Estadística Bayesiana, algoritmos MCMC y modelos jerárquicos

Mario Enrique Carranza Barragán
Dra. Leticia Ramírez Ramírez

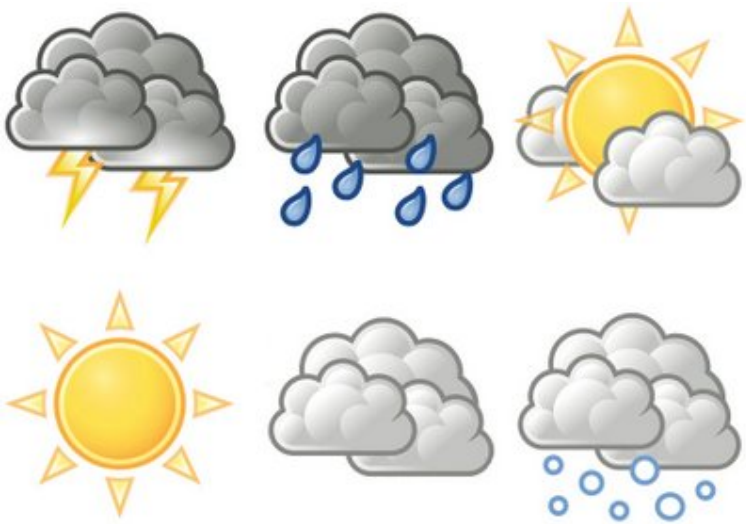
Grupo Bimbo/ Centro de Investigación en Matemáticas

18, 20 y 25 de abril de 2023

mario.carranza@grupobimbo.com/ mario.carranza@cimat.mx

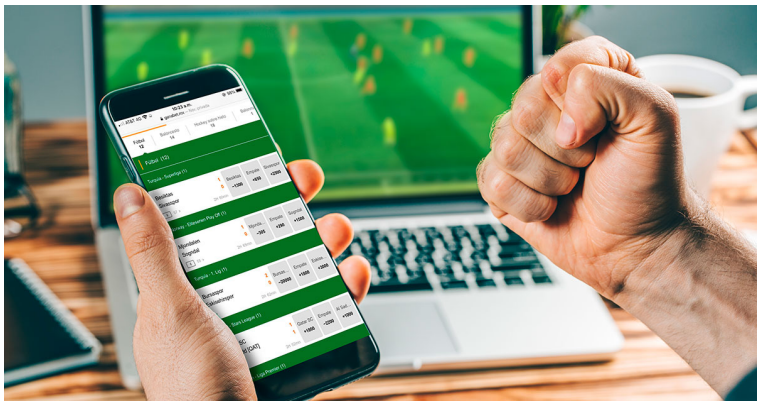
Tabla de contenidos

La probabilidad subjetiva/ Apuestas justas



(a) Clima

(b) Política



(c) Momios/ Pagos de la casa

(d) Opciones financieras

Inferencia estadística

El problema de la inferencia estadística

En general tenemos:

- Datos X
- Cantidades desconocidas θ

Como estadísticos (frecuentistas o Bayesianos), postulamos un modelo de probabilidad para los datos

$$p(x|\theta).$$

Con el paradigma Bayesiano

Además:

- θ debe tener una distribución de probabilidad que refleje nuestra incertidumbre inicial acerca de su valor.
- X es conocido, así que debemos condicionar en su valor observado, x.

Así, nuestro conocimiento de θ queda descrito en su distribución final $p(\theta|x)$. El Teorema de Bayes nos dice cómo encontrarla:

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = \frac{p(\theta)p(x|\theta)}{\int p(\tilde{\theta})p(x|\tilde{\theta})d\tilde{\theta}}.$$

Obteniendo la distribución posterior

Notamos que en

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\int p(\tilde{\theta})p(x|\tilde{\theta})d\tilde{\theta}}$$

el denominador $p(x) = \int p(\tilde{\theta})p(x|\tilde{\theta})d\tilde{\theta}$ no depende de θ. Es común escribir

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

Obtener la constante de normalización, también llamada densidad de la predictiva previa en los datos observados, en general no es sencillo.

El reto de la inferencia Bayesiana: Cálculos (integrales)

El problema son las densidades marginales para cada uno de los parámetros de interés

$$p(\theta_i|x) = \int p(\theta|x)d\theta_{-i} \text{ con } \theta_{-i} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p)$$

así como la constante de normalización de la distribución final.
Bajo el enfoque frecuentista el problema de los parámentos de estorbo se puede resolver con la llamada función de verosimilitud perfil.

La posterior de hoy es la previa de mañana

Supongamos un modelo $p(x|\theta)$, una previa $p(\theta)$ y una muestra aleatoria $x = (x_1..., x_m)$ de donde se obtiene $p(\theta|x)$. Luego de un tiempo se obtienen más observaciones $x_b = (x_{b1}, ..., x_{bn})$.
Asumiendo intercambiabilidad (o independencia condicional) es equivalente para obtener la distribución posterior $p(\theta|x, x_b)$ tanto:

- Usar la posterior $p(\theta|x)$ como previa del modelo y la verosimilitud basada en x_b.
- Usar $p(\theta)$ y la verosimilitud usando todas las muestra (x, x_b) como una sola muestra.

Útil cuando lidiamos con pruebas o estimación de naturaleza secuencial o de aprendizaje continuo.

Problemas de inferencia como problemas de decisión

La teoría de inferencia Bayesiana nos indica que debemos tratar los problemas de inferencia dentro del marco de problemas de desición con incertidumbre.

- $\mathcal{A}$: las posibles acciones o decisiones.
- $\mathcal{E}$: los posibles estados de naturaleza.
- $\mathcal{C}$: conjunto de consecuencias.

Cada $(a_i, e_j) \in \mathcal{A} \times \mathcal{E}$ tiene asociada una única consecuencia c_{ij}.
Debe cumplirse que existe una relación de preferencia en $\mathcal{C}$ tal que $c_1, c_2 \in \mathcal{C}$ sólo se cumple una de las siguientes:

$$c_1 \succ c_2, c_1 \sim c_2, c_1 \prec c_2.$$

Podemos condensar estas preferencias en una función de utilidad $U(c_{ij}) = U(a_i, e_j)$ o perdida $L(c_{ij}) = L(a_i, e_j)$.

Problemas de inferencia como problemas de decisión

- Nuestro conocimiento sobre la probabilidad de ocurrencia de los eventos $\mathcal{E}$ a través de una medida de probabilidad($\mathcal{E}$ es $\sigma-$álgebra).
- Se usan apuestas para definir qué es probabilidad.
- En la teoría se requiere conocer dos consecuencias, c^*(la mejor) y c_* (la peor).
- En los contextos de inferencia, dado un modelo del fenómeno la información del estado de naturaleza está contenido en el conocimiento del parámetro θ.

Resultado principal de la teoría de la decisión Bayesiana

Los problemas bajo incertidumbre se resuelven minimizando la perdida (maximizando la utilidad) esperada posterior.

$$a^* = \min_a E_{\theta|D} L(a, \theta)$$

Podemos decir que...

Bajo el enfoque frecuentista:

- El estimador es la v.a.
- Una v.a. toma distintos valores y su probabilidad es el límite de su frecuencia relativa.
- Cada problema de inferencia tiene una metodología propia.
- Los problemas numéricos usualmente son de optimización.
- Hay más trabajo sobre el problema de validación.

Bajo el enfoque Bayesiano:

- El parámetro es la v.a.
- Una v.a. es cualquier cantidad desconocida (incertidumbre).
- Existe una única receta para resolver cualquier problema de inferencia.
- Los problemas numéricos usualmente son de integración.
- Es más natural el problema de predicción.

Problemas de inferencia

Bajo el enfoque frecuentista:

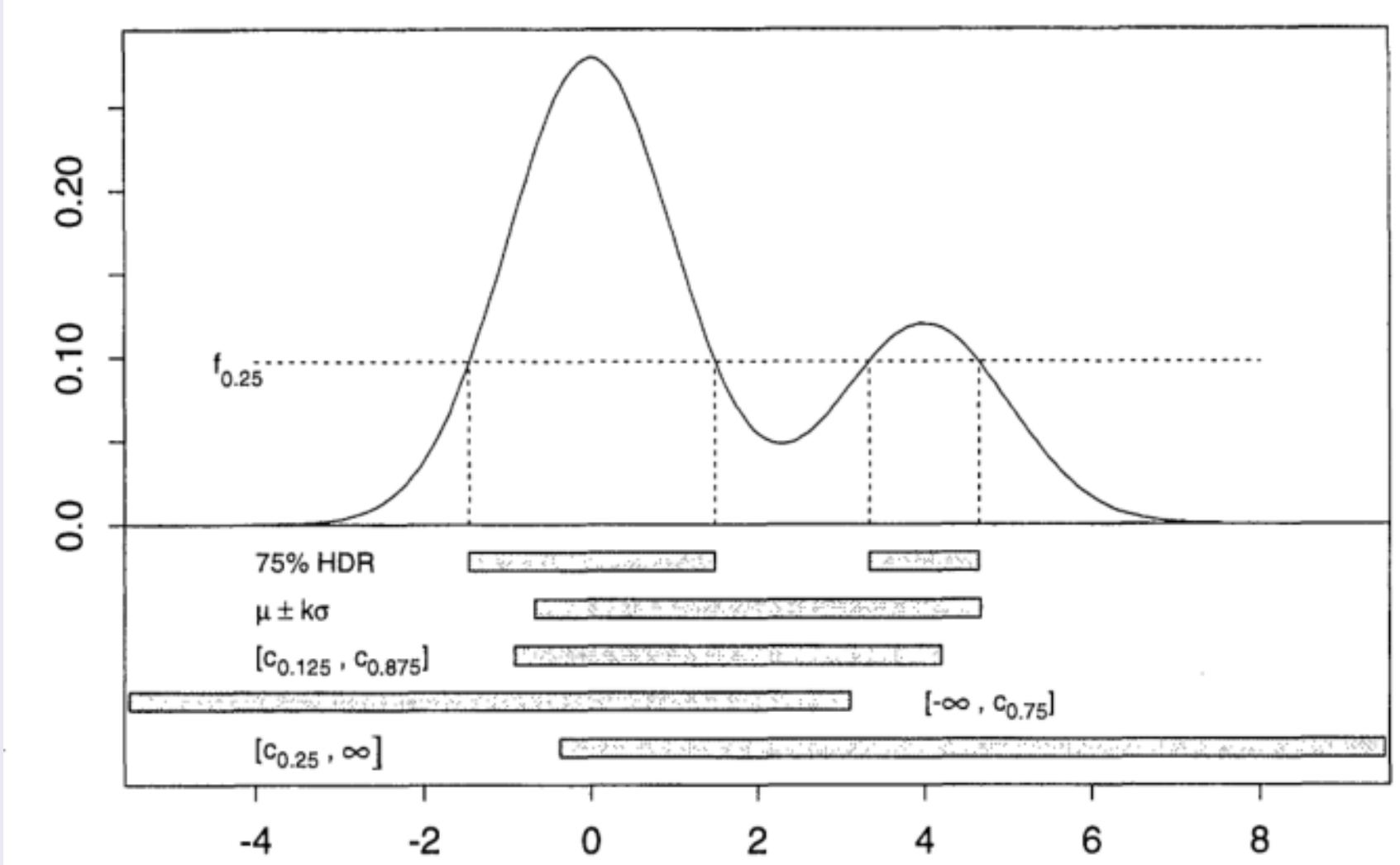
- Estimadores puntuales: Método de momentos y máxima verosimilitud (consistencia)
- Estimación por intervalos: Intervalos de confianza- Cantidades pivotales (Wald)
- Pruebas de hipótesis (Estilo Neyman-Pearson): Razón de verosimilitud generalizada (Wilks)

Bajo el enfoque Bayesiano:

- Estimadores puntuales:
  $\mathcal{A} = \{a; a \in \Theta\}$
  $L_1(a, \theta) = (a - \theta)^2 \rightarrow$ Media,
  $L_2(a, \theta) = |a - \theta| \rightarrow$ Mediana
- Estimación por intervalos:
  $\mathcal{A} = \{B; B \subseteq \Theta\}$
  $L(B, \theta) = \lambda(B) + k\mathbb{1}_{B^c}(\theta)$
  Intervalos de máxima densidad y de colas iguales.
- Pruebas de hipótesis (Estilo NP): $\mathcal{A} = \{H_0, H_1\}$,
  $L(H_i, H_j) = c_{ij}$ con $i, j = 0, 1$.
  Factor de Bayes.

Region de máxima densidad, colas iguales, etc.

Región o intervalos de credibilidad Bayesianos



Distribuciones iniciales objetivas

Interpretaciones de las distribuciones iniciales objetivas

- Las distribuciones iniciales objetivas son representaciones de ignorancia.
- Una distribución inicial objetiva es aquella que provee poca información en relación a la aportada por el experimento.
- Manifiesta la poca información *a priori* sobre una magnitud o al menos se actúa como si se fuese ignorante sobre dicha magnitud.

Principio de razón insuficiente

- Cuando a priori no se conoce nada sobre θ, asúmase que la incial $\pi(\theta)$ es una distribución uniforme.
- Desconocer la probabilidad de eventos mutuamente excluyentes y conocer que estos tienen la misma probabilidad son dos estados de conocimiento muy distintos.
- La distribución uniforme continua no es invariante bajo reparametrización.

Otras alternativas

Asignar una distribución plana (quizás impropia) un parámetro no implica que sus transformaciones tengan distribuciones planas. Jeffreys propone una distribución que no cambia ante transformaciones en que el parámetro es invariante (en especial la escala).

Análisis conjugado

Definición

Si la distribución posterior $p(\theta|x)$ pertenece a la misma familia de distribuciones que la distribución previa $p(\theta)$, se dice que tanto la previa como la posterior son distribuciones conjugadas y se dice que la previa es conjugada previa para el modelo (o verosimilitud) $p(x|\theta)$.

Ejemplo de análisis conjugado

Binomial (Bernoulli) o series de volados

Suponga que $\boldsymbol{X} = (X_1, ..., X_n)$ es una muestra de distribución Binomial(m, θ) , y que θ tiene una previa Beta(α, β). Entonces la distribución posterior de θ dado $\boldsymbol{x}$ es Beta$(\alpha + mn, (\sum_{i=1}^{n} x_i) + \beta)$. Recordemos que la función de densidad de una distribución Beta es

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Ejemplo de análisis conjugado

Veamos que

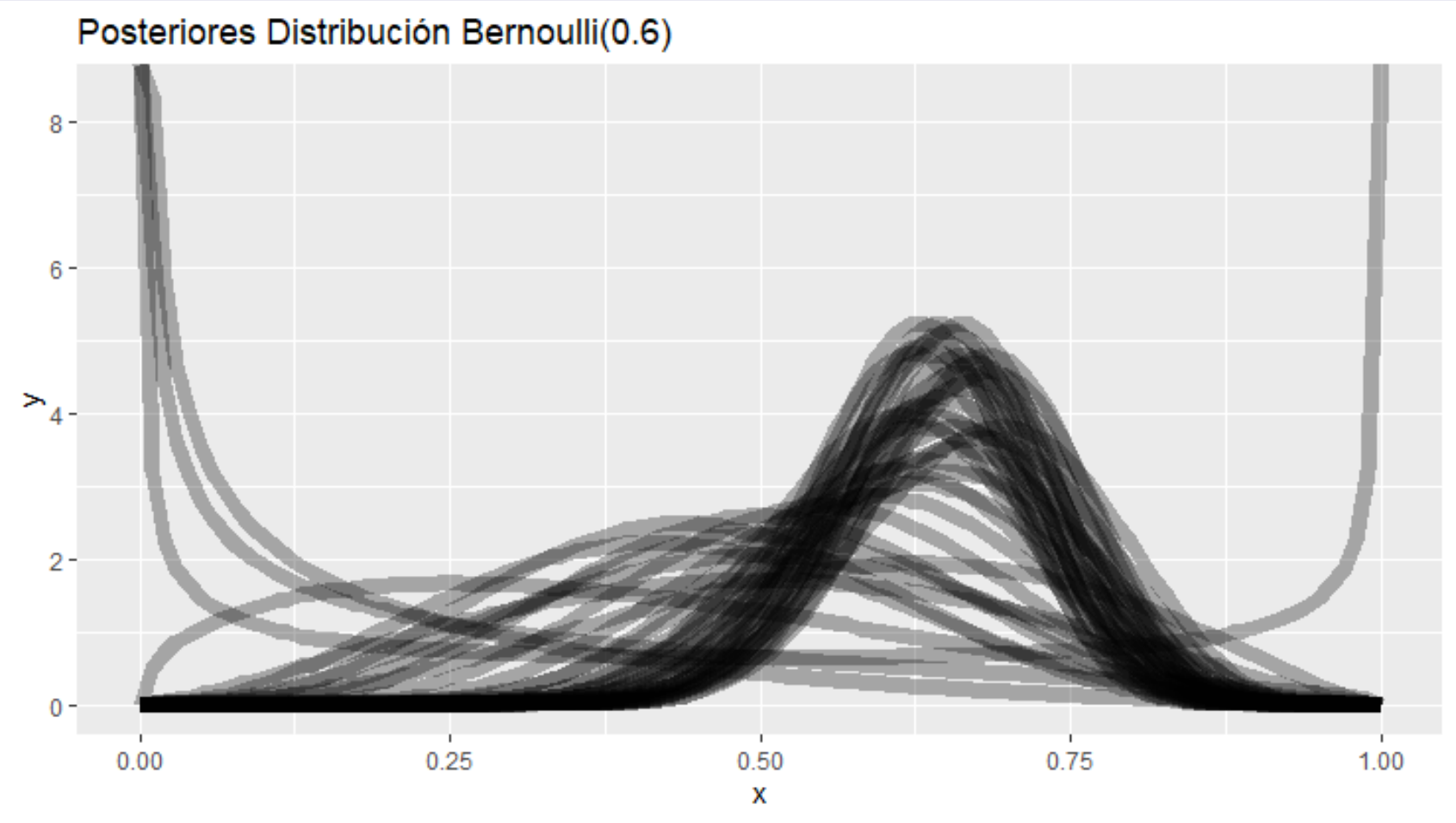
$$
\begin{aligned}
p(\theta|m, \boldsymbol{X}) &\propto p(\boldsymbol{X}|m, \theta)p(\theta) \\
&\propto \prod_{i=1}^{n} \left[\binom{m}{x_i} \theta^{x_i}(1-\theta)^{m-x_i} \right] \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \left[\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{nm-\sum_{i=1}^{n} x_i} \right] \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{\sum_{i=1}^{n} x_i+\alpha-1}(1-\theta)^{nm-\sum_{i=1}^{n} x_i+\beta-1}
\end{aligned}
$$

