clase importante de medidas de probabilidad aleatorias que también se estudiaron son las mezclas de distribuciones (aleatorias). Los procesos anteriormente mencionados son medidas de probabilidad aleatorias discretas c.s., lo cual es un inconveniente si los datos se modelan como realizaciones de distribuciones continuas; sin embargo, utilizar mezclas con estos procesos y un kernel absolutamente continuo, genera medidas de probabilidad aleatorias continuas, con lo que se solventa este problema. Estas medidas permiten hacer análisis más complejos, ya que permiten introducir, además, información acerca del número de grupos de la mezcla. Un caso de particular atención es el proceso geométrico, ya que tiene una estructura más sencilla a las mezclas de procesos Dirichlet, pero igualmente útil. Un problema que resuelven este tipo de modelos de mezclas es el de estimación de densidades.

Por otro lado, en el ámbito computacional, en este trabajo también se estudiaron los métodos MCMC. Los modelos de mezclas, debido a su estructura, resultan demasiado complejos como para obtener analíticamente las distribuciones posteriores necesarias y es por esta razón que los métodos MCMC son de gran importancia. Estos métodos permiten obtener numéricamente ya sean estimaciones puntuales, predicciones e inclusive las densidades posteriores y predictivas necesarias. Específicamente se estudiaron distintos métodos para estimación de densidades, basados en el esquema de urnas de Pólya y en un truncamiento de la representación *stick–breaking*; asimismo se estudió la implementación del proceso geométrico a

través del *Gibbs sampler*.

La implementación de estos métodos para casos específicos de las distribuciones de la medida base y del kernel también es parte importante, ya que de esta manera, además de permitir su uso, es posible explorar su comportamiento, como se hizo en el último capítulo.

En este aspecto, como parte fundamental de este trabajo, se hicieron poco más de mil simulaciones con datos de cuatro muestras provenientes de modelos de datos específicos, así como alrededor de 300 más para el conjunto de datos de las galaxias, datos que se utilizan frecuentemente para estudiar modelos de mezclas en el ámbito bayesiano. Con base en estos resultados se pueden concluir varios puntos:

- Con respecto a los métodos, el método *Gibbs sampler* para urnas de Pólya resulta ser el menos eficiente, esto se puede atribuir a que, como se ha mencionado, los pesos resultantes ocasionan que los parámetros que definen al grupo (media y varianza en los métodos programados) tarden muchas iteraciones en cambiar. Sin embargo, incluir el llamado paso de aceleración mejora considerablemente su eficiencia.

  El *Gibbs sampler* por bloques, por otro lado, resulta una buena implementación, principalmente para el proceso Dirichlet y el Poisson–Dirichlet con el parámetro σ cercano a cero, en otras palabras, el parámetro σ ocasiona que su devianza estimada sea mayor; este efecto resulta mayor que el truncamiento de la medida *stick–breaking*.

  En cuanto a los métodos diseñados para distribuciones no conjugadas, los tres, en general, son aceptables. Sin embargo, el método Metropolis–Hastings y *Gibbs sampler* no es una buena elección si se tienen interés en hacer inferencias sobre el número de grupos de la mezcla debido a que su distribución es platicúrtica y con un soporte considerablemente mayor a los demás métodos (incluyendo el *Gibbs sampler* para urnas de Pólya). Los otros dos métodos, por el contrario son eficientes en este aspecto, tomando en cuenta que para Metropolis–Hastings para distribuciones no conjugadas es necesario utilizar un valor para r moderadamente grande y para el *Gibbs sampler* con parámetros auxiliares, por el contrario, es mejor utilizar un valor para l entre uno o dos.

- Considerando los tres procesos que se utilizaron, resulta interesante observar que el proceso Dirichlet, por lo general, es más eficiente que los otros dos. El proceso σ–estable normalizado puede ser mejor opción cuando las observaciones son tales que no se distinguen grupos a simple vista. Por último, se pudiera pensar que el proceso Poisson–Dirichlet de dos parámetros es más recomendable que los dos anteriores, sin embargo, los resultados parecen no apoyar esta idea. Un estudio más detallado de sus momentos podría ayudar a comprenderlo mejor.

- Una variable aleatoria de interés en los modelos de mezclas es la que indica el número de grupos. Lo recomendable es incluir la información que se tenga acerca de ellos, principalmente si no se quieren actualizar los parámetros del proceso elegido.

- En relación a la estimación de densidades, se puede observar que el número de modas no tiene relación con el número de grupos, ejemplo de esto son los resultados del modelo bimodal: en varias simulaciones la distribución posterior del número de grupos tiene moda en valores mayores a uno, sin embargo, sólo el *Gibbs sampler* para urnas de Pólya tiene densidad estimada bimodal.

Con respecto a la actualización de los parámetros de los procesos, es importante mencionar que en este trabajo se implementó su actualización de una forma simple, en especial para el proceso Poisson–Dirichlet de dos parámetros, que, sin embargo, no se encontró más que una referencia (bastante más elaborada y posiblemente no tan precisa, cabe mencionar). Asimismo, se extendieron los algoritmos diseñados para distribuciones no conjugadas para poder utilizar el proceso Poisson–Dirichlet de dos parámetros.

Por otro lado, la simulaciones realizadas utilizando el proceso geométrico son comparables con los demás métodos. Es importante, sin embargo, hacer un análisis previo para determinar los mejores parámetros; esto se debe a que cada parámetro tiene más relevancia.

Otra aportación de este trabajo son los programas de cómputo hechos. Estos pueden utilizarse como en este trabajo, para estimar densidades; sin embargo, pueden fácilmente ser utilizados en otro tipo de modelos, como los modelos lineales, modelos dinámicos, análisis de conglomerados y otros; tomando en cuenta que todos ellos serán no paramétricos. También pueden ser modificados sin ninguna dificultad para el caso en que no exista conjugamiento, excepto el *Gibbs sampler* para urnas de Pólya. Asimismo, la generalización al caso multivariado puede hacerse fácilmente. En el caso del proceso geométrico, por otro lado, gracias a la rapidez con que corre el programa, se podría hacer un estudio más detallado para determinar de mejor manera la influencia de sus parámetros e incluso detectar alguna estructura de agrupamiento, si existe, entre otras cosas.

– A –
Tablas de las simulaciones

En este apéndice se presentan las tablas de las configuraciones de los parámetros utilizados para correr las simulaciones descritas en el Capítulo IV (Sección 1), así como sus resultados (Sección 2).

§ 1 Configuración de parámetros

Las siguientes tablas muestran los parámetros necesarios para correr los métodos estudiados. Se muestran en primer lugar los parámetros de la medida base para cada conjunto de datos, así como los valores para el número de grupos, con el cual se calcula $\mathbb{E}[K_n]$ *a priori*. En seguida se muestran los valores de $\mathbb{E}[K_n]$ *a priori* para cada uno de los procesos: Dirichlet, σ–estable normalizado y Poisson–Dirichlet de dos parámetros, así como el parámetro adicional que requiere el *Gibbs sampler* por bloques; estos dependen del número de observaciones, por lo que se necesitan dos conjuntos: la Tabla 2 está calculada para 100 observaciones, mientras que la Tabla 3 lo está para 82. Los parámetros del proceso geométrico no dependen del tamaño de muestra, por lo que sólo se presenta la Tabla 4.

	Medida base				
Modelo	ω	t	a	b	k
Separado	1.65718	433.35051	2.00	8.66701	$\{2, 4, 8\}$
Platicúrtico	−0.05874	134.68417	2.00	2.69368	$\{2, 5, 10\}$
Bimodal	−0.18462	19.05820	2.00	0.38116	$\{2, 8\}$
Leptocúrtico	0.09517	20.62737	2.00	0.41255	$\{2, 8\}$
Galaxias	21.72550	630.36145	2.00	12.60723	$\{3, 6, 12\}$

Tabla 1: Parámetros de la medida base para los distintos modelos de datos y para el conjunto de datos reales, así como valores para calcular $\mathbb{E}[K_n]$ *a priori*.

			Número de componentes					
k	α_k	$\mathcal{G}a(\alpha; a, b)$	N_1	*error*	N_2	*error*	N_3	*error*
2	0.20469	$(0.02095\,, 0.10235)$	13	2.25×10^{-7}	18	3.28×10^{-11}	30	0.00000
4	0.67954	$(0.23088\,, 0.33977)$	23	9.03×10^{-7}	34	4.30×10^{-11}	30	1.61×10^{-9}
5	0.94759	$(0.44897\,, 0.47380)$	29	6.94×10^{-7}	42	5.94×10^{-11}	30	3.38×10^{-7}
8	1.86664	$(1.74217\,, 0.93332)$	48	7.00×10^{-7}	69	8.57×10^{-11}	30	1.57×10^{-3}
10	2.57220	$(3.30810\,, 1.28610)$	62	7.97×10^{-7}	90	8.09×10^{-11}	30	2.82×10^{-2}
—	1.00000	$(0.00001\,, 0.00001)$	30	7.45×10^{-7}	43	9.09×10^{-11}	30	7.45×10^{-7}

(A) Proceso Dirichlet

			Número de componentes			
k	σ_k	$\mathcal{B}e(\sigma; a, b)$	N_1	*error*	N_2	*error*
2	0.13665	$(3.08753\,, 19.50741)$	70	9.17×10^{-7}	30	1.43×10^{-4}
4	0.27860	$(10.92021\,, 28.27633)$	150	3.33×10^{-3}	30	0.15601
5	0.32538	$(13.95930\,, 28.94226)$	150	2.53×10^{-2}	30	0.44226
8	0.42557	$(20.38131\,, 27.51068)$	150	0.43208	30	1.61793
10	0.47391	$(23.15686\,, 25.70696)$	150	0.99314	30	10.57354
—	0.50000	$(1.00000\,, 1.00000)$	150	1.38890	30	3.07660

(B) Proceso σ–estable normalizado

TABLA 2: Parámetros adicionales para las simulaciones para un tamaño de muestra $n = 100$; los últimos renglones, indicados con —, corresponden a las distribuciones iniciales no informativas.

					Número de componentes			
k	σ_k	θ_k	$\mathcal{B}e(\sigma; a, b)$	$\mathcal{G}a(\theta; c, d)$	N_1	*error*	N_2	*error*
2	0.08990	0.06823	(1.38124 , 13.98265)	(0.00233 , 0.03412)	34	9.62×10^{-7}	30	2.87×10^{-6}
	0.04439	0.13646	(0.33224 , 7.15203)	(0.00931 , 0.06823)	20	6.24×10^{-7}	30	1.88×10^{-9}
4	0.17623	0.22651	(4.94060 , 23.09415)	(0.02565 , 0.11326)	150	1.41×10^{-5}	30	1.70×10^{-2}
	0.08443	0.45302	(1.22097 , 13.23982)	(0.10262 , 0.22651)	58	8.64×10^{-7}	30	2.68×10^{-4}
5	0.20228	0.31586	(6.32591 , 24.94685)	(0.04989 , 0.15793)	150	2.02×10^{-4}	30	7.52×10^{-2}
	0.09593	0.63173	(1.56812 , 14.77785)	(0.19954 , 0.31586)	84	9.16×10^{-7}	30	2.59×10^{-3}
8	0.25411	0.62221	(9.37870 , 27.52909)	(0.19357 , 0.31111)	150	1.13×10^{-2}	30	0.66857
	0.11853	1.24443	(2.35835 , 17.53803)	(0.77430 , 0.62221)	150	6.13×10^{-6}	30	0.10292
10	0.27764	0.85740	(10.85878 , 28.25235)	(0.36757 , 0.42870)	150	4.82×10^{-2}	30	1.34790
	0.12894	1.71480	(2.76739 , 18.69537)	(1.47027 , 0.85740)	150	8.41×10^{-5}	30	0.39166
—	0.50000	1.00000	(1.00000 , 1.00000)	(0.00001 , 0.00001)	150	2.88096	30	5.54913

(c) Proceso Poisson–Dirichlet

Tabla 2: Parámetros adicionales para las simulaciones para un tamaño de muestra $n = 100$, continuación; los últimos renglones, indicados con —, corresponden a las distribuciones iniciales no informativas.

			Número de componentes					
k	α_k	$\mathcal{G}a(\alpha; a, b)$	N_1	*error*	N_2	*error*	N_3	*error*
3	0.45305	$(0.10263\,, 0.22653)$	18	8.12×10^{-7}	26	7.29×10^{-11}	30	6.89×10^{-13}
6	1.31239	$(0.86118\,, 0.65619)$	36	8.05×10^{-7}	52	9.33×10^{-11}	30	2.41×10^{-5}
12	3.64129	$(6.62951\,, 1.82065)$	82	9.55×10^{-7}	120	9.45×10^{-11}	30	0.25480
—	1.00000	$(0.00001\,, 0.00001)$	30	6.11×10^{-7}	43	7.46×10^{-11}	30	6.11×10^{-7}

(A) Proceso Dirichlet

			Número de componentes			
k	σ_k	$\mathcal{B}e(\sigma; a, b)$	N_1	*error*	N_2	*error*
3	0.22842	$(7.82342\,, 26.42598)$	150	1.40×10^{-4}	30	2.61×10^{-2}
6	0.38010	$(17.53226\,, 28.59270)$	150	0.123007	30	0.914122
12	0.53728	$(26.17739\,, 22.54465)$	150	1.834654	30	2.863625
—	0.50000	$(1.00000\,, 1.00000)$	150	1.242604	30	2.459754

(B) Proceso σ–estable normalizado

TABLA 3: Parámetros adicionales para las simulaciones para un tamaño de muestra $n = 82$; los últimos renglones, indicados con —, corresponden a las distribuciones iniciales no informativas.

					Número de componentes			
k	σ_k	θ_k	$\mathcal{B}e(\sigma; a, b)$	$\mathcal{G}a(\theta; c, d)$	N_1	*error*	N_2	*error*
3	0.14732	0.15102	(3.55387 , 20.56955)	(0.01140 , 0.07551)	118	9.72×10^{-7}	30	1.67×10^{-3}
	0.07156	0.30204	(0.87932 , 11.40854)	(0.04561 , 0.15102)	38	9.17×10^{-7}	30	9.82×10^{-6}
6	0.23290	0.43746	(8.08874 , 26.64236)	(0.09569 , 0.21873)	150	1.88×10^{-3}	30	0.23654
	0.10978	0.87493	(2.03585 , 16.50940)	(0.38275 , 0.43746)	127	9.76×10^{-7}	30	1.67×10^{-2}
12	0.31135	1.21376	(13.03990 , 28.84218)	(0.73661 , 0.60688)	150	0.19048	30	2.26634
	0.14489	2.42753	(3.44536 , 20.33391)	(2.94645 , 1.21376)	150	1.15×10^{-3}	30	1.12407
—	0.50000	1.00000	(1.00000 , 1.00000)	(0.00001 , 0.00001)	150	2.68277	30	3.75571

(c) Proceso Poisson–Dirichlet

Tabla 3: Parámetros adicionales para las simulaciones para un tamaño de muestra $n = 82$, continuación; los últimos renglones, indicados con —, corresponden a las distribuciones iniciales no informativas.

