EST-46115: Modelación Bayesiana

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Diagnósticos MCMC.
Objetivo. Estudiaremos diagnósticos típicos de métodos de simulación Markoviana (por Metropolis-Hastings o HMC) con el objetivo de poder identificar problemas y posibles soluciones para simulaciones deficientes.
Lectura recomendada: Para una revisión de los diagnósticos de MCMC busca la sección 11.4 de [3]. El Capítulo 3 de [1] tiene una revisión muy bonita de series de tiempo que es útil para algunos términos que estudiaremos. El artículo de Roy [5] tiene un breve resumen sobre diagnósticos. Además, se mencionan algunas técnicas modernas de diagnóstico como los métodos basados en los principios de Stein o distancias adecuadas para espacios de probabilidad.

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 de convergencia 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:

- Los valores simulados `no estén influenciados` por el valor inicial (arbitrario) y deben explorar todo el rango de la posterior.
- Debemos tener `suficientes simulaciones` de tal manera que las estimaciones sean precisas y estables.
- 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 *mencionaremos* particularidades del método de simulación HMC.

En general, el problema es doble:

1. Determinar si la cadena de Markov ha alcanzado el estado estacionario;
2. Determinar si los estimadores Monte Carlo convergen a los valores esperados.

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.

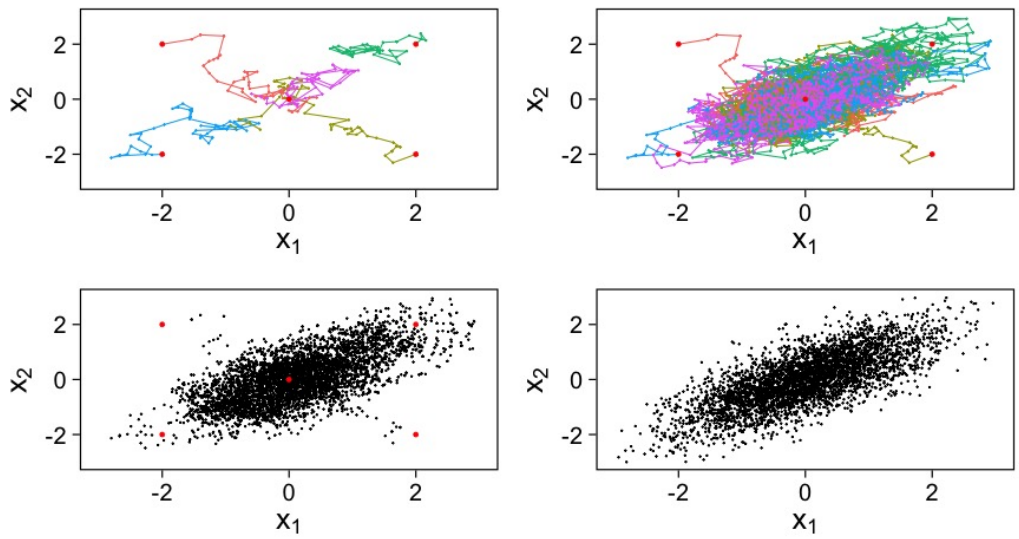


FIGURA 1. *Distintas cadenas de Markov.*

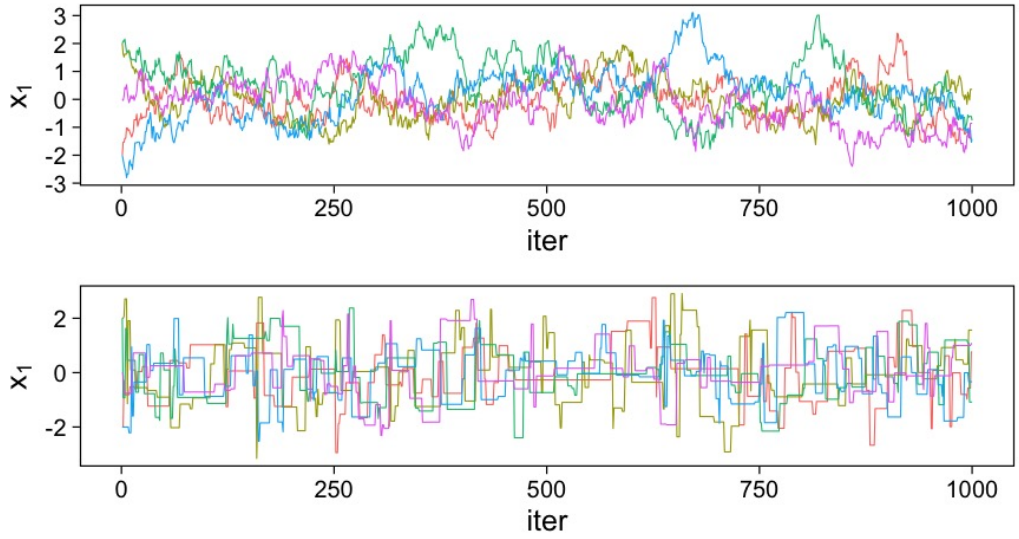


FIGURA 2. *Trayectorias de simulación para* X_1.

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).

2.1.1. Datos: Cantantes de ópera

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

```
# A tibble: 20 × 2
  voice.part estatura_cm
  <fct>            <dbl>
1 Tenor 1            178
2 Tenor 2            173
3 Tenor 1            165
# ... with 17 more rows
#  Use ‘print(n = ...)‘ to see more rows
```

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

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

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

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

Gelman et al. [3] 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 % de simulaciones (en especial para métodos basados en Metropolis-Hastings).

2.2. Monitoreo de mezcla dentro y entre cadenas

Podemos utilizar `todas` las simulaciones como si vinieran de una sola cadena (argumentando por estacionariedad)

```
cadenas.cantantes ▷
  unnest(cadenas) ▷
  filter(iter > 100) ▷
  summarise(.estimate = mean(sigma), .variance = var(sigma), .error_mc = sqrt
      (.variance/n()))
```

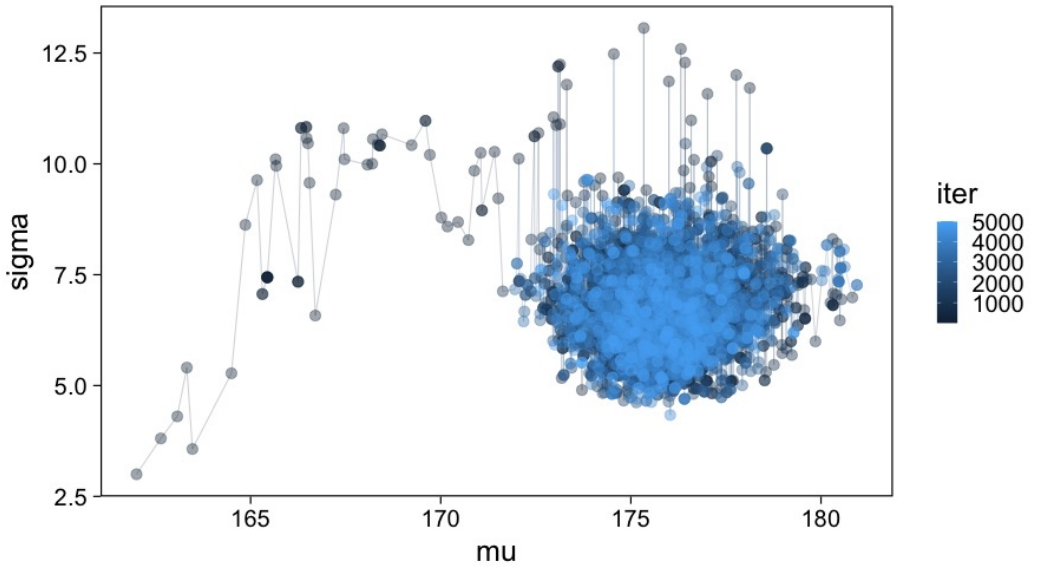


FIGURA 3. *Cadena de Markov simulando de la posterior como distribución objetivo.*

