Estadística Bayesiana, algoritmos MCMC y modelos jerárquicos

Mario Enrique Carranza Barragán

Centro de Investigación en Matemáticas, A.C.

10-12 de junio de 2019

mario.carranza@cimat.mx

Tabla de contenidos

Inferencia estadística

El problema de la inferencia estadistica

En general tenemos:

- Datos X
- Cantidades desconocidas θ

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

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

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.
- Se proponen estadísticos ingeniosos que no siempre cumplen el principio de verosimilitud.

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.
- Siempre se debe cumplir el principio de verosimilitud.

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.

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 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

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

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

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

Segundo 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 $x_i \sim \text{Unif}(0, \theta)$ 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,

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

Tercer ejemplo de análisis conjugado

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

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

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

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

Así,

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

Cuarto ejemplo de análisis conjugado

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

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

Así,

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

Quinto ejemplo de análisis conjugado

Distribución Normal Gama Inversa

Suponga que

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

y además

$$\sigma^2 \mid \alpha, \beta \sim \Gamma^{-1}(\alpha, \beta).$$

Decimos que

$$(\mu, \sigma^2) \sim \text{N-}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta)$$

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 N-$\Gamma^{-1}(\mu, \lambda, \alpha, \beta)$ 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

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

$$\equiv N\Gamma^{-1}\left(\mu_0', \lambda', \alpha', \beta'\right)$$

Mezclas finitas de conjugadas... ¿son conjugadas?

Si para la distribución de $X|\Theta$ se tienen dos previas conjugadas con fd (fdp) $p_1(\theta)$ y $p_2(\theta)$, se tiene que

$$p_3(\theta|\alpha) = \alpha p_1(\theta) + (1-\alpha)p_2(\theta).$$

Observemos que

$$\begin{aligned} p_3(\theta|\alpha, \vec{x}) &\propto p_3(\theta|\alpha)p(\vec{x}|\theta, \alpha) \\ &\propto (\alpha p_1(\theta) + (1-\alpha)p_2(\theta))p(\vec{x}|\theta) \\ &= \alpha p_1(\theta)p(\vec{x}|\theta) + (1-\alpha)p_2(\theta)p(\vec{x}|\theta) \\ &\propto \alpha p_1^*(\theta|\vec{x}) + (1-\alpha)p_2^*(\theta|\vec{x}) \\ &= p_3^*(\theta|\vec{x}, \alpha). \end{aligned}$$

La posterior pertenece a la misma familia que de combinaciones lineales de familias de previas conjugadas.

Mezclas infinitas de conjugadas... ¿son conjugadas?

Consideremos el modelo $p(\theta|\gamma)$ donde γ es el hiperparámetro. Supongamos que la previa es de la forma

$$p(\theta) = \int \Gamma p(\theta|\gamma) dF(\gamma),$$

donde F es una función de distribución para γ.

$$\begin{aligned} p(\theta|\vec{x}) &\propto p(\vec{x}|\theta)p(\theta) \\ &= p(\vec{x}|\theta) \int_{\Gamma} p(\theta|\gamma) dF(\gamma) \\ &= \int_{\Gamma} p(\vec{x}|\theta)p(\theta|\gamma) dF(\gamma) \\ &\propto \int_{\Gamma} p^*(\theta|\gamma, \vec{x}) dF(\gamma) \\ &\propto p^*(\theta|\vec{x}). \end{aligned}$$

La posterior pertenece a la misma familia que la previas.

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 Poisson(λ) con previa Gama(α, β) (y por tanto posterior Gamma(α', β') $\equiv$ Gamma($a + \sum x_i, b + n$)).
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,

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

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.

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

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:
  $\mathcal{A} = \{a; a \in \Theta\}$
  $L_1(a, \theta) = (a - \theta)^2$,
  $L_2(a, \theta) = |a - \theta|$
- Estimación por intervalos:
  $\mathcal{A} = \{B; B \subseteq \Theta\}$
  $L(B, \theta) = \lambda(B) + k\mathbb{1}_{B^c}(\theta)$
- Pruebas de hipótesis (Estilo Neyman-Pearson):
  $\mathcal{A} = \{H_0, H_1\}$, $L(H_i, H_j) = c_{ij}$ con $i = 0, 1$ y $j = 0, 1$.

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.

Ejemplo

Si no se tiene información sobre θ, tampoco se tiene información sobre $\log\theta$, pero una distrribución inicial sobre θ no corresponde a una distribución uniforme para $\log\theta$. Esto es, para $\phi = g(\theta)$ y $\pi(\theta) = 1$,

$$\pi(\theta) \neq \pi(\phi) = \left| \frac{d}{d\phi} g^{-1}(\phi) \right|.$$

Distribuciones iniciales objetivas invariantes

Medida de Haar

Buscar una estructura invariante en el problema e imponerla a la distribución inicial.

Ejemplo

Sea X una v.a. cuya fd es $f(x-\theta), x>0$ y ésta es invariante al parametro de localización θ. Esto significa que $Y = X + a$ se distribuye como $f(y-\phi)$ con $\phi = \theta + a$. Dado que el modelo es invariante a la localización, la distribución inicial también debe serlo.

$$\pi(\theta) = \pi(\theta - a), \forall a; \Rightarrow \pi(\theta) = 1.$$

Una inicial invariante objetiva para un parámetro de localización es la distribución uniforme.

Distribuciones iniciales objetivas invariantes

Ejemplo

Sea X una v.a. cuya función de densidad es $\frac{1}{\sigma} f(\frac{x}{\sigma})$, es invariante al parámetro de escala, σ. Esto significa que $Y = cX$ tiene la misma distribución que X, pero con un difrente parámetro de escala. Dado que la densidad es invariante a la escala, la distribución inicial también debe serlo:

$$\pi(\sigma) = \frac{1}{c}\pi\left(\frac{\sigma}{c}\right), c > 0; \Rightarrow \pi(\sigma) = \sigma^{-1}.$$

Una inicial invariante objetiva para el log de un parámetro de localización es la distribución uniforme.

Distribuciones inicial de Jeffreys

Este método fue creado por Jeffreys (1946) y se basa en la información de Fisher (para modelos regulares) dada por

$$h_\theta(\theta) = -\int_{\mathcal{X}} f(x|\theta)\frac{d^2}{d\theta^2}\log f(x|\theta)dx \ \theta \in \Theta \subset \mathbb{R}.$$

La distribución inicial de Jeffreys está definida como

$$\pi_\theta(\theta) \propto h_\theta(\theta)^{1/2}.$$

Jeffreys justificó su método por el hecho de que éste satisface el requerimiento de invarianza ante reparametrizaciones.

$$\pi_\theta(\theta)n \propto h_\theta(\theta)^{1/2} = \left\{ h_\gamma(\gamma(\theta)) \left(\frac{d}{d\theta}\gamma(\theta) \right)^2 \right\}^{1/2} = \pi_\gamma(\gamma(\theta)) \left| \frac{d}{d\theta}\gamma(\theta) \right|.$$

donde $\gamma(\theta)$ es una transformación uno a uno de θ.

Ejemplo de distribución inicial de Jeffreys

Consideremos $X_i \sim \text{Bernoulli}(\theta)$ con $\theta \in [0,1]$. Notemos que

$$
\begin{aligned}
h(\theta) &= -E_\theta \left[\frac{d^2}{d\theta^2} \log f(x|\theta) \right] \\
&= -E_\theta \left[\frac{d^2}{d\theta^2} \log \left(\theta^x (1-\theta)^{1-x} \right) \right] \\
&= -E_\theta \left[\frac{d^2}{d\theta^2} x \log \theta + (1-x) \log(1-\theta) \right] \\
&= -E_\theta \left[x \frac{-1}{\theta^2} + (1-x) \frac{-1}{(1-\theta)^2} \right] \\
&= \theta \frac{1}{\theta^2} + (1-\theta) \frac{1}{(1-\theta)^2} \\
&= \frac{1}{\theta} + \frac{1}{(1-\theta)} = \frac{1}{\theta(1-\theta)} = \theta^{-1}(1-\theta)^{-1}
\end{aligned}
$$

Ejemplo de distribución inicial de Jeffreys

Así, la previa de Jeffreys

$$\pi_\theta(\theta)n \propto h_\theta(\theta)^{1/2} = (\theta^{-1}(1-\theta)^{-1})^{1/2} = \theta^{-1/2}(1-\theta)^{-1/2}.$$

Notamos que se trata de un kernel Beta

$$p(\theta) \propto \theta^{1/2-1}(1-\theta)^{1/2-1}$$

por lo que

$$\theta \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$$

Distribuciones inicial de Jeffreys (general)

La inicial de Jeffreys se puede generalizar a parámetros multidimensionales definiéndola como la raíz cuadrada del determinante de la matriz de información de Fisher:

$$\pi(\boldsymbol{\theta}) \propto |\boldsymbol{H}(\boldsymbol{\theta}|^{1/2},$$

donde el elemento típico de $\boldsymbol{H}(\boldsymbol{\theta})$ esta dado por

$$[\boldsymbol{H}(\boldsymbol{\theta})]_{ij} = -\int_{\mathcal{X}} f(x|\boldsymbol{\theta}) \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x|\boldsymbol{\theta}) dx$$

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

¿Es difícil obtener la verosimilitud marginal o predictiva previa?