$\mathbb{E}[\lambda]$	$\mathrm{Var}[\lambda]$	$\mathcal{B}e(\lambda; a, b)$
0.5	0.083	(1.0 , 1.0)
0.5	0.125	(0.5 , 0.5)
0.9	0.005	(15.3 , 1.7)
0.5	0.005	(24.5 , 24.5)
0.1	0.005	(1.7 , 15.3)

TABLA 4: Parámetros para el *Gibbs sampler* para el proceso geométrico

§ 2 RESULTADOS DE LAS SIMULACIONES

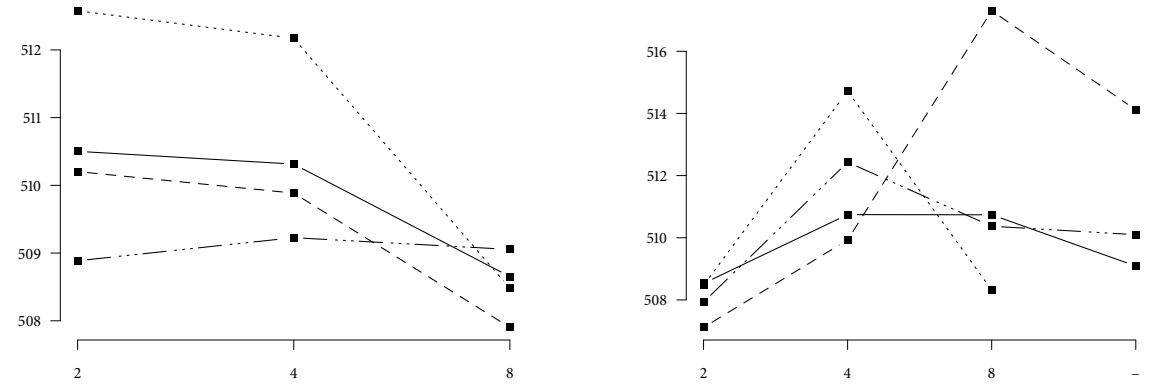
En las siguientes figuras y tablas se muestran los resultados de las simulaciones para cada uno de los modelos de datos y para el conjunto de datos reales; el orden de éstas es el siguiente:

1. *Gráficas de devianzas estimadas.* En estas gráficas se muestran, para cada simulación, las devianzas estimadas calculadas como en (IV.1) utilizando $\hat{h}$ dada por (IV.3). Se muestran dos gráficas por cada método de estimación de densidades: en la primera se dejaron fijos los parámetros de los procesos de la medida de probabilidad aleatoria, mientras que en la segunda se les asignó una distribución inicial. Los puntos aparecen unidos con una línea para identificar cada proceso. (Figuras 1, 3, 5, 7 y 9.)

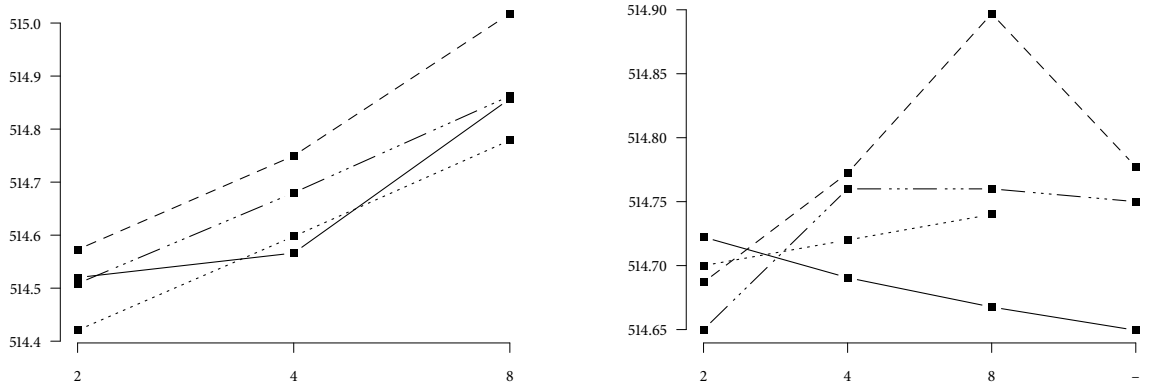
2. *Distribución posterior del número de grupos.* Tomando la simulación con menor devianza para cada método con un mismo $\mathbb{E}[K_n]$ *a priori*, se grafica su correspondiente distribución posterior de K_n. Cada par de barras corresponde a un mismo $\mathbb{E}[K_n]$, pero en la primera los parámetros del proceso se dejaron fijos, mientras que para la segunda se les asignó una distribución inicial; la última barra corresponde a la distribución inicial no informativa. (Figuras 2, 4, 6, 8 y 10.)

3. *Tiempo de autocorrelación integrado.* Se muestran tablas para $\hat{\tau}_D$ y $\hat{\tau}_{K_n}$. Estas se construyeron tomando la simulación con menor devianza para cada método de estimación de densidades para un mismo $\mathbb{E}[K_n]$ *a priori*. (Tablas 5, 7, 9, 11 y 13.)

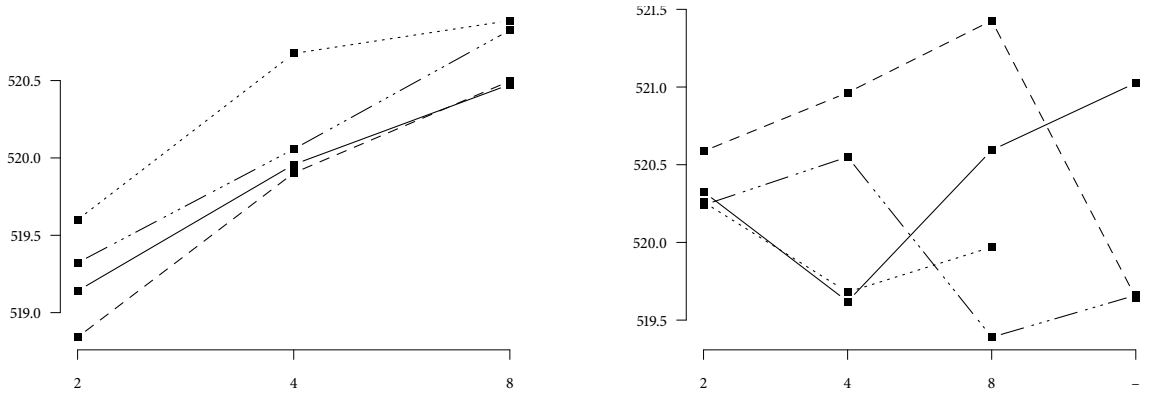
4. *Resultados del Gibbs sampler para el proceso geométrico.* Debido a la estructura del proceso geométrico, no es posible compararlo de la misma manera que los demás; por tanto, sus resultados se muestran en una tabla por separado. (Tablas 6, 8, 10, 12 y 14.)



(A) *Gibbs sampler* para urnas de Pólya

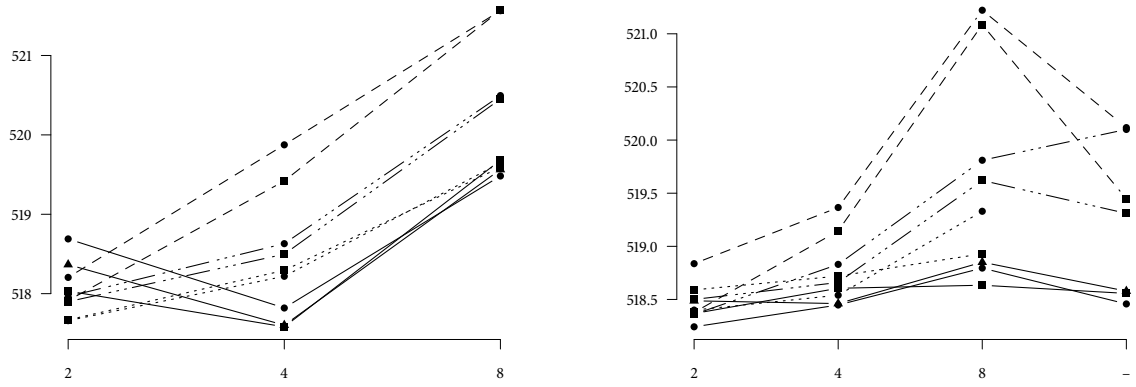


(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración

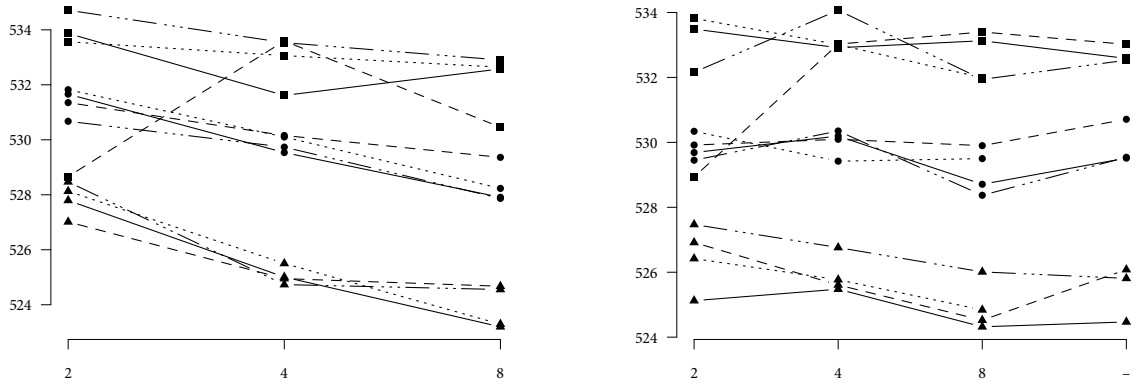


(C) Metropolis–Hastings y *Gibbs sampler*

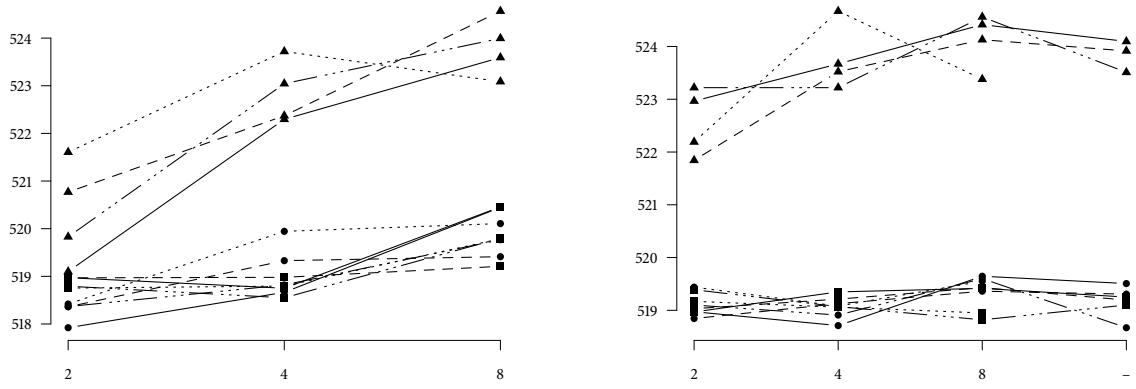
FIGURA 1: Devianzas estimadas, modelo separado; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



(D) *Gibbs sampler* por bloques con N_1 (●), N_2 (▲) y N_3 (■) componentes para el proceso Dirichlet y N_1 (●) y N_2 (■) componentes para el resto, de acuerdo a la Tabla 2.

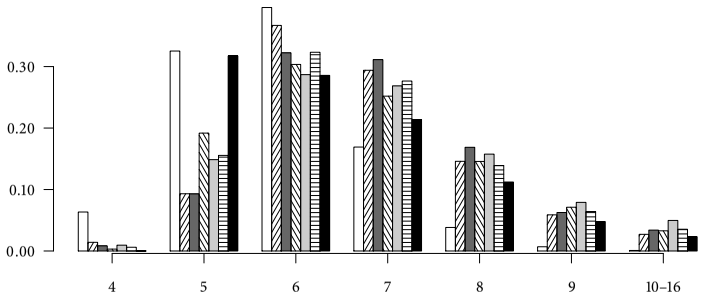


(E) Metropolis–Hastings para distribuciones no conjugadas con 1 (■), 5 (●) y 15 (▲) actualizaciones de K_i.

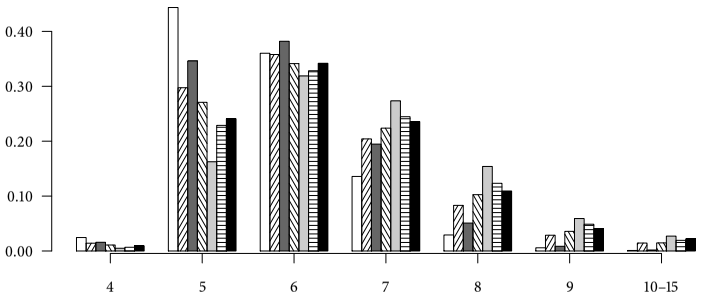


(F) *Gibbs sampler* con parámetros auxiliares, número de parámetros: 1 (■), 2 (●) y 10 (▲).

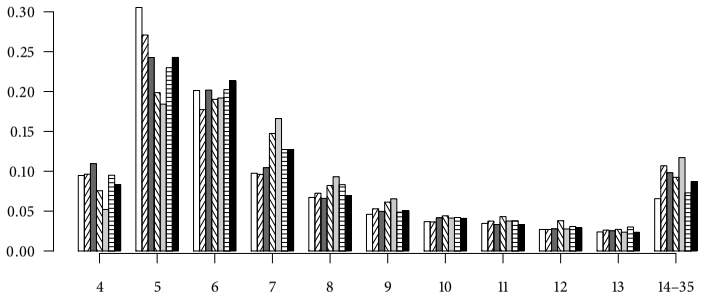
FIGURA 1: Devianzas estimadas, modelo separado, continuación; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



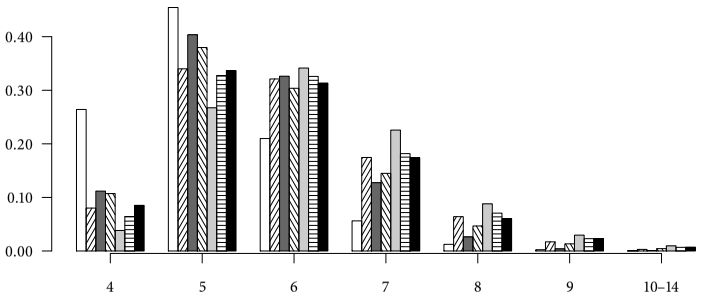
(A) *Gibbs sampler* para urnas de Pólya



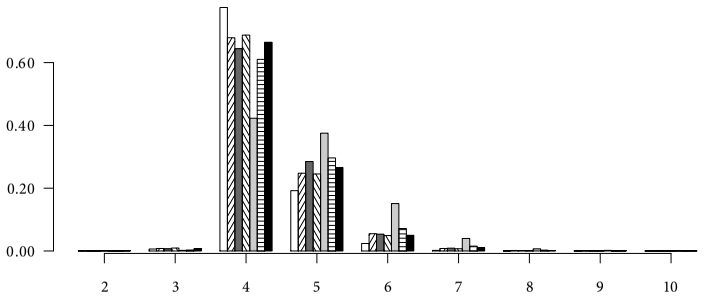
(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración



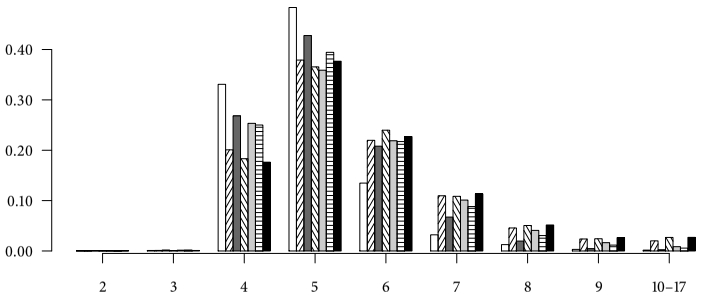
(C) Metropolis–Hastings y *Gibbs sampler*



(D) *Gibbs sampler* por bloques



(E) Metropolis–Hastings para distribuciones no conjugadas



(F) *Gibbs sampler* con parámetros auxiliares

Figura 2: Distribución posterior de K_n, modelo separado; las texturas indican los valores iniciales para $\mathbb{E}[K_n]$: □ y ▨ 2, ■ y ▧ 4 y ▤ y ▤ 8; en las segundas se actualizaron los parámetros de los procesos y, asimismo, ■ corresponde a la distribución inicial no informativa para estos.