Notamos que la densidad posterior es proporcional al kernel de una distribución Beta con parámetros $\sum_{i=1}^{n} x_i + \alpha$ y $nm - \sum_{i=1}^{n} x_i + \beta$. Por lo que concluimos que

$$
\theta|\boldsymbol{X} \sim \text{Beta}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + nm - \sum_{i=1}^{n} x_i \right)
$$

Gráficas de las densidades posteriores

Ejemplo Beta-Binomial



Ejemplo de análisis conjugado

Binomial Negativa

Suponga que $\boldsymbol{X} = (X_1, ..., X_n)$ es una muestra de distribución BinNeg(m, θ) , y que θ tiene una previa Beta(α, β). Muestre que la distribución posterior de θ dado $\boldsymbol{x}$ es Beta$(\alpha + mn, (\sum_{i=1}^{n} x_i) + \beta)$. Recordemos que la función de densidad de una distribución Beta es

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Ejemplo de análisis conjugado

Veamos que

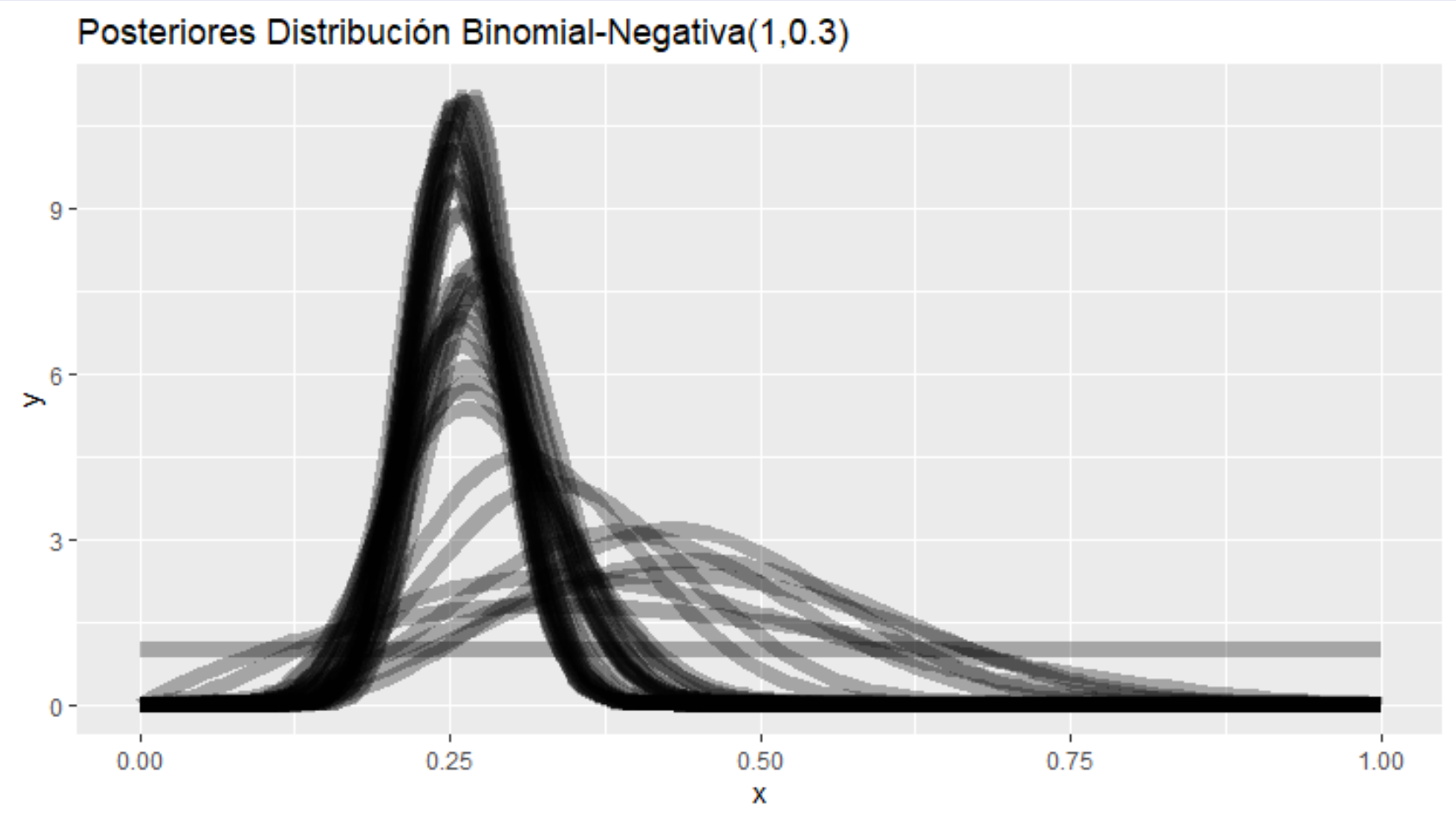
$$
\begin{aligned}
p(\theta|k, \boldsymbol{X}) &\propto p(\boldsymbol{X}|k, \theta)p(\theta) \\
&\propto \prod_{i=1}^{n} \left[\binom{k + x_i - 1}{k} \theta^k (1-\theta)^{x_i} \right] \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \left[\theta^{nk} (1-\theta)^{\sum_{i=1}^n x_i} \right] \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{nk+\alpha-1}(1-\theta)^{\sum_{i=1}^n x_i + \beta - 1}
\end{aligned}
$$

Notamos que la densidad posterior es proporcional al kernel de una distribución Beta con parámetros $nk + \alpha$ y $\sum_{i=1}^n x_i + \beta$. Por lo que concluimos que

$$
\theta|\boldsymbol{X} \sim \text{Beta}\left(\alpha + kn, \beta + \sum_{i=1}^{n} x_i \right)
$$

Gráficas de las densidades posteriores

Ejemplo Beta-BinomialNegativa



Tercer ejemplo de análisis conjugado

Veamos que la distribución Gama es conjugada al modelo Poisson.

$$\text{Gama}(a, b) \xrightarrow{\lambda} \text{Pois}(\lambda)$$

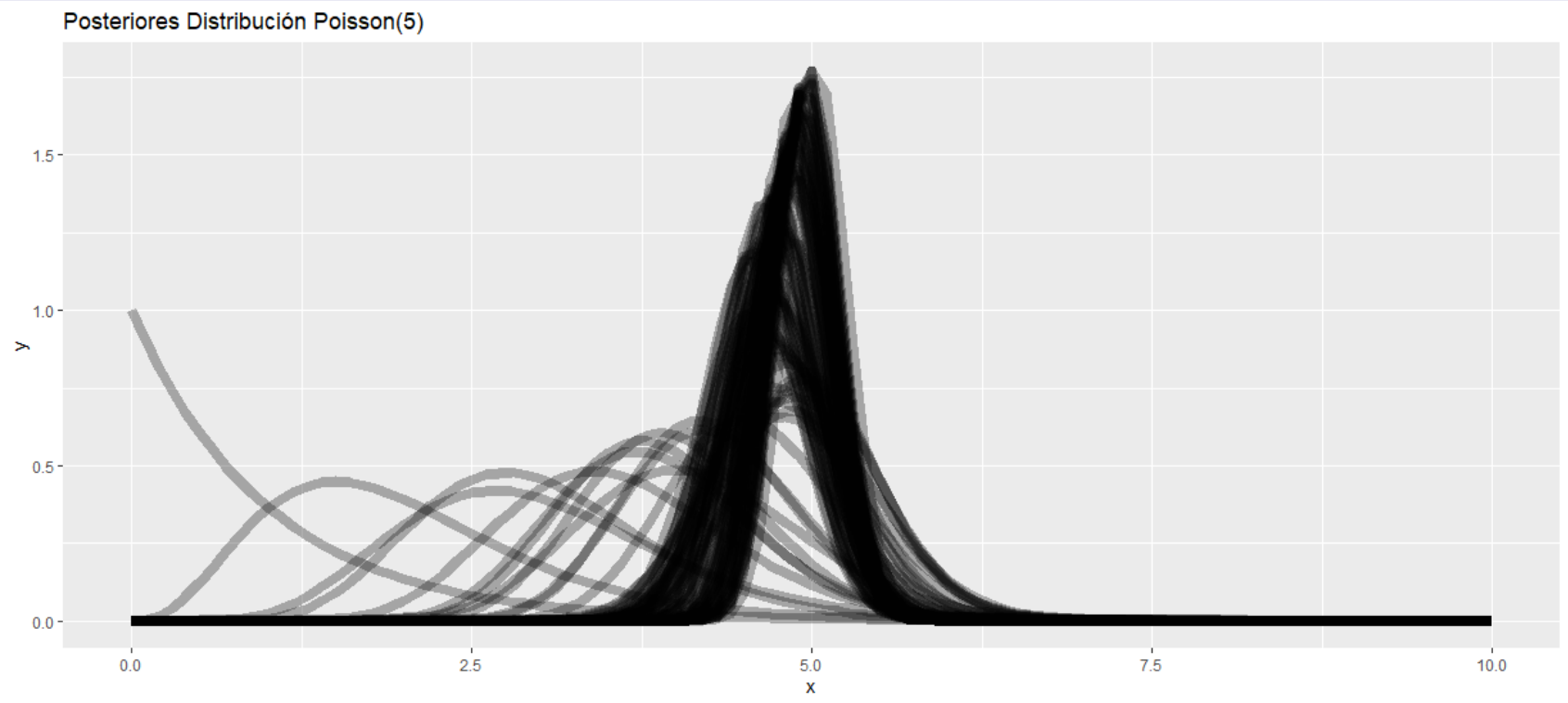
$$\begin{aligned} p(\theta|\vec{x}) &= \frac{(b\theta)^a e^{-\theta b}}{\theta\Gamma(a)} \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \\ &\propto \theta^{a-1} e^{-b\theta} \theta^{\sum x_i} e^{-n\theta} = \theta^{a+\sum x_i - 1} e^{-\theta(b+n)} \\ &= \theta^{a+\sum x_i - 1} e^{-\theta(b+n)} \end{aligned}$$

Así,

$$\theta|\vec{x} \sim \text{Gama}(a + \sum x_i, b + n)$$

Gráficas de las densidades posteriores

Ejemplo Gama-Poisson



Cuarto ejemplo de análisis conjugado

Veamos que la distribución Gama es conjugada al modelo Exponecial (Gama(1,λ)).

$$\text{Gama}(a, b) \xrightarrow{\theta} \text{Exp}(\theta)$$

$$p(\theta|\vec{x}) \propto \theta^{a-1} \exp(-b\theta) \prod_{i=1}^{n} \theta \exp\{-x_i\theta\}$$

$$= \theta^{a-1} \exp(-b\theta)\, \theta^n \exp\left\{-\sum_{i=1}^{n} x_i\theta\right\}$$

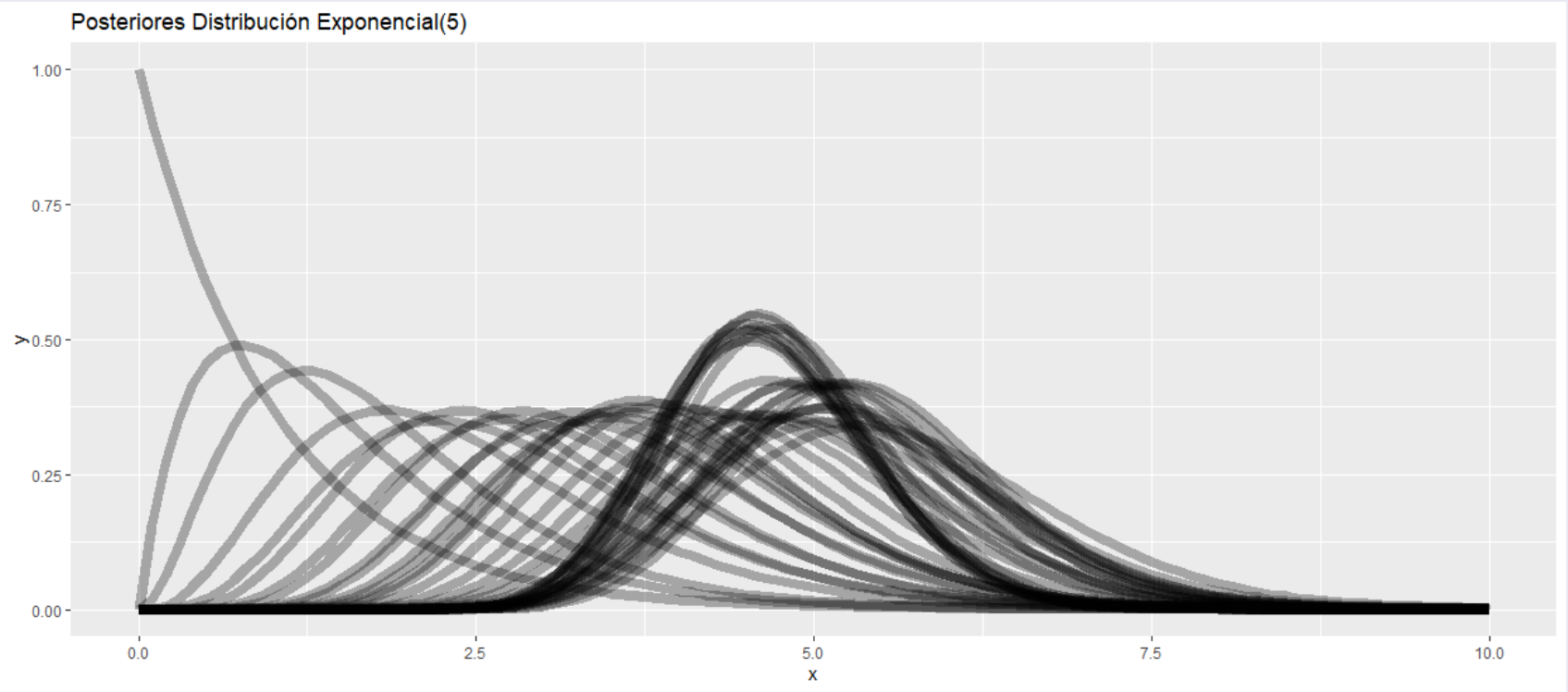
$$= \theta^{(a+n)-1} \exp\left\{-\left(b + \sum_{i=1}^{n} x_i\right)\theta\right\}$$

Así,

$$\theta|\vec{x} \sim \text{Gama}\left(a + n, b + \sum x_i\right)$$

Gráficas de las densidades posteriores

Ejemplo Gama-Exponencial



Quinto ejemplo de análisis conjugado

Podemos demostrar que la distribución Pareto(α, β) es previa conjugada del modelo uniforme en el intervalo $(0, \theta)$.
Si $\theta \sim$ Pareto(a, b),

$$p(\theta) = \frac{ab^a}{(\theta + b)^{a+1}} 1_{(0,\infty)}(\theta)$$

notemos que la densidad puede escribirse equivalentemente

$$p(\theta) = \frac{ab^a}{(\theta)^{a+1}} 1_{(b,\infty)}(\theta).$$

Continuando con el ejemplo

Si $x_i \sim \text{Uniforme}(0, \theta)$ entonces

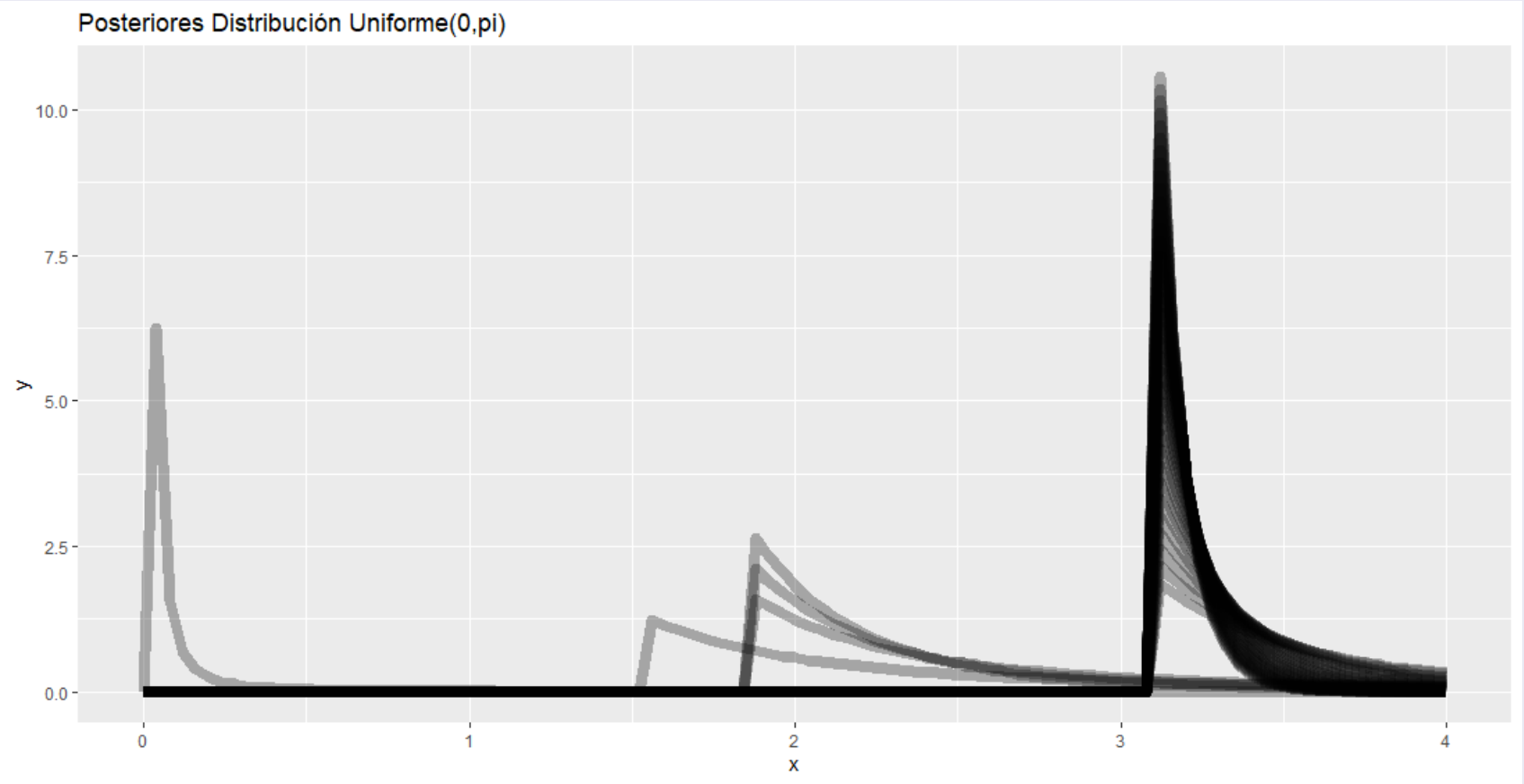
$$\begin{aligned} p(\theta|x_1, ..., x_n) &\propto \frac{ab^a}{(\theta)^{a+1}} 1_{(b,\infty)}(\theta) \prod_{i=1}^{m} \frac{1}{\theta - 0} 1_{(0,\theta)}(x_i) \\ &= \frac{ab^a}{(\theta)^{a+1}} 1_{(b,\infty)}(\theta) \frac{1}{\theta^n} 1_{(\max_i x_i, \infty))}(\theta) \\ &\propto \frac{1}{(\theta)^{a+1}} 1_{(b,\infty)}(\theta) \frac{1}{\theta^n} 1_{(\max_i x_i, \infty))}(\theta) \\ &= \frac{1}{(\theta)^{a+n+1}} 1_{(\max\{b, \max_i x_i\}, \infty))}(\theta) \end{aligned}$$