Recordemos que

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Podríamos intentar

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,\infty)}(\theta^{(i)}) \approx \int_{-\infty}^{\infty} \mathbb{1}_{(-\infty,\infty)}(\theta)p(\theta|D)d\theta = 1$$

Supongamos $\theta^{(1)}, ..., \theta^{(n)} \sim \Theta|D$

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{p(D|\theta)} \approx \int_{-\infty}^{\infty} \frac{p(\theta|D)}{p(D|\theta)}d\theta$$

$$= \int_{-\infty}^{\infty} \frac{p(D|\theta)p(\theta)/p(D)}{p(D|\theta)}d\theta = \int_{-\infty}^{\infty} \frac{p(\theta)}{p(D)}d\theta = (p(D))^{-1}$$

Supongamos $\theta^{(1)}, ..., \theta^{(n)} \sim \Theta$

$$\frac{1}{n}\sum_{i=1}^{n} p(D|\theta) \approx \int_{-\infty}^{\infty} p(D|\theta)p(\theta)d\theta$$

$$= \int_{-\infty}^{\infty} p(D|\theta)p(\theta)\frac{p(D)}{p(D)}d\theta = p(D)\int_{-\infty}^{\infty} \frac{p(D|\theta)p(\theta)}{p(D)}d\theta$$

$$= p(D)\int_{-\infty}^{\infty} p(\theta|D)d\theta = p(D)$$

Ejemplo de calculo de factor de Bayes

Supongamos que tenemos una muestra aleatoria $X_1, ..., X_n \sim \text{Unif}(0, \theta)$ 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:

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

Correcciones al factor de Bayes

Dos posibles correcciones

- Factor de Bayes Posterior (Aitkin 1991):

$$\overline{p(x|\mathcal{M}_j)} = \int p(x|\theta_k)\pi(\theta_k|x)d\theta$$

- Factor de Bayes Posterior (O'Hagan 1995):
  Dividir en entrenamiento x_t y prueba x_v.

$$\overline{p(x|\mathcal{M}_j)} = \int p(x_v|\theta_k)\pi(\theta_k|x_t)d\theta$$

  ¿Cómo elegir x_t?

Paradoja de Lindley

Sea H hipótesis simple y x los datos, puede ocurrir simultáneamente:

i. Un test de significancia (Fisheriano) para H diga que x es significativo, es decir, el p-valor calculado sea bajo.

ii. La probabilidad posterior de H sea alta, dado x y con una probabilidad inicial baja para H.

El ejemplo de Lindley (1957) $x_1, ..., x_n \sim N(\theta, \sigma^2)$ con σ^2 conocida (por simplicidad $\sigma^2 = 1$).

$$H : \theta = \theta_0$$

Elección de previa (no informativa). La probabilidad inicial de la hipótesis nula c. Repartimos uniformemente $1 - c$ en el intervalo I que contiene θ_0.

$$P(\theta = \theta_0) = \frac{cL(\theta)}{cL(\theta) + (1 - c) \int_I L(\theta) d\theta}$$

La cantidad pivotal $\bar{x} \sim N(\theta, 1/\sqrt{n})$ permite calcular el p-valor.

Criterios de selección de modelos Bayesianos

Criterios globales y tipo *Leave One Out* (LOO)

Tipo	Criterio	Definición
Global	MLIK	$p(\boldsymbol{y}\|\mathcal{M}_m)$
Global	F. de Bayes	$p(\boldsymbol{y}\|\mathcal{M}_0)/p(\boldsymbol{y}\|\mathcal{M}_1)$
Global	Devianza	$D(\theta) - 2\log(p(y\|\theta)) + C$
Global	DIC	$\text{DIC} = -2\log p(y\|\hat{\theta}_{\text{Bayes}}) + 2p_{\text{DIC}}.$ $p_{\text{DIC}} = 2\left(\log p(y\|\hat{\theta}_{\text{Bayes}}) - \text{E}_{\text{post}}(\log p(y\|\theta))\right)$
Global	WAIC	$\text{WAIC j} = -2\sum_{i=1}^{n}\log\int p(y_i \mid \theta)p_{\text{post}}(\theta)d\theta$ $+2p_{\text{WAIC j}}$, con $j = \{1, 2\}$
LOO	CPO	$CPO_i = \pi(y_i\|y_{-i})$
LOO	PIT	$PIT_i = \pi(Y_i^{\text{nuevo}} \leq y_i\|y_i)$
LOO	PIT discr.	$PIT_i^{\text{ajustado}} = PIT_i - 0 - 5 * CPO_i$

$$p_{\text{WAIC 1}} = 2\sum_{i=1}^{n}\left(\log(\text{E}_{\text{post}}p(y_i|\theta)) - \text{E}_{\text{post}}(\log p(y_i|\theta))\right),$$
$$p_{\text{WAIC 2}} = \sum_{i=1}^{n}\text{Var}_{\text{post}}(\log p(y_i|\theta)) \text{ y } \hat{\theta}_{Bayes} = \text{E}_{\text{post}}(\theta|y).$$

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 $u \sim U(0,1)$
3. si $u \leq \alpha(\theta^*, \theta^{(t)})$ hacer $\theta^{(t+1)} = \theta^*$; en caso contrario, hacer $\theta^{(t+1)} = \theta^{(t)}$

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

Ecuación de balance detallado

La cadena de Markov de MH satisface la ecuación de balance detallado pues si $\theta^* \neq \theta^{(t)}$ cumple que si $\alpha(\theta^*, \theta^{(t)}) < 1$ entonces necesariamente $\alpha(\theta^{(t)}, \theta^{(t)}) = 1$, de donde se tiene que

$$\alpha(\theta^{(t)}, \theta^*) Q(\theta^* | \theta^{(t)}) p(\theta^{(t)} | x) = \alpha(\theta^*, \theta^{(t)}) Q(\theta^{(t)} | \theta^*) p(\theta^* | x).$$

Para el caso $\theta^* = \theta^{(t)}$ y la relación anterior se tiene que

$$(1 - r(\theta^{(t)})) p(\theta^{(t)} | x) = (1 - r(\theta^*)) p(\theta^* | x).$$

Por lo que se cumple la EBD para $\theta^* \neq \theta^{(t)}$ y $\theta^* = \theta^{(t)}$.

Ecuación de balance detallado

Por cumplir la EBD se tiene que f es la densidad invariante de la cadena. Pues , para todo conjunto medible B,

$$\begin{aligned}\int_{\Theta} K(\theta^*, B)p(\theta^*|x)d\theta^* &= \int_{\Theta}\int_{B} K(\theta^*, z)p(\theta^*|x)dzd\theta^* \\ &= \int_{\Theta}\int_{B} K(z, \theta^*)p(z|x)dzd\theta^* \\ &= \int_{B} p(z|x)dz,\end{aligned}$$

pues $\int K(z, \theta^*)d\theta^* = 1$.

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 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 $X_i \sim \text{Bi}(k, p)$ 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,

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

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

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.