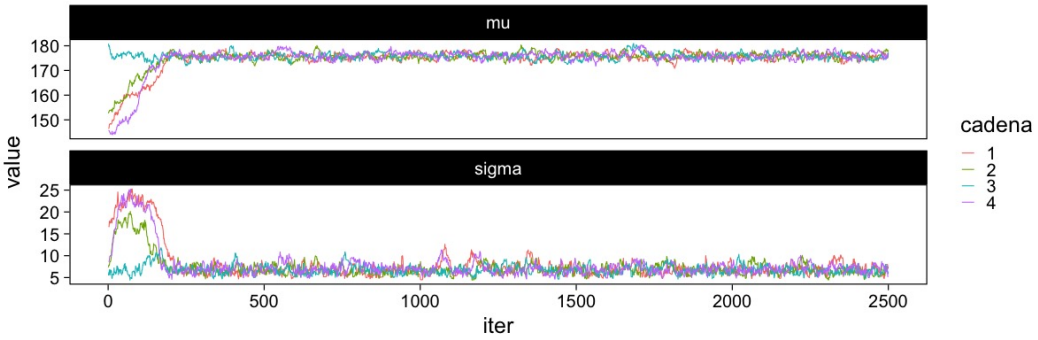


FIGURA 4. *Trayectorias con dependencias iniciales.*

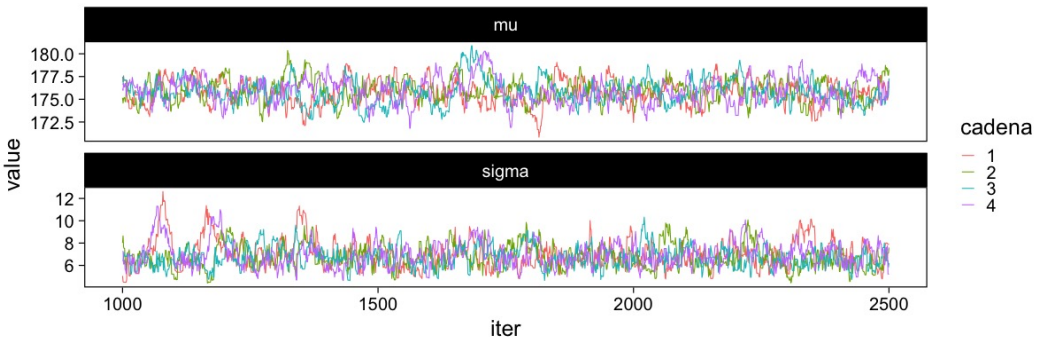


FIGURA 5. *Trayectorias estacionarias.*

```
# A tibble: 1 × 3
  .estimate .variance .error_mc
      <dbl>     <dbl>     <dbl>
1      7.11      4.14    0.0208
```

Sin embargo, al calcular la varianza como si fueran 4 cadenas independientes vemos que nuestro estimador del error Monte Carlo es mucho mas elevado de lo que esperamos ¿por qué?

```
cadenas.cantantes ▷
 unnest(cadenas) ▷
 filter(iter > 100) ▷
 group_by(cadena) ▷
 summarise(media = mean(sigma), varianza = var(sigma)) ▷
 summarise(.estimate = mean(media), .error_mc = sd(media))
```

```
# A tibble: 1 × 2
  .estimate .error_mc
      <dbl>     <dbl>
1      7.11     0.272
```

Al inspeccionar cada cadena tenemos los siguientes resúmenes

```
cadenas.cantantes ▷
 unnest(cadenas) ▷
 filter(iter > 100) ▷
 group_by(cadena) ▷
 summarise(media = mean(sigma), varianza = var(sigma))
```

```
# A tibble: 4 × 3
  cadena media varianza
  <fct>  <dbl>    <dbl>
1 1       7.34     7.74
2 2       6.93     2.35
3 3       6.82     1.13
4 4       7.35     5.12
```

Podemos partir cada cadena a la mitad y calcular nuestra estimación del error Monte Carlo. Ahora tenemos 8 cadenas que *esperamos* sean `estacionarias` (*idénticamente distribuidas*).

```
# A tibble: 1 × 2
  .estimate .error_mc
      <dbl>     <dbl>
1      7.11     0.418
```

Nota cómo está sucediendo algo contraintuitivo. Tenemos mas observaciones (pasamos de 1 cadena a 8) y el error Monte Carlo no decrece. Lo cual indica que nuestras cadenas realmente no han terminado de converger y tienen comportamiento distinto aunque en promedio parecen estar cercanas.