	$\mathbb{E}[K_n]$						
Método	2		4		8		—
G. s. para urnas de Pólya	77.46	193.88	52.66	80.62	47.69	55.23	28.43
G. s. para urnas de Pólya con paso de aceleración	0.53	0.50	0.50	0.50	0.50	0.50	0.50
M. H. y *Gibbs sampler*	1.91	2.19	2.10	2.06	1.94	2.02	1.84
Gibbs sampler por bloques	3.29	1.94	1.90	2.00	1.51	1.66	1.65
M. H. para distr. no conjugadas	7.99	5.13	5.42	4.83	3.92	5.08	6.02
G. s. con parámetros auxiliares	1.27	1.49	1.29	1.11	2.09	2.07	0.99

(A) Tiempo de autocorrelación integrado para la devianza, $\hat{\tau}_D$

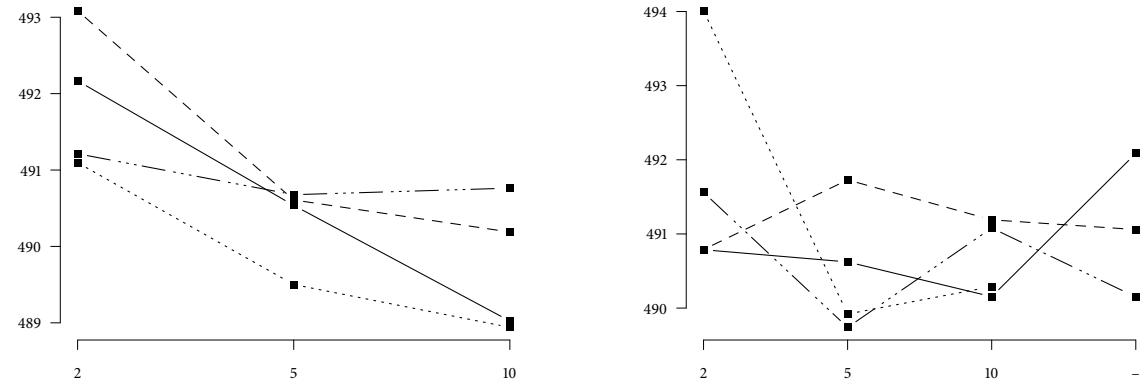
	$\mathbb{E}[K_n]$						
Método	2		4		8		—
G. s. para urnas de Pólya	285.67	119.08	168.96	117.71	103.91	224.41	517.69
G. s. para urnas de Pólya con paso de aceleración	2.71	2.66	2.35	2.93	2.07	3.17	3.86
M. H. y *Gibbs sampler*	12.80	14.84	10.76	13.18	10.36	11.14	10.94
Gibbs sampler por bloques	35.10	19.39	13.73	15.10	7.18	15.86	18.11
M. H. para distr. no conjugadas	5.74	7.83	5.98	6.64	6.95	6.68	6.07
G. s. con parámetros auxiliares	10.15	10.49	7.01	9.23	7.23	9.89	9.15

(B) Tiempo de autocorrelación integrado para el número de grupos, $\hat{\tau}_{K_n}$

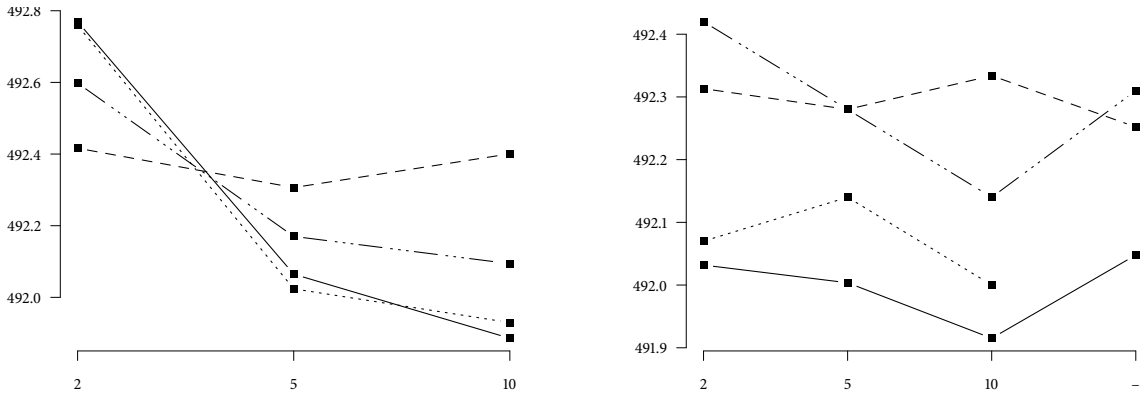
TABLA 5: Tiempo de autocorrelación integrado, modelo separado

	Hiperparámetros				
	$(1.0, 1.0)$	$(0.5, 0.5)$	$(15.3, 1.7)$	$(24.5, 24.5)$	$(1.7, 15.3)$
Devianza	535.12	529.81	535.39	555.56	529.83
$\hat{\tau}_D$	7.36	37.24	59.77	1.13	3.26

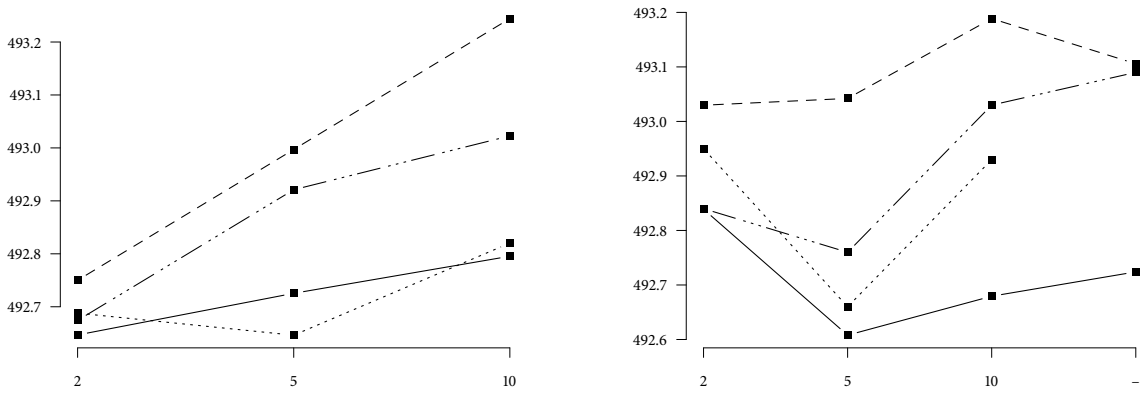
TABLA 6: Resultados del *Gibbs sampler* para el proceso geométrico, modelo separado



(A) *Gibbs sampler* para urnas de Pólya

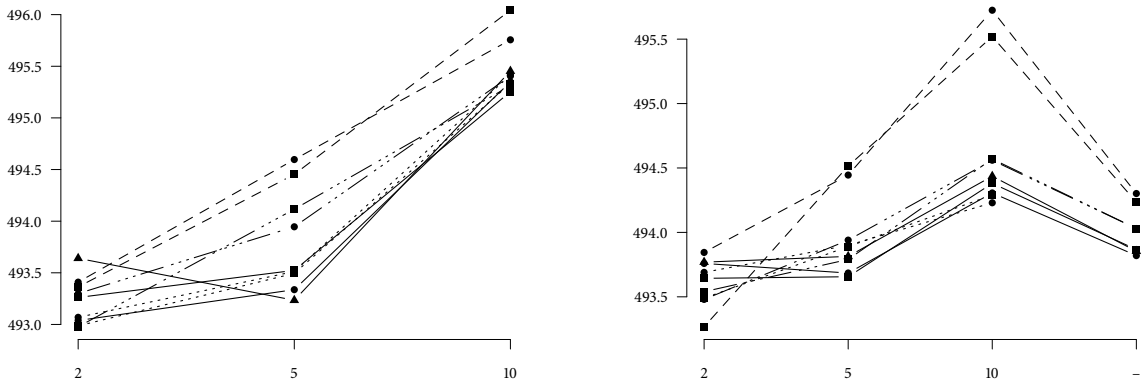


(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración

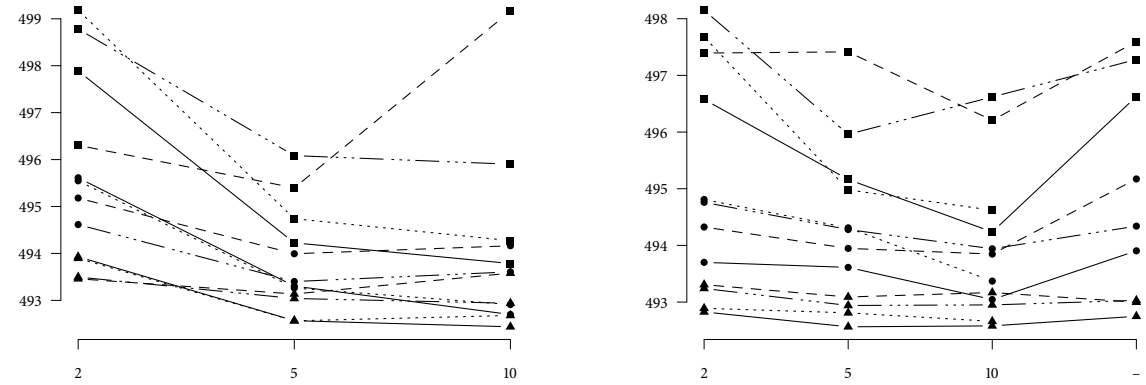


(C) Metropolis–Hastings y *Gibbs sampler*

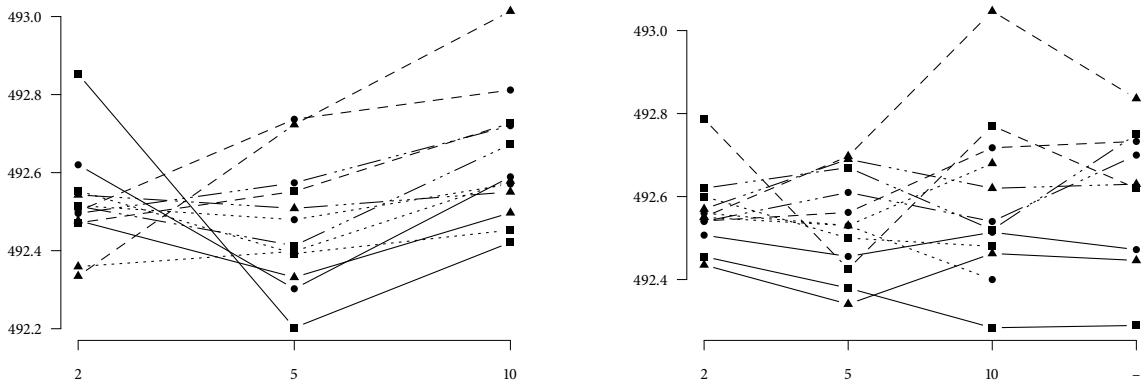
FIGURA 3: Devianzas estimadas, modelo platicúrtico; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



(D) *Gibbs sampler* por bloques con N_1 (●), N_2 (▲) y N_3 (■) componentes para el proceso Dirichlet y N_1 (●) y N_2 (■) componentes para el resto, de acuerdo a la Tabla 2.

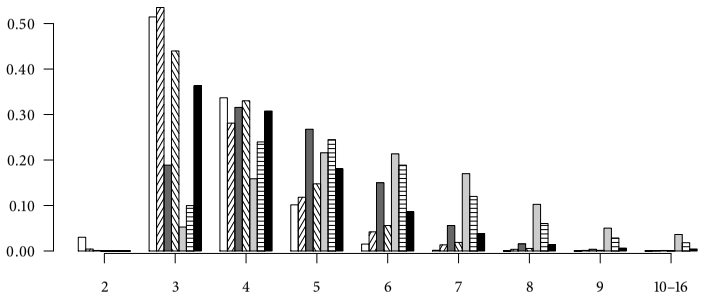


(E) Metropolis–Hastings para distribuciones no conjugadas con 1 (■), 5 (●) y 15 (▲) actualizaciones de K_i.

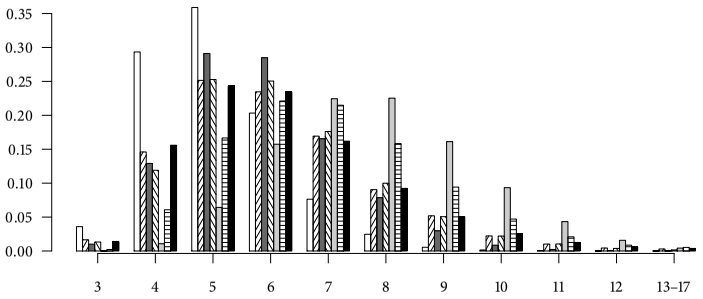


(F) *Gibbs sampler* con parámetros auxiliares, número de parámetros: 1 (■), 2 (●) y 10 (▲).