Así,

$$\theta|\vec{x} \sim \text{Pareto}(a + n, \max\{b, \max_i x_i\}$$

Gráficas de las densidades posteriores

Ejemplo Pareto-Uniforme



Sexto ejemplo de análisis conjugado (Dos parámetros)

Distribución Normal Gama Inversa

Suponga que

$$\mu \mid \sigma^2, \mu_0, \lambda \sim \text{Normal}(\mu_0, \sigma^2/\lambda)$$

y además

$$\sigma^2 \mid \alpha, \beta \sim \text{GamaInversa}(\alpha, \beta).$$

Decimos que

$$(\mu, \sigma^2) \sim \text{Normal} - \text{GamaInversa}(\mu_0, \lambda, \alpha, \beta)$$

Su función de densidad es

$$f(\mu, \sigma^2 \mid \mu_0, \lambda, \alpha, \beta) = \frac{\sqrt{\lambda}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{2\beta + \lambda(\mu - \mu_0)^2}{2\sigma^2}\right)$$

Ejemplo de análisis conjugado con dos parámetros

Supongamos que contamos con una muestra aleatoria $X_1, ..., X_n$ independiente e idénticamente distribuida Normal(μ, σ^2). Veamos que la distribución Normal-GamaInversa$(\mu, \lambda, \alpha, \beta)$ es distribución previa conjugada del modelo Normal(μ, σ^2).

$$
\begin{aligned}
& p(\mu, \sigma^2 | \vec{x}) \\
& \propto \prod_{i=1}^{n} \left[\frac{1}{\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2} \right] \right] \frac{1}{\sigma} \left(\frac{1}{\sigma^2} \right)^{\alpha+1} \exp\left(-\frac{2\beta + \lambda(\mu - \mu_0)^2}{2\sigma^2} \right) \\
& \propto \frac{1}{\sigma^n} \left[\exp\left[-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2} \right] \right] \frac{1}{\sigma} \left(\frac{1}{\sigma^2} \right)^{\alpha+1} \exp\left(-\frac{2\beta + \lambda(\mu - \mu_0)^2}{2\sigma^2} \right) \\
& \propto \frac{1}{\sigma} \left(\frac{1}{\sigma^2} \right)^{\overbrace{\alpha + \frac{n}{2}}^{\alpha'} + 1} \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \mu)^2 + 2\beta + \lambda(\mu - \mu_0)^2}{2\sigma^2} \right]
\end{aligned}
$$

Quinto ejemplo de análisis conjugado

Podemos concentrarnos en el exponente

$$\exp\left[-\frac{\sum x_i^2 - 2\sum x_i\mu + n\mu^2 + \lambda\mu^2 - 2\mu\mu_0 + \mu_0^2 + 2\beta}{2\sigma^2}\right]$$

$$= \exp\left[-\frac{(n+\lambda)\mu^2 - 2\mu(\sum x_i + \mu_0) + \sum x_i^2 + \mu_0^2 + 2\beta}{2\sigma^2}\right]$$

$$= \exp\left[-\frac{(n+\lambda)\left(\mu^2 - 2\mu\frac{(\sum x_i + \mu_0)}{n+\lambda}\right) + \sum x_i^2 + \mu_0^2 + 2\beta}{2\sigma^2}\right]$$

Siguiendo el ejemplo

$$= \exp\left[-\frac{1}{2\sigma^2}\left[\overbrace{(n+\lambda)}^{\lambda'}\left(\mu - \overbrace{\frac{\sum x_i + \mu_0}{n+\lambda}}^{\mu_0'}\right)^2 + \overbrace{\frac{-(\sum x_i + \mu_0)^2}{n+\lambda} + \sum x_i^2 + \mu_0^2 + 2\beta}^{2\beta'} \right] \right]$$

Por lo tanto

$$\mu, \sigma^2 | \vec{X} \sim N\Gamma^{-1}\left(\frac{\sum x_i + \mu_0}{n+\lambda}, n+\lambda, \alpha + \frac{n}{2}, \beta + \frac{\sum x_i^2 + \mu_0^2}{2} + \frac{-(\sum x_i + \mu_0)^2}{2(n+\lambda)^2} \right)$$

$$\equiv \text{Normal-GamaInversa}\left(\mu_0', \lambda', \alpha', \beta'\right)$$

Densidades predictivas

Para una muestra aleatoria independiente idénticamente distribuida $X_1, ..., X_n$ con función de densidad $p(x|\theta)$ con una previa para θ con densidad $p(\theta)$, su densidad predictiva para una nueva observación puede calcularse mediante

$$p(x) = \int p(x|\theta)p(\theta)d\theta$$

La distribución predictiva posterior es simplemente

$$\begin{aligned} p(x_{n+1}|X_1, ..., X_n) &= \int p(x_{n+1}|\theta, X_1, ..., X_n)p(\theta|\boldsymbol{x})d\theta \\ &= \int p(x_{n+1}|\theta)p(\theta|X_1, ..., X_n)d\theta \end{aligned}$$

Esto se extiende de igual manera a funciones de probabilidad.

Ejemplo de densidad predictiva

Consideremos el caso de modelo Binomial(n,θ) con previa Beta(α, β) (posterior Beta(α', β') $\equiv$ Beta($\alpha + \sum x_i, \beta + nm - \sum x_i$)). La integral

$$
\begin{aligned}
P(k) &= \int_0^1 P_{\text{Binomial}}(x|\theta) P_{\text{Beta}}(\theta|\alpha', \beta') d\theta \\
&= \int_0^1 \binom{m}{x} \theta^x (1-\theta)^{m-x} \frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')} \theta^{\alpha'-1}(1-\theta)^{\beta'-1} d\theta \\
&\propto \frac{m!}{x!(m-x)!} \int_0^1 \underbrace{\theta^{\alpha'+x-1}(1-\theta)^{\beta'+m-x-1}}_{\text{Es kernel Beta}} d\theta \\
&= \frac{\Gamma(m+1)}{\Gamma(x+1)\Gamma(m-x+1)} \left(\frac{\Gamma(\alpha'+\beta'+m)}{\Gamma(\alpha'+x)\Gamma(\beta'+m-x)} \right)^{-1} \\
&\propto \frac{\Gamma(\alpha'+x)\Gamma(\beta'+m-x)}{\Gamma(x+1)\Gamma(m-x+1)}
\end{aligned}
$$

Así, $X|\vec{X} \sim$ Beta-Binomial $\left(x|m, \alpha', \beta'\right)$

Ejemplo de densidad predictiva

Consideremos el caso de modelo Poisson(λ) con previa Gama(α, β) (y por tanto posterior Gamma(α', β') $\equiv$ Gamma($\alpha + \sum x_i, \beta + n$)).
Resolviendo la integral

$$
\begin{aligned}
P(k) &= \int_0^\infty P_{\text{Poisson}}(k|\lambda) P_{\text{Gama}}(\lambda|\alpha', \beta') d\lambda \\
&= \int_0^\infty \frac{e^{-\lambda}\lambda^k}{k!} \frac{(\beta')^m}{\Gamma(\alpha')} \lambda^{\alpha'-1} e^{-\lambda\beta'} d\lambda \\
&= \frac{(\beta')^m}{\Gamma(\alpha')k!} \int_0^\infty e^{-\lambda(\beta'+1)} \lambda^{k+\alpha'-1} d\lambda \\
&= \frac{(\beta')^{\alpha'}}{\Gamma(\alpha')k!} \int_0^\infty \underbrace{e^{-\lambda(\beta'+1)} \lambda^{k+\alpha'-1}}_{\text{Es kernel Gama}(k+\alpha', \beta'+1)} d\lambda
\end{aligned}
$$

Ejemplo de densidad predictiva

$$P(k) = \frac{(\beta')^{\alpha'}}{\Gamma(\alpha')k!}\Gamma(k+\alpha')\left(\frac{1}{\beta'+1}\right)^{k+\alpha'} \underbrace{\int_0^\infty \ldots d\lambda}_{1}$$

$$= \frac{\Gamma(k+\alpha')}{\Gamma(\alpha')k!}\frac{(\beta')^m}{(\beta'+1)^m(\beta'+1)^k}$$

$$= \frac{(k+\alpha'-1)!}{(\alpha'-1)!k!}\left(\frac{\beta'}{\beta'+1}\right)^{\alpha'}\left(\frac{1}{\beta'+1}\right)^{k}$$

Así,

$$X|\vec{X} \sim \text{BinomialNegativa}\left(k\middle|\frac{1}{1+\beta'}, \alpha'\right)$$

Factor de Bayes

El factor de Bayes es la razón de dos verosimilitudes marginales de dos hipótesis a contrastar.
La probabilidad posterior $\mathrm{P}(M|D)$ del modelo M dados los datos D esta dada por el Teorema de Bayes

$$\mathrm{P}(M|D) = \frac{\mathrm{P}(D|M)\mathrm{P}(M)}{\mathrm{P}(D)}$$

El término $\mathrm{P}(D|M)$ representa la probabilidad de que los datos D se produzcan bajo el supuesto del modelo M.

Factor de Bayes

Para un problema de selección de modelos en que debemos escoger entre dos modelos basado en los datos D, la plausibilidad de ambos modelos M_1 y M_2, parametrizados por el vector de parámetros θ_1 y θ_2, se contrasta por el factor de Bayes K

$$K = \frac{\mathrm{P}(D|M_1)}{\mathrm{P}(D|M_2)} = \frac{\int \mathrm{P}(\theta_1|M_1)\mathrm{P}(D|\theta_1, M_1)\, d\theta_1}{\int \mathrm{P}(\theta_2|M_2)\mathrm{P}(D|\theta_2, M_2)\, d\theta_2} = \frac{\mathrm{P}(M_1|D)}{\mathrm{P}(M_2|D)} \frac{\mathrm{P}(M_2)}{\mathrm{P}(M_1)}.$$

Si los dos modelos son igualmente probables inicialmente, tal que $\mathrm{P}(M_1) = \mathrm{P}(M_2)$, entonces el factor de Bayes es igual a la razón de las probabilidades posteriores de M_1 y M_2.

Como interpretar el factor de Bayes

Una tabla sugerida por Kass & Raftery (1995)

2 ln K	K	Contundencia de la evidencia
De 0 a 2	De 1 a 3	No merece más que una breve mención
De 2 a 6	De 3 a 20	Positiva
De 6 a 10	De 20 a 150	Fuerte
> 10	> 150	Muy Fuerte

Ejemplo de cálculo de factor de Bayes

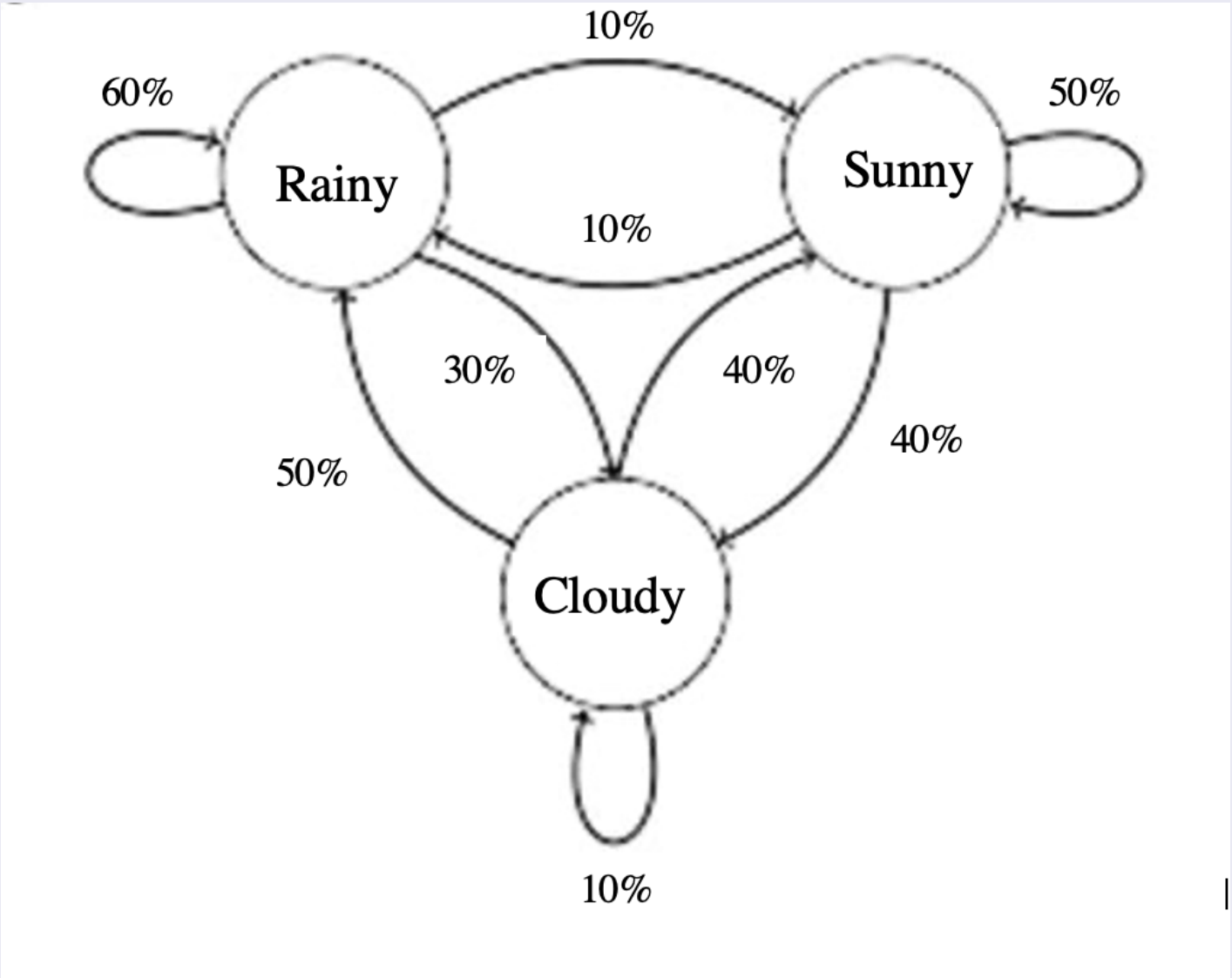
Supongamos que tenemos una muestra aleatoria $X_1, ..., X_n \sim \text{Uniforme}(0, \theta)$ y que deseamos contrastar las hipótesis

$$H_0 : \theta < 1 \text{ vs. } H_1 : \theta \geq 1$$

Podemos empezar con un análisis conjugado suponiendo que $\theta \sim \text{Pareto}(a, b)$ y así $\theta|\vec{x} \sim \text{Pareto}(a + n, \max\{b, \max_i x_i\})$. Podemos calcular el factor de Bayes:

$$\begin{aligned} K &= \frac{P(H_0|D)}{P(H_1|D)} \frac{P(H_1)}{P(H_0)} = \frac{\int_0^1 P(\theta|D, H_0)d\theta \int_1^\infty P(\theta|H_1)d\theta}{\int_1^\infty P(\theta|D, H_1)d\theta \int_0^1 P(\theta|H_0)d\theta} \\ &= \frac{(F_{\text{Pareto}(a+n,\max\{b,\max_i x_i\}}(1))(1 - F_{\text{Pareto}(a,b)}(1))}{(1 - F_{\text{Pareto}(a+n,\max\{b,\max_i x_i\})}(1))(F_{\text{Pareto}(a,b)}(1))} \end{aligned}$$