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. Recuerda que la varianza de un estimador nos ayuda a construir intervalos en el sentido frecuentista.

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 `estacionariedad` (de haberse logrado cada mitad de una cadena deberían de poseer las mismas característica 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[†].

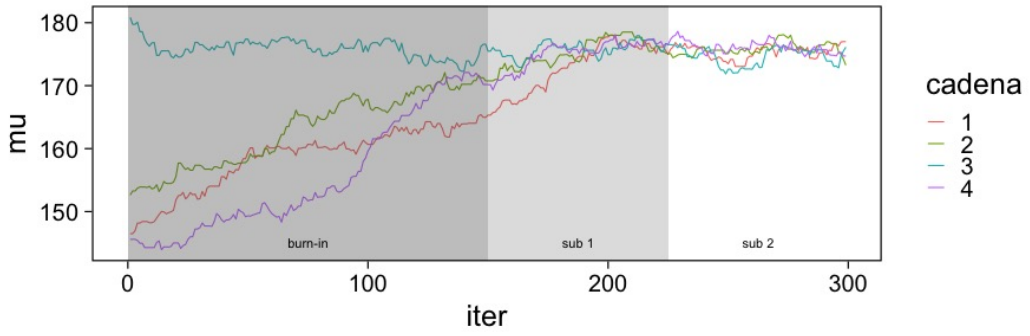


FIGURA 6. *Separación de simulaciones para cálculo de* $\hat{R}$.

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 $\phi = f(\theta)$.

> Ejemplos de esto puede ser en un modelo Beta-Binomial donde nos interesa la tasa de éxitos θ pero necesitamos monitorear $\phi = \log(\theta/1-\theta)$. Otra situación puede ser el caso de un modelo normal con varianza desconocida σ^2 y necesitamos monitorear $\phi = \log \sigma^2$.

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

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

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\phi}_{\cdot j} - \bar{\phi}_{\cdot\cdot})^2, \quad \text{donde} \quad \bar{\phi}_{\cdot\cdot} = \frac{1}{m}\sum_{j=1}^{m}\bar{\phi}_{\cdot j}. \tag{4}$$

> 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 del `estimador agregado de varianza`

$$\hat{\mathbb{V}}(\phi)^+ = \frac{n-1}{n}W + \frac{1}{n}B\,. \tag{5}$$

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 $n \to \infty$.

