EST-46115: Modelación Bayesiana

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Diagnósticos MCMC.
Objetivo. Que veremos.
Lectura recomendada: Referencia.

1. INTRODUCCIÓN

El avance en poder computacional ha permitido la proliferación de métodos Bayesianos. El poder generar cadenas de Markov es múltiples procesadores nos ayuda a relajar los requisitos y ayuda a explotar los recursos computacionales disponibles.

Cuando generamos una muestra de la distribución posterior usando MCMC, sin importar el método (Metrópolis, Gibbs, HMC), buscamos que:

1. Los valores simulados `no estén influenciados` por el valor inicial

(arbitrario) y deben explorar todo el rango de la posterior.

1. Debemos tener suficientes simulaciones de tal manera que las

estimaciones sean precisas y estables.

1. Queremos tener métodos y resúmenes informativos que nos ayuden diagnosticar

correctamente el desempeño de nuestas simulaciones.

En la `práctica` intentamos cumplir lo más posible estos objetivos. Debemos de tener un criterio para considerar cadenas de longitud finita y evaluar la calidad de las simulaciones.

Primero estudiaremos diagnósticos generales para métodos que utilicen MCMC y después estudiaremos particularidades del método de simulación HMC.

2. DIAGNÓSTICOS GENERALES

Una forma que tenemos de evaluar la (o identificar la falta de) convergencia es considerar distintas secuencias independientes.

2.1. Monitoreo de convergencia

`Burn-in e iteraciones iniciales`. En primer lugar, en muchas ocasiones las condiciones iniciales de las cadenas las escogemos de tal forma que que son .[a]típicos.[en] relación a la posterior.

> Estrategias de selección de puntos iniciales pueden ser valores aleatorios de la previa o perturbaciones aleatorias a estimadores MLE.

Correr varias cadenas en puntos dispersos tienen la ventaja de explorar desde distintas regiones de la posterior. Eventualmente, esperamos que todas las cadenas mezclen bien y representen realizaciones independientes del mismo proceso estócastico (Markoviano).

Para contrarrestar la dependencia en los distintos puntos iniciales se descarta parte de la cadena en un periodo inicial (periodo de calentamiento).