Ejemplo de modelo definiendo probabilidades de transición



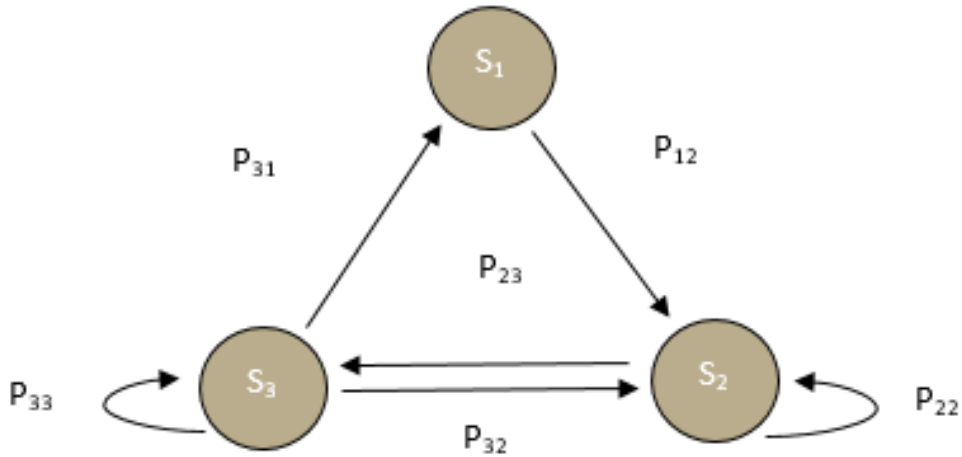
Matriz de trasición

Construcción de la matriz de transición

Succeeding State

Initial State	S_1	S_2	S_3
S_1	0	P_{12}	0
S_2	0	P_{22}	P_{23}
S_3	P_{31}	P_{32}	P_{33}

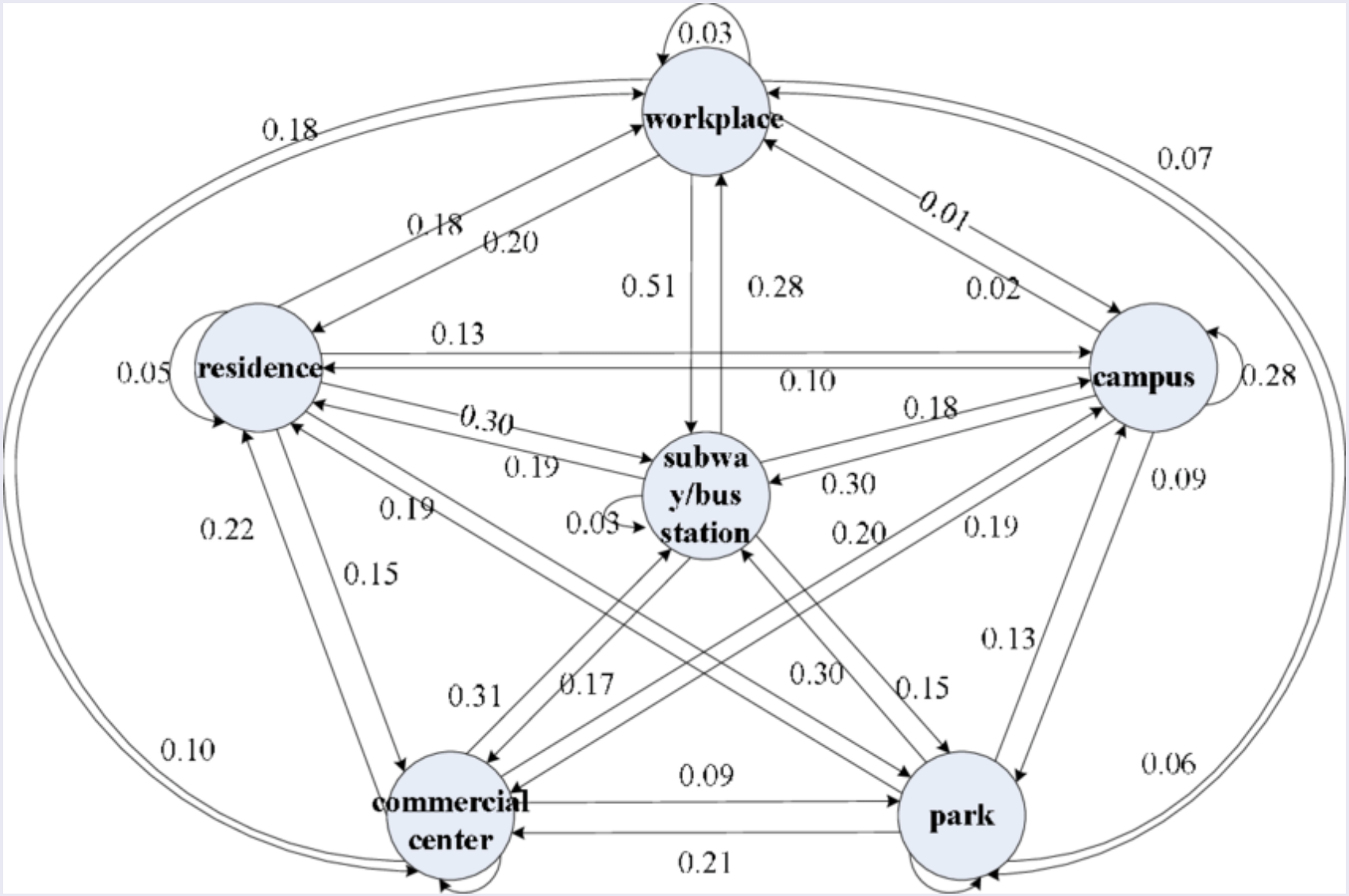
Transition Matrix



Transition Diagram

Otro ejemplo

Diagrama de transición ejemplo bicicletas



Distribuciones de equilibrio y límite

- Una distribución de equilibrio o estacionaria es una distribución π tal que si la distribución sobre los estados en el paso k es π entonces también la distribución sobre los estados en el paso $k+1$ es π. Es decir

$$\pi = \pi P$$

.

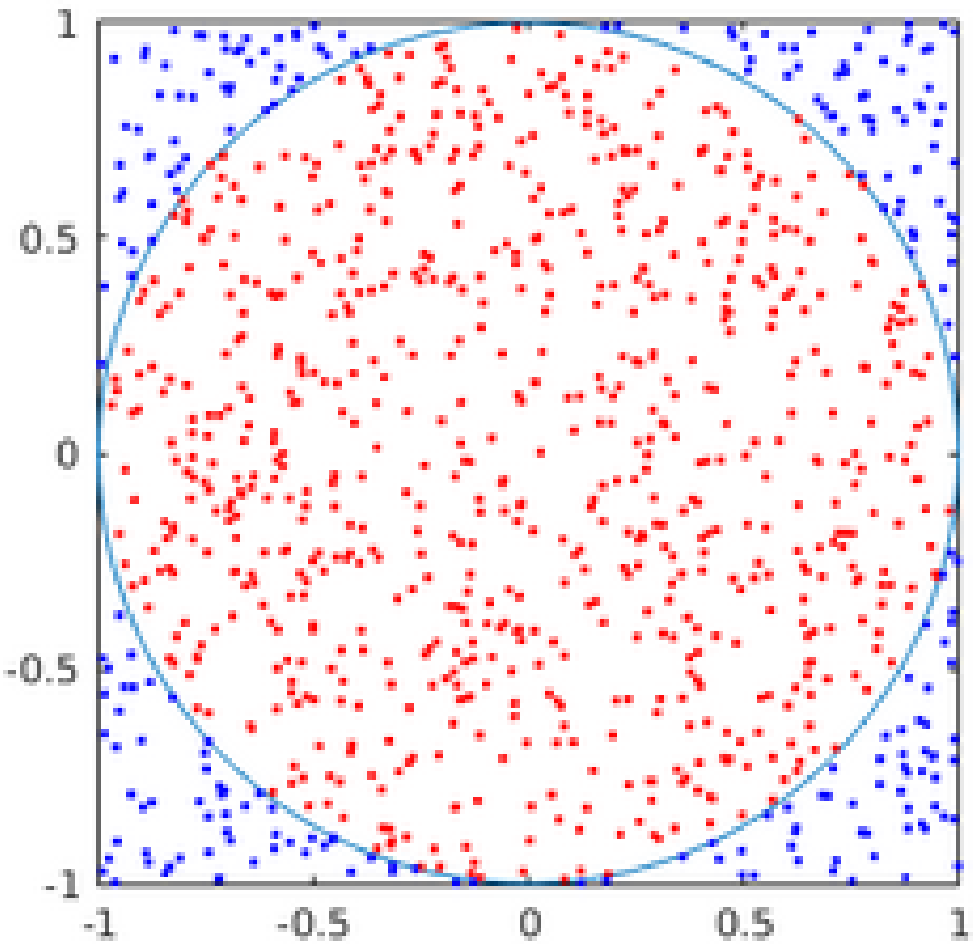
- Una distribución límite es una distribución π que no importa cuál sea la distribución inicial, la distribución sobre los estados converge a π a medida que el número de pasos llega a infinito:

$$\lim_{k \to \infty} \pi^{(0)} P^k = \pi$$

.

Método de Monte Carlo

Ejemplo: Estimar Pi con Monte Carlo



Método de Monte Carlo

Ley de grandes números para calcular cualquier momento

- La Ley Fuerte de los Grandes Números justifica este método. Si $X^{(1)}, X^{(2)}, \ldots$ es una sucesión infinita de variables independientes e identicamente distribuidas, con $\mathrm{E}(X) = \mu < \infty$ entonces siendo

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{\infty} X^{(i)}, \text{ tenemos } \mathrm{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1 \text{ o bien } \overline{X}_n \xrightarrow{\text{c.s.}} \mu$$

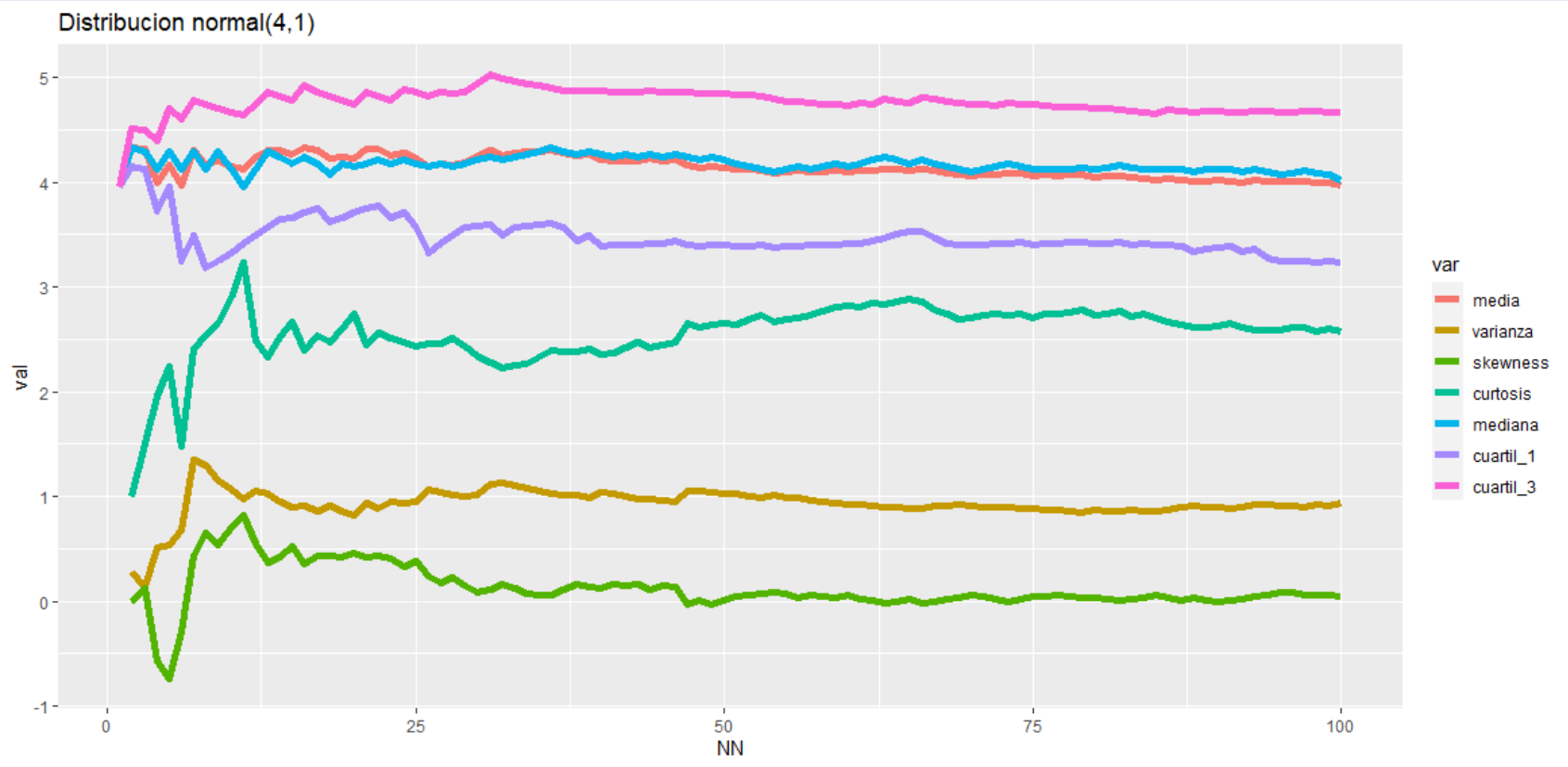
- Gracias a la Ley del Estadístico Inconciente es claro que siendo $g(\cdot)$ una función también tenemos

$$\frac{1}{n}\sum_{i=1}^{\infty} g(X^{(i)}) \xrightarrow{\text{c.s.}} E(g(X))$$

- Notemos que $g(\cdot)$ puede ser cualquier polinomio. Podemos construir cualquier momento.