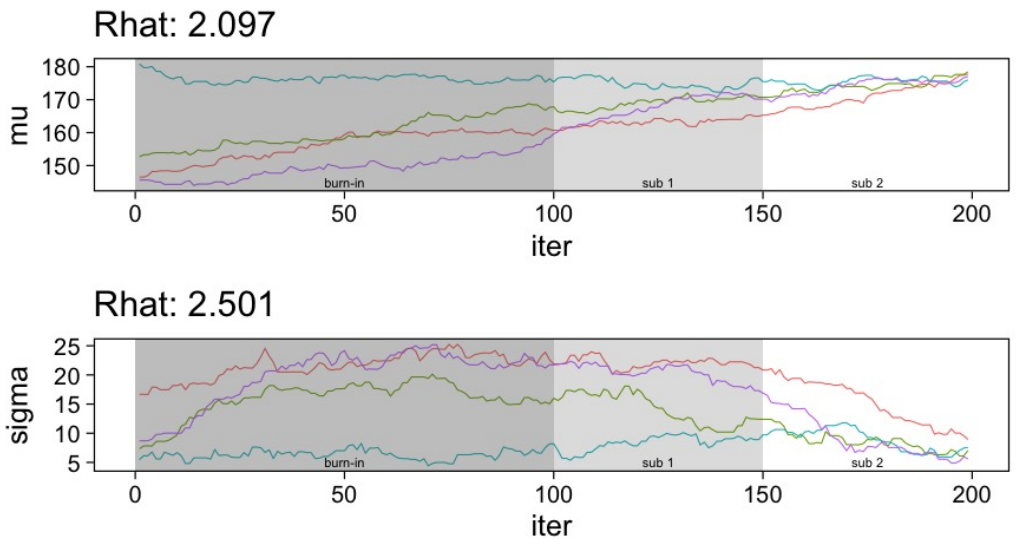


FIGURA 7. *Diágnostico de reducción de escala. Sugerencia: generar mas simulaciones.*

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 en el articulo de Vehtari et al. [6].

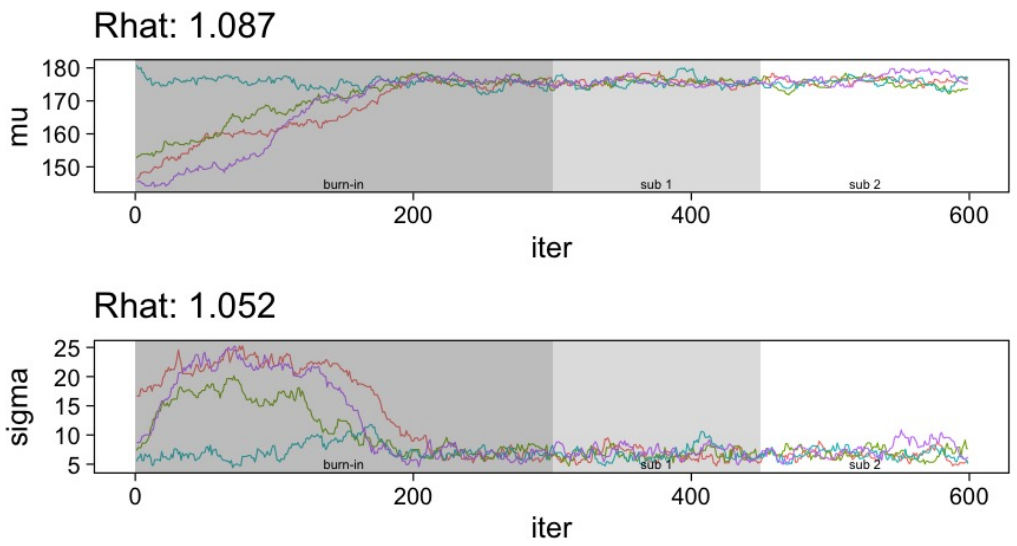


FIGURA 8. *Diágnostico de reducción de escala. Observaciones: parece estar bien.*

Existe una versión multivariada de este estadístico que busca resumir la misma información a lo largo de todos los componentes de $\theta \in \mathbb{R}^p$. Es prácticamente el mismo estimador con las versiones multivariadas de varianza dentro de cada cadena Σ_W y la varianza agregada Σ_{V^+}. Es decir, el estadístico se calcula como

$$\hat{R}_p = \max_{a \in \mathbb{R}^p} \frac{a^\top \Sigma_{V^+} a}{a^\top \Sigma_W a}\,, \tag{6}$$

el cual, se puede probar, está relacionado con el valor propio mas grande de la matriz $\Sigma_W^{-1}\Sigma_{V^+}/n$ [5].

2.3. Número efectivo de simulaciones

Queremos que los recursos que hemos asignado a generar simulaciones sean representativos de la distribución objetivo. Si las n simulaciones dentro de cada cadena en verdad son realizaciones independientes entonces la estimación de B sería un estimador insesgado de $\mathbb{V}(\phi)$.

En esta situación tendríamos $n \times m$ realizaciones de la distribución que queremos simular. Sin embargo, la correlación entre las muestras hacen que B sea mayor que $\mathbb{V}(\phi)$ en promedio.