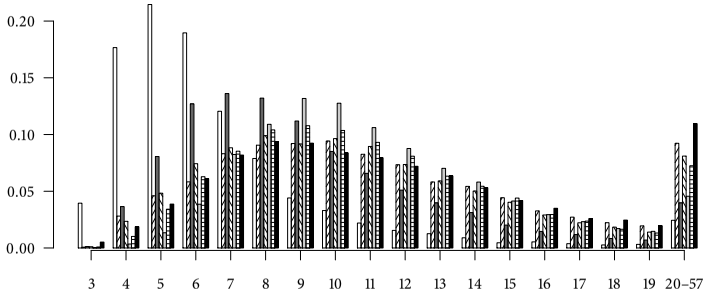
FIGURA 3: Devianzas estimadas, modelo platicúrtico, continuación; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



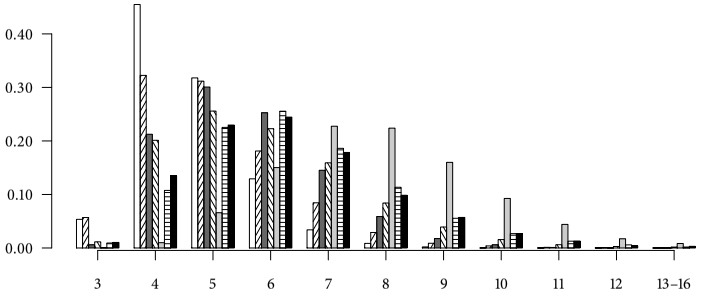
(A) *Gibbs sampler* para urnas de Pólya



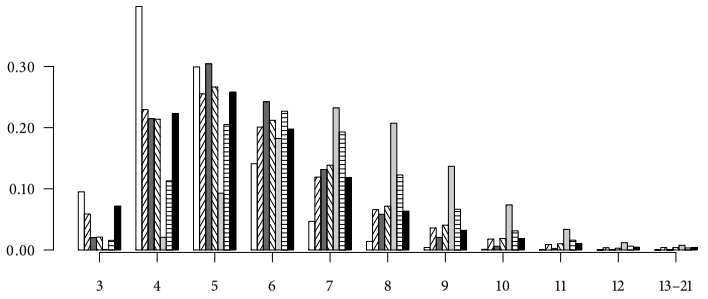
(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración



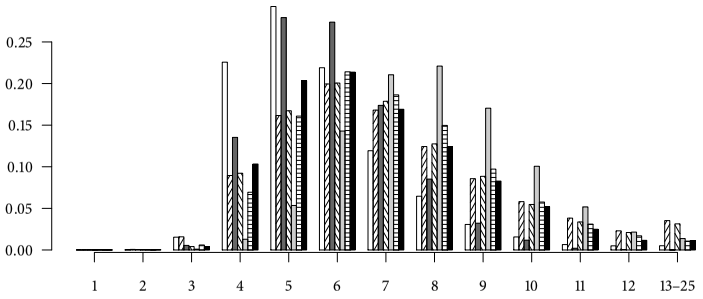
(C) Metropolis–Hastings y *Gibbs sampler*



(D) *Gibbs sampler* por bloques



(E) Metropolis–Hastings para distribuciones no conjugadas



(F) *Gibbs sampler* con parámetros auxiliares

FIGURA 4: Distribución posterior de K_n, modelo platicúrtico; las texturas indican los valores iniciales para $\mathbb{E}[K_n]$: □ y ▨ 2, ■ y ▧ 5 y ▤ y ▤ 10; en las segundas se actualizaron los parámetros de los procesos y, asimismo, ■ corresponde a la distribución inicial no informativa para estos.

	$\mathbb{E}[K_n]$						
Método	2		5		10		—
G. s. para urnas de Pólya	546.57	148.91	93.27	139.05	53.72	58.83	147.22
G. s. para urnas de Pólya con paso de aceleración	2.67	2.21	1.32	1.52	0.71	1.05	1.75
M. H. y *Gibbs sampler*	0.91	1.08	1.01	1.19	1.07	1.16	1.28
Gibbs sampler por bloques	7.86	9.17	1.85	2.61	1.23	2.57	3.40
M. H. para distr. no conjugadas	1.02	1.40	1.30	1.13	1.22	1.20	1.79
G. s. con parámetros auxiliares	0.50	0.72	0.55	0.85	0.56	0.85	1.03

(A) Tiempo de autocorrelación integrado para la devianza, $\hat{\tau}_D$

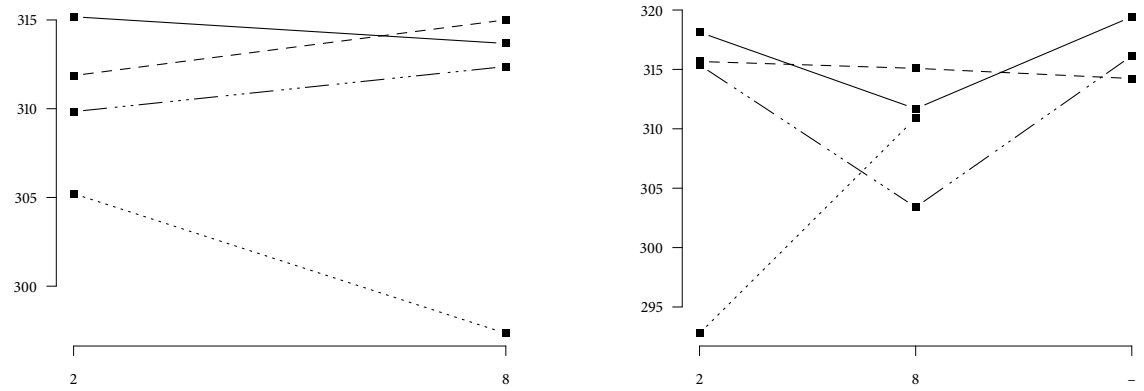
	$\mathbb{E}[K_n]$						
Método	2		5		10		—
G. s. para urnas de Pólya	116.84	8.71	27.40	10.27	52.73	39.37	36.49
G. s. para urnas de Pólya con paso de aceleración	9.13	7.17	5.75	8.33	3.29	5.23	7.80
M. H. y *Gibbs sampler*	9.82	15.47	7.91	9.34	6.38	7.88	9.74
Gibbs sampler por bloques	42.68	28.44	22.73	22.01	12.87	22.08	25.49
M. H. para distr. no conjugadas	10.30	11.03	6.74	9.21	4.86	8.02	15.54
G. s. con parámetros auxiliares	10.71	9.05	6.96	9.51	4.16	8.38	9.17

(B) Tiempo de autocorrelación integrado para el número de grupos, $\hat{\tau}_{K_n}$

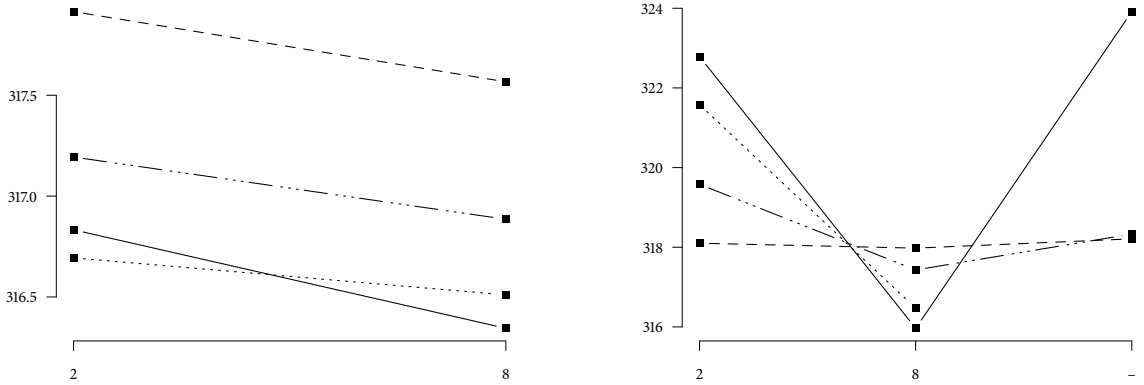
TABLA 7: Tiempo de autocorrelación integrado, modelo platicúrtico

	Hiperparámetros				
	$(1.0, 1.0)$	$(0.5, 0.5)$	$(15.3, 1.7)$	$(24.5, 24.5)$	$(1.7, 15.3)$
Devianza	497.71	497.89	500.71	500.64	497.35
$\hat{\tau}_D$	3.10	2.79	3.64	1.98	2.21

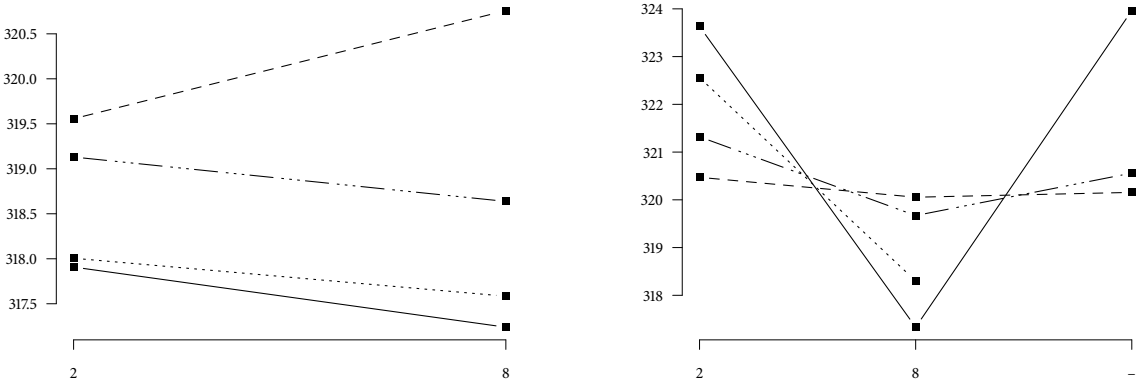
TABLA 8: Resultados del *Gibbs sampler* para el proceso geométrico, modelo platicúrtico



(A) *Gibbs sampler* para urnas de Pólya

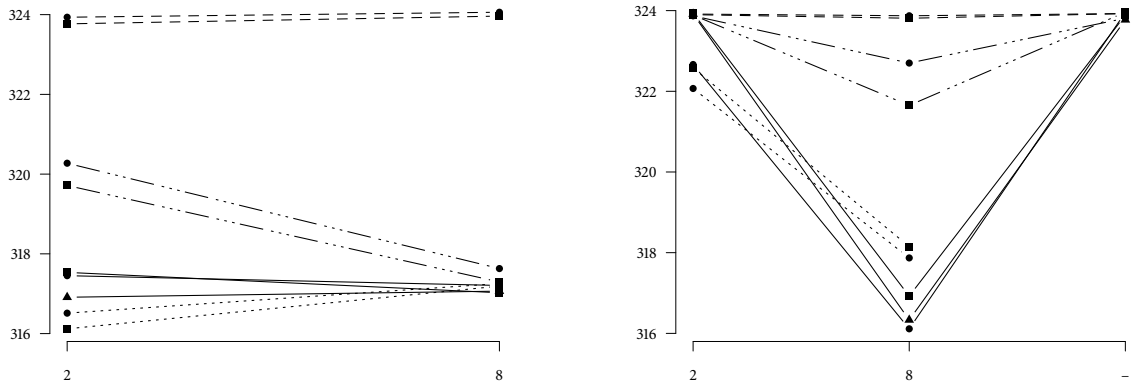


(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración

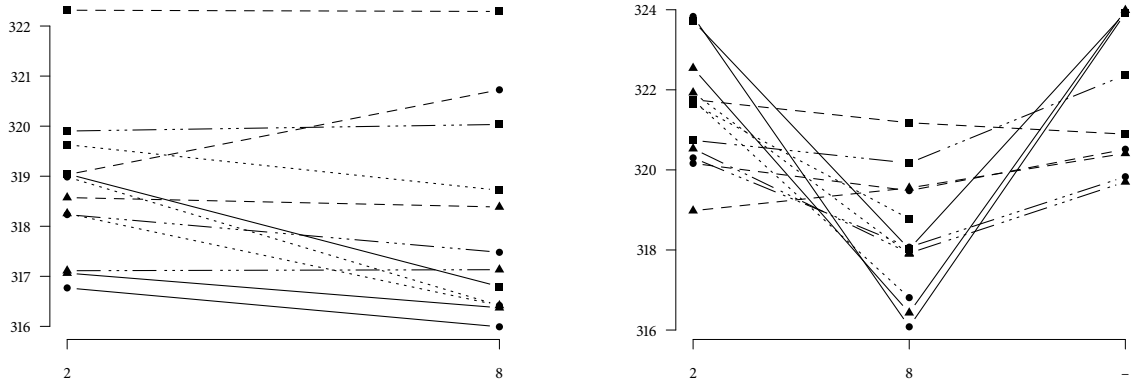


(C) Metropolis–Hastings y *Gibbs sampler*

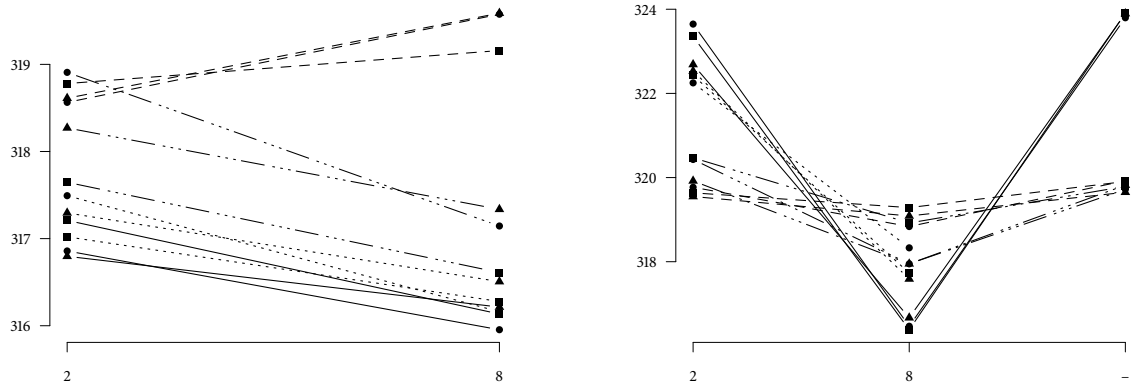
FIGURA 5: Devianzas estimadas, modelo bimodal; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



(D) *Gibbs sampler* por bloques con N_1 (●), N_2 (▲) y N_3 (■) componentes para el proceso Dirichlet y N_1 (●) y N_2 (■) componentes para el resto, de acuerdo a la Tabla 2.

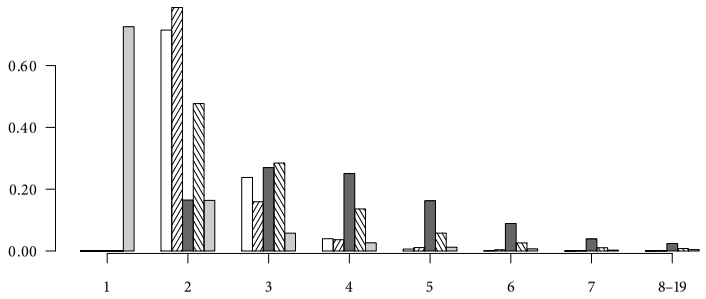


(E) Metropolis–Hastings para distribuciones no conjugadas con 1 (■), 5 (●) y 15 (▲) actualizaciones de K_i.

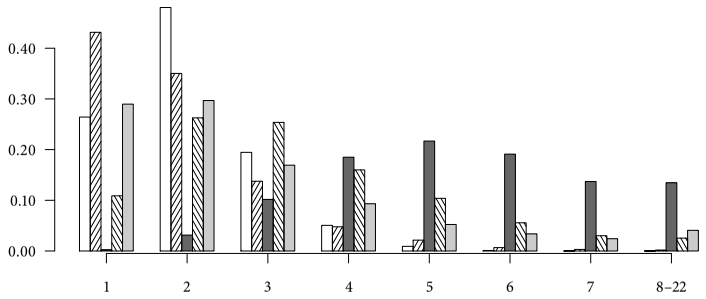


(F) *Gibbs sampler* con parámetros auxiliares, número de parámetros: 1 (■), 2 (●) y 10 (▲).