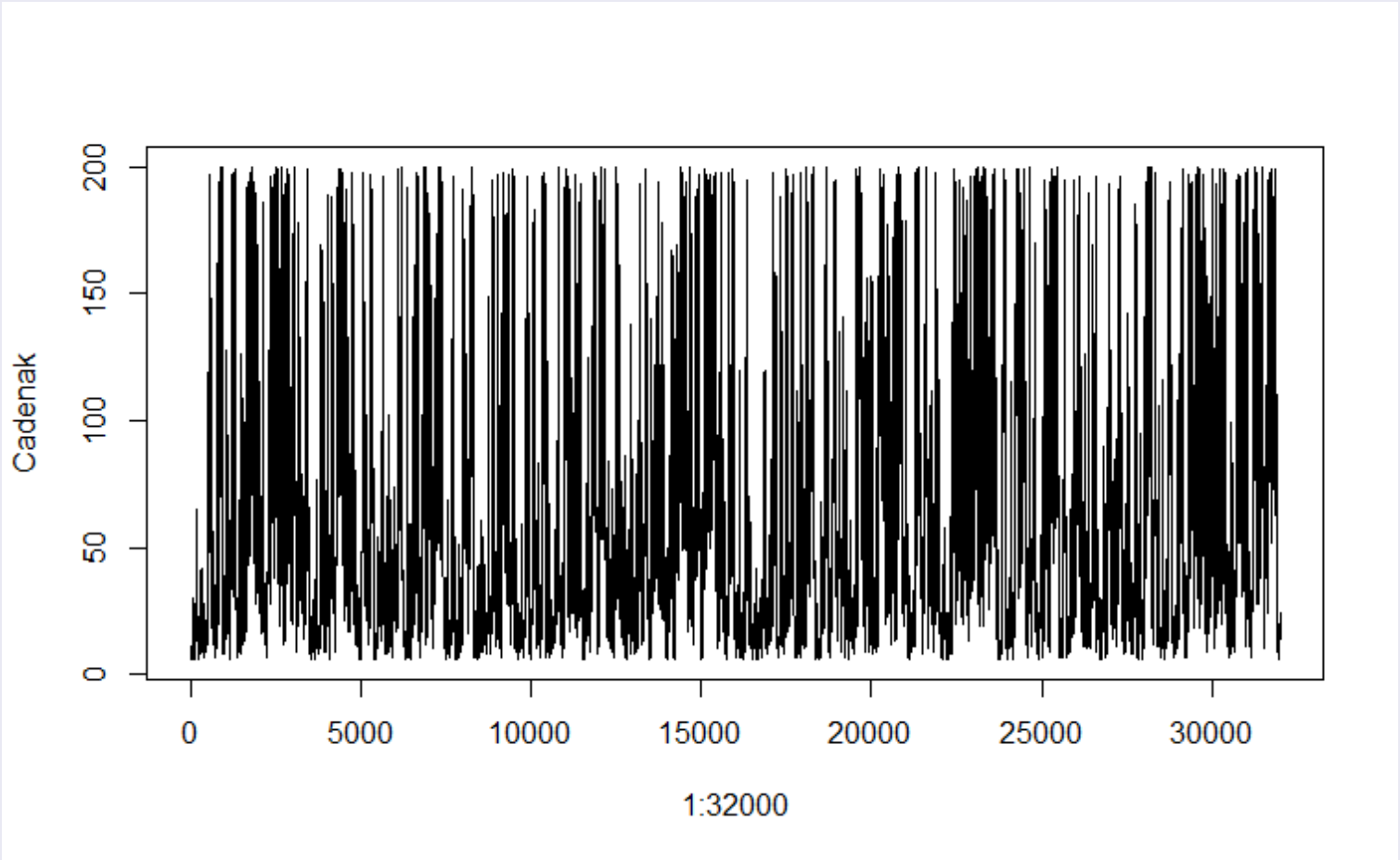


Figure 1: Traza de la cadena de K

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.

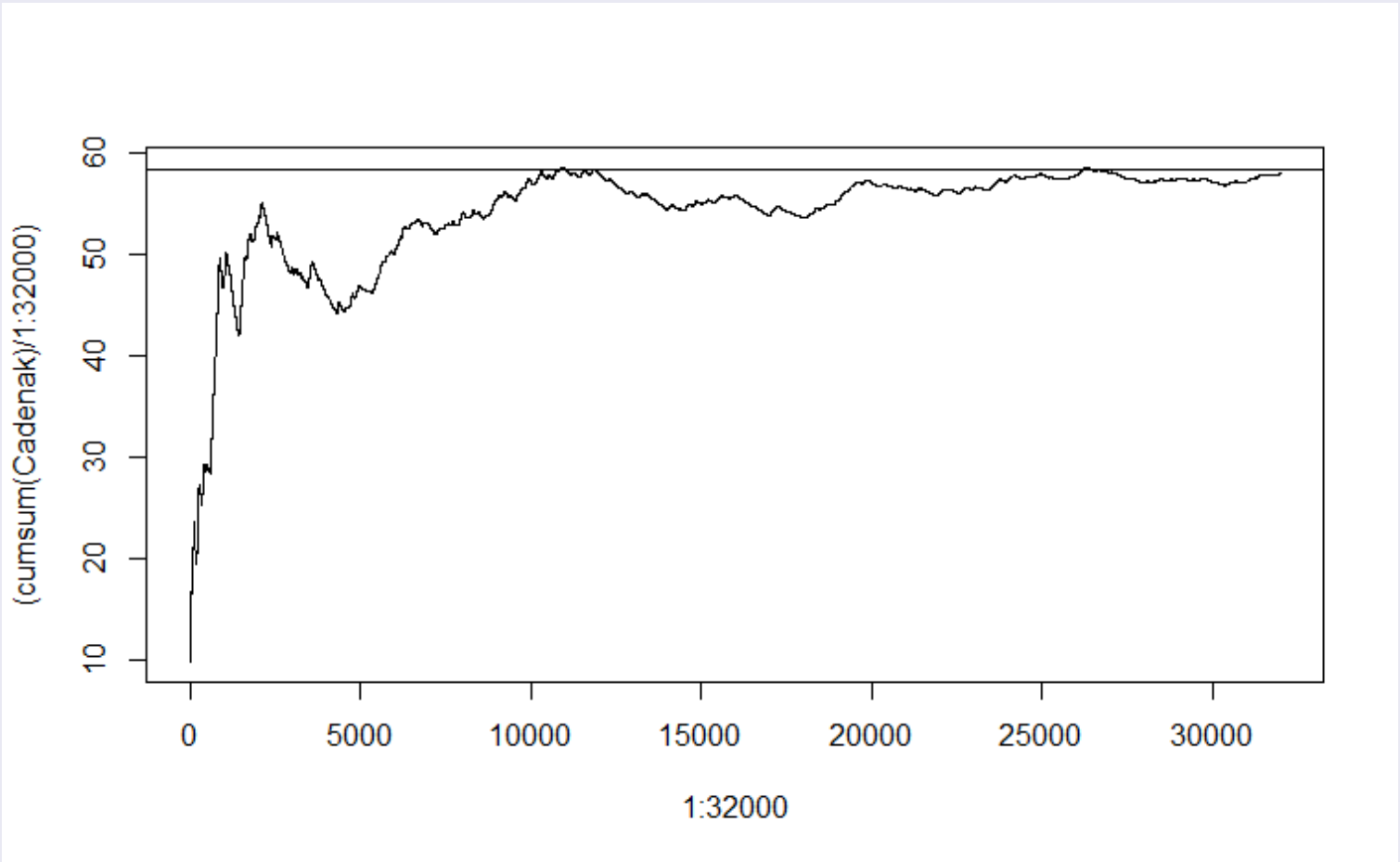


Figure 2: Promedios ergódicos de la cadena de K

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.

