

EST-24107: Simulación

Profesor: Alfredo Garbuno Iñigo — Otoño, 2022 — Diagnósticos MCMC.

Objetivo. Estudiaremos diagnósticos típicos de métodos de simulación Markoviana 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 [6]. El Capítulo 3 de [3] tiene una revisión muy bonita de series de tiempo que es útil para algunos términos.

1. INTRODUCCIÓN

El avance en poder computacional ha permitido la proliferación de métodos Bayesianos. El poder generar cadenas de Markov en 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:

- 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 nuestras 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.

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 *“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 estocástico (Markoviano).

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

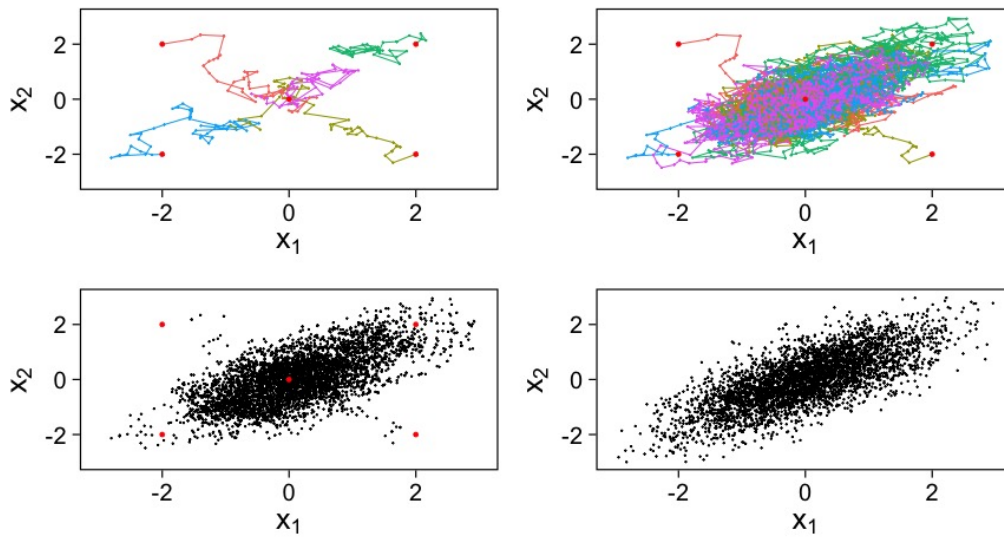
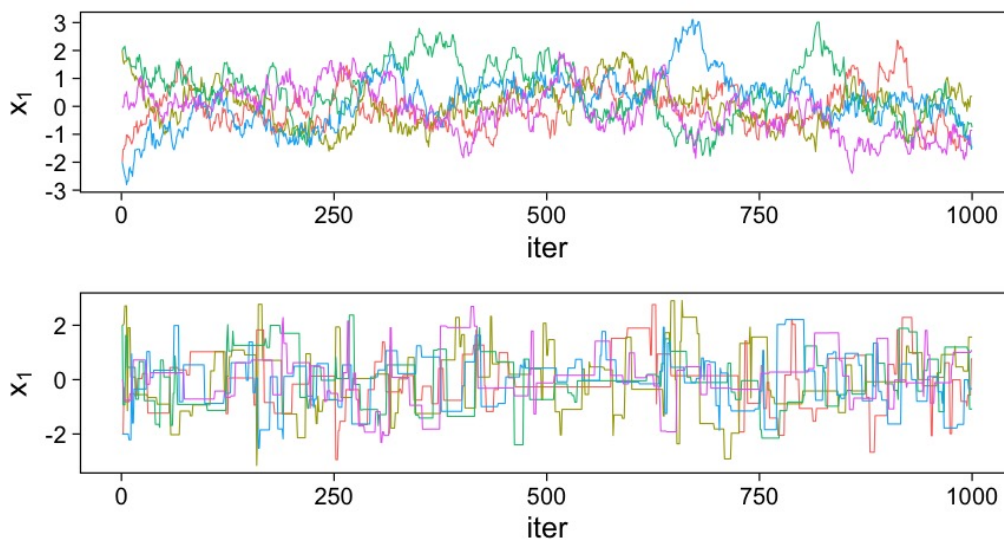


FIGURA 1. Distintas cadenas de Markov.

FIGURA 2. Trayectorias de simulación para X_1 .