Una manera para definir el tamaño efectivo de simulaciones es por medio del estudio del estimador

$$\bar{\phi}_{\cdot\cdot} \approx \mathbb{E}(\phi)\,. \tag{7}$$

Del cual podemos derivar que

$$\mathbb{V}(\bar{\phi}_{\cdot\cdot}) = \frac{\mathbb{V}(\phi)}{m \cdot n}\,.$$

El problema es que la correlación en las cadenas implica el denominador $(m \cdot n)$ realmente sea una fracción del total de muestras, digamos λ. De tal forma que el número efectivo de

simulaciones es

$$N_{\text{eff}} = \lambda \cdot (m\, n)\,,$$

donde

$$\lambda = \frac{1}{\sum_{t=-\infty}^{\infty} \rho_t} = \frac{1}{1 + 2\sum_{t=1}^{\infty} \rho_t}\,.$$

El término ρ_t denota la **auto-correlación** con diferencia en t unidades de tiempo.

`Definición (autocorrelación)`: La autocovarianza y autocorrelación de una serie temporal **estacionaria** $\{Y_t : t = 0, \ldots\}$ están definidas (respectivamente) como

$$C_\tau = \mathbb{E}[(Y_{t+\tau} - \mu)(Y_t - \mu)], \qquad \rho_\tau = \frac{\mathbb{E}[(Y_{t+\tau} - \mu)(Y_t - \mu)]}{\sigma^2}\,. \tag{8}$$

`Definición (estimador de autocorrelación)`: La función de autocorrelación se estima utilizando

$$\hat{C}_\tau = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau}(Y_{t+\tau} - \hat{\mu})(Y_t - \hat{\mu}), \qquad \hat{\rho}_\tau = \frac{\hat{C}_\tau}{\hat{C}_0}\,. \tag{9}$$

`Definición (variograma)`: El variograma de una serie temporal **estacionaria** $\{Y_t : t = 0, \ldots\}$ está definido como

$$V_\tau = \mathbb{E}[(Y_{t+\tau} - Y_t)^2]\,. \tag{10}$$

Nota que $V_\tau = C_0 - C_\tau$.

Regresando a nuestro contexto... para estimar ρ_t partimos de nuestro estimador $\hat{\mathbb{V}}(\phi)^+$; y utilizamos el **variograma** V_t para `cada retraso` t

$$V_t = \frac{1}{m(n-t)}\sum_{j=1}^{m}\sum_{i=t+1}^{n}(\phi_{i,j} - \phi_{i-t,j})^2\,.$$

Utilizando la igualdad $\mathbb{E}(\phi_i - \phi_{i-t})^2 = 2(1-\rho_t)\mathbb{V}(\phi)$, podemos estimar

$$\hat{\rho}_t = 1 - \frac{V_t}{2\,\hat{\mathbb{V}}(\phi)^+}\,.$$

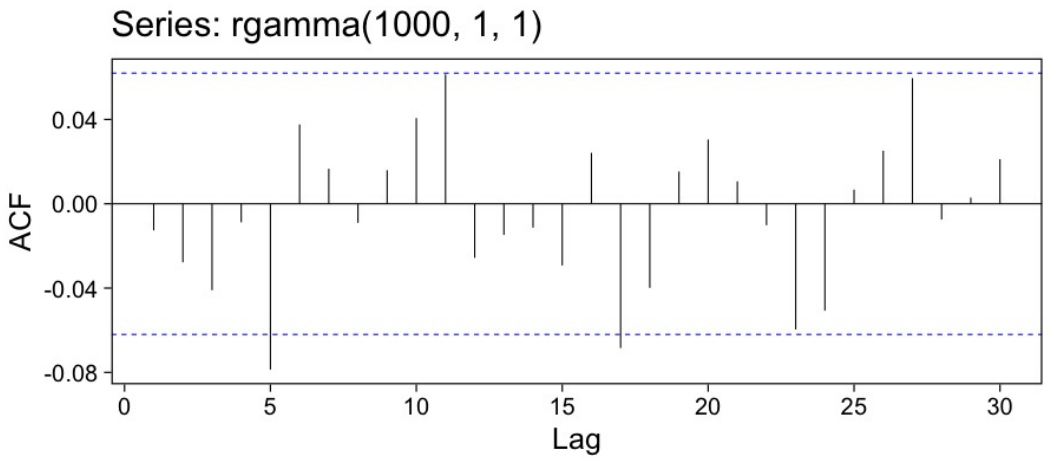
> La mayor dificultad que presenta el estimador es considerar **todos** los retrasos posibles. Eventualmente agotaremos la longitud de las cadenas para ello. Por otro lado, para t eventualmente grande nuestros estimadores del variograma V_t serán muy ruidosos (¿por qué?). En la práctica truncamos la serie de acuerdo a las observaciones [4]. La serie tiene la propiedad de que para cada par $\rho_{2t} + \rho_{2t+1} > 0$. Por lo tanto, la serie se trunca cuando observamos $\hat{\rho}_{2t} + \hat{\rho}_{2t+1} < 0$ para dos retrasos sucesivos.