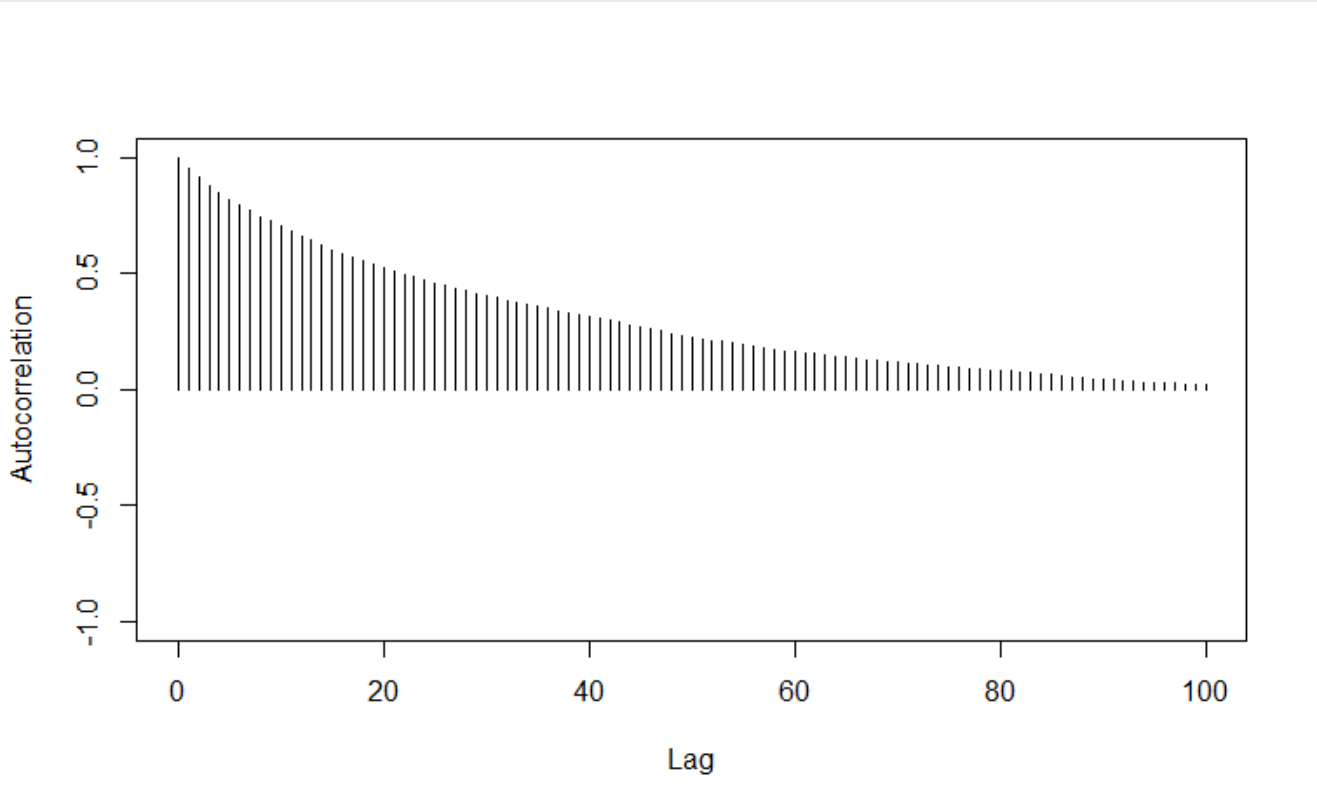


Figure 3: Gráfica de autocorrelación de K

Primer ejemplo: Modelo no jerárquico

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

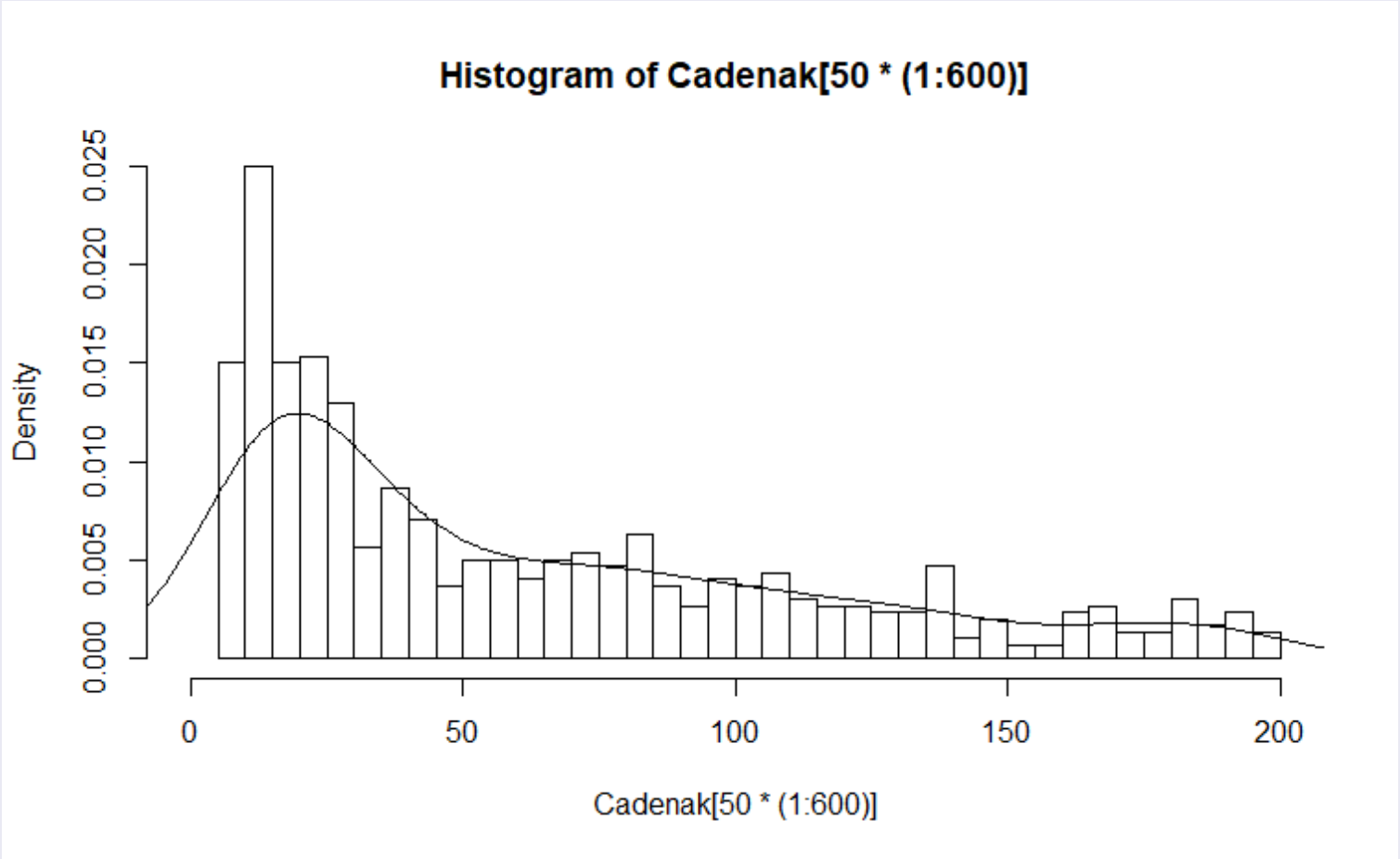


Figure 4: Histograma y estimación de la densidad mediante kernels de K

Primer ejemplo: Modelo no jerárquico

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

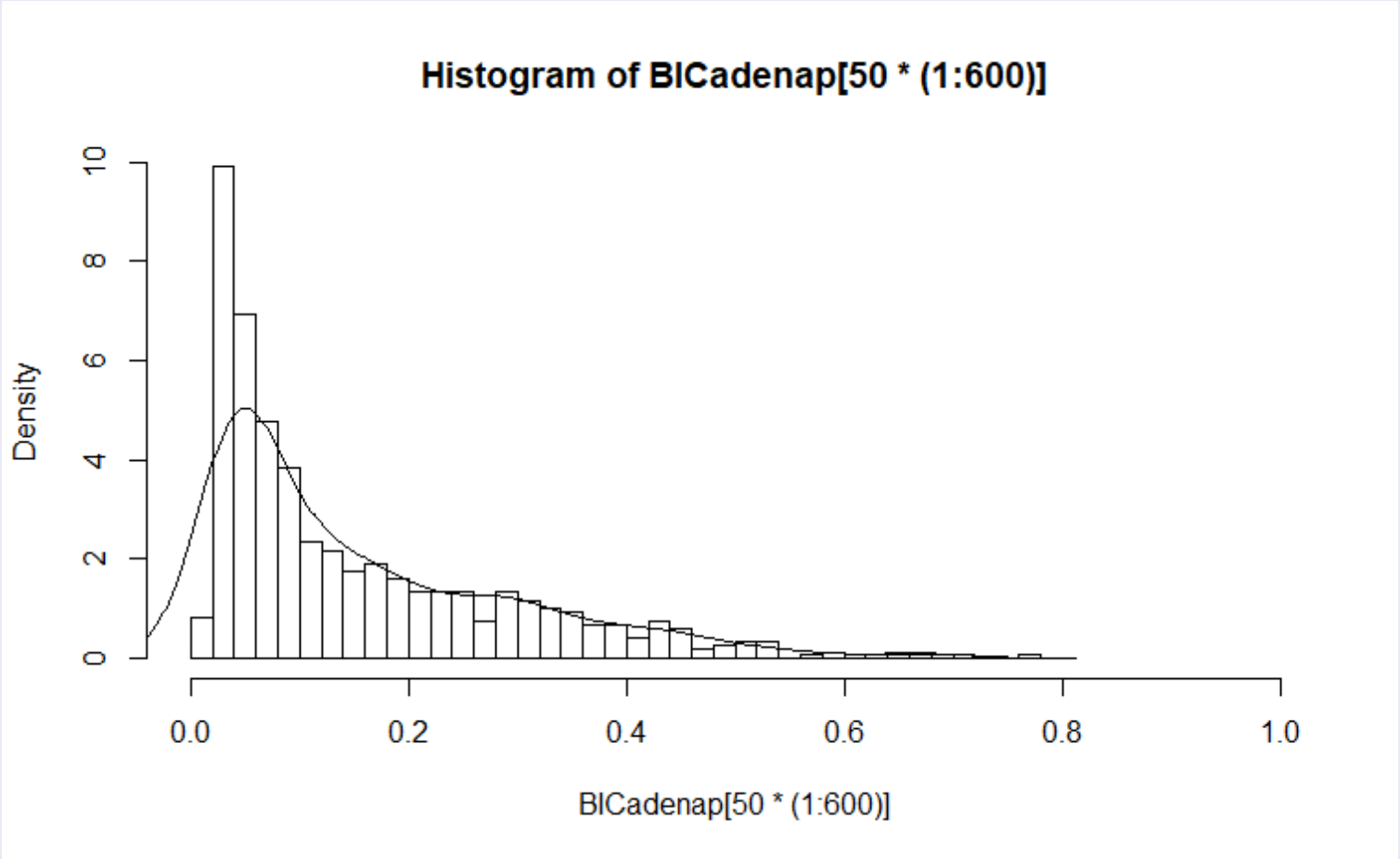


Figure 5: Histograma y estimación de la densidad mediante kernels de P

```
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

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

Cuantificación de la incertidumbre (UQ)

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

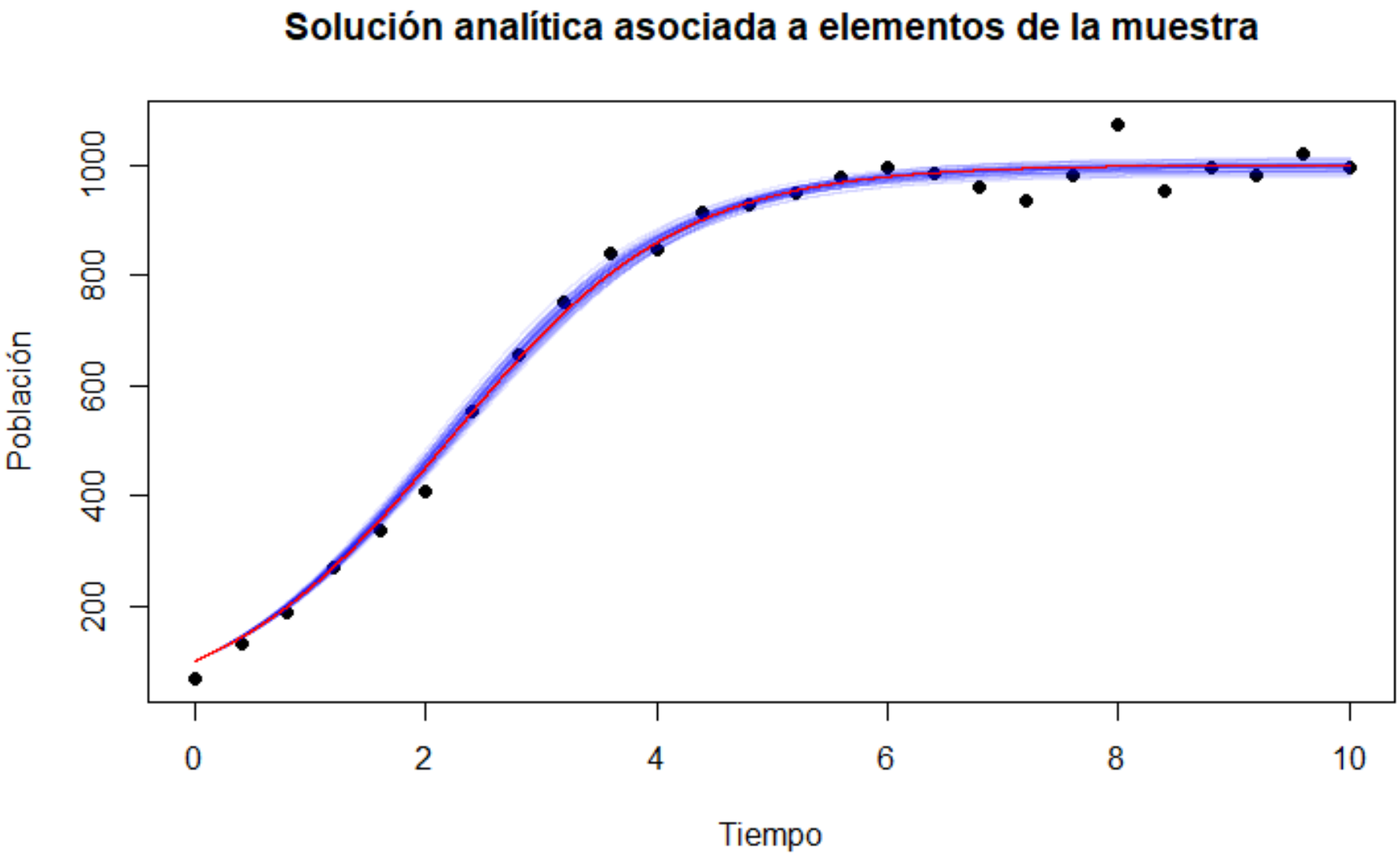
$$\epsilon \sim \mathrm{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.1 & 0 \\ 0 & 10 \end{bmatrix}\right)$$

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

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

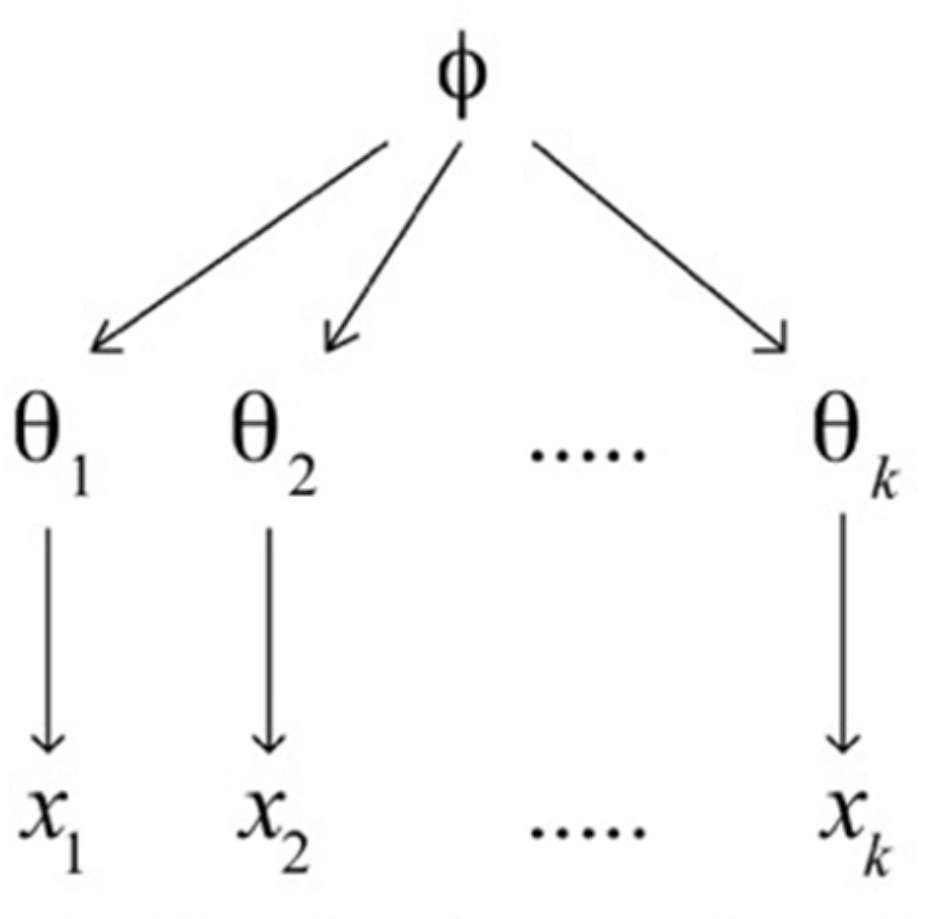
$$\alpha(\theta, \theta^*) = \frac{p(\theta^*)Q(\theta^{(i-1)}|\theta^*)}{p(\theta^{(i-1)})Q(\theta^*|\theta^{(i-1)})} = \frac{p(\theta^*)}{p(\theta^{(i-1)})}.$$

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

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

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

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,

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

Una prueba de hipótesis interesante

Sería interesante saber si vale la pena emplear un ZIP o si podemos quedarnos con un modelo Poisson ordinario.
El contraste de hipótesis es de la forma

$$H_1 : \pi = 0 \text{ vs. } H_2 : \pi > 0$$

Suponiendo que tenemos una muestra (previa) $(\pi^{(1)}, \lambda^{(1)}), ..., (\pi^{(M)}, \lambda^{(M)})$ podemos aproximar el factor de Bayes

$$k = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\mathcal{L}(D|\pi = 0)}{\int_0^1 \mathcal{L}(D|\pi) f_\pi(\pi) d\pi} \approx \frac{\frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(D|\pi = 0, \lambda^{(i)})}{\frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(D|\pi^{(i)}, \lambda^{(i)})}$$