2.1.1. Datos: Cantantes de ópera

```

1 ## Datos: cantantes de ópera -----
2 set.seed(3413)
3 cantantes <- lattice::singer >
4   mutate(estatura_cm = round(2.54 * height)) >
5   filter(str_detect(voice.part, "Tenor")) >
6   select(voice.part, estatura_cm) >
7   sample_n(20)

```

```

1   voice.part estatura_cm
2 139      Tenor 1      178
3 150      Tenor 2      173
4 140      Tenor 1      165

```

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

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

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

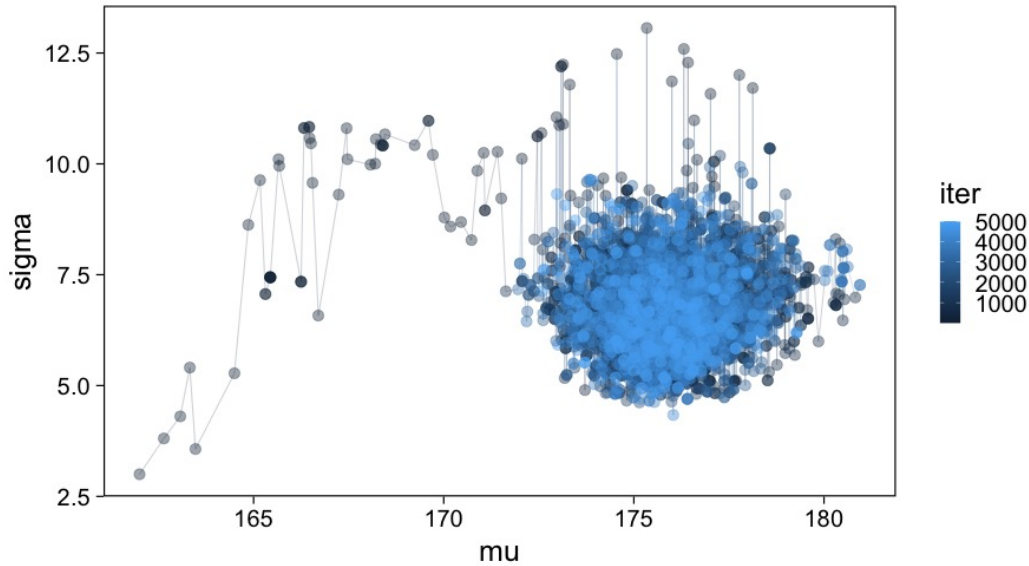


FIGURA 3. Cadena de Markov simulando de la posterior como distribución objetivo.

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

Gelman et al. [6] 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

Gelman y diversos de sus coautores han desarrollado un diagnóstico numérico para evaluar implementaciones de MCMC al considerar múltiples cadenas. Aunque éste estadístico se ha

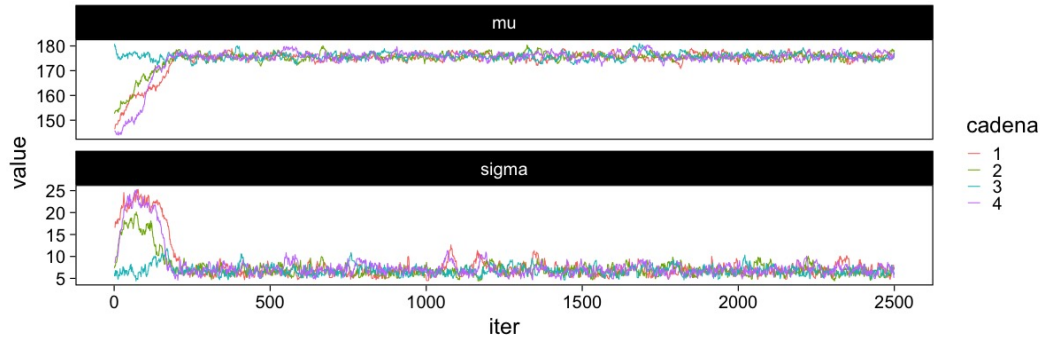


FIGURA 4. Trayectorias con dependencias iniciales.