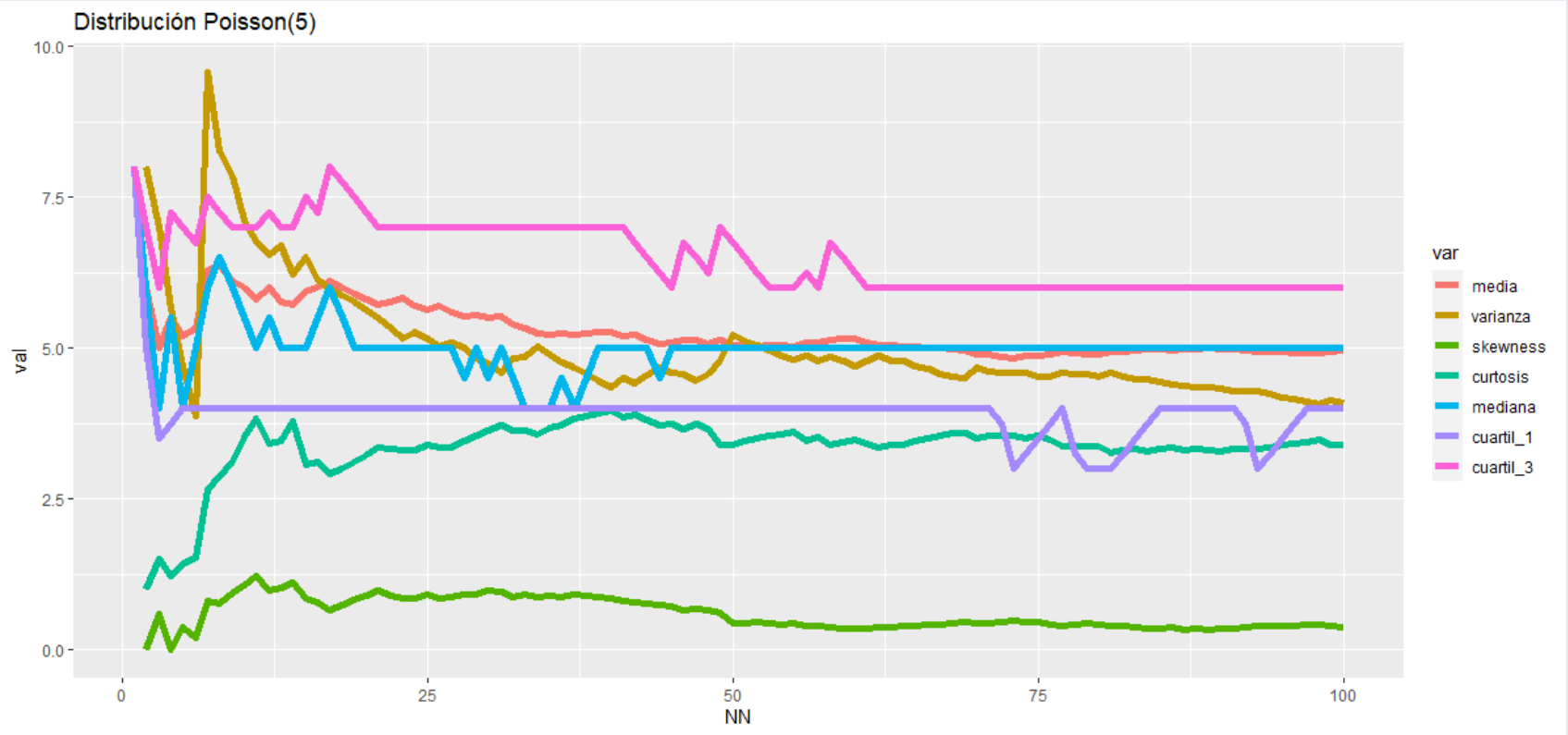
Método de Monte Carlo

Aproximar momentos y otros atributos de la distribución normal



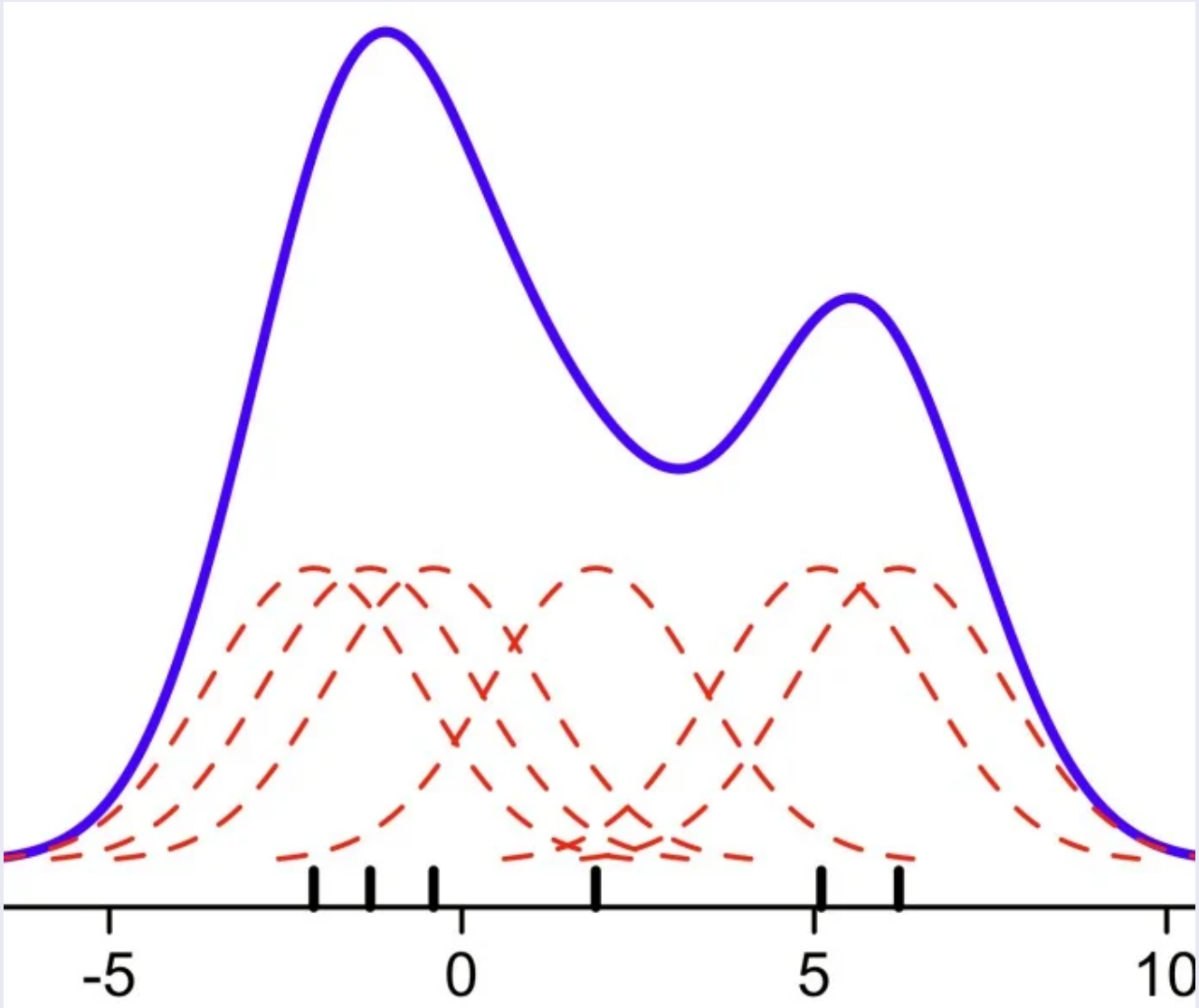
Método de Monte Carlo

Aproximar momentos y otros atributos de la distribución Poisson



Estimar integrales no es lo único que podemos hacer

Estimación no paramétrica de la f. de densidad (kernel Gaussiano)



Condiciones del teorema ergódico para MCMC

- Una cadena de Márkov se dice homogénea si

$$P(X_n = j|X_{n-1} = i) = P(X_1 = j|X_0 = i)$$

para todo n y para cualquier i, j.

- El periodo de un estado $x \in E$ se define como:

$$d(x) = \text{mcd}\{n : P_{x,x}^{(n)} > 0\}$$

donde mcd denota el máximo común divisor.
Si $d(x) = 1$ diremos que x es un estado aperiódico. Una cadena de Márkov se dice aperiódica si todos sus estados son aperiódicos.

- Una cadena de Márkov se dice irreducible si desde cualquier estado de E se puede acceder a cualquier otro.

Resultados del teorema ergódico para MCMC

Sea $\theta^{(1)}, \theta^{(2)}, \ldots$ una cadena de Markov homogénea, irreducible y aperiódica, con espacio de estados Θ y distribución de equilibrio $p(\theta|x)$. Entonces, conforme $t \to \infty$:

- $\theta^{(t)} \xrightarrow{d} \theta$, donde $\theta \sim p(\theta|x)$
- $\frac{1}{t}\sum_{i=0}^{t} g(\theta^{(i)}) \to \mathrm{E}(g(\theta)|x)$
  con g una función medible de esperanza finita.

Monte Carlo vía cadenas de Markov

Algortimos

Son dos los algoritmos más famosos e importantes :

- Metropolis-Hastings
- Muestreador de Gibbs (caso particular de MH)

Históricamente, el primero fue el Muestreador de Gibbs.
Además, resulta sumamente conveniente para los modelos jerárquicos.
¡Es aún más conveniente si usamos iniciales semi-conjugadas!

Metrolis-Hastings

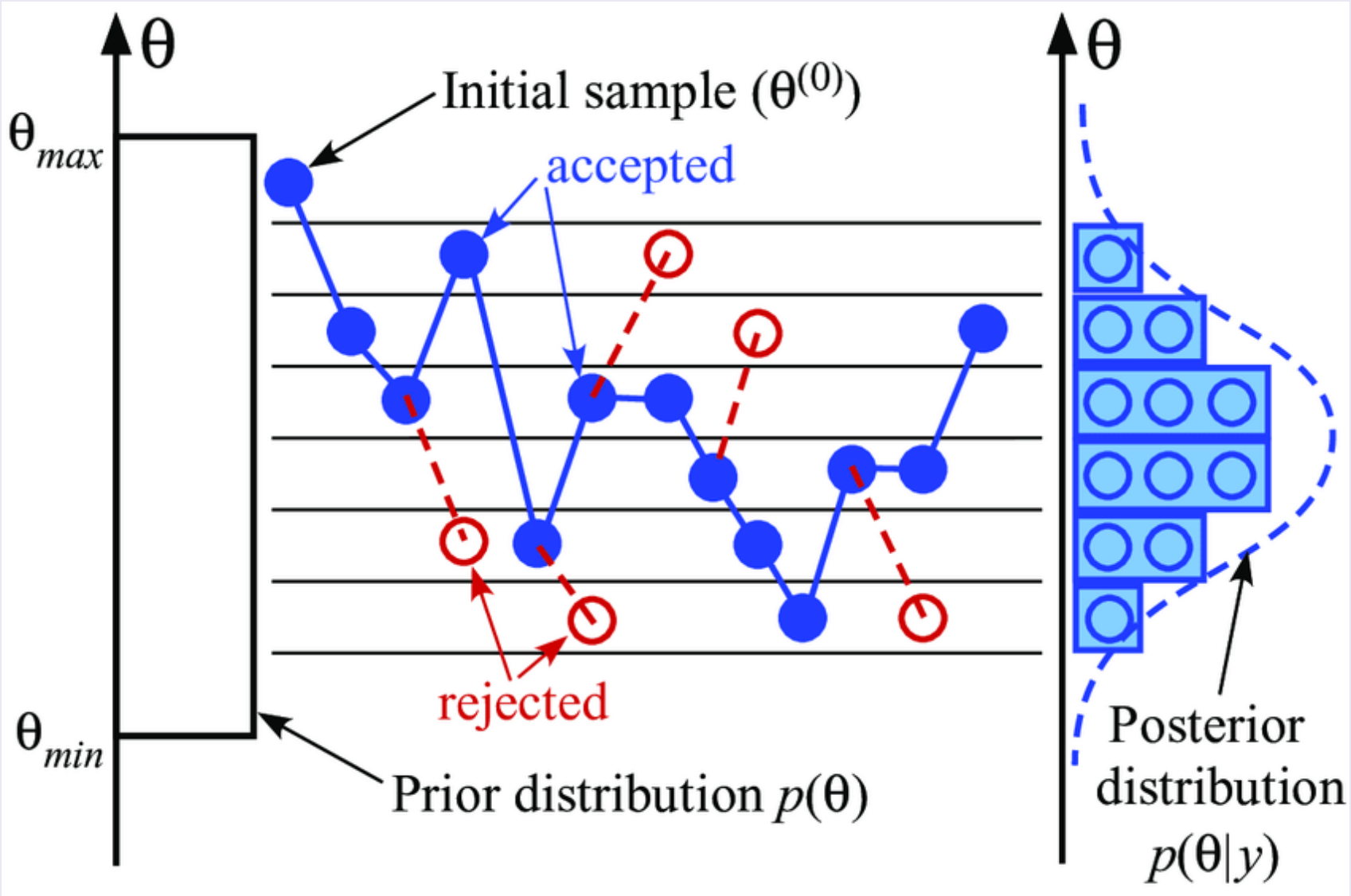
Supongamos que nos interesa simular de una distribución con densidad $p(\theta|x)$. Sea $Q(\theta^*|\theta)$ una distribución de transición (arbitraria) y definimos

$$\alpha(\theta^*, \theta) = \min\left\{\frac{p(\theta^*|x)Q(\theta|\theta^*)}{p(\theta|x)Q(\theta^*|\theta)}, 1\right\}$$

Algoritmo Dado un valor inicial $\theta^{(0)}$, la t-ésima iteración consiste en:

1. generar una observación θ^* de $Q(\theta^*|\theta^{(t)})$;
2. generar variable $u \sim \text{Uniforme}(0, 1)$
3. si $u \leq \alpha(\theta^*, \theta^{(t)})$ hacer $\theta^{(t+1)} = \theta^*$; en caso contrario, hacer $\theta^{(t+1)} = \theta^{(t)}$

Diagrama de Metropolis-Hastings



Ecuación de balance detallado

El kernel de transición esta dado por

$$K(\theta^*, \theta^{(t)}) \begin{cases} Q(\theta^*|\theta^{(t)})\alpha(\theta^*, \theta^{(t)}) & \text{si } \theta^* \neq \theta^{(t)} \\ Q(\theta^*|\theta^{(t)})\alpha(\theta^*, \theta^{(t)}) + (1 - r(\theta^{(t)})) & \text{si } \theta^* = \theta^{(t)} \end{cases}$$

con $r(\cdot) = \int Q(\cdot|y)\alpha(\cdot, y)dy$. Una cadena de Markov con kernel de transición K satisface la ecuación de balance detallado si existe una función f con la que cumple

$$K(\theta^{(t)}, \theta^*)f(\theta^{(t)}) = K(\theta^*, \theta^{(t)})f(\theta^*)$$

para todo $(\theta^{(t)}, \theta^*)$.

Preguntas sobre Metropolis-Hastings

- ¿Qué puede decir del algoritmo Metropolis-Hastings cuando $Q(\theta|\theta^*) = Q(\theta)$, es decir, no depende del estado actual?
- ¿Qué puede decir del algoritmo Metropolis-Hastings cuando $Q(\theta|\theta^*) = Q(\theta^*|\theta)$, es decir, Q es simétrica?
- Es posible que obtengamos vectores simulados θ^* donde $p(\theta^*) = 0$ y por lo tanto nos estaríamos saliendo del dominio de estas variables aleatorias. ¿Es esto un problema? ¿Por qué?

Muestrador de Gibbs

Definimos a la *densidad condicional completa* de θ_i dado θ_{-i}, como

$$p(\theta_i|\theta_{-i}, x) = \frac{p(\theta_i, \theta_{-i}|x)}{p(\theta_{-i}|x)} = \frac{p(\theta|x)}{\int p(\theta|x)d\theta_i}.$$

Las densidades condicionales completas

$$p(\theta_1|\theta_2, ..., \theta_k, x)$$

$$\vdots$$

$$p(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k, x) \ (i = 2, ..., k-1)$$

$$\vdots$$

$$p(\theta_k|\theta_1, ..., \theta_{k-1}, x)$$

¡Pueden identificarse fácilmente al inspeccionar la forma de $p(\theta|x)$!

El algoritmo GS

De hecho, para cada $i = 1, ..., k$,

$$p(\theta_i|\theta_{-i}, x) \propto p(\theta|x),$$

donde $p(\theta|x) = p(\theta_1, ..., \theta_k|x)$ es vista sólo como función de θ_i. Dado un valor inicial $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_k^{(0)})$, el algoritmo de Gibbs simula una cadena de Markov en la que $\theta^{(t+1)}$ se obtiene a partir de $\theta^{(t)}$ de la siguiente manera:

generar una observación $\theta_1^{(t+1)}$ de $p(\theta_1|\theta_2^{(t)}, ..., \theta_k^{(t)}, x)$;
generar una observación $\theta_2^{(t+1)}$ de $p(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_k^{(t)}, x)$;
$\vdots$
generar una observación $\theta_k^{(t+1)}$ de $p(\theta_k|\theta_1^{(t+1)}, ..., \theta_{k-1}^{(t+1)}, x)$;

GS es caso particular de MH

Supongamos que las distribuciones condicionales completas son todas conocidas y fáciles de simular. Si en el algoritmo de Metropolis-Hastings hacemos

$$Q(\theta_i|\theta) = p(\theta_i|\theta_{-i}, x)$$

destacando que como en este salto nos quedamos con los mismos valores para los θ_{-i} y entonces $\theta^*_{-i} = \theta_{-i}$ tendremos que

$$\frac{p(\theta^*)}{p(\theta)}\frac{Q(\theta|\theta^*)}{Q(\theta^*|\theta)} = \frac{p(\theta^*_i, \theta^*_{-i})}{p(\theta_i, \theta_{-i})}\frac{p(\theta_i|\theta^*_{-i})}{p(\theta^*_i|\theta_{-i})} = \frac{p(\theta^*_i|\theta_{-i})p(\theta_{-i})}{p(\theta^*_i|\theta_{-i})p(\theta_{-i})}\frac{p(\theta_i|\theta^{-i})}{p(\theta_i|\theta_{-i})} = 1$$

Por lo tanto

$$\alpha(\theta_i, \theta) = \min\left\{\frac{p(\theta^*)}{p(\theta)}\frac{Q(\theta|\theta^*)}{Q(\theta^*|\theta)}, 1\right\} = 1$$

es decir los valores "candidatos" se eligen con probabilidad uno.

Criterios de convergencia e independencia

Los índices empíricos para verificar convergencia a la distribución estacionaria e independencia de las muestras son

- Graficar promedios por ventana/ ergódicos
- Graficar las trazas
- Gráficas de autocorrelación

¡Podemos elegir que elementos de la cadena tomamos para alcanzar estos objetivos!

Primer ejemplo: Modelo no jerárquico

Sea $X_i \sim \text{Binomial}(k, p)$ cond. independientes $i = 1, 2, ..., n$, pero ambos k y p son desconocidos. Los datos son : $\{4, 3, 1, 6, 6, 6, 5, 5, 5, 1\}(n = 10)$. Notemos

$$p(x) = \binom{k}{x} p^x (1-p)^{k-x}$$

Una propuesta de previa que parece razonable,

$$K \sim \text{Uniforme}(1, ..., 200)$$

$$P \sim \text{Beta}(\alpha, \beta)$$

Parece razonable no asumir, a priori, independencia entre estas dos variables aleatorias. Podemos tomar $\alpha, \beta = 1$. Recordemos

$$p(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \propto p^{\alpha-1}(1-p)^{\beta-1}$$

Primer ejemplo: Modelo no jerárquico

$$
\begin{aligned}
&p(K, P|\boldsymbol{X}) \\
&\propto p(\boldsymbol{X}|K, P)p(K, P) \\
&\propto \prod_{i=1}^{n}\left[\binom{k}{x_i} p^{x_i}(1-p)^{k-x_i}\mathbb{1}_{\{1,\ldots,k\}}(x_i)\right] p^{\alpha-1}(1-p)^{\beta-1}\mathbb{1}_{\{1,\ldots,100\}}(k) \\
&\propto \prod_{i=1}^{n}\left[\frac{k!}{(k-x_i)!(x_i)!}\right] p^{\sum x}(1-p)^{nk-\sum x} p^{\alpha-1}(1-p)^{\beta-1}\mathbb{1}_{\{\max_i\{x_i\},\ldots,100\}}(k) \\
&= \prod_{i=1}^{n}\left[\frac{k!}{(k-x_i)!(x_i)!}\right] p^{\sum x+\alpha-1}(1-p)^{nk-\sum x+\beta-1}\mathbb{1}_{\{\max_i\{x_i\},\ldots,100\}}(k)
\end{aligned}
$$

De este modo, es fácil obtener las distribuciones condicionales completas para implementar un Muestreador de Gibbs.

Primer ejemplo: Modelo no jerárquico

$$p(P|\boldsymbol{X}, K) \propto p^{\sum x+\alpha-1}(1-p)^{nk-\sum x+\beta-1}$$

Se trata de un Kernel Beta, por lo tanto podemos simular fácilmente

$$P|\boldsymbol{X}, K \sim \text{Beta}(\sum x + \alpha, nk - \sum x + \beta)$$

En nuestro caso particular $\alpha, \beta = 1$

$$P|\boldsymbol{X}, K \sim \text{Beta}(\sum x + 1, nk - \sum x + 1)$$

Por otro lado,

$$p(K|\boldsymbol{X}, P) \propto \prod_{i=1}^{n} \left[\frac{k!}{(k-x_i)!(x_i)!}\right] (1-p)^{nk} \mathbb{1}_{\{\max_i\{x_i\},\ldots,100\}}(k)$$

Al tratarse de una distribución discreta, podemos hacer una lotería estandarizada pesada según la densidad condicional completa.