Si denotamos por T el **tiempo de paro** (el máximo número de rezagos que podemos considerar), el estimador para el número efectivo de simulaciones es

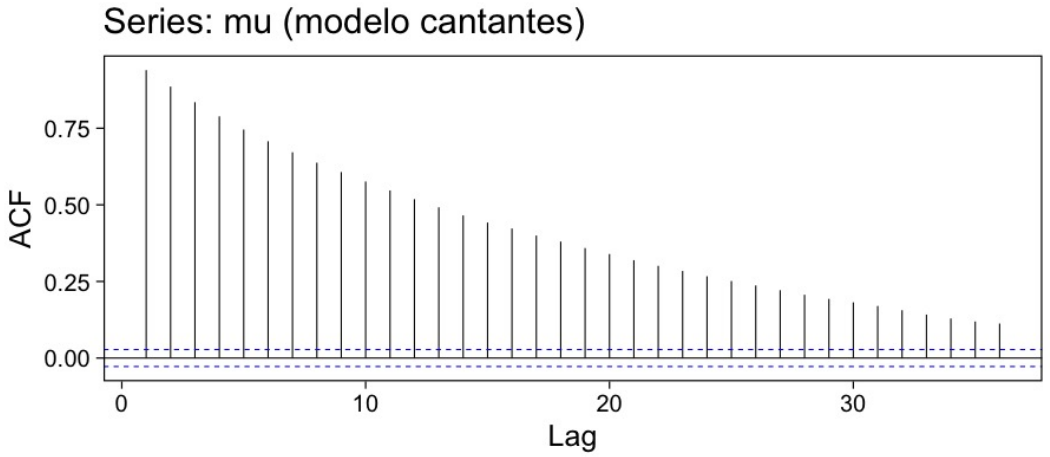
$$\widehat{\text{ESS}} = \frac{m\, n}{1 + 2\sum_{t=1}^{T}\hat{\rho}_t}\,.$$

El `tamaño efectivo de simulaciones` nos ayuda a monitorear lo siguiente. Si las simulaciones fueran independientes N_{eff} sería el número total de simulaciones; sin embargo, las simulaciones de MCMC suelen estar correlacionadas, de modo que cada iteración de MCMC es menos informativa que si fueran independientes.

Por ejemplo si graficaramos simulaciones independientes, esperaríamos valores de autocorrelación chicos:



Sin embargo, los valores que simulamos tienen el siguiente perfil de autocorrelación:




```
     mu    sigma   accept
0.02921 0.04822 0.69000
```

Listing 1. *Fracción* ESS/nm *y tasas de aceptación para la simulación de la posterior los cantantes de ópera.*

```
     mu    sigma   accept
0.08824 0.14556 0.38020
```

Listing 2. *Fracción* ESS/nm *y tasas de aceptación para la simulación (calibrada) de la posterior los cantantes de ópera.*