```
cat("model
    {# Especifica la estructura del modelo
    for(i in 1:N){
    X[i] ~ dpois(lambda)
    I[i] ~ dbern(1-pi)
    Y[i] ~ dnorm(X[i]*I[i],10000000)
    }# Define la distribucion inicial
    pi ~  dbeta(1,1)
    lambda ~ dgamma(3, 0.005)
    }", fill=TRUE, file="ZIP.jags")
ZIP.fit <- jags(data=Datos, inits=mezclas.inits,
                    parameters.to.save=mezclas.params,
                    n.chains=1, n.iter=10000,
                    n.burnin=0,n.thin=1,
                    DIC=F,model.file="ZIP.jags")
```

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)

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

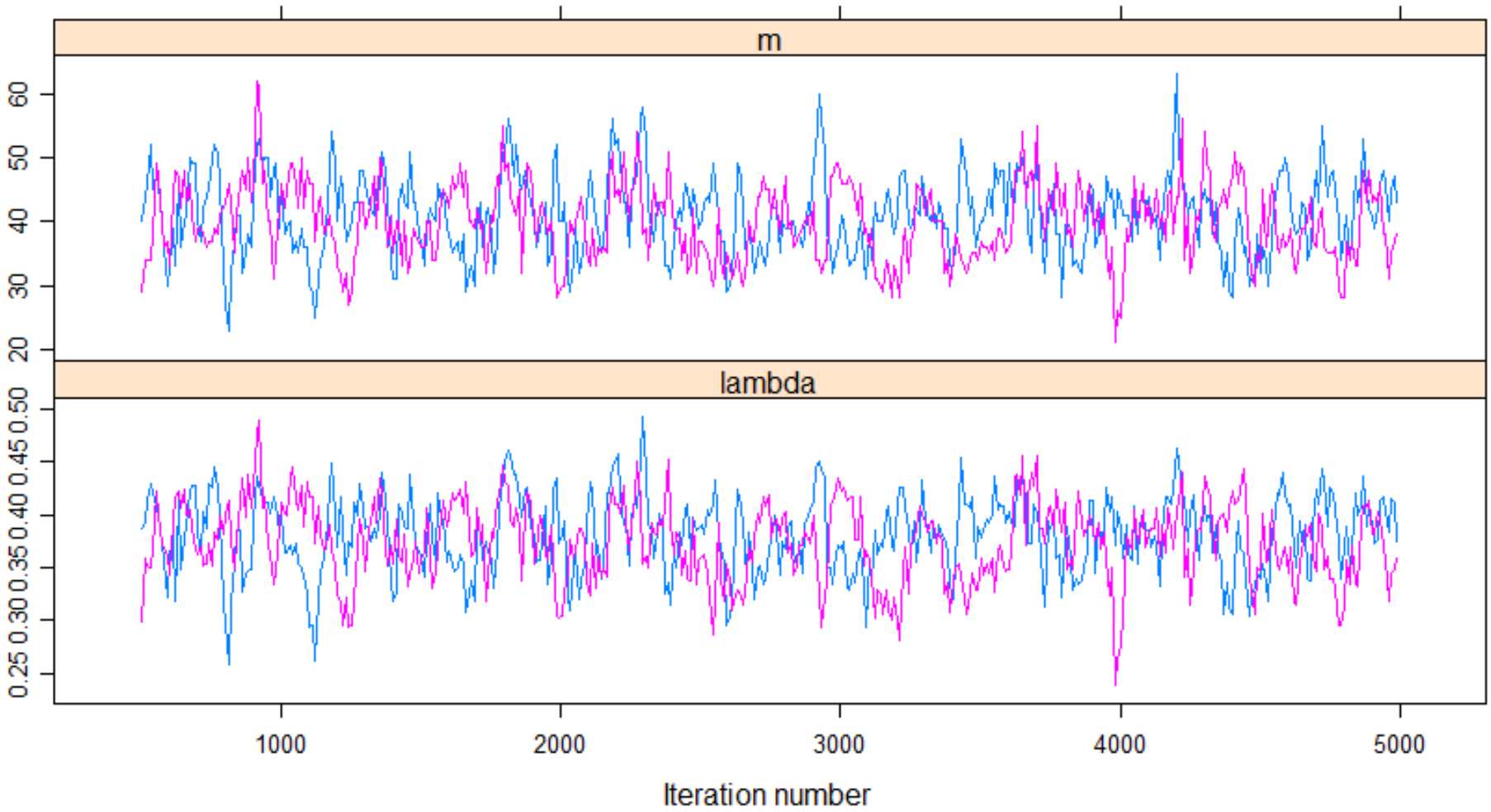
$$m \sim \text{Poisson}(\phi)$$

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