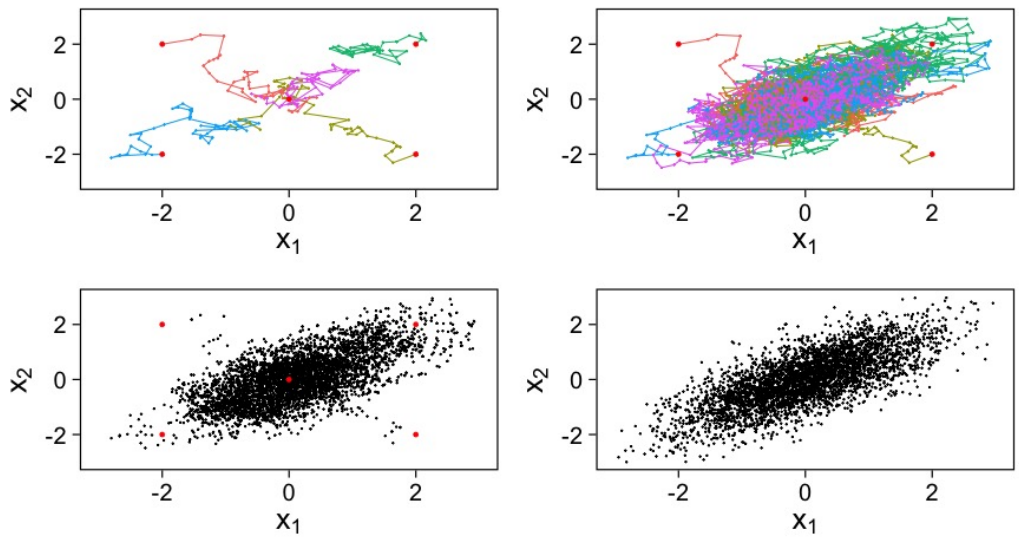

FIGURA 1. *Distintas cadenas de Markov.*

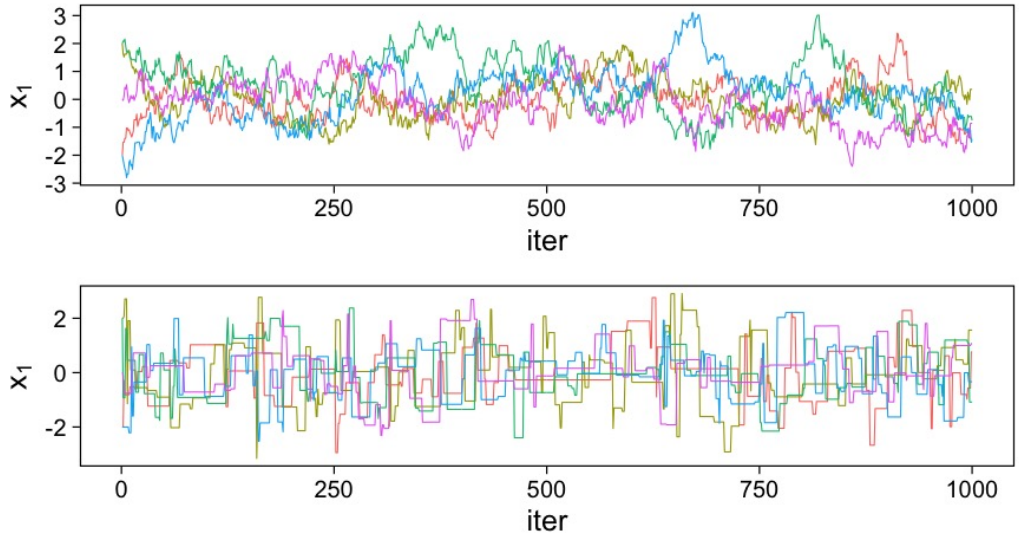

FIGURA 2. *Trayectorias de simulación para* X_1.

2.1.1. Datos: Denotamos por x_i la estatura de tenores (cantantas de ópera). Asumimos un modelo Normal con parámetros poblacionales no observados: μ y σ. El modelo previo lo asumimos como

$$\mu|\sigma \sim \text{Normal}\left(\mu_0, \frac{\sigma}{n_0}\right), \tag{1}$$

$$\sigma^{-1} \sim \text{Gamma}(a_0, b_0)\,. \tag{2}$$

```
## Datos: cantantes de opera -----------------------
set.seed(3413)
cantantes ← lattice::singer %>%
  mutate(estatura_cm = round(2.54 * height)) %>%
  filter(str_detect(voice.part, "Tenor")) %>%
  sample_n(20)
```

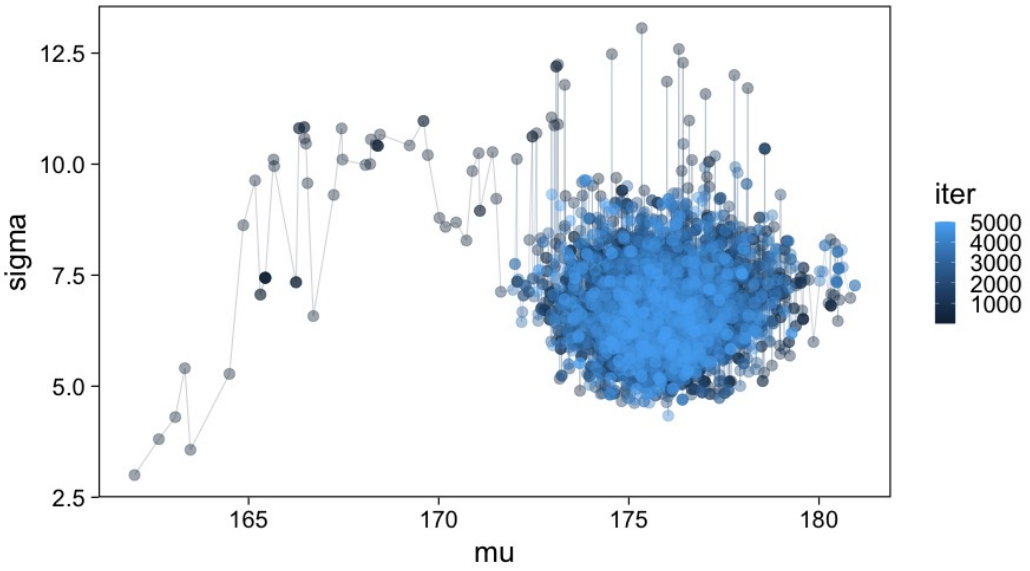

FIGURA 3. *Cadena de Markov con distribución objetivo la posterior.*

En esta simulación es evidente[†] que necesitamos descartar una parte inicial de la simulación.