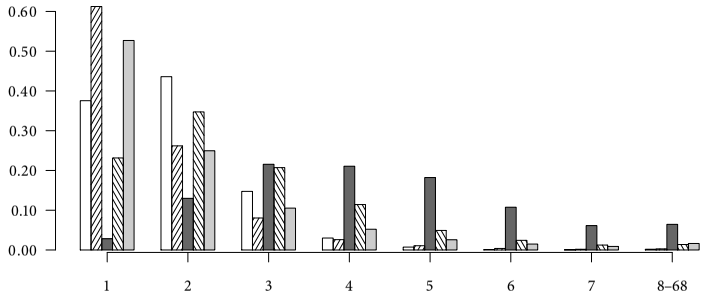
FIGURA 5: Devianzas estimadas, modelo bimodal, continuación; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



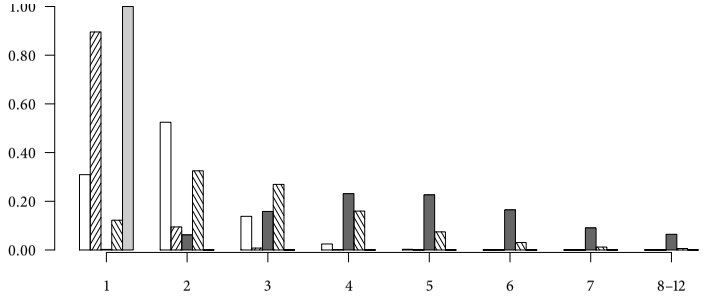
(A) *Gibbs sampler* para urnas de Pólya



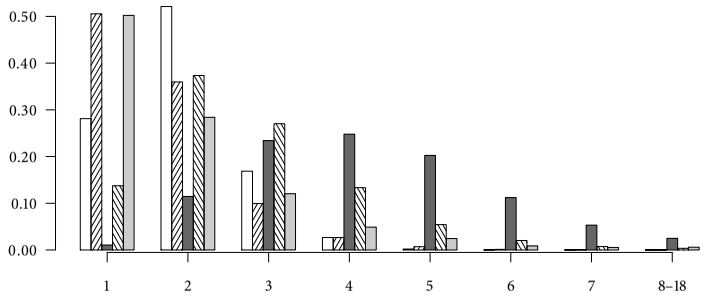
(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración



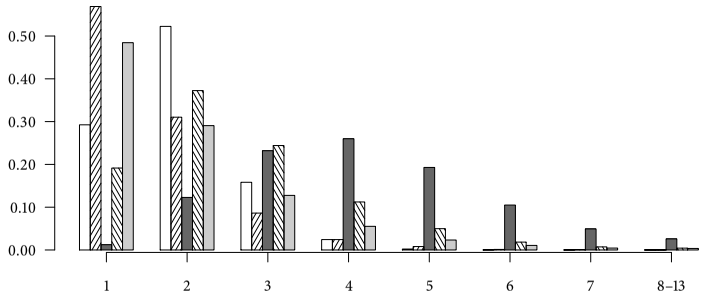
(C) Metropolis–Hastings y *Gibbs sampler*



(D) *Gibbs sampler* por bloques



(E) Metropolis–Hastings para distribuciones no conjugadas



(F) *Gibbs sampler* con parámetros auxiliares

FIGURA 6: Distribución posterior de K_n, modelo bimodal; las texturas indican los valores iniciales para $\mathbb{E}[K_n]$: □ y ▨ 2 y ■ y ▧ 8; en las segundas se actualizaron los parámetros de los procesos y, asimismo, ▭ corresponde a la distribución inicial no informativa para estos.

	$\mathbb{E}[K_n]$				
Método	2		8		—
G. s. para urnas de Pólya	1613.55	2.77	22.67	2.47	2.76
G. s. para urnas de Pólya con paso de aceleración	7.57	11.92	1.80	4.39	9.10
M. H. y *Gibbs sampler*	7.71	9.21	3.99	6.94	6.75
Gibbs sampler por bloques	32.46	23.25	2.49	10.72	0.50
M. H. para distr. no conjugadas	16.26	8.01	3.86	7.45	7.89
G. s. con parámetros auxiliares	8.52	8.61	2.76	6.52	6.44

(A) Tiempo de autocorrelación integrado para la devianza, $\hat{\tau}_D$

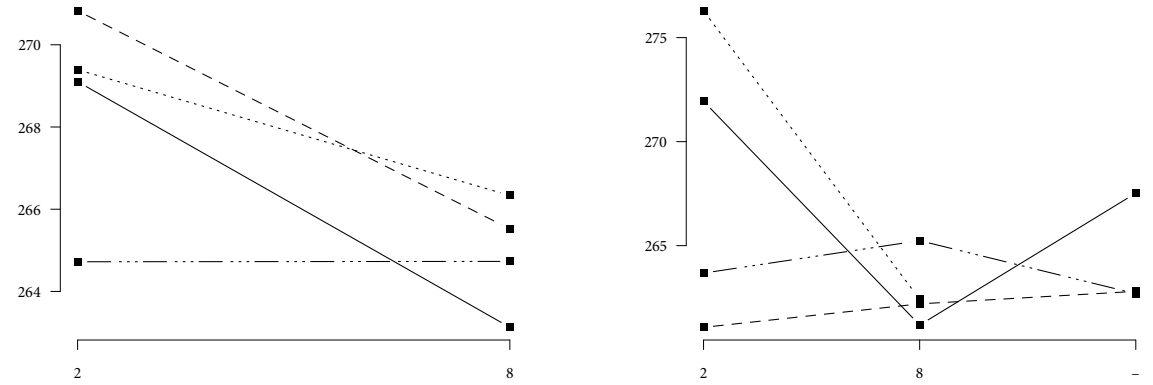
	$\mathbb{E}[K_n]$				
Método	2		8		—
G. s. para urnas de Pólya	2.01	2.20	1.51	1.83	3.04
G. s. para urnas de Pólya con paso de aceleración	12.69	21.45	2.80	7.49	7.66
M. H. y *Gibbs sampler*	9.04	9.10	2.05	6.32	8.33
Gibbs sampler por bloques	60.60	212.40	17.63	20.02	0.50
M. H. para distr. no conjugadas	25.47	26.91	2.41	8.10	15.51
G. s. con parámetros auxiliares	17.84	15.62	2.15	8.69	11.69

(B) Tiempo de autocorrelación integrado para el número de grupos, $\hat{\tau}_{K_n}$

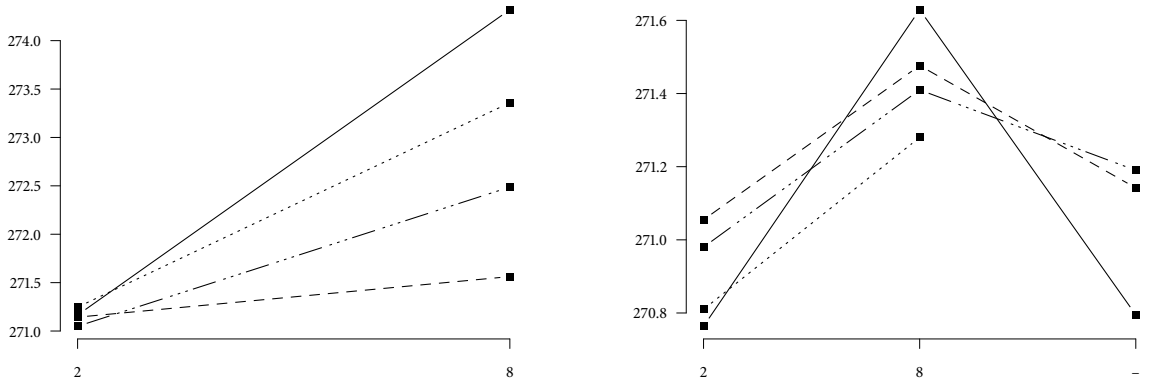
TABLA 9: Tiempo de autocorrelación integrado, modelo bimodal

	Hiperparámetros				
	$(1.0, 1.0)$	$(0.5, 0.5)$	$(15.3, 1.7)$	$(24.5, 24.5)$	$(1.7, 15.3)$
Devianza	320.08	320.99	321.60	316.92	319.67
$\hat{\tau}_D$	4.43	4.47	0.89	0.50	0.50

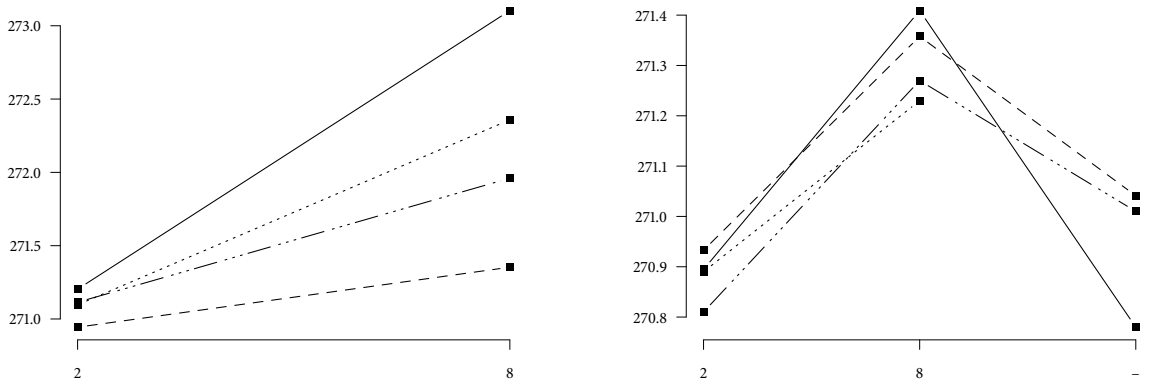
TABLA 10: Resultados del *Gibbs sampler* para el proceso geométrico, modelo bimodal



(A) *Gibbs sampler* para urnas de Pólya

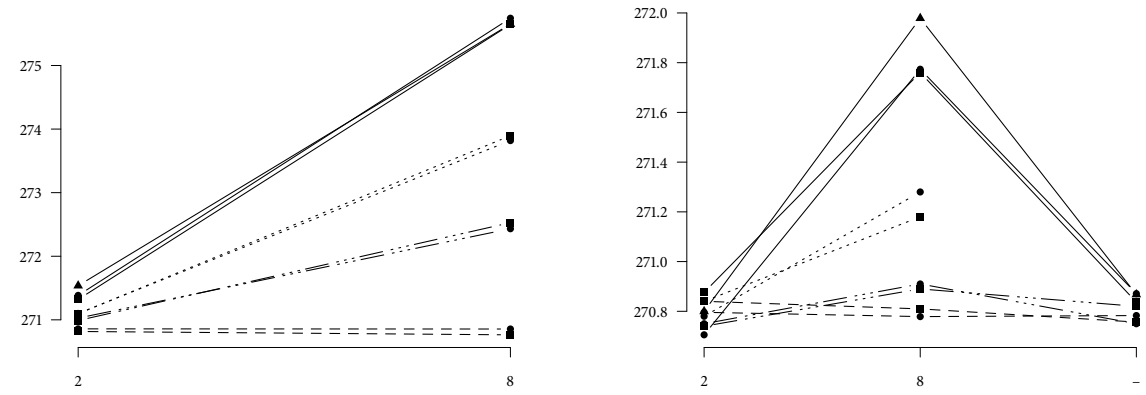


(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración

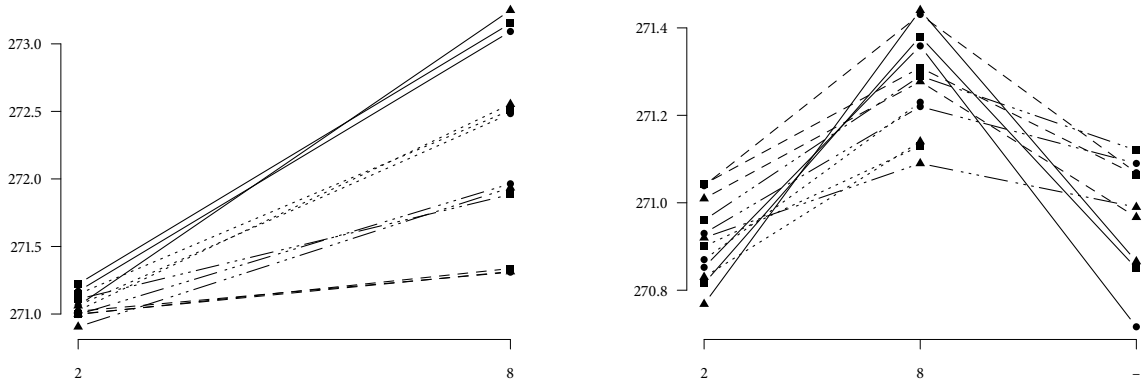


(C) Metropolis–Hastings y *Gibbs sampler*

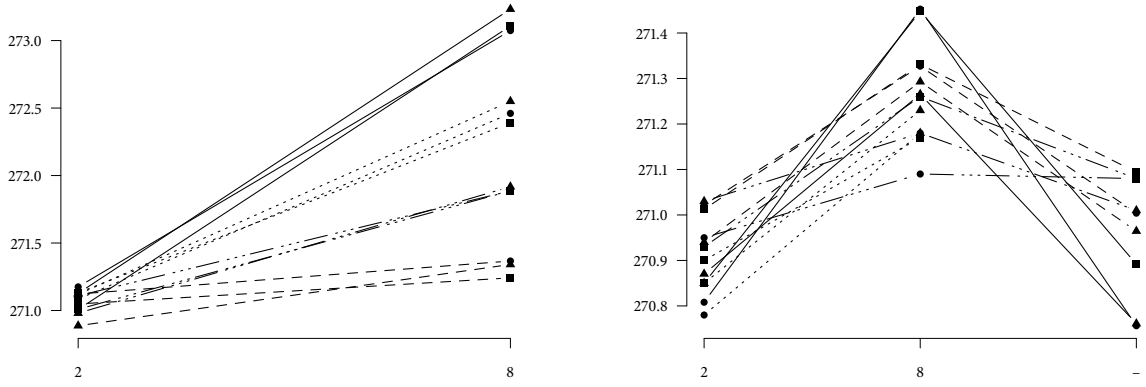
FIGURA 7: Devianzas estimadas, modelo leptocúrtico; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



(D) *Gibbs sampler* por bloques con N_1 (●), N_2 (▲) y N_3 (■) componentes para el proceso Dirichlet y N_1 (●) y N_2 (■) componentes para el resto, de acuerdo a la Tabla 2.

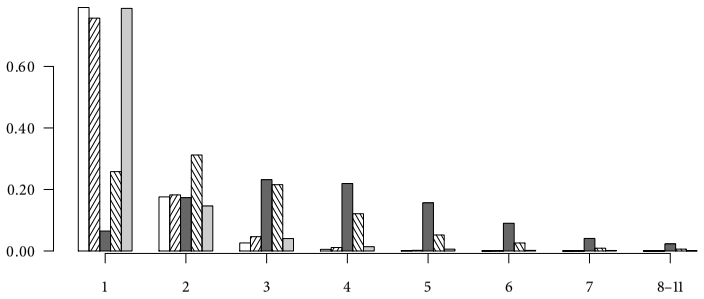


(E) Metropolis–Hastings para distribuciones no conjugadas con 1 (■), 5 (●) y 15 (▲) actualizaciones de K_i.