```
cat("model
    {
    # Especificamos la estructura del modelo
    for(i in 1:N){
    X[i] ~ dpois(theta[i])
    theta[i] ~ dgamma(m,((1-lambda)/lambda)**(-1))
    }
    #especificamos las distribuciones iniciales
    m ~dpois(40)
    lambda ~ dbeta(1,1)
    }",
    fill=TRUE, file="BinNegDC.jags")
```

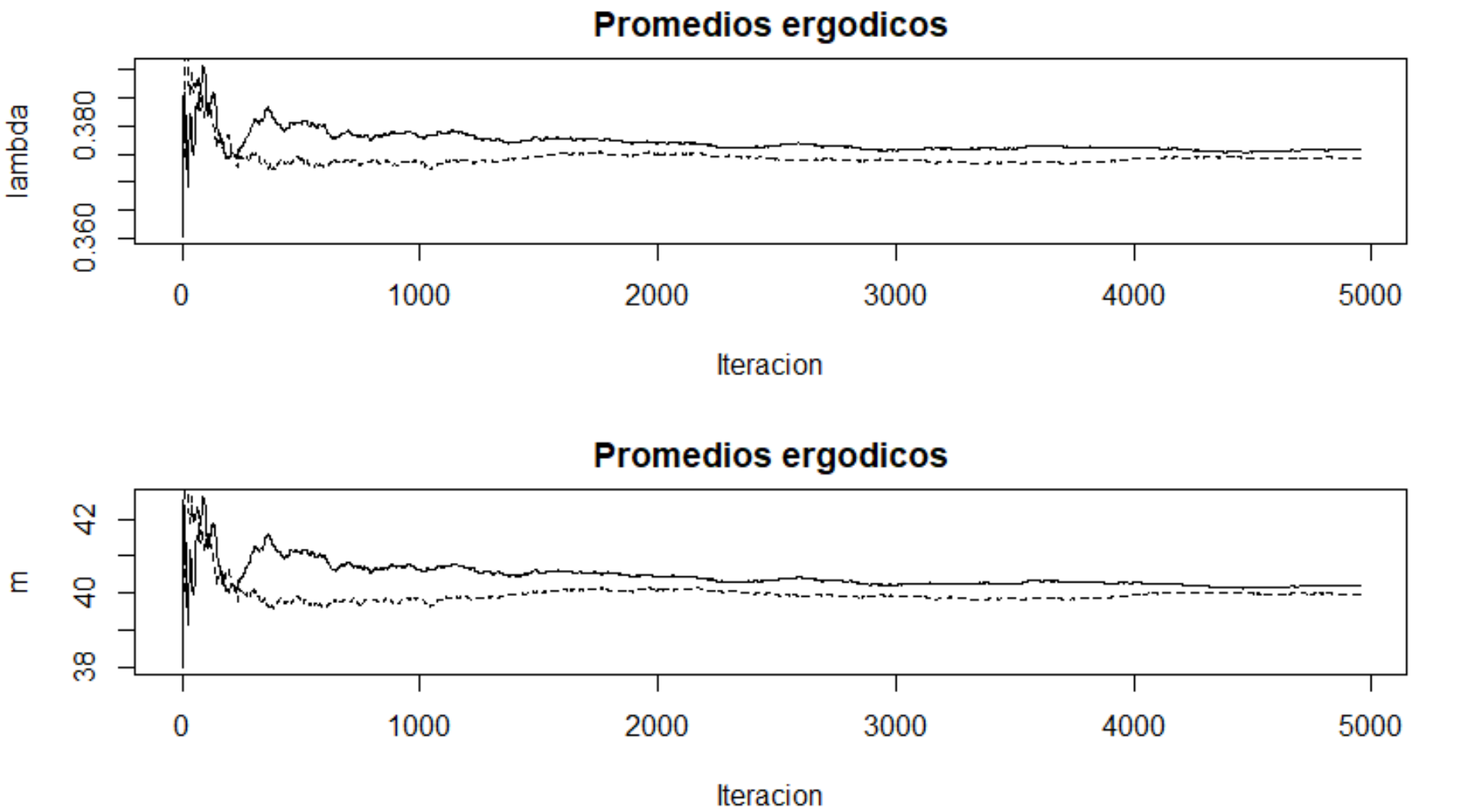
Gráficas

Trazas de m y λ

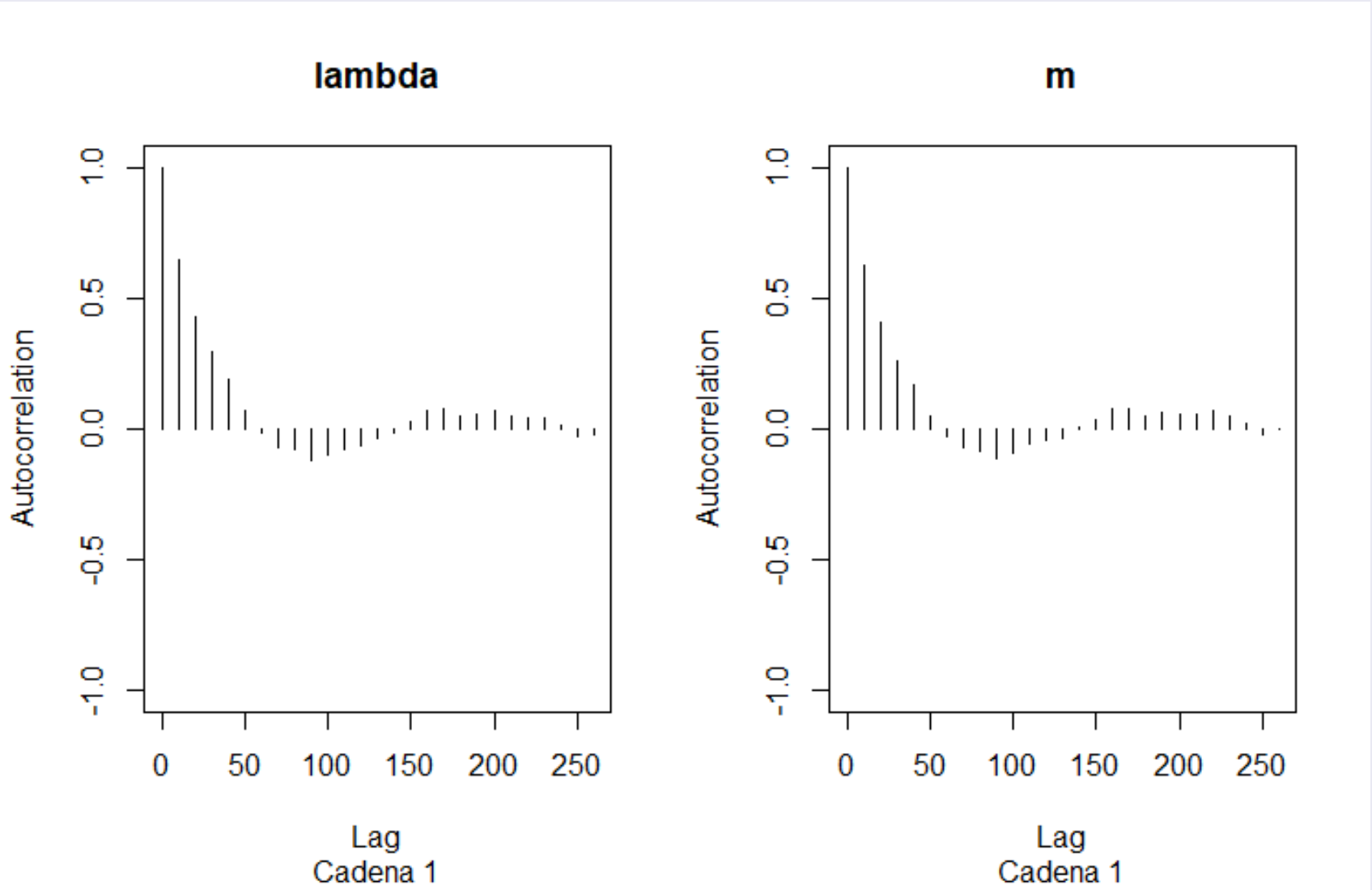


Gráficas

Promedios ergódicos de m y λ

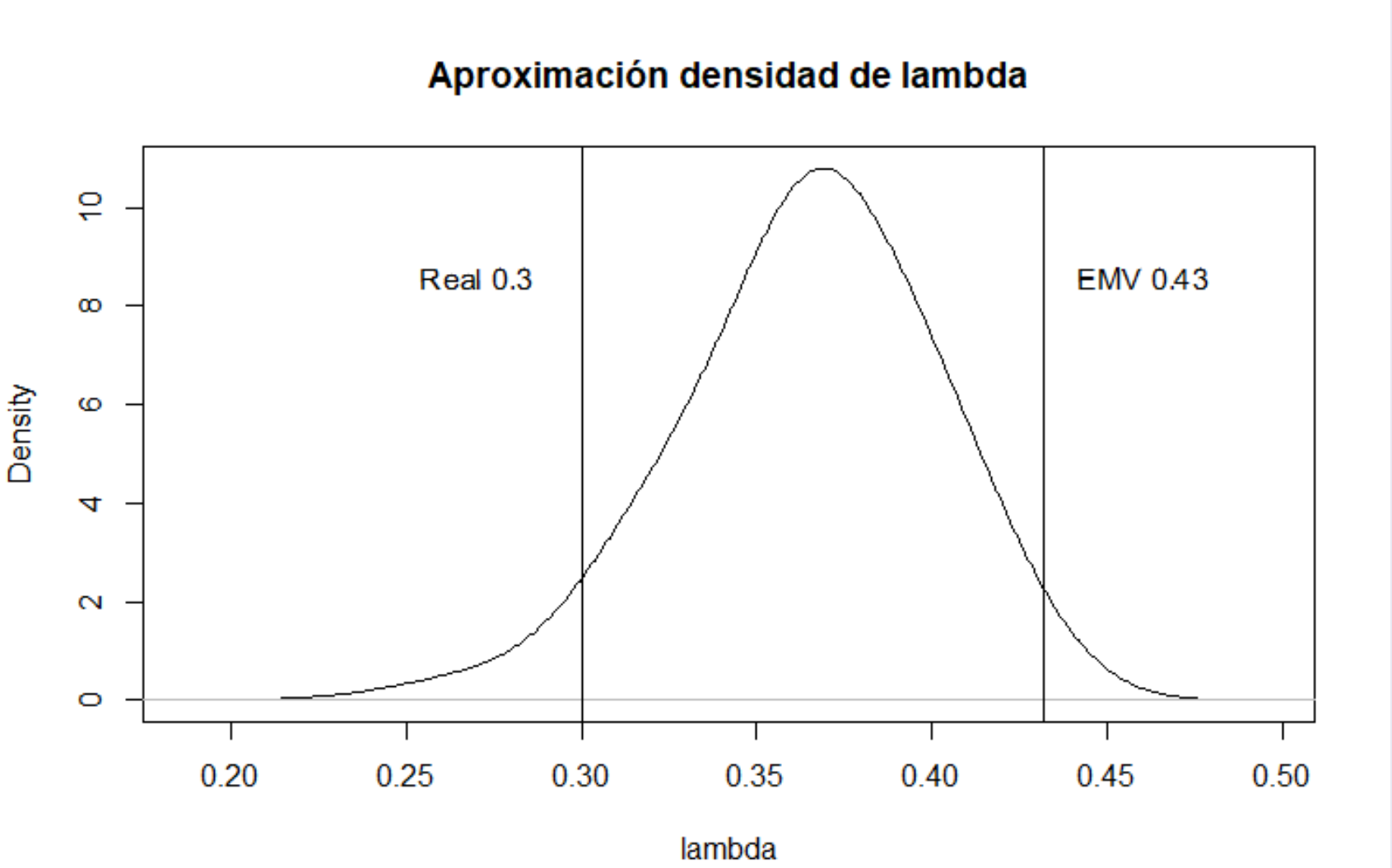


Gráfica de autocorrelación



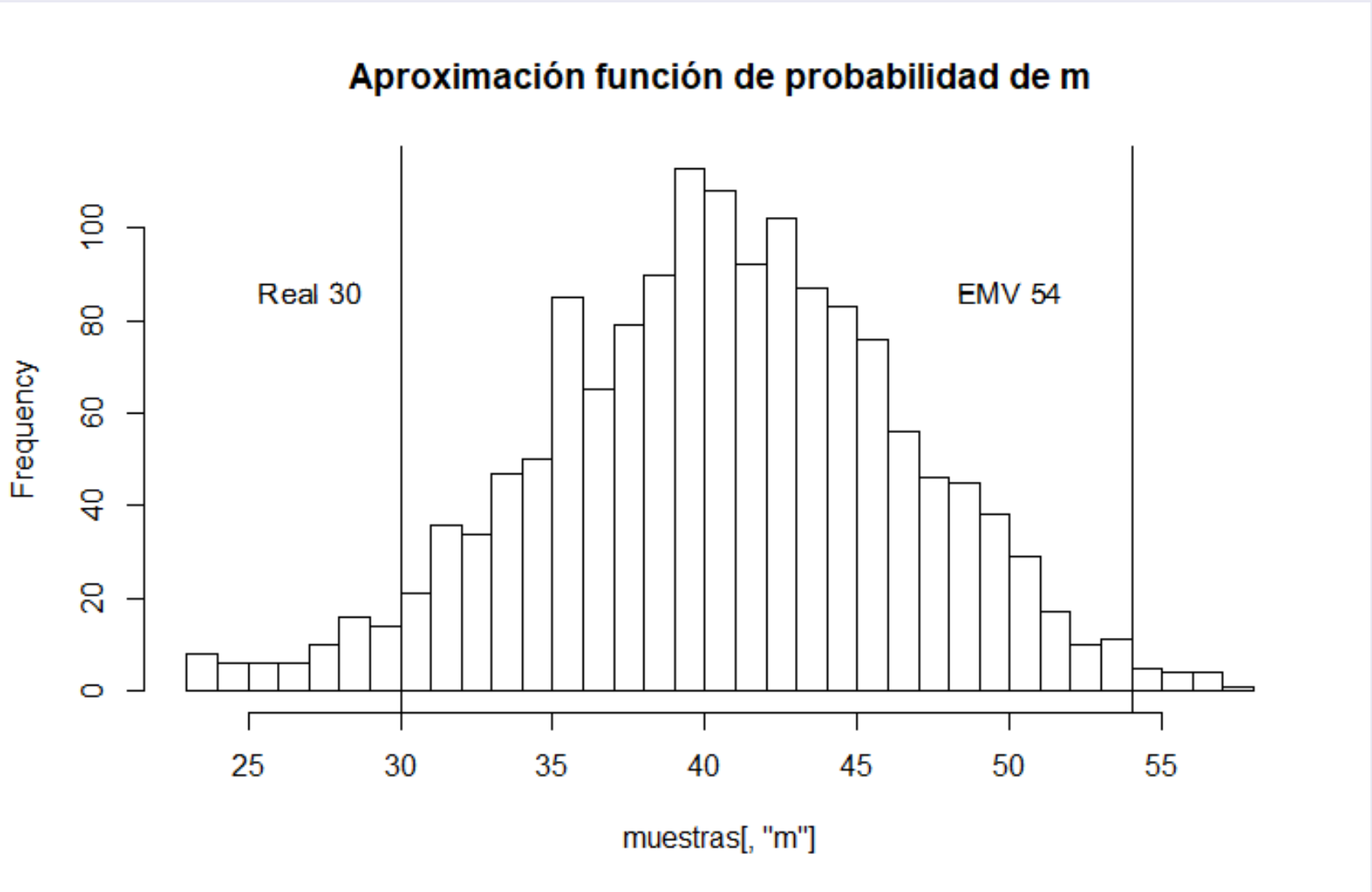
Gráficas

Para λ



Gráficas

Para m