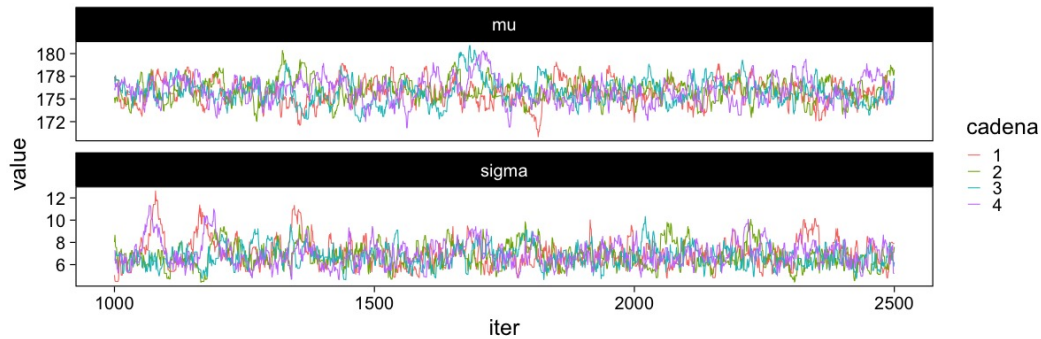


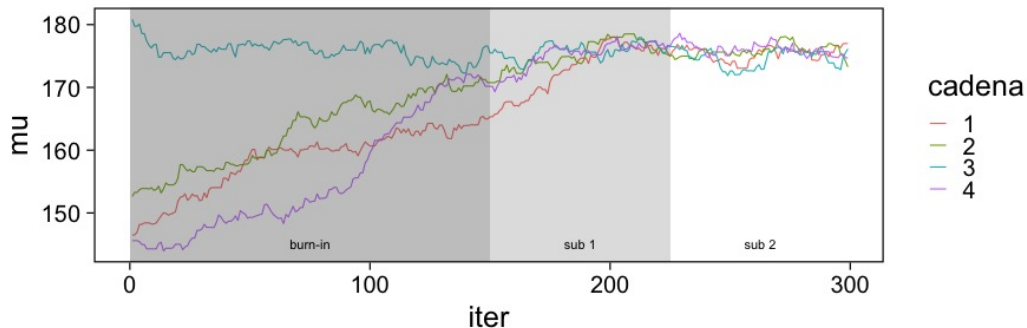
FIGURA 5. Trayectorias estacionarias.

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 **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 $i = 1, \dots, n$, y $j = 1, \dots, m$. Calculamos B y W , la variabilidad **entre** (*between*) y **dentro** (*within*) cadenas, respectivamente, por medio de

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

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

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 de

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

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 $n \rightarrow \infty$, 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 $n \rightarrow \infty$. Esto es,

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

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

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 incorporar transformaciones que nos permitan generar un buen estimador. Puedes leer mas de esto aqui: [?].

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 $n \cdot m$ 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.

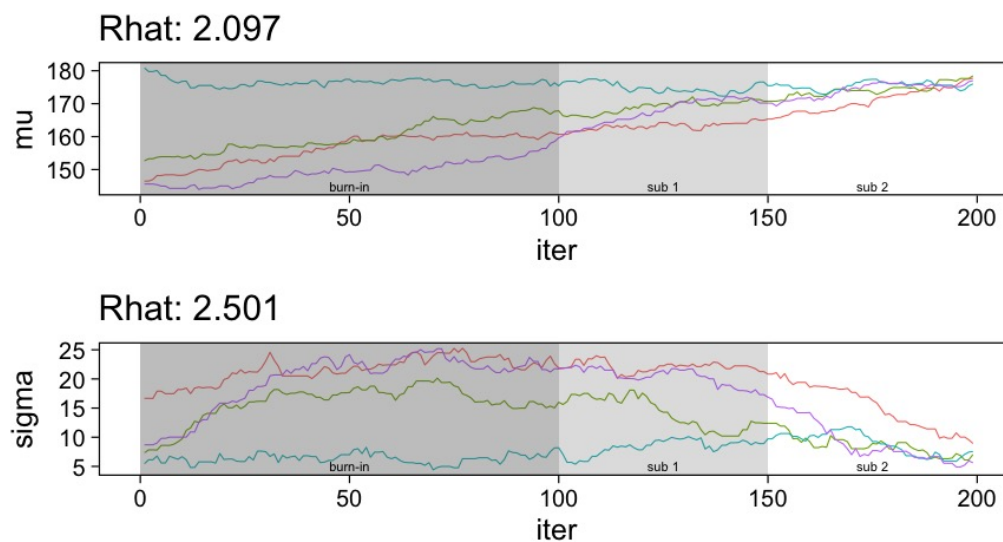


FIGURA 7. Diagnóstico de reducción de escala. Sugerencia: generar mas simulaciones.