Ejemplo de modelo no jerárquico

La cadena sólo se comporta de forma un poco extraña antes de los primeros 2000 elementos. Este análisis sugeriría tomarlo como el Bur-in.

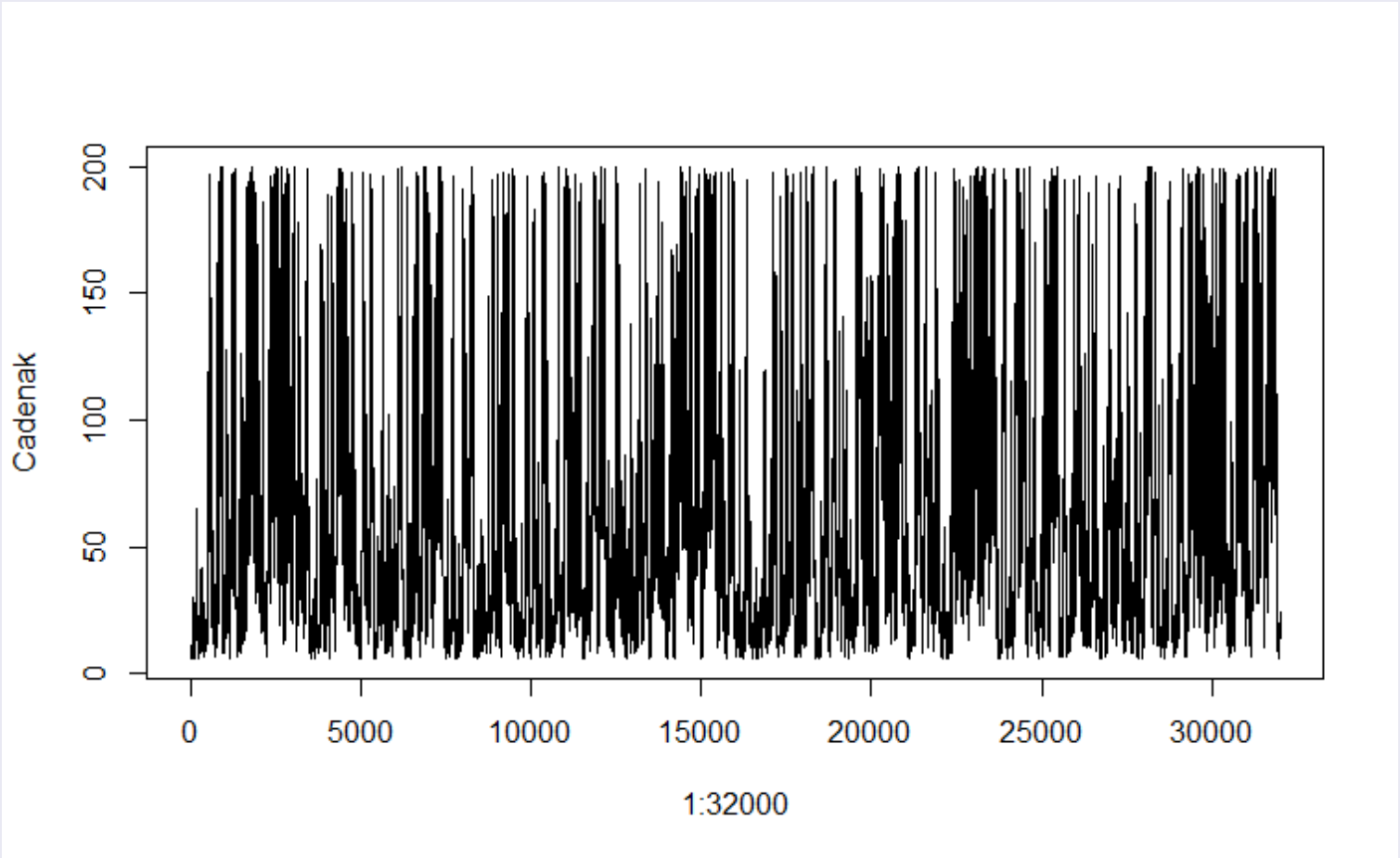


Figure 1: Traza de la cadena de K

Primer ejemplo: Modelo no jerárquico

Notamos que los promedios ergódicos de la cadena comienza a estabilizarse a los 10000 elementos. Nuestro Burn-in debe ser al menos tal.

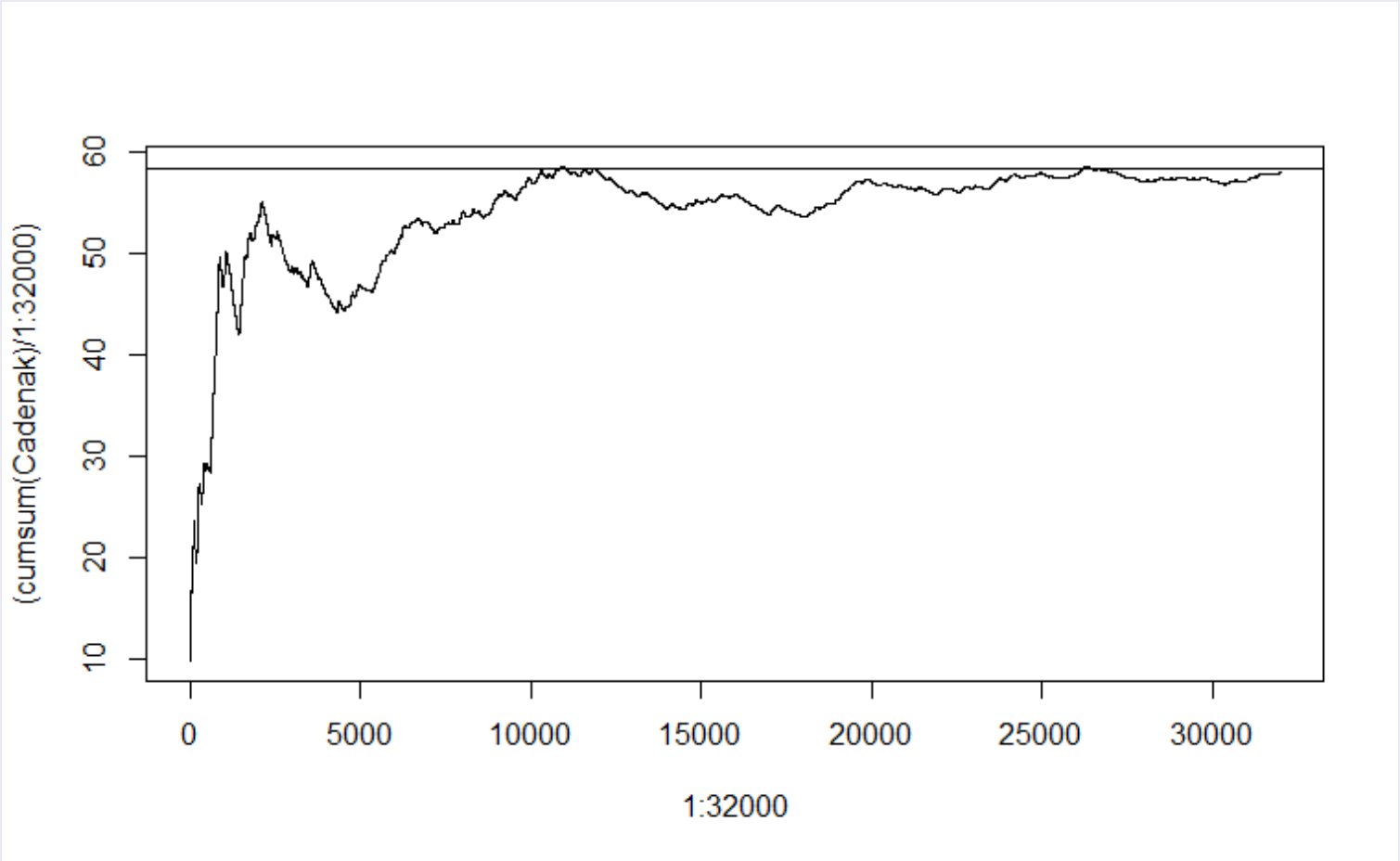


Figure 2: Promedios ergódicos de la cadena de K

Primer ejemplo: Modelo no jerárquico

Las observaciones están muy correlacionadas con sus observaciones siguientes. Elegimos tomar saltos de 50 elementos para reducir la correlación y no quedarnos con muestras tan pequeñas.

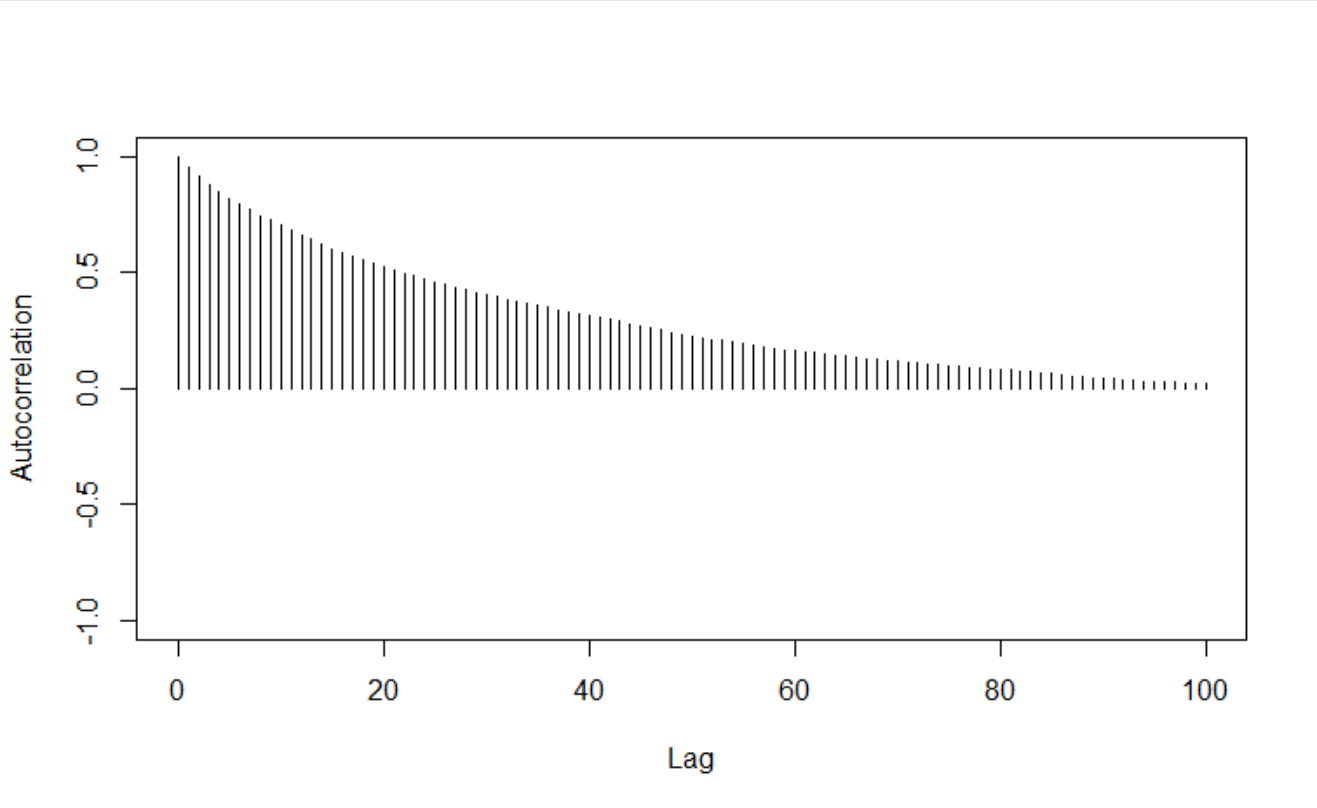


Figure 3: Gráfica de autocorrelación de K

Primer ejemplo: Modelo no jerárquico

Es más probable que sean pocos ($P(K < 42 = 0.5)$)

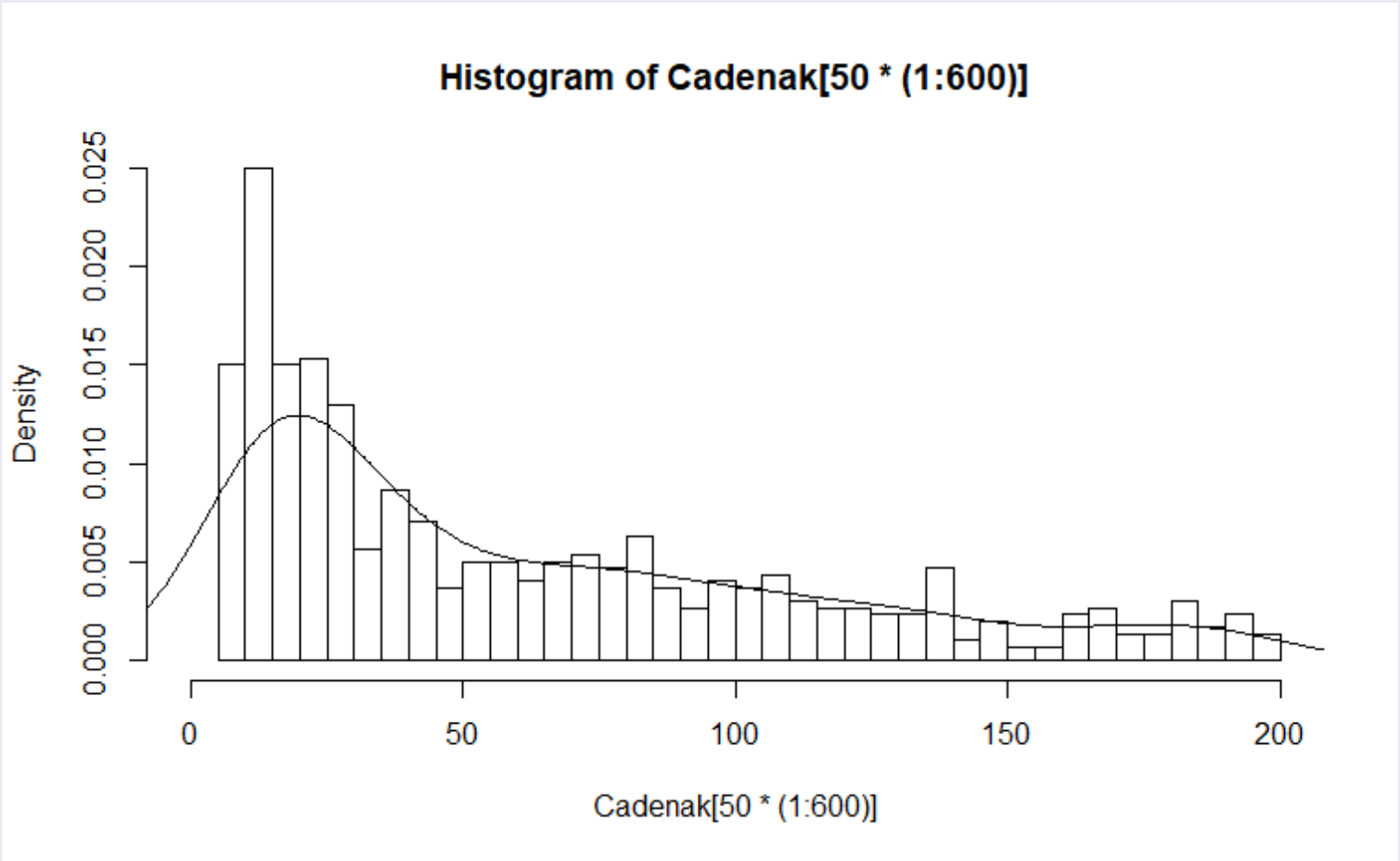


Figure 4: Histograma y estimación de la densidad mediante kernels de K

Primer ejemplo: Modelo no jerárquico

Notemos que $P(P < 0.1020 = 0.5)$.