Gelman et al. [1] recomiendan descartar la mitad de las iteraciones de cada una de las cadenas que se simularon. Para problemas en dimensiones altas, incluso se podría esperar descartar hasta un 80\Metropolis-Hastings).

2.2. Monitoreo de mezcla dentro y entre cadenas

Gelman y diversos de sus coatures han desarollado un diagnóstico numérico para evaluar implementaciones de MCMC al considerar múltiples cadenas. Aunque éste estadístico se ha ido refinando con los años, su desarrollo muestra un entendimiento gradual de éstos métodos en la práctica. La medida $\hat{R}$ se conoce como el **factor de reducción potencial de escala**.

El estadístico $\hat{R}$ pretende ser una estimación de la posible reducción en la longitud de un intervalo de confianza si las simulaciones continuaran infinitamente.

La $\hat{R}$ estudia de manera simultánea `la mezcla` de todas las cadenas (cada cadena, y fracciones de ella, deberían de haber transitado el soporte de la distribución objetivo) y

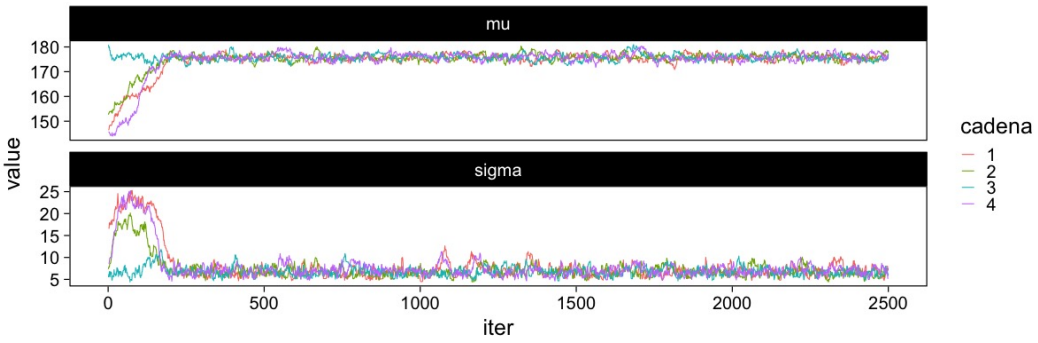

FIGURA 4. *Trayectorias con dependencias iniciales.*

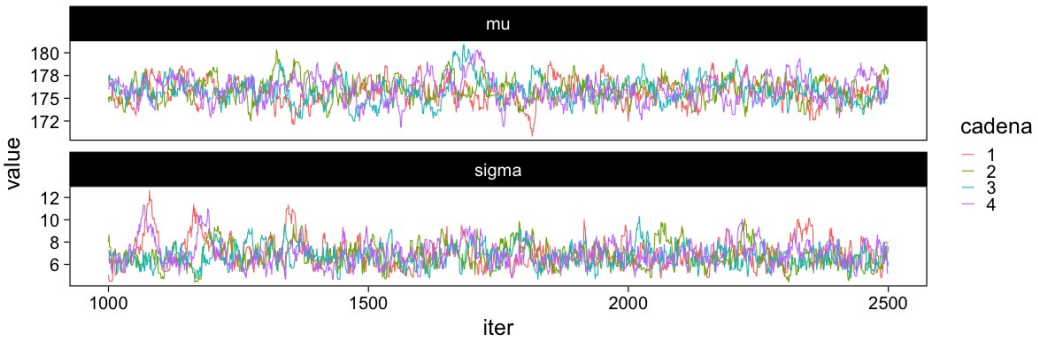

FIGURA 5. *Trayectorias estacionarias.*

`estacionalidad` (de haberse logrado cada mitad de una cadena deberían de poseer las mismas estadísticas).