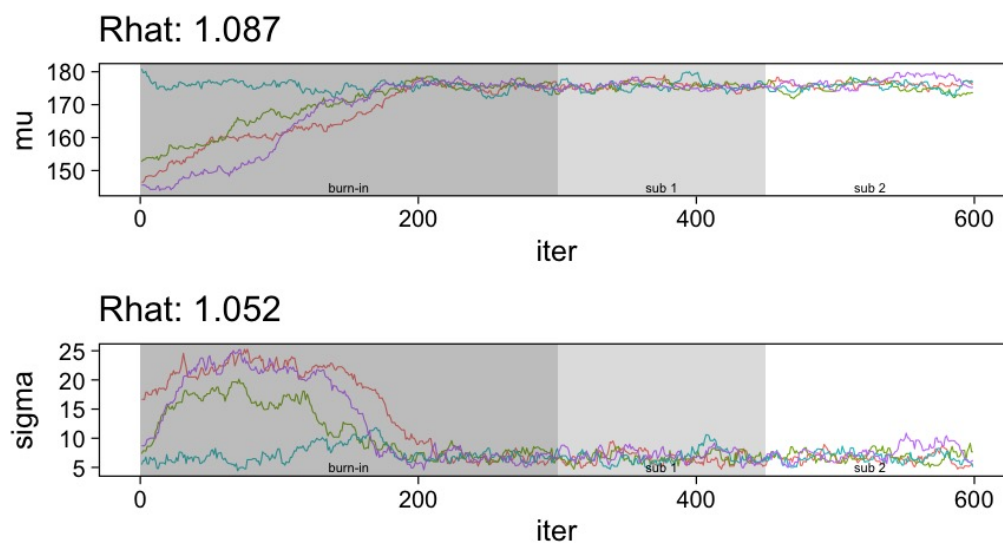


FIGURA 8. Diagnóstico de reducción de escala. Observaciones: parece estar bien.

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

$$\bar{\phi}_{..} \approx \mathbb{E}(\phi). \quad (6)$$

Del cual podemos derivar que

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

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

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

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** $\{Y_t : t = 0, \dots\}$ están definidas (respectivamente) como

$$C_\tau = \mathbb{E}[(Y_{t+\tau} - \mu)(Y_t - \mu)], \quad \rho_\tau = \frac{\mathbb{E}[(Y_{t+\tau} - \mu)(Y_t - \mu)]}{\sigma^2}. \quad (7)$$

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

$$\hat{C}_\tau = \frac{1}{T - \tau} \sum_{t=1}^{T-\tau} (Y_{t+\tau} - \hat{\mu})(Y_t - \hat{\mu}), \quad \hat{\rho}_\tau = \frac{\hat{C}_\tau}{\hat{C}_0}. \quad (8)$$

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

$$V_\tau = \mathbb{E}[(Y_{t+\tau} - Y_t)^2]. \quad (9)$$

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

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

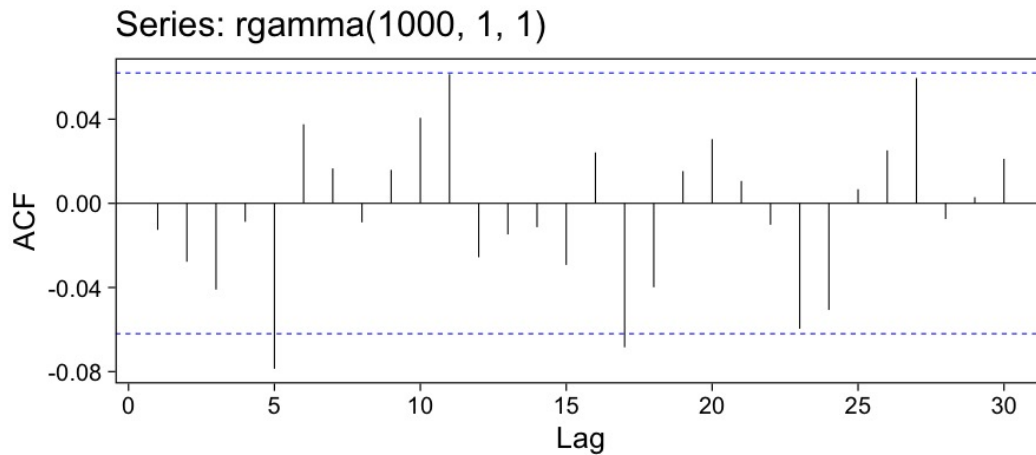
La mayor dificultad que presenta el estimador es considerar **todos** los retrasos posibles. Eventualmente agotaremos la longitud de las cadenas para ello. Por otro lado, para t eventualmente grande nuestros estimadores del variograma V_t serán muy ruidosos (¿por qué?). En la práctica truncamos la serie de acuerdo a las observaciones [?]. 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 estimador para el número efectivo de simulaciones es