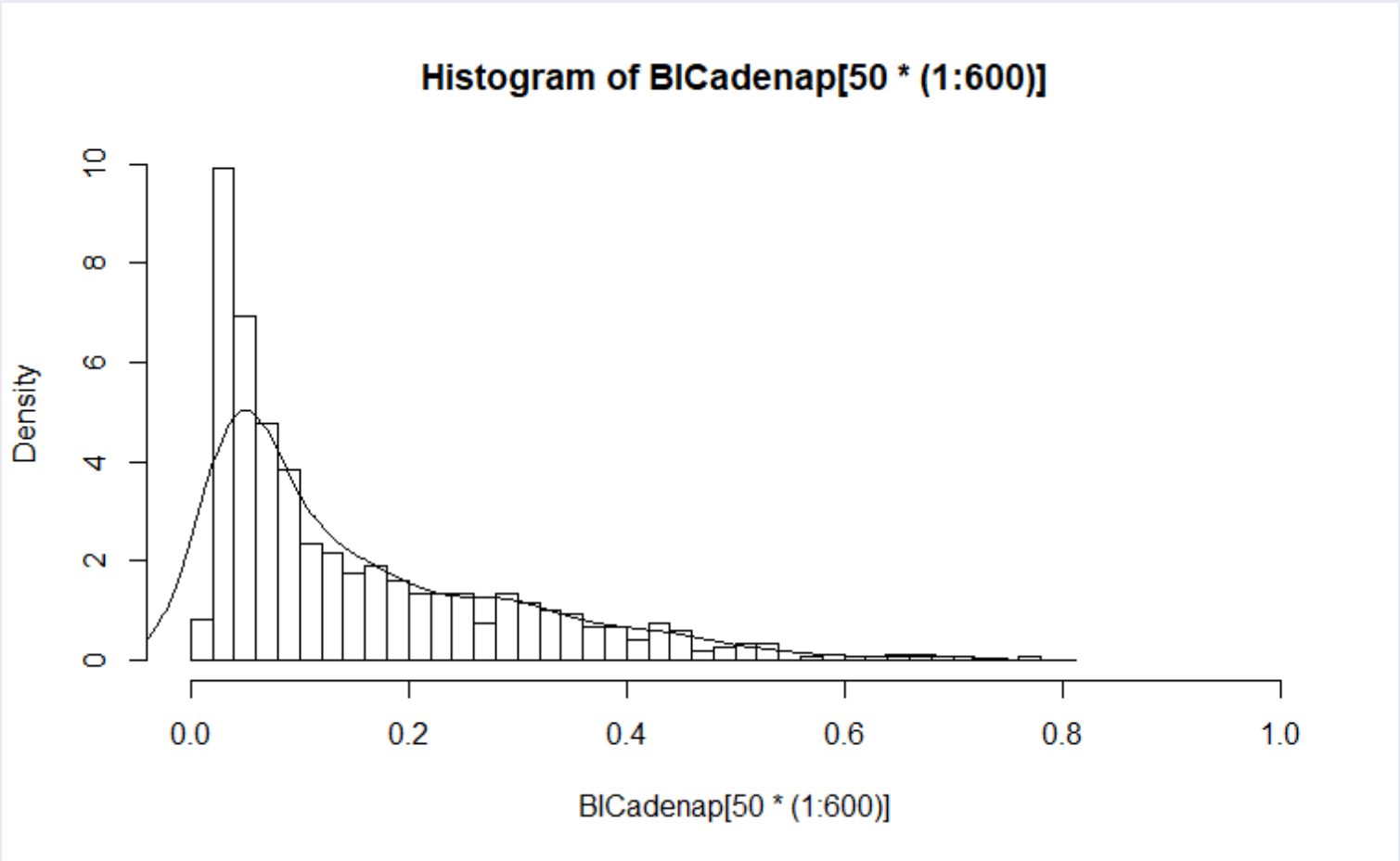


Figure 5: Histograma y estimación de la densidad mediante kernels de P

```
Resumen para la muestra de K

 Min.      1st Qu.   Median     Mean     3rd Qu.     Max.
   6.00    18.00     42.00      62.06    95.25       200.00

  Var.         SD.
  2738.37      52.3295

 Resumen para la muestra de P

 Min.      1st Qu.   Median     Mean     3rd Qu.     Max.
0.01628   0.04495   0.10207    0.15576  0.23050     0.70839

 Var.         SD.
 0.02001      0.14146
```

Ejemplo Modelo no jerárquico

Supongamos que, en el contexto antes descrito, interesa contrastar

$$H_0 : N \leq 50 \text{ vs. } H_1 : N > 50$$

Tenemos que, *a priori*, $\mathrm{P}(H_1) = 1/4$ y $\mathrm{P}(H_2) = 3/4$. Podemos aproximar el factor de Bayes

$$K = \frac{\mathrm{P}(D|M_1)}{\mathrm{P}(D|M_2)} = \frac{\mathrm{P}(M_1|D)}{\mathrm{P}(M_2|D)} \frac{\mathrm{P}(M_2)}{\mathrm{P}(M_1)} = \frac{0.5824139}{1 - 0.5824139} \frac{3}{1} = 4.184147.$$

Ejemplo Metropolis-Hastings (No adaptativo y no Gibbs)

Sea el modelo de crecimiento logístico $\frac{dX}{dt} = \theta_1 X(\theta_2 - X)$ con $X(0) = X_0$. Suponga que tenemos observaciones y_i para $X(t_i), t_1 < t_2 < ... < t_n$, con ruido Gaussiano aditivo independiente, esto es

$$y_i = X(t_i) + \epsilon_i; \epsilon_i \sim N(0, \sigma), i = 1, 2, ..., n.$$

Simule datos con $X(0) = 100$, $\theta_1 = 1$, $\theta_2 = 1000$, $\sigma = 30$, $n = 26$ equiespaciados en $t \in [0, 10]$. ¿Cómo hacer inferencia bayesiana para los parámetros θ_1, θ_2?

Consideremos la ecuación diferencial

$$\frac{dX}{dt} = \theta_1 X(t)(\theta_2 - X(t)), X(0) = X_0$$

Sabemos que esta ecuación tiene una solución analítica

$$X(t) = \frac{\theta_2 X_0 e^{\theta_1 t}}{\theta_2 + X_0(e^{\theta_1 t} - 1)}.$$

Proponemos una previa ligeramente informativa

$$\theta_1 \sim \text{Gama}(2, 2)$$

y

$$\theta_2 \sim \text{Normal}(1000, 100^2)$$

independientes.

La distribución posterior es

$$p(\theta_1, \theta_2|\vec{y_i}) \propto p(\vec{y_i}|\theta_1, \theta_2)p(\theta_1, \theta_2)$$

$$\propto \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{n}\left(y_i - \frac{\theta_2 X_0 e^{\theta_1 t}}{\theta_2 + X_0(e^{\theta_1 t_i}-1)}\right)^2}{\sigma^2}\right]$$

$$\times \left(\theta_1 e^{-2\theta_1}\right)\left(\exp\left\{-\frac{1}{2}\frac{(\theta_2 - 1000)^2}{100^2}\right\}\right)$$

Podemos implementar un algoritmo de Metropolis-Hastings. Proponemos una distribución de tipo caminata aleatoria para la propuesta

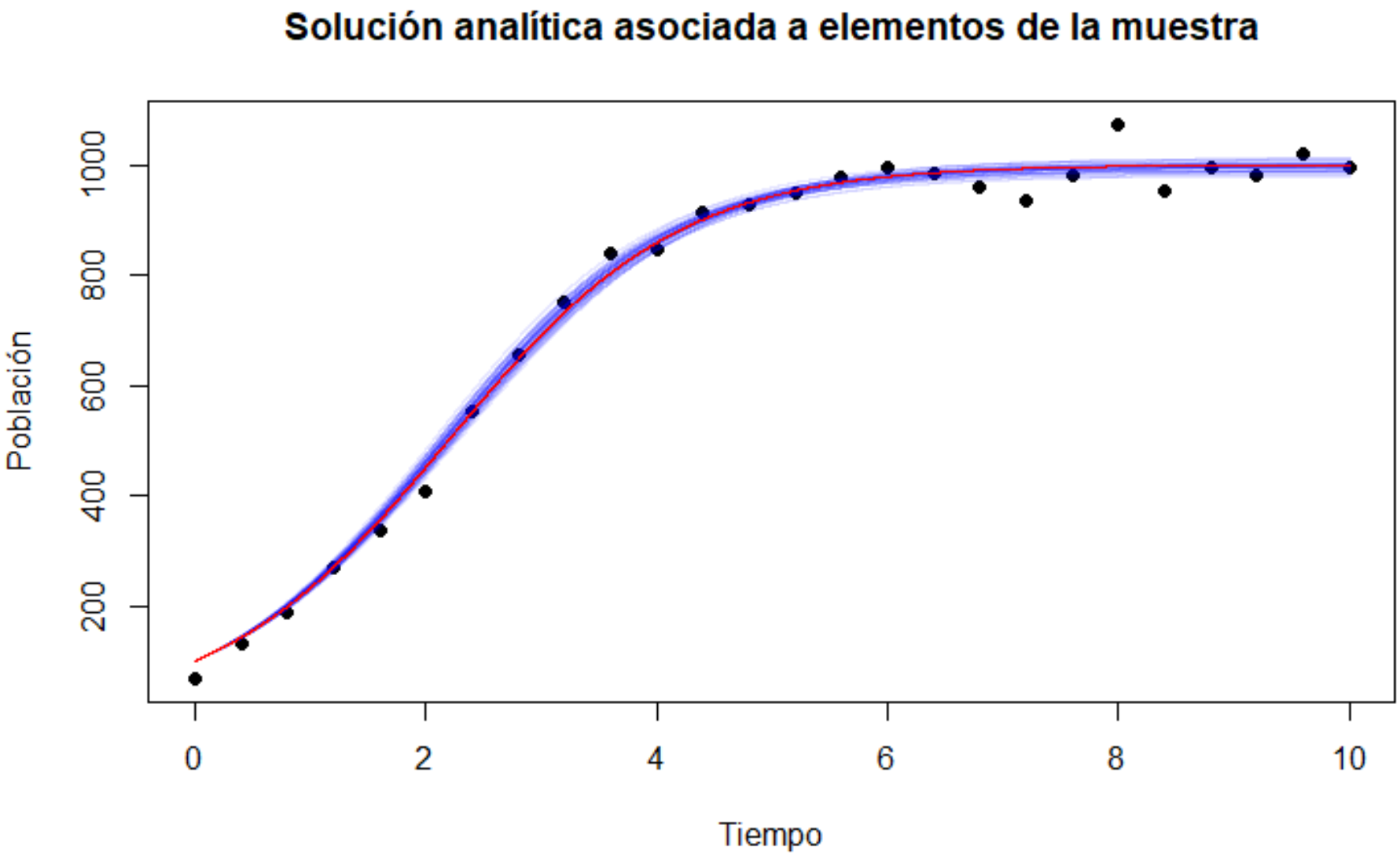
$$\epsilon \sim \mathrm{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.1 & 0 \\ 0 & 10 \end{bmatrix}\right)$$

$$(\theta_1^*, \theta_2^*) = \left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right) + \epsilon.$$

Notemos que por tratarse de una distribución de propuesta simétrica $Q(\theta^*|\theta) = Q(\theta|\theta^*)$ tenemos que

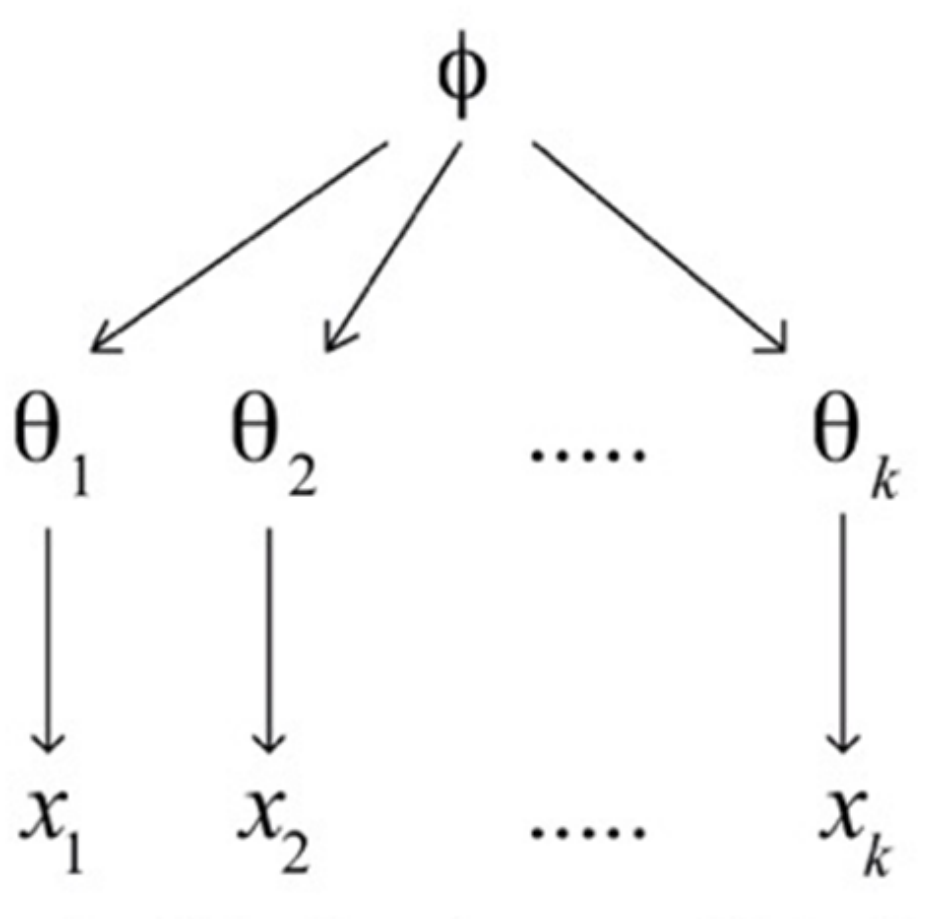
$$\alpha(\theta, \theta^*) = \frac{p(\theta^*)Q(\theta^{(i-1)}|\theta^*)}{p(\theta^{(i-1)})Q(\theta^*|\theta^{(i-1)})} = \frac{p(\theta^*)}{p(\theta^{(i-1)})}.$$

Proponemos puntos iniciales $\theta_1^{(0)} = 0.99$ y $\theta_2^{(0)} = 999$
Tras correr el MH, registramos un porcentaje de aceptación del 7.28%.


Solución analítica asociada a elementos de la muestra
Población
200
400
600
800
1000
0
2
4
6
8
10
Tiempo

Esquema de modelos jerárquicos (intuición visual)

Esquema de modelo jerárquico



Modelos jerárquicos (lineales)

Motivación y casos de aplicación

Las aplicaciones en las que surgen:

- Datos longitudinales: medidas repetidas y de curvas de crecimiento.
- Datos recabados por estratos. Covariables a nivel de colectivos (hospitales, escuelas).
- Meta-análisis: combinar información de estudios inter-relacionados.

Variables latentes

Definición y tratamiento frecuentista

- Son variables aleatorias (presentan variabilidad) pero nunca son observadas.
- Problema con interpretación frecuentista de la probabilidad. ¿Frecuencias relativas de algo que nunca se observa?
- Solución: ¡Son observaciones perdidas!
  Herramientas: Algoritmo EM y predictores lineales (BLUP's).

Variables latentes

Tratamiento Bayesiano e inferencia vs. modelo

- Más natural desde el paradigma Bayesiano. ¡Aparecen todo el tiempo! Aunque no se observen podemos postular una distribución (hiperpárametros).
- Dilemas filosóficos. Linea entre elementos del modelo y las herramientas de inferencia es difusa.
  Fundamentalmente conceptual, no técnico.

Ejemplos de variables latentes

Ejemplos donde aparecen variables latentes

1. Como parte del modelo
   - Efectos aleatorios/ Modelos mixtos
   - Análisis de factores
   - Modelos jerárquicos
   - Mezclas de distribuciones
2. Como solución a problemas en la práctica
   - Observaciones faltantes
   - Observaciones censuradas
3. Como herramientas de inferencia
   - Ampliación del modelo
   - Modelación de datos categóricos

Estructura general de los modelos jerárquicos

Planteamiento No Bayesiano

Un modelo jerárquico tiene la siguiente estructura

Nivel I.(Observaciones)

$$p(x|\theta) = p(x_1, ..., x_k|\theta_1, ..., \theta_k) = \prod_{i=1}^{k} p(x_i|\theta_i).$$

Nivel II.(Parámetros/ Variables latentes u observaciones perdidas)

$$p(\theta; \phi) = p(\theta_1, ..., \theta_k; \phi) = \prod_{i=1}^{k} p(\theta_i; \phi).$$

Nivel III. (Hiperparámetros)

$$\phi$$

Estructura general de los modelos jerárquicos

Planteamiento Bayesiano

Nivel I.(Observaciones)

$$p(x|\theta) = p(x_1, ..., x_k|\theta_1, ..., \theta_k) = \prod_{i=1}^{k} p(x_i|\theta_i).$$

Nivel II.(Parámetros)

$$p(\theta|\phi) = p(\theta_1, ..., \theta_k|\phi) = \prod_{i=1}^{k} p(\theta_i|\phi).$$

Nivel III. (Hiperparámetros)

$$p(\phi)$$

El problema con la función de verosimilitud

Esquema de modelos con variables latentes y su función de verosimilitud

El modelo es

$$Y \sim f(y|X = x, \phi)$$

$$X \sim g(x|\theta).$$

Y la función de verosimilitud es

$$L(\theta, \phi; y) = \int \int ... \int f(y|x, \phi) g(x|\theta) dx_1 dx_2 ... dx_p$$

$$= \int f(y|x, \phi) g(x|\theta) dx.$$

Primer ejemplo de modelo jerárquico

En algunos problemas de conteo, las observaciones que registran cero son muy superiores a los que uno esperaría observar con un modelo discreto clásico (por ejemplo Poisson). Una forma de resolver este problema es con modelos cero inflado, donde la probabilidad de obtener cero puede fijarse arbitrariamente.