La estrategia es descartar la `primera mitad` de cada cadena. El resto lo volvemos a dividir en dos y utilizamos cada fracción como si fuera una cadena independiente[†].

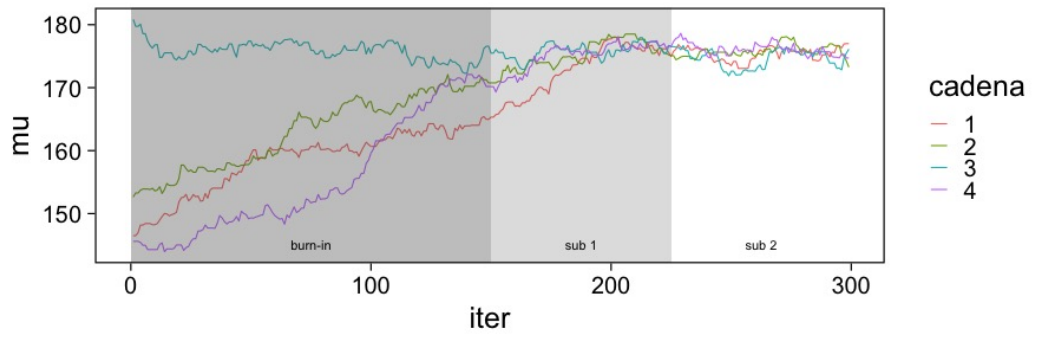

FIGURA 6. *Separación de simulaciones para cálculo de* $\hat{R}$.

Denotemos por m el número de cadenas simuladas y por n el número de simulaciones dentro de cada cadena. Cada una de las `cantidades escalares de interés` las denotamos por ϕ. Éstas pueden ser los parámetros originales θ o alguna otra cantidad derivada $\phi = f(\theta)$.

Ahora denotemos por ϕ_{ij} las simulaciones que tenemos disponibles con $i = 1, \ldots, n$, y $j = 1, \ldots, m$. Calculamos B y W, la variabilidad `entre` (*between*) y `dentro` (*within*) cadenas,

respectivamente, por medio de

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2, \quad \text{con} \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\phi_{ij} - \bar{\phi}_{\cdot j})^2, \quad \text{donde} \quad \bar{\phi}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\phi_{ij}, \tag{3a}$$

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\phi}_{\cdot j} - \bar{\phi}_{\cdot\cdot})^2, \quad \text{donde} \quad \bar{\phi}_{\cdot\cdot} = \frac{1}{m}\sum_{j=1}^{m}\bar{\phi}_{\cdot j}. \tag{3b}$$

La varianza entre cadenas, B, se multiplica por n dado que ésta se calcula por medio de promedios y sin este factor de corrección no reflejaría la variabilidad de las cantidades de interés ϕ.

La varianza de ϕ se puede estimar por medio de

$$\hat{\mathbb{V}}(\phi)^+ = \frac{n-1}{n}W + \frac{1}{n}B\,. \tag{4}$$

Este estimador `sobre-estima` la varianza pues los puntos iniciales pueden estar sobre-dispersos, mientras que es un `estimador insesgado` una vez que se haya alcanzado el estado estacionario (realizaciones de la distribución objetivo)

Por otro lado, la varianza estimada por W será un sub-estimador pues podría ser el caso de que cada cadena no ha tenido la oportunidad de recorrer todo el soporte de la distribución. En el límite $n \to \infty$, el valor esperado de W aproxima $\mathbb{V}(\phi)$.

Se utiliza como diagnostico el factor por el cual la escala de la distribución actual de ϕ se puede reducir si se continua con el procedimiento en el límite $n \to \infty$. Esto es,

$$\hat{R} = \sqrt{\frac{\hat{\mathbb{V}}(\phi)^+}{W}}\,,$$

por construcción converge a 1 cuando $n \to \infty$.