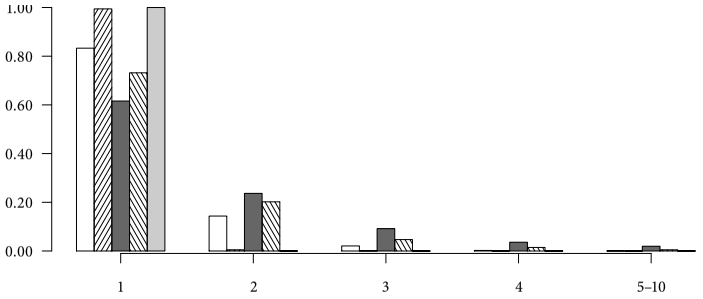


(F) *Gibbs sampler* con parámetros auxiliares, número de parámetros: 1 (■), 2 (●) y 10 (▲).

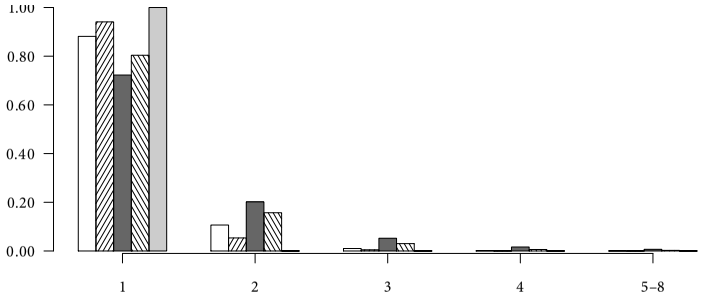
FIGURA 7: Devianzas estimadas, modelo leptocúrtico, continuación; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



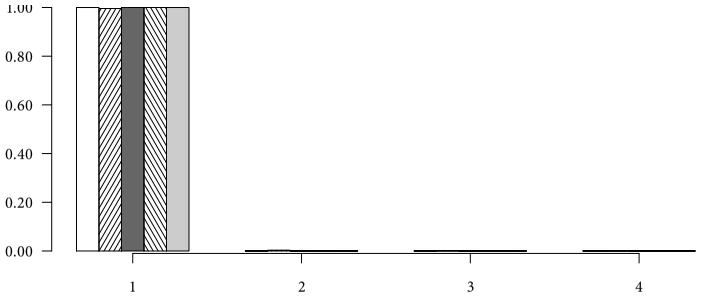
(A) *Gibbs sampler* para urnas de Pólya



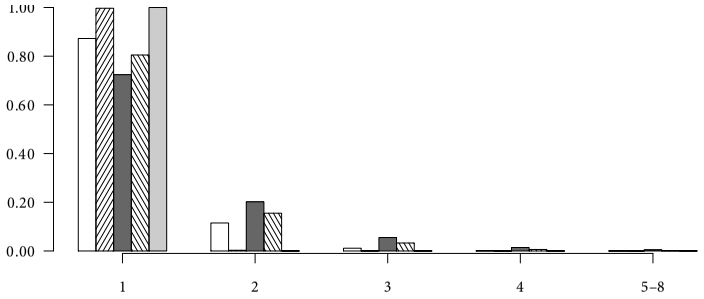
(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración



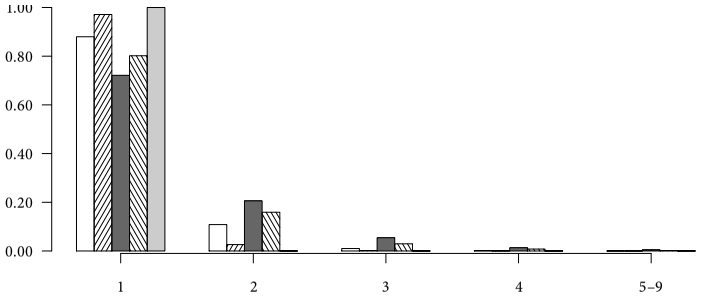
(C) Metropolis–Hastings y *Gibbs sampler*



(D) *Gibbs sampler* por bloques



(E) Metropolis–Hastings para distribuciones no conjugadas



(F) *Gibbs sampler* con parámetros auxiliares

FIGURA 8: Distribución posterior de K_n, modelo leptocúrtico; las texturas indican los valores iniciales para $\mathbb{E}[K_n]$: □ y ▨ 2 y ■ y ▧ 8; en las segundas se actualizaron los parámetros de los procesos y, asimismo, ▭ corresponde a la distribución inicial no informativa para estos.

	$\mathbb{E}[K_n]$				
Método	2		8		—
G. s. para urnas de Pólya	6.98	1006.77	303.61	78.51	1.54
G. s. para urnas de Pólya con paso de aceleración	0.50	0.50	0.50	0.50	0.50
M. H. y *Gibbs sampler*	0.50	0.50	0.53	0.50	0.50
Gibbs sampler por bloques	0.50	0.50	0.50	0.50	0.50
M. H. para distr. no conjugadas	0.50	0.50	0.54	0.50	0.50
G. s. con parámetros auxiliares	0.50	0.50	0.50	0.50	0.50

(A) Tiempo de autocorrelación integrado para la devianza, $\hat{\tau}_D$

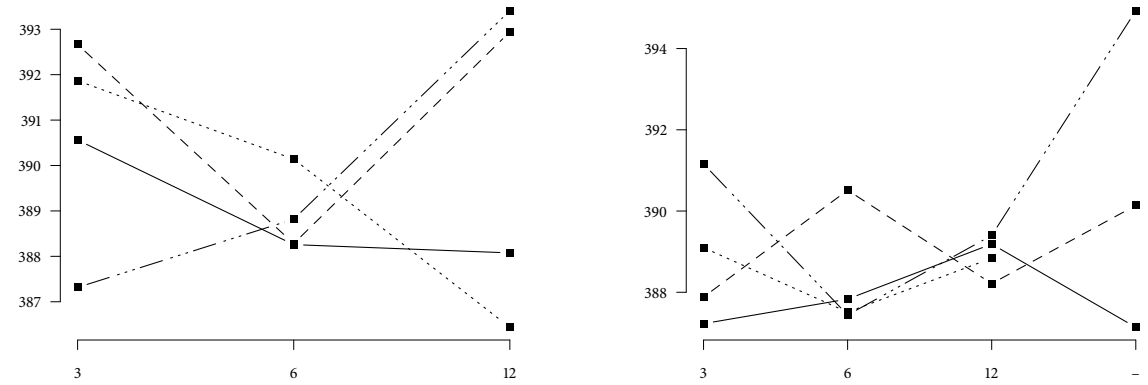
	$\mathbb{E}[K_n]$				
Método	2		8		—
G. s. para urnas de Pólya	2.74	3.66	297.78	498.44	1.62
G. s. para urnas de Pólya con paso de aceleración	1.83	2.56	1.48	2.39	0.50
M. H. y *Gibbs sampler*	2.10	1.18	1.31	1.81	0.50
Gibbs sampler por bloques	0.50	7.02	0.50	0.50	0.50
M. H. para distr. no conjugadas	1.33	1.66	1.13	1.28	0.50
G. s. con parámetros auxiliares	2.43	3.55	1.03	1.63	0.50

(B) Tiempo de autocorrelación integrado para el número de grupos, $\hat{\tau}_{K_n}$

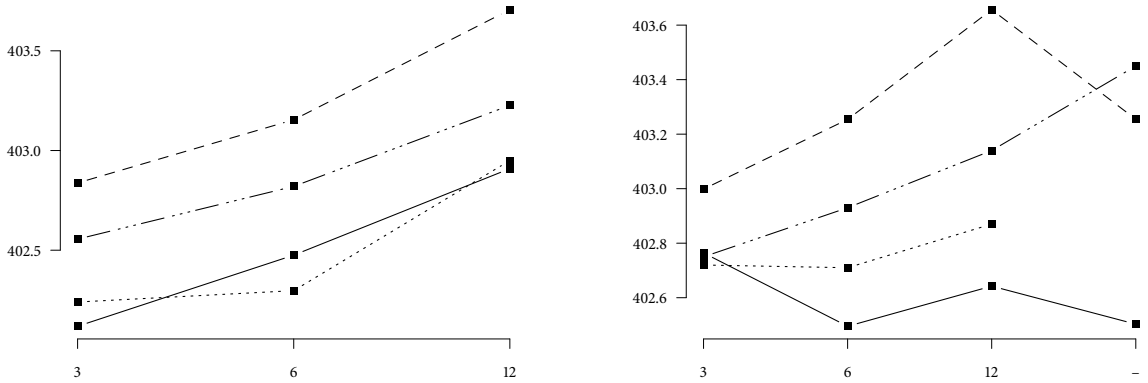
TABLA 11: Tiempo de autocorrelación integrado, modelo leptocúrtico

	Hiperparámetros				
	$(1.0, 1.0)$	$(0.5, 0.5)$	$(15.3, 1.7)$	$(24.5, 24.5)$	$(1.7, 15.3)$
Devianza	272.77	271.92	272.96	281.16	285.14
$\hat{\tau}_D$	1.33	2.28	0.80	1.21	3.57

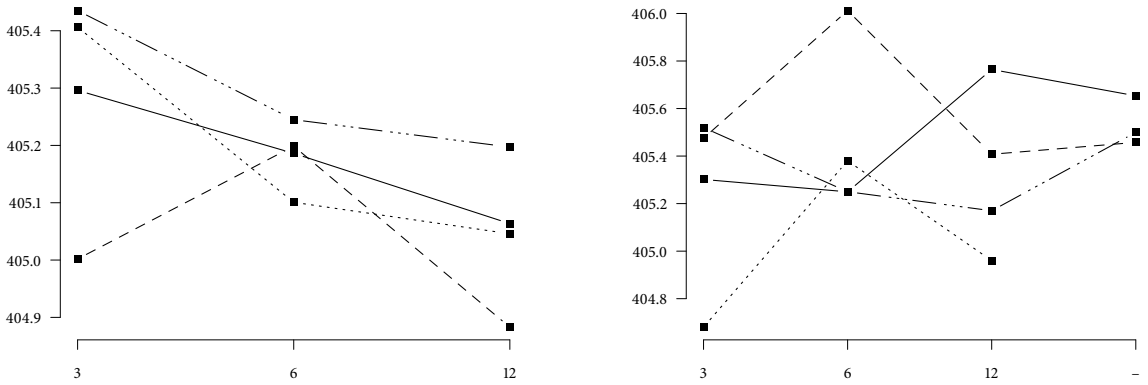
TABLA 12: Resultados del *Gibbs sampler* para el proceso geométrico, modelo leptocúrtico



(A) *Gibbs sampler* para urnas de Pólya

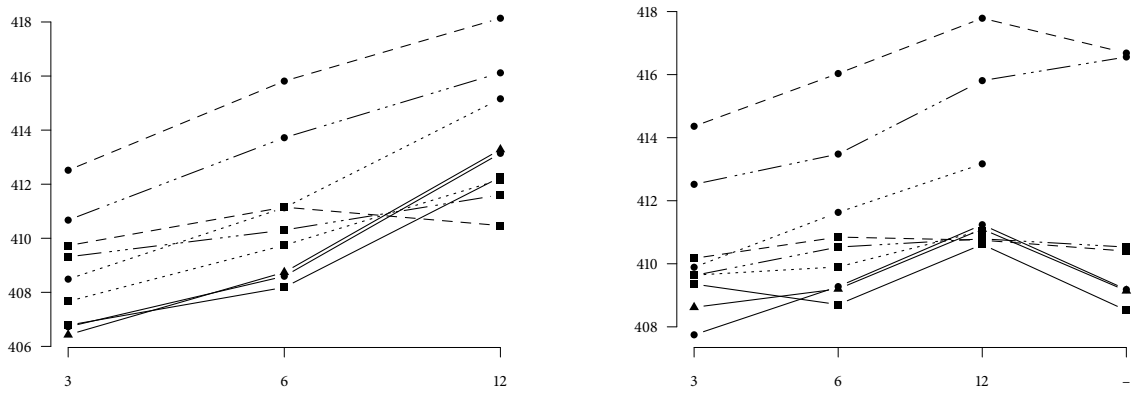


(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración

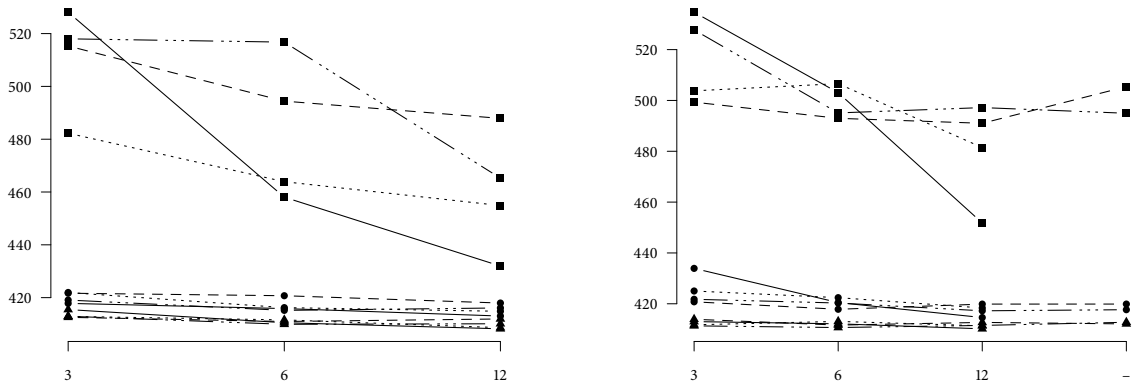


(C) Metropolis–Hastings y *Gibbs sampler*

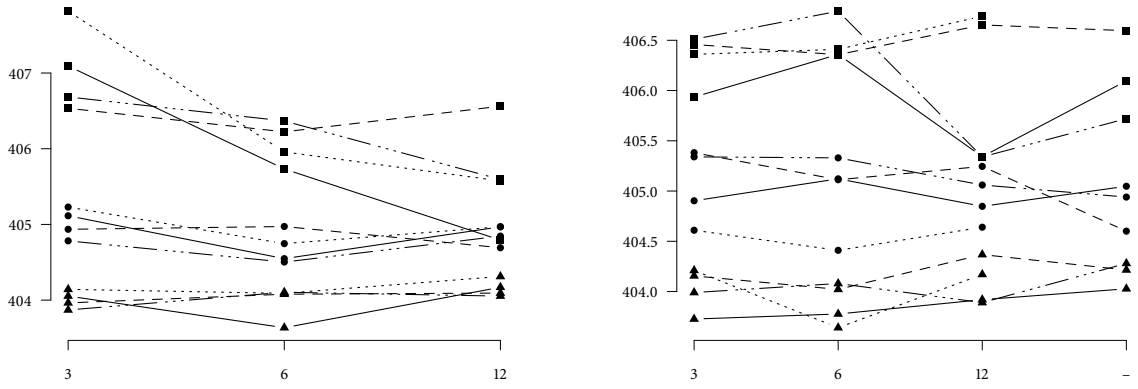
FIGURA 9: Devianzas estimadas, datos de las galaxias; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



(D) *Gibbs sampler* por bloques con N_1 (●), N_2 (▲) y N_3 (■) componentes para el proceso Dirichlet y N_1 (●) y N_2 (■) componentes para el resto, de acuerdo a la Tabla 2.