```
         Mean      SD  Naive SE Time-series SE
lambda   0.36 0.03481 0.0007987       0.001585
m       40.24 5.86493 0.1345507       0.235272

2. Quantiles for each variable:

           2.5%     25%     50%     75%   97.5%
lambda   0.2895  0.3371  0.3603  0.3844  0.4252
m       29.0000 36.0000 40.0000 44.0000 52.0000
```

Sobre JAGS

¿Qué es JAGS (BUGS)?

- JAGS (Just Another Gibbs Sampler) es un programa para hacer análisis Bayesiano. Basta especificar el modelo.
- JAGS se encarga de implementar el muestreador de Gibbs.
- Genera automáticamente indicadores de convergencia e independencia.
- Calcula resúmenes numéricos y densidades aproximadas.
- Existen paquetes de R que llaman a JAGS desde R. Así podemos trabajar las muestras en R.
- Puede descargarse desde <http://mcmc-jags.sourceforge.net/>.

Conclusiones

- El marco Bayesiano permite tratar los problemas usuales de inferencia de forma unificada. A veces tiene más sentido plantear una función de perdida que un resumen de la posterior.
- Utilidad del análisis conjugado. Posibilita muchas cuentas incluyendo predictivas y factores de Bayes. Puede hacerse flexible con las mezclas.
- Las previas objetivas (análisis de referencia) puede utilizarse para comparar contra otras previas. Puede complicarse. Siempre tratar hacer, al menos, análisis de sensibilidad.
- El factor de Bayes es lo que usamos para comparar modelos. Su calculo pude ser complicado o inestable.

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 no requiere evaluar explícitamente la función de verosimilitud. Sin embrago, puede ser muy costoso computacionalmente. Una alternativa son las aproximaciones analíticas (aproximación de Laplace) 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

Centro de Investigación en Matemáticas, A.C.

10-12 de junio de 2019

mario.carranza@cimat.mx