Por ejemplo, la función de probabilidad de un modelo Poisson cero inflado (ZIP) sería

$$\Pr(y_i = 0) = \pi + (1 - \pi)e^{-\lambda}$$

$$\Pr(y_i = k) = (1 - \pi)\frac{\lambda^k e^{-\lambda}}{k!}, \qquad k \geq 1$$

Visto como un modelo jerárquico introduciendo la variables latentes x_i

Nivel I.(Observaciones)

$$p(y_i|x_i,\lambda,\pi) = (\mathbb{1}[y_i=0])^{x_i}\left(\frac{\lambda^{y_i}e^{-\lambda}}{y_i!}\right)^{1-x_i}.$$

Nivel II.(Parámetros/ Variables latentes u observaciones perdidas)

$$p(x_i|\pi,\lambda) \propto (\pi)^{x_i}(1-\pi)^{1-x_i}.$$

Nivel III. (Hiperparámetros)

$$p(\lambda,\pi) \propto \lambda^{\alpha-1}e^{-\beta\lambda}(\pi)^{a-1}(1-\pi)^{b-1}$$

Calculando las condicionales completas (pasos muestreador de Gibbs)

La densidad generalizada conjunta posterior es

$$p(\vec{x}, \lambda, \pi|\vec{y}) \propto \lambda^{\alpha-1} e^{-\beta\lambda} (\pi)^{a-1} (1-\pi)^{b-1} \prod_{i=1}^{n} \left[(\pi \mathbb{1}[y_i = 0])^{x_i} \left((1-\pi) \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \right)^{1-x_i} \right]$$

Notemos que los X_i por ser binarias, dado todos los demás parámetros conocidos solo pueden tener distribución Bernoulli. Lo que falta es obtener la constante de normalización, que es

$$\frac{1}{\pi \mathbb{1}[y_i = 0] + (1-\pi)\frac{\lambda^{y_i} e^{-\lambda}}{y_i!}} = \frac{1}{(\pi \mathbb{1}[y_i = 0] + (1-\pi)\frac{\lambda^{y_i} e^{-\lambda}}{y_i!})^{x_1} (\pi \mathbb{1}[y_i = 0] + (1-\pi)\frac{\lambda^{y_i} e^{-\lambda}}{y_i!})^{1-x_1}}$$

Así,

$$X_i|X_{-i}, \lambda, \pi \sim \text{Bern}\left(\frac{\pi \mathbb{1}[y_i = 0]}{\pi \mathbb{1}[y_i = 0] + (1-\pi)\frac{\lambda^{y_i} e^{-\lambda}}{y_i!}}\right)$$

Es fácil ver que

$$\pi|\vec{X}, \lambda \sim \text{Beta}\left(a + \sum_{i=1}^{n} x_i, b + n - \sum_{i=1}^{n} x_i\right)$$

Para λ basta ver que si sabemos cuales observaciones viene del proceso Poisson basta hacer el análisis con las y_i cuyas $x_i = 0$. Sabemos como se actualizan los hiperparámetros, $\lambda|\boldsymbol{X} \sim \text{Gama}\left(\alpha + \sum_{i=1}^{n} x_i, \frac{\beta}{1+n\beta}\right)$. Por lo tanto,

$$\lambda|\vec{X}, \pi \sim \text{Gama}\left(\alpha + \sum_{i=1}^{n}[(1-x_i) * y_i], \frac{\beta}{1 + (n - \sum_{i=1}^{n} x_i)\beta}\right).$$

Segundo ejemplo de modelo jerárquico

La distribución Binomial-Negativa se puede expresa como una distribución marginal de la distribución Poisson-Gamma. Si

$$\text{Binomial-Negativa}(x|\lambda, m) = \binom{m+x-1}{x}\lambda^x(1-\lambda)^m$$

entonces

$$\text{Binomial-Negativa}(x|\lambda, m) = \int_0^\infty \text{Poisson}(x|\theta)\text{Gamma}\left(\theta\middle|m, \frac{1-\lambda}{\lambda}\right)d\theta.$$

Planteamiento del modelo Bayesiano

Sea $X_1, ..., X_n$ una m.a. con distribución Binomial-Negativa$(x|\lambda, m)$ con $0 < \lambda < 1$ y $m \in \mathbb{Z}$ desconocidas. Debemos proponer una distribución inicial conjunta para m, λ, por ejemplo

$$\lambda \sim \text{Beta}(\alpha, \beta)$$

$$m \sim \text{Poisson}(\phi)$$

con $m \perp\!\!\!\perp \lambda$.

Nuestro modelo de juguete en la estructura general de los modelos jerárquicos

Nivel I.(Observaciones)

$$X_i|\theta_i \sim \text{Poisson}(\theta_i) \text{ con } X_i \perp\!\!\!\perp X_j \text{ si } i \neq j.$$

Nivel II.(Parámetros)

$$\Theta_i|m,\lambda \sim \text{Gamma}\left(m, \frac{1-\lambda}{\lambda}\right) \text{ con } \Theta_i \perp\!\!\!\perp \Theta_j \text{ si } i \neq j.$$

Nivel III. (Hiperparámetros)

$$\lambda \sim \text{Beta}(\alpha, \beta)$$

$$m \sim \text{Poisson}(\phi)$$

con $m \perp\!\!\!\perp \lambda$.

Sobre STAN

¿Qué es STAN?

- STAN es un programa para hacer análisis Bayesiano. Basta especificar el modelo.
- STAN se encarga de implementar el algoritmo MCMC.
- Genera automáticamente indicadores de convergencia e independencia.
- Calcula resúmenes numéricos y densidades aproximadas.
- Existen paquetes de R que llaman a STAN. Podemos especificar el modelo (RStan) o usar interfaces de alto nivel o con especificación de fórmula (RStanArm).

Modelos lineales generalizados (MLG)

El MLG consiste de tres elementos:

- Una función de distribución f, perteneciente a la familia exponencial.
- Un predictor lineal $\eta = \mathbf{X}\beta$.
- Una función de enlace g tal que

$$\mathbb{E}(\mathbf{Y}) = \mu = g^{-1}(\eta)$$

Regresión Logística

El modelo de regresión logística puede plantearse como

$$Y_i \sim \text{Binomial}(p_i, n_i), \text{ para } i = 1, \ldots, m,$$

donde

$$p_i = \text{E}\left(\frac{Y_i}{n_i}\middle| X_i\right).$$

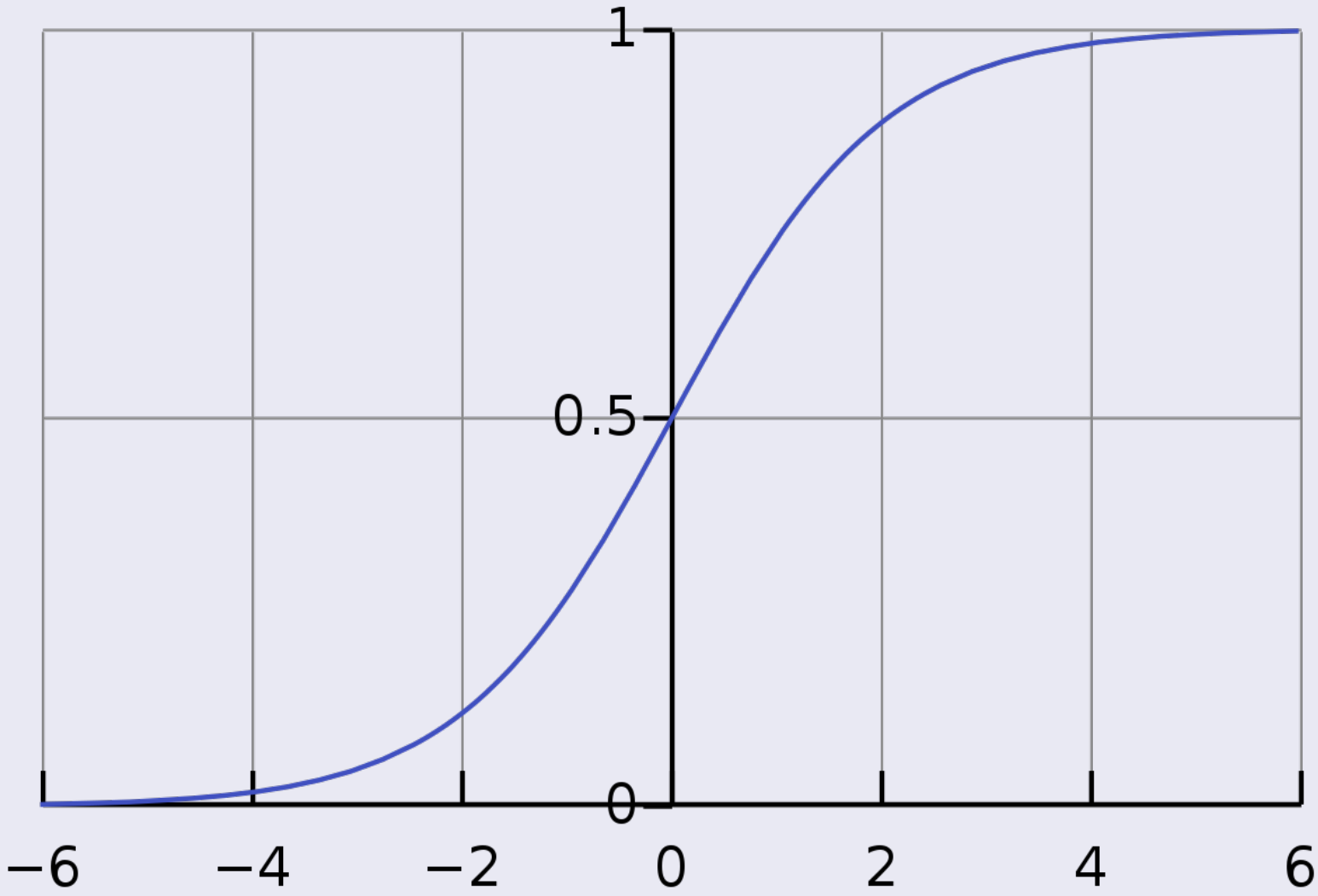
Los logits de las probabilidades binomiales desconocidas (los logaritmos de la razón de momios) son modeladas como una función lineal de los X_i.

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

o bien

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}$$

Curva logística



Motivación

La investigación del microbioma humano tiene como objetivo comprender cómo las comunidades de microbiomas interactuar con su huésped, responder a su entorno e influir en la salud del sistema.

Las tecnologías de secuenciación de alto rendimiento han permitido a los investigadores caracterizar la composición del microbioma mediante la cuantificación de la riqueza, la diversidad y la abundancia.

Sin embargo, las interacciones ambientales con el microbioma son complejas y desafían nuestra comprensión de la función de la comunidad y su impacto en la salud del individuo o el sistema.

Por otro lado, conocimiento de las relaciones entre la composición microbiana y otras covariables pueden ayudar a los investigadores a diseñar intervenciones personalizadas para ayudar mantener una comunidad de microbioma saludable

Regresión Dirichlet-multinomial

Un enfoque popular para modelar la relación entre los datos microbianos y las covariables es el modelo de regresión multinomial (DM) de Dirichlet, ya que

- maneja la estructura de composición de los datos del microbioma,
- acomoda la sobredispersión inducida por la heterogeneidad de la muestra y las proporciones variables entre las muestras y
- permite la introducción de covariables.

Regresión Dirichlet-multinomial: Modelo sin covariables

Sea $y_i = (y_{i,1}, \ldots, y_{iK})$ los conteos de taxa con la siguiente distribucón multinomial

$$y_i \sim \text{Multinomial}\,(y_i^* \mid p_i)$$

con $y_i^* = \sum_{k=1}^{K} y_{i,k}$, y p_i definido en el K-dimensional simplejo:

$$S^{K-1} = \left\{ (p_{i,1}, \ldots, p_{i,K}) : p_{i,k} \geq 0, \forall k, \quad \sum_{k=1}^{K} p_{i,k} = 1 \right\}.$$

Para considerar la posible sobredispersión, especificamos una previa conjugada para las probabilidades de conteos de taxa como

$$p_i \sim \text{Dirichlet}\,(\gamma_i)$$

donde γ_i es el vector K-dimensional $\gamma_i = (\gamma_{i,k} > 0, \forall k \in K)$.

Distribución posterior

Si y_i se describen con el modelo Dirichlet-multinomial(γ_i), entonces

$$p(p_i|y = y_i) \propto p_{i,1}^{y_{i,1}} \cdots p_{i,K}^{y_{i,K}} p_{i,1}^{\gamma_{i,1}-1} \cdots p_{i,K}^{\gamma_{i,K}-1} = p_{i,1}^{y_{i,1}+\gamma_{i,1}-1} \cdots p_{i,K}^{y_{i,K}+\gamma_{i,K}-1}$$

que es el kernel de una distribución Dirichlet($y_i + \gamma_i$).

```
#datos simulados
p<-c(.2,.3,.5)
set.seed(10)
dat<-rmultinom(1,size=100,prob=p)
dat

#previa
library(Compositional)
library(MCMCpack)
gamma<-c(1,1,1)
previa <- ddirichlet(c(.5,.5,.5), gamma)
muestra_previa <- rdirichlet(100, gamma)
bivt.contour(muestra_previa)

#posterior
muestra_posterior <- rdirichlet(50, dat+gamma)
bivt.contour(muestra_posterior)

#Estimador puntual: media posterior
(dat+gamma)/sum(dat+gamma)
```

Regresión Dirichlet-multinomial: Modelo con covariables

Supongamos que para cada sitio i observamos P covariables

$$\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,P})^\top.$$

Para incorporar el efecto de covariables sobre las composiciones, consideramos

$$\log(\gamma_{i,k}) = \lambda_{i,k} = \alpha_k + \boldsymbol{x}_i^\top \varphi_k$$

donde $\varphi_k = (\varphi_{k,1}, \ldots, \varphi_{k,P})^\top$ representa la relación de las P covariables con el k-ésimo taxón composicional, y α_k es el intercepto de cada taxón.

Reparametrización de la distribución Dirichlet considerando dos parámetros

Definimos los parámetros $\mu_k = \mathrm{E}[y_k]$ y $\theta_i = \sum_{k=1}^{K} \gamma_{i,k}$ para modelar la "dispersión" (precisión). Esta parametrización se denota $y_i \sim \mathcal{D}_a(\boldsymbol{\mu}_i, \theta_i)$ y tiene fdp

$$f(y_i \mid \boldsymbol{\mu}_i, \theta_i) = \frac{1}{B(\boldsymbol{\mu}_i \theta_i)} \prod_{k=1}^{K} y_k^{(\mu_{i,k}\theta_i - 1)}$$

donde $\boldsymbol{\mu}_i \in (0,1)^K$ and $\theta_i > 0$.
Bajo esta parametrización $\mathrm{E}[y_k] = \mu_k$, $\mathrm{VAR}[y_k] = [\mu_k (1-\mu_k)]/(\theta+1)$, and $\mathrm{COV}[y_i, y_j] = -\mu_i \mu_j/(\theta+1)$.
Para convertir $\boldsymbol{\mu}_i$ and θ_i a la versión original $\boldsymbol{\gamma}_i$, definimos $\gamma_{i,k} = \mu_{i,k}\theta$ y $\gamma_{0,i} = \theta_i$.

Regresión Dirichlet-multinomial: Modelo con covariables

Con la formulación del modelo Dirichlet considerando dos parámetros, el modelo de regresión es

$$\mathbf{Y}_i \mid \mathbf{x}_i \sim \mathcal{D}_a(\boldsymbol{\mu}_i, \theta_i)$$

Además para cada componente k

$$\lambda_{i,k} = \mathbf{x}_i^\top \boldsymbol{\beta}_k,$$

donde para uno de los componentes $\tilde{k}$, todos los elementos de $\boldsymbol{\beta}$ son iguales a 0, con el fin de garantizar la identificabilidad del modelo (grados de libertad del modelo es entonces $K - 1$).

Regresión Dirichlet-multinomial: Modelo con covariables

Entonces

$$\mu_{i,k} = \frac{\exp\left(\lambda_{i,k}\right)}{\sum_{k=1}^{K} \exp\left(\lambda_{i,k}\right)}.$$

Aplicando nuevamente la función exponencia e introduciendo el vector de coeficientes $\boldsymbol{\delta}$, podemos expresar a θ_i como

$$\theta_i = \exp\left(\mathbf{z}_i^\top \boldsymbol{\delta}\right)$$

Obtenemos:

$$f\left(\mathbf{y}_i \mid \left(\mu_{i,1}, \mu_{i,2}, \ldots, \mu_{c,i}\right)^\top, \theta_i\right) = \\ = f\left(\mathbf{y}_i \mid \boldsymbol{\mu}_i, \theta_i\right),$$

y la verosimilitud:

$$L\left(\{\mathbf{y}_i; i = 1, \ldots, n\}\right) = \prod_{i=1}^{n} f\left(\mathbf{y}_i \mid \boldsymbol{\mu}_i, \theta_i\right).$$

Regresión Dirichlet-multinomial: MCMC

$$
\begin{aligned}
\mathbf{y}_i \mid \boldsymbol{x}_i, \boldsymbol{z}_i &\sim D\left(\boldsymbol{\mu}_i, \theta_i\right), \\
\mu_{i,k} &= \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k\right), \\
\beta_{i,k} &\sim N\left(0, \sigma_k^2\right), \quad j = 1, \ldots, P_{\boldsymbol{\beta}}, \quad \forall k \neq \tilde{k}, \\
\beta_{i,\tilde{k}} &= 0, \quad j = 1, \ldots, P_{\boldsymbol{\beta}}, \\
\delta &\sim N\left(0, \omega_k^2\right), \quad j = 1, \ldots, P_{\boldsymbol{\gamma}}.
\end{aligned}
$$

con $\theta_{i,k} \equiv \theta = \delta$.

Ejemplo usando Stan (en R)

Los datos de Aitchison (2003) corressponde a la composición de muestras sanguíneas de

- Albumin,
- Pre-Albumin,
- Globulin A, and
- Globulin B

en relación a dos tipos de enfermedades (A y B).

14 de los pacientes presentan la enfermedad A, 16 la enfermedad B y 6 no están clasificados.

Conclusiones

- El marco Bayesiano permite tratar los problemas usuales de inferencia de forma unificada. Partimos de la densidad posterior.
- Utilidad del análisis conjugado. Posibilita muchas cuentas incluyendo predictivas y factores de Bayes.
- Las previas objetivas (análisis de referencia) puede utilizarse para comparar contra otras previas.

Conclusiones

- Los modelos jerárquicos, y en general cualquier modelo al que introduzcamos variables latentes, son muy flexibles.
- Bajo el enfoque Bayesiano las variables latentes pueden tratarse como parámetros de estorbo.
- MCMC puede ser muy costoso computacionalmente, en especial cuando hay alta autocorrelación. Una alternativa son las aproximaciones analíticas (aproximación de Laplace e INLA) o de integracón numérica.
- Los métodos MCMC simulan muestras de la distribucipón posterior. Trabajar con ellas para hacer la inferencia es sencillo.

FIN

Estadística Bayesiana, algoritmos MCMC y modelos jerárquicos

Mario Enrique Carranza Barragán
Dra. Leticia Ramírez Ramírez

Grupo Bimbo/ Centro de Investigación en Matemáticas

18, 20 y 25 de abril de 2023

mario.carranza@grupobimbo.com/ mario.carranza@cimat.mx