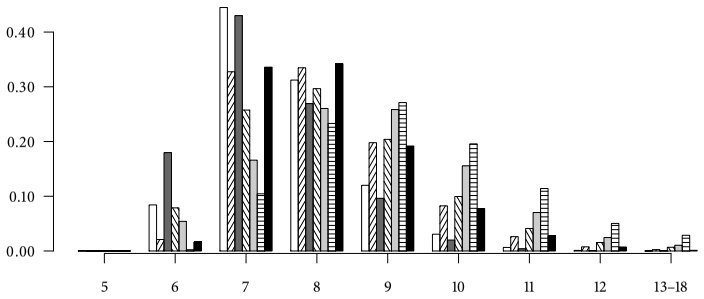


(E) Metropolis–Hastings para distribuciones no conjugadas con 1 (■), 5 (●) y 15 (▲) actualizaciones de K_i.

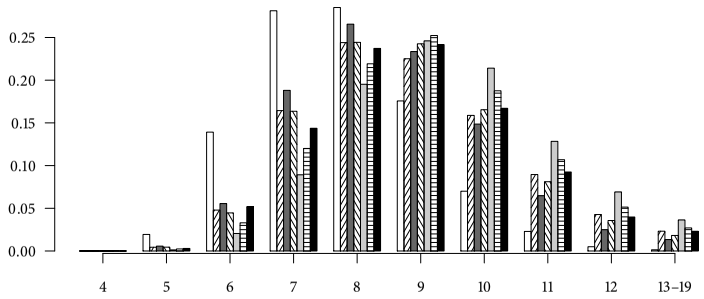


(F) *Gibbs sampler* con parámetros auxiliares, número de parámetros: 1 (■), 2 (●) y 10 (▲).

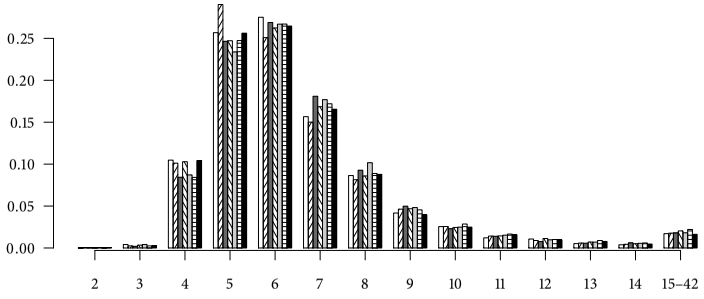
FIGURA 9: Devianzas estimadas, datos de las galaxias, continuación; los distintos tipos de línea corresponden a uno de los siguientes procesos: —— Dirichlet, – – – σ–estable normalizado, — ··· $\mathcal{P}_{D_\sigma}$ y ····· $\mathcal{P}_{D_\theta}$: mientras que los valores sobre el eje X corresponden a los valores *a priori* para $\mathbb{E}[K_n]$, el etiquetado con — corresponde a la distribución inicial no informativa.



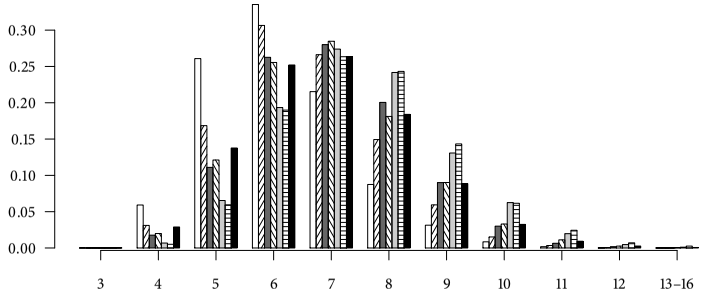
(A) *Gibbs sampler* para urnas de Pólya



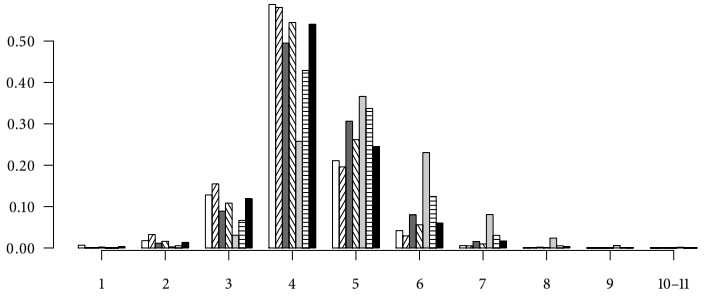
(B) *Gibbs sampler* para urnas de Pólya con paso de aceleración



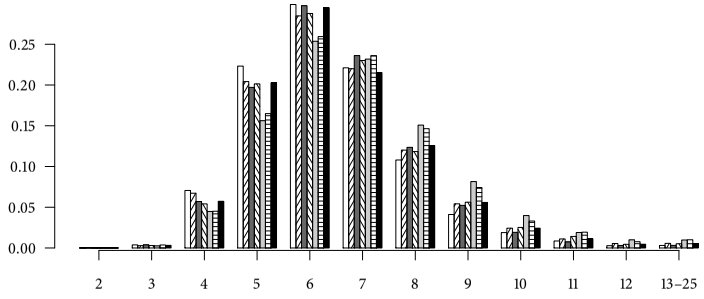
(C) Metropolis–Hastings y *Gibbs sampler*



(D) *Gibbs sampler* por bloques



(E) Metropolis–Hastings para distribuciones no conjugadas



(F) *Gibbs sampler* con parámetros auxiliares

Figura 10: Distribución posterior de K_n, datos de las galaxias; las texturas indican los valores iniciales para $\mathbb{E}[K_n]$: □ y ▨ 3, ■ y ▩ 6 y ▭ y ▤ 12; en las segundas se actualizaron los parámetros de los procesos y, asimismo, ■ corresponde a la distribución inicial no informativa para estos.

	$\mathbb{E}[K_n]$						
Método	3			6		12	—
G. s. para urnas de Pólya	10.02	21.61	348.31	3.00	3.16	1.64	15.34
G. s. para urnas de Pólya con paso de aceleración	0.50	0.50	0.50	0.50	0.50	0.50	0.50
M. H. y *Gibbs sampler*	0.91	1.20	1.21	0.96	1.02	1.14	0.95
Gibbs sampler por bloques	1.00	0.80	0.91	0.79	0.70	0.68	0.90
M. H. para distr. no conjugadas	10.76	12.33	6.55	7.75	3.31	6.96	8.98
G. s. con parámetros auxiliares	0.74	0.75	0.69	0.75	0.72	0.86	0.78

(A) Tiempo de autocorrelación integrado para la devianza, $\hat{\tau}_D$

	$\mathbb{E}[K_n]$						
Método	3			6		12	—
G. s. para urnas de Pólya	46.09	26.03	44.60	170.02	252.03	86.16	12.13
G. s. para urnas de Pólya con paso de aceleración	1.58	2.26	1.74	2.24	1.67	2.25	2.18
M. H. y *Gibbs sampler*	6.11	6.42	5.32	4.36	4.98	5.76	5.95
Gibbs sampler por bloques	18.64	7.63	7.52	7.16	10.77	5.57	10.57
M. H. para distr. no conjugadas	6.50	9.58	4.26	9.08	3.27	4.26	6.04
G. s. con parámetros auxiliares	4.06	3.24	3.14	3.20	2.30	3.06	3.15

(B) Tiempo de autocorrelación integrado para el número de grupos, $\hat{\tau}_{K_n}$

TABLA 13: Tiempo de autocorrelación integrado, datos de las galaxias

	Hiperparámetros				
	$(1.0, 1.0)$	$(0.5, 0.5)$	$(15.3, 1.7)$	$(24.5, 24.5)$	$(1.7, 15.3)$
Devianza	440.11	449.89	527.98	440.51	438.15
$\hat{\tau}_D$	148.59	760.58	350.26	6.11	7.03

TABLA 14: Resultados del *Gibbs sampler* para el proceso geométrico, datos de las galaxias

– B –
Programas de cómputo

Para hacer el análisis comparativo se hicieron programas de cómputo para cada método. En este apéndice se proporciona, en la primera sección, una descripción de cada uno de ellos. Además, con la finalidad de simplificar el uso de este conjunto de programas se programó una interfaz gráfica; en la segunda sección se presenta una guía de uso de esta interfaz.

§ 1 Conjunto de programas

Los programas realizados se pueden dividir en tres grupos:

1. Los programas con prefijo `run_` realizan una simulación utilizando alguno de los métodos para estimación de densidades estudiados:

   `run_npbm_polya`. *Gibbs sampler* para urnas de Pólya

   `run_npbm_polyaacc`. *Gibbs sampler* para urnas de Pólya con paso de aceleración

   `run_npbm_blocked`. *Gibbs sampler* por bloques

   `run_npbm_mh`. Metropolis–Hastings para distribuciones no conjugadas

   `run_npbm_mhgibbs`. Metropolis–Hastings y *Gibbs sampler*

   `run_npbm_gibbsaux`. *Gibbs sampler* con parámetros auxiliares

   `run_npbm_gibbsgeo`. *Gibbs sampler* para el proceso geométrico

   Todos estos requieren de un archivo con extensión `mcf` como argumento, el cual contiene toda la información necesaria para realizar la simulación.

2. Aquellos con prefijo `do_` se utilizan para realizar tareas previas o posteriores a la simulación:

   `do_npbm_config`. Crea un archivo con extensión `mcf` para realizar una simulación el cual, posteriormente, se utiliza para realizar otros cálculos

   `do_npbm_post_kn`. Calcula la distribución posterior de K_n

   `do_npbm_deviance`. Calcula la devianza estimada por iteración de acuerdo a (IV.2)

   `do_npbm_density`. Calcula la densidad estimada

do_npbm_posterior. Calcula la distribución posterior de los parámetros de los procesos: α, σ, (σ, θ) o λ; si es que fueron actualizados durante la simulación

do_npbm_histogram. Calcula el histograma para un archivo determinado

do_npbm_plot. Crea un gráfico utilizando los resultados de do_npbm_density o do_npbm_posterior

do_npbm_export. Gran parte de los programas anteriores crean archivos en formato binario; este programa convierte todos los archivos binarios en archivos de texto plano para una configuración específica

Cada uno de ellos requiere de uno o más argumentos; para conocer cuáles son se puede ejecutar el programa en una consola de comandos, por ejemplo, sin ninguno.

3. Por último, los programas con prefijo show_ muestran los resultados de algunos de los programas anteriores:

   show_npbm_post_kn. Muestra la distribución posterior de K_n, calculada por do_npbm_post_kn

   show_npbm_deviance. Muestra la devianza estimada de acuerdo a (IV.3), para ello se debió correr previamente do_npbm_deviance

   show_npbm_iat. Muestra $\hat{\tau}_D$ y $\hat{\tau}_{K_n}$

Además se tienen otros programas. La interfaz gráfica, que se explica a continuación, es el programa npbm_interface; éste, a su vez, requiere de otros programas: do_npbm_convert, do_npbm_preview, do_npbm_to_bin, show_npbm_random_list y show_npbm_range. Asimismo se incluye una serie de bibliotecas de funciones, son todos aquellos archivos con extensión so.

§ 2 INTERFAZ GRÁFICA

Todos los programas funcionan a través de la consola de comandos. Sin embargo, para integrarlos y facilitar su uso se programó una interfaz gráfica. Con esta interfaz se pueden realizar las acciones típicas de crear, editar y eliminar «configuraciones». Una configuración se refiere a un archivo con extensión mcf con toda la información necesaria para realizar una corrida, para un conjunto específico de datos.

Esencialmente, los pasos a seguir para utilizar la interfaz son los siguientes:

1. Crear un «espacio de trabajo», i.e., seleccionar un conjunto de datos y un directorio para guardar todos los archivos resultantes
2. Agregar configuraciones según se requiera
3. Correr cada una de las configuraciones
4. Exportar los resultados y gráficos obtenidos

La ventana principal de la interfaz se muestra en la Figura 1 y está formada por los siguientes elementos:

1. Una lista con todas las configuraciones creadas
2. Una ventana web que muestra un resumen de la configuración seleccionada en (1)
3. Una ventana web que muestra los resultados de la corrida para configuración seleccionada
4. Una consola que muestra información acerca de la tarea que se esté ejecutando
5. Una pestaña que contiene opciones para los gráficos

6–8. Elementos comunes a las interfaces gráficas: barra de menús, dos barras de herramientas y barra de estado, respectivamente

Todas las opciones de las barras de herramientas se encuentran en los menús *Archivo* o *Configuraciones*, por tal razón, únicamente se hará referencia a los menús y sus opciones.

2.1 Espacio de trabajo

Lo primero que se debe de hacer para trabajar es crear un «espacio de trabajo». Un espacio de trabajo está formado por dos elementos: un archivo de texto plano con los datos que se deseen trabajar (cada renglón representa una observación unidimensional) y un directorio de trabajo en el cual se crearán todos los archivos necesarios para realizar las corridas y presentar sus resultados.

Para crear un espacio de trabajo se tiene la opción *Archivo* | *Nuevo espacio de trabajo...*. Al seleccionarla se abrirá una ventana como la que se muestra en la Figura 2. En el campo Datos se debe especificar la ruta del archivo de datos, mientras que en el campo Directorio el directorio de trabajo (se recomienda que sea uno vacío). Cuando se tengan ambas rutas y sean válidas se generará una vista previa de los datos, i.e., aparecerá su histograma, algunas estadísticas descriptivas y la lista de datos.

También es posible abrir algún espacio de trabajo existente o guardar el espacio de trabajo actual de acuerdo a las siguientes opciones:

- Para abrir un espacio de trabajo existente, la opción *Archivo* | *Abrir espacio de trabajo...*;
- Si se desea guardar el espacio de trabajo actual, la opción *Archivo* | *Guardar espacio de trabajo*;

El tipo de archivo que puede abrirse o guardarse debe tener extensión mws; este tipo de archivo únicamente contiene la lista de configuraciones, por lo que los resultados no serán guardados. Es recomendable, por tanto, conservar el directorio de trabajo con su contenido.