2.4. Relación con error Monte Carlo

Conocer el número efectivo de simulaciones nos permite calcular el error estándar de una aproximación Monte Carlo por medio de expresiones como

$$\mathsf{SE}(\pi(f)) \approx \left(\frac{\hat{\mathbb{V}}_\pi(f)}{\widehat{\mathsf{ESS}}} \right)^{\frac{1}{2}}. \tag{11}$$

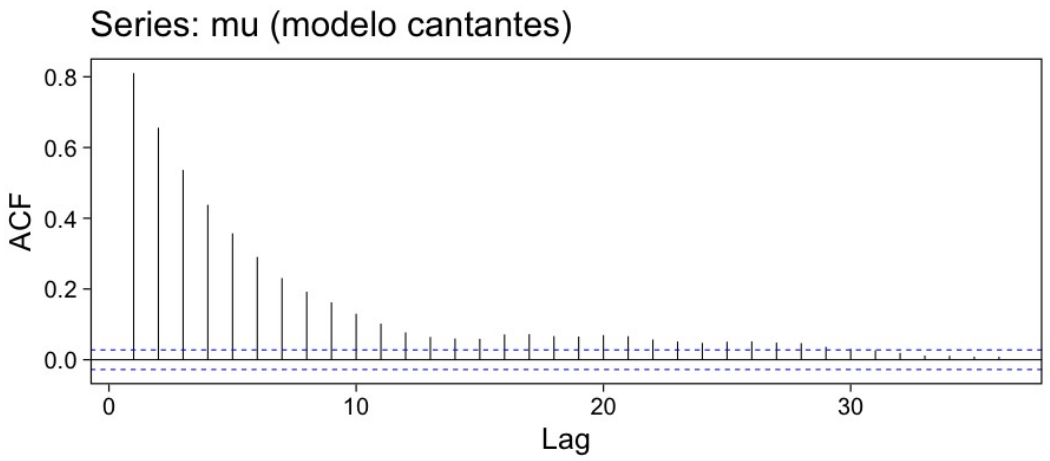


FIGURA 9. *Perfil de correlación para la simulación calibrada.*

Geyer [4] menciona que realizar estimaciones puntuales sin medidas de incertidumbre es el *deber* ser de un analista/estadístico. Esto es, independientemente de si la técnica es Bayesiana o frecuentista. El artículo de Flegal et al. [2] también discute la importancia de reportar errores estándar.

2.5. Adelgazamiento de cadenas

El método de simulación por medio de cadenas de Markov es computacionalmente intensivo. Sin embargo, los estimadores de varianza pueden ser muy volátiles (debido a altas correlaciones entre muestras). Adelgazar la cadena tiene como objetivo buscar quedarse con muestras de `alta calidad` para realizar estimaciones Monte Carlo. Se puede utilizar el ESS como una noción de cuántas muestras preservar.

3. CONCLUSIONES

- Ambos estadísticos asumen la existencia de un teorema de límite central Markoviano.
- En la práctica, las condiciones teóricas para garantizar su existencia (CLT) se necesitan probar caso a caso. Sin embargo, no es común que estos no existan bajo un mecanismo de muestreo tipo Metropolis-Hastings.
- Existen otras alternativas para diagnosticar un buen comportamiento de la cadena de Markov. Por ejemplo, se pueden utilizar pruebas de hipótesis para diferencias en medias con dos pedazos de cadenas.
- Para casos multivariados se pueden ajustar los análisis univariados por medio de pruebas de hipótesis múltiples con sus ajustes correspondientes (tipo Bonferroni, por nombrar uno).

REFERENCIAS

[1] N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2015. 1

[2] J. M. Flegal, M. Haran, and G. L. Jones. Markov chain monte carlo: Can we trust the third significant figure? *Statistical Science*, pages 250–260, 2008. 11

[3] 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. 1, 3

[4] C. J. Geyer. Introduction to Markov Chain Monte Carlo. *Handbook of Markov Chain Monte Carlo*, pages 3–48, 2002. 9, 11

[5] V. Roy. Convergence diagnostics for Markov chain Monte Carlo, oct 2019. 1, 8

[6] 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. . 7