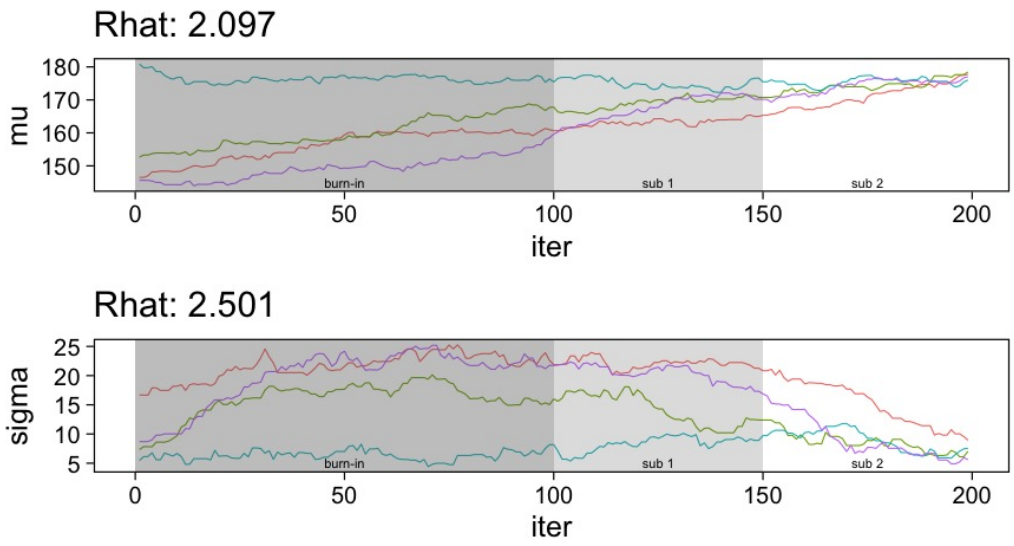

FIGURA 7. *Diágnostico de reducción de escala. Sugerencia: generar mas simulaciones.*

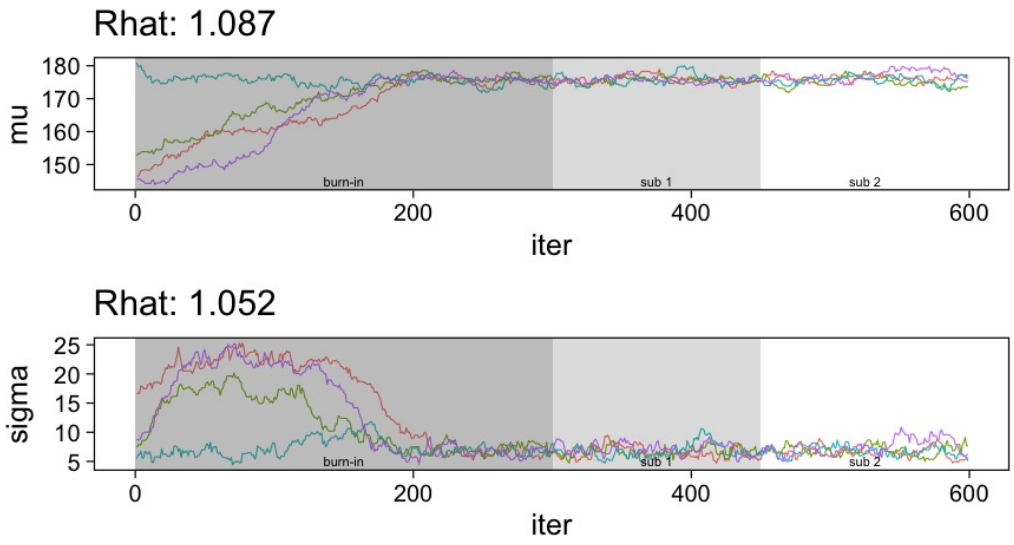

FIGURA 8. *Diágnostico de reducción de escala. Observaciones: parece estar bien.*

Problemas con $\hat{R}$. El estimador de reducción de escala funciona bien para monitorear estimadores y cantidades de interés basados en medias y varianzas, o bien, cuando la distribución es simétrica y cercana a una Gaussiana. Es decir, colas ligeras. Sin embargo, para percentiles, o distribuciones lejos del supuesto de normalidad no es un buen indicador. Es por esto que también se recomienda incorprorar transformaciones que nos permitan generar un buen estimador. Puedes leer mas de esto aqui: [2].

REFERENCIAS

[1] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*, volume 2. CRC press Boca Raton, FL, 2014. 3

[2] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-Normalization, Folding, and Localization: An Improved R^ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2), jun 2021. ISSN 1936-0975. . 6