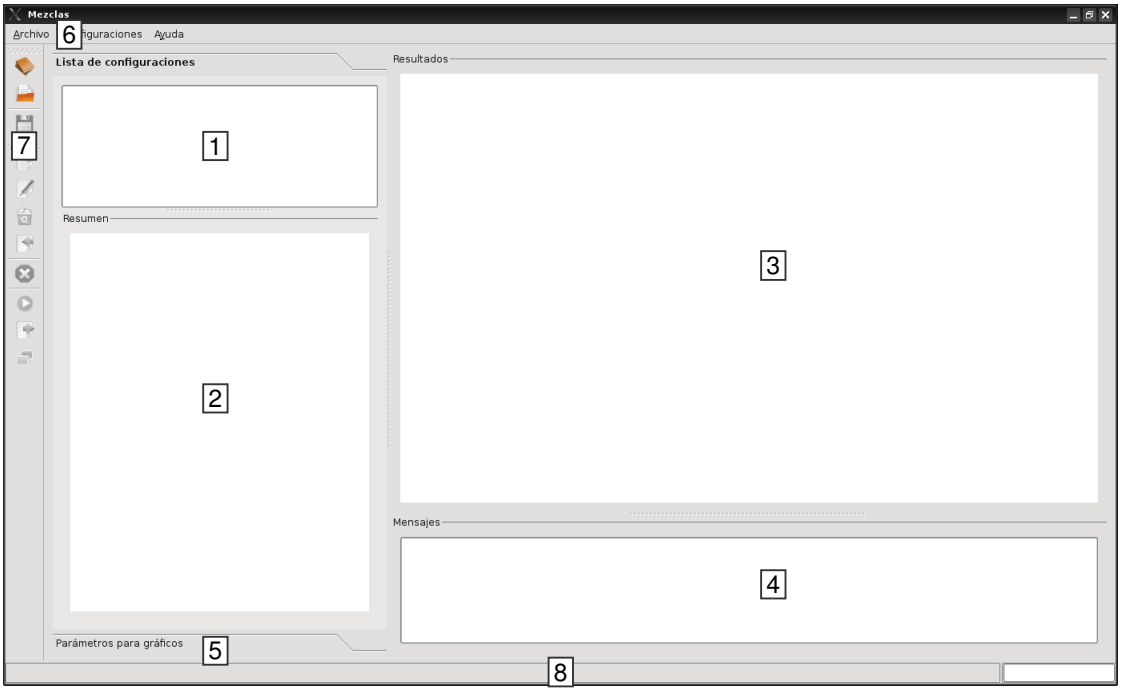


FIGURA 1: Ventana principal de la interfaz gráfica

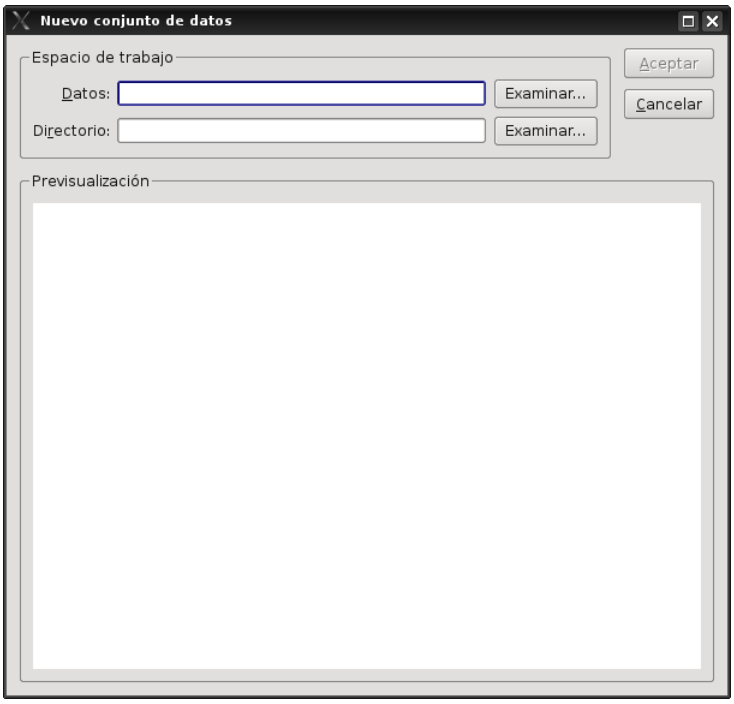
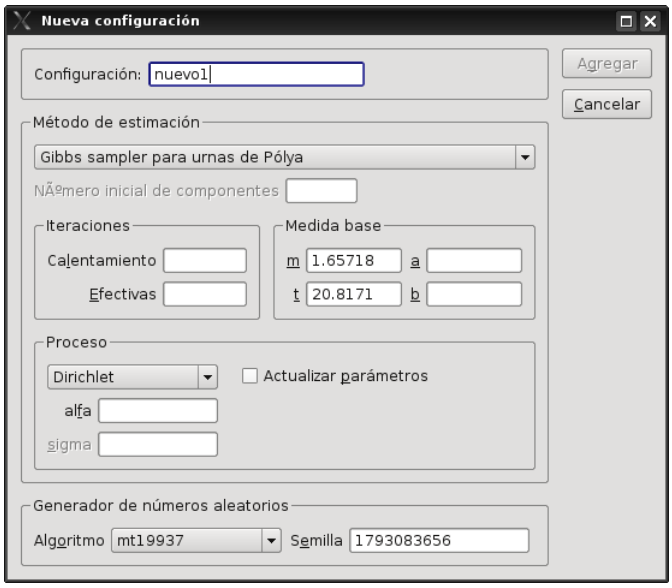
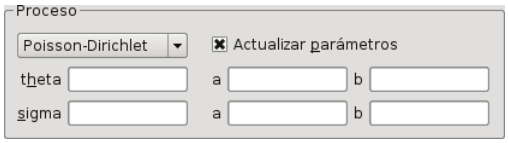


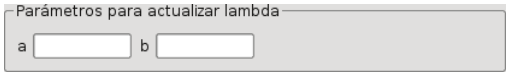
FIGURA 2: Ventana para crear un nuevo espacio de trabajo



(A) Ventana principal



(B) Actualización de parámetros de los procesos



(C) Actualización de λ

FIGURA 3: Ventana y opciones para agregar/editar configuraciones

2.2 CONFIGURACIONES

Una vez creado el espacio de trabajo se pueden agregar configuraciones utilizando la opción *Configuraciones | Nueva configuración...*. Esta acción mostrará una ventana como la que se muestra en la Figura 3a.

En esta ventana se podrá agregar toda la información que se necesita para realizar una simulación. El campo Configuración es un nombre que servirá para identificar la configuración en el resto del programa. En el grupo Método de estimación se dan todos los parámetros de la simulación, de acuerdo al método seleccionado en el cuadro combinado:

- Si el método es *Gibbs sampler* por bloques, Metropolis–Hastings, *Gibbs sampler* con parámetros auxiliares o *Gibbs sampler* para el proceso geométrico se activará en campo debajo del cuadro combinado en el cual se debe indicar el valor del parámetro auxiliar (véase la Sección IV.3).

- En seguida, en el grupo Iteraciones se debe especificar cuántas iteraciones se harán en la simulación, tanto las del periodo de calentamiento como las del muestreo efectivo.

- En el grupo Medida base se especifican los parámetros de la medida base[1] (m, t, a, b); de manera predeterminada se dan valores para m y t: el punto medio del rango y el rango de los datos, respectivamente. Para estos mismos campos, al dar un clic derecho aparece un menú contextual que muestra dos valores para cada valor, además de los ya mencionados se puede elegir la media y la varianza muestrales, respectivamente.

- Con excepción del *Gibbs sampler* para el proceso geométrico, debajo de los campos mencionados se puede elegir el proceso subyacente a la medida de probabilidad aleatoria: Dirichlet, σ–estable normalizado o Poisson–Dirichlet de dos parámetros. Al seleccionar alguno se pueden realizar dos acciones: dar un valor a sus parámetros o incluirlos en el esquema de simulación (Sección III.2). Para la segunda acción se debe activar la casilla Actualizar parámetros, lo cual mostrará dos campos adicionales para cada parámetro (Figura 3b).

- Si se selecciona el *Gibbs sampler* para el proceso geométrico, en lugar de hacer lo anterior se deben dar valores para la distribución inicial de λ (Sección III.1.4), Figura 3c.

- De manera opcional, en la última parte, grupo Generador de números aleatorios se puede elegir algún algoritmo generador de números aleatorios junto con un valor para la semilla. Los algoritmos disponibles están tomados directamente de las bibliotecas GNU SCIENTIFIC LIBRARY[2].

[1] En toda la tesis se manejó ω como el primer parámetro de la medida base, sin embargo, en la interfaz se cambió a m por razones prácticas.

[2] Puede consultarse la documentación de estas bibliotecas para mayor referencia, por ejemplo, en las Secciones 9–11 de `http://www.gnu.org/software/gsl/manual/html_node/Random-Number-Generation.html`.

Es importante mencionar que se podrá agregar la nueva configuración si todos los campos contienen un valor válido, e.g., que al menos se realice una iteración de muestreo efectivo, que los parámetros de la medida base sean no negativos (excepto m), etc.

Modificación de configuraciones

Una vez agregada la configuración, en la ventana principal, en (1), aparecerá una entrada con el nombre dado. Al seleccionar el nombre, en (2) aparecerá un resumen de la configuración.

Si se desea modificar alguna configuración de la lista se puede hacer de dos maneras: seleccionar el nombre de la configuración y, posteriormente, seleccionar la opción *Configuraciones | Editar configuración...*, o simplemente dar doble clic sobre el nombre. En ambos casos se abrirá la ventana de la Figura 3a.

Además, es posible eliminar alguna configuración seleccionando la opción *Configuraciones | Borrar configuración*, seleccionando previamente el nombre de la configuración.

Por último, se puede importar una configuración a través de archivos con extensión `mcf` con una configuración válida. Para hacerlo se debe seleccionar la opción *Configuraciones | Importar configuración...*. Si el archivo seleccionado es válido se agregará a la lista de configuraciones.

2.3 Corridas

Después de agregar todas las configuraciones deseadas se pueden realizar las «corridas». Se entiende por una corrida las siguientes acciones: la simulación, el cálculo de la densidad estimada y distribuciones posteriores de los parámetros de los procesos (si aplica), el cálculo de la distribución posterior de K_n (si aplica), de la devianza estimada, de $\hat{\tau}_D$ y de $\hat{\tau}_{Kn}$.

Es posible seleccionar varias configuraciones para correrlas secuencialmente. Después de seleccionarlas en (1), sólo es necesario seleccionar la opción *Configuraciones | Correr seleccionados*. Durante este proceso se activará la opción *Configuraciones | Cancelar* para detener las corridas si es necesario. Además, en (4) se mostrará el progreso de cada corrida.

Al terminar cada corrida se crea un reporte con los resultados. Para verlo en (3) se debe seleccionar en (1) el nombre de la configuración.

2.4 Resultados

Además de ver el reporte en el programa, es posible exportar tanto los datos de la corrida como los gráficos. Para ambas tareas se deben seleccionar las configuraciones que se deseen exportar y elegir qué se desea exportar:

- Para exportar los resultados se debe seleccionar la opción *Configuraciones | Exportar resultados...*. Esto mostrará un diálogo en el cual se debe elegir el directorio en donde se quieren guardar los archivos de texto plano.

FIGURA 4: Parámetros para hacer gráficos

- Si se desean exportar los gráficos se debe seleccionar la opción *Configuraciones* | *Exportar gráficos...*. De manera similar a lo anterior se debe elegir el directorio en donde se quieren guardar los gráficos.

La pestaña (5) contiene opciones para realizar los gráficos (Figura 4). En esta se observan dos grupos: Densidad y Posteriores; en cada uno se elige el intervalo a graficar, en función del rango de los datos, y el tamaño de la partición del mismo, para la densidad estimada y las distribuciones posteriores, respectivamente. Estos valores se deben fijar antes de hacer las corridas, de lo contrario se deberá dar clic en el botón Actualizar gráficos para que los cambios tengan efecto. El valor predeterminado para Intervalo es 120 % y para Partición es 200.

Para exportar los gráficos se utiliza el programa GNU PLOT. Por esta razón en el grupo Opciones gnuplot para exportar se tienen dos campos: Terminal debe ser una terminal válida de GNU PLOT y Opciones son las opciones de la terminal seleccionada. De manera predeterminada se crearán archivos POSTSCRIPT a color de tamaño 15×10 cm.

Bibliografía

Aldous, D. J. (1985), Exchangeability and related topics, *in* «École d'Été de Probabilités de Saint-Flour XIII», Lecture notes in mathematics, Springer-Verlag.

Antoniak, C. E. (1974), «Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems», *The Annals of Statistics* **2**(6), 1152–1174.

Blackwell, D. y MacQueen, J. B. (1973), «Ferguson distributions via Pólya urn schemes», *The Annals of Statistics* **1**(2), 353–355.

Ferguson, T. S. (1973), «A Bayesian analysis of some nonparametric problems», *The Annals of Statistics* **1**(2), 209–230.

Freedman, D. A. (1963), «On the asymptotic behavior of Bayes' estimates in the discrete case», *The Annals of Mathematical Statistics* **34**(4), 1386–1403.

Fuentes-García, R., Mena, R. H. y Walker, S. G. (2010), «A new Bayesian nonparametric mixture model», *Communications in Statistics: Simulation and Computation* **39**(4), 669–682.

Geman, S. y Geman, D. (1984), «Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images», *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.

Gilks, W., Best, N. G. y Tan, K. K. G. (1995), «Adaptive rejection Metropolis sampling within Gibbs sampling», *Applied Statistics* **44**(4), 455–472.

Hastings, W. K. (1970), «Monte Carlo sampling methods using Markov chains and their applications», *Biometrika* **57**(1), 97–109.

Ishwaran, H. y James, L. F. (2001), «Gibbs sampling methods for stick-breaking priors», *Journal of the American Statistical Association* **96**(453), 161–173.

James, L. F., Lijoi, A. y Prünster (2006), «Conjugacy as a distinctive feature of the Dirichlet process», *Scandinavian Journal of Statistics* **33**(1), 105–120.

Kalli, M., Griffin, J. y Walker, S. (2009), «Slice sampling mixture models», *Statistics and Computing* . **URL:** *http://dx.doi.org/10.1007/s11222-009-9150-y*

Kingman, J. F. C. (1975), «Random discrete distributions», *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **37**, 1–22.

Lijoi, A., Mena, R. H. y Prünster, I. (2007*a*), «Bayesian nonparametric estimation of the probability of discovering a new species», *Biometrika* **94**, 769–786.

Lijoi, A., Mena, R. H. y Prünster, I. (2007*b*), «Controlling the reinforcement in Bayesian nonparametric mixture models», *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(4), 715–740.

Lo, A. Y. (1984), «On a class of Bayesian nonparametric estimates: I. Density estimates», *The Annals of Statistics* **12**(1), 351–357.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. y Teller, A. H. (1953), «Equation of state calculations by fast computing machines», *The Journal of Chemical Physics* **21**(6), 1087–1092.

Neal, R. M. (2000), «Markov chain sampling methods for Dirichlet process mixture models», *Journal of Computational and Graphical Statistics* **9**(2), 249–265.

Pitman, J. (1996), Some developments of the Blackwell–MacQueen urn scheme, *in* T. S. Ferguson, L. S. Shapley y J. B. MacQueen, eds, «Statistics, Probability and Game Theory; papers in honor of David Blackwell», Vol. 30 of *Lecture Notes–Monograph Series*, Institute of Mathematical Statistics, pp. 245–267.

Pitman, J. (2002), *Combinatorial stochastic processes*, Ecole d'eté de probabilités de Saint-Flour XXXII - 2002, Springer.

Pitman, J. y Yor, M. (1997), «The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator», *The Annals of Probability* **25**(2), 855–900.

Regazzini, E., Lijoi, A. y Prünster, I. (2003), «Distributional results for means of normalized random measures with independent increments», *The Annals of Statistics* **31**(2), 560–585.

Richardson, S. y Green, P. J. (1997), «On Bayesian analysis of mixtures with an unknown number of components», *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(4), 731–792.

Robert, C. P. y Casella, G. (2005), *Monte Carlo statistical methods*, Springer Texts in Statistics, 2 edn, Springer-Verlag New York, Inc.

Sato, K. (1999), *Lévy processes and infinitely divisible distributions*, Cambridge University Press.

Schervish, M. J. (1995), *Theory of Statistics*, Springer-Verlag, New York, NY.

Tierney, L. (1994), «Markov chains for exploring posterior distributions», *The Annals of Statistics* **22**(4), 1701–1728.

West, M. (1992), Hyperparameter estimation in Dirichlet process mixture models, Technical report, Institute of Statistics and Decision Sciences, Duke University.