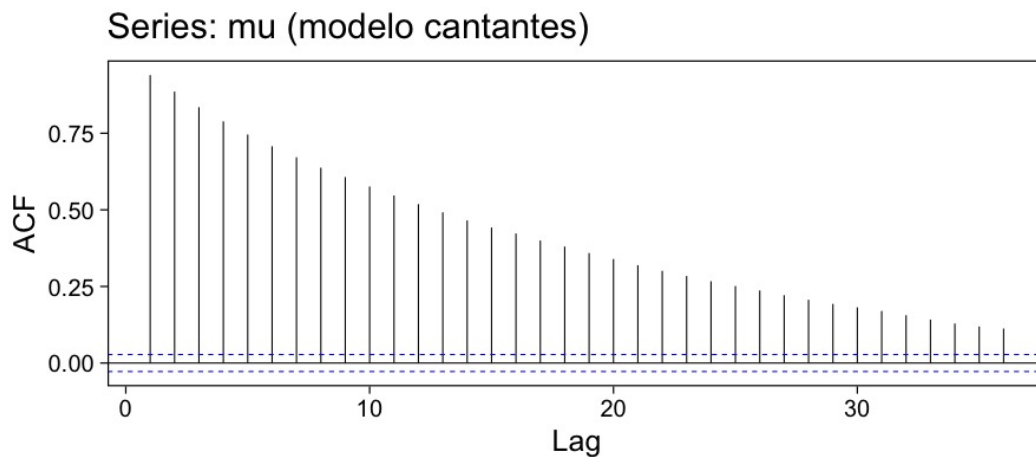
$$\hat{N}_{\text{eff}} = \frac{m n}{1 + 2 \sum_{t=1}^T \hat{\rho}_t}.$$

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 graficáramos simulaciones independientes, esperaríamos valores de autocorrelación chicos:



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



Podemos utilizar la función `posterior::ess_basic`.

```
1 # A tibble: 1 × 5
2   Neff    N.M  ratio sigma accept
3   <dbl> <int>  <dbl> <dbl>  <dbl>
4 1    441  5000 0.0882 0.146  0.380
```

LISTING 1. Fracción N_{eff}/nm para la simulación de la posterior los cantantes de ópera.

```
1 # A tibble: 1 × 5
```


	Neff	N.M	ratio	sigma	accept
	<dbl>	<int>	<dbl>	<dbl>	<dbl>
1	441	5000	0.0882	0.146	0.380

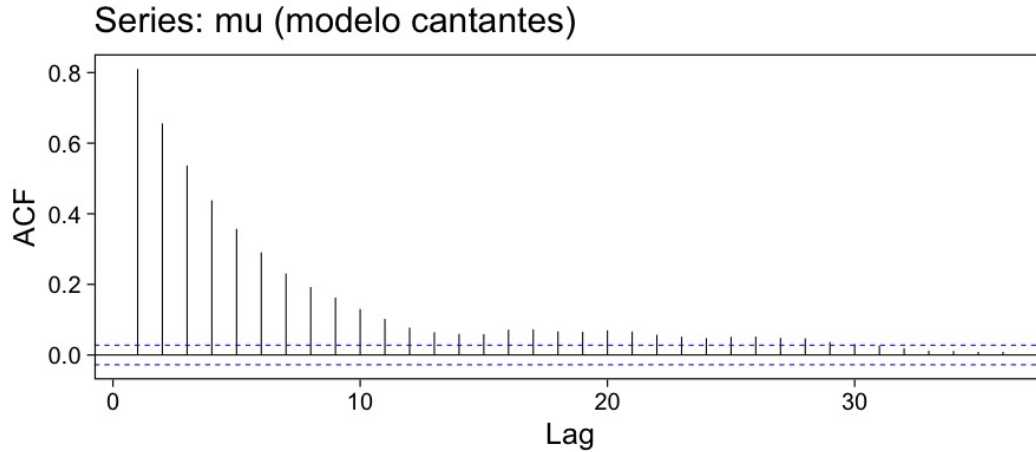
LISTING 2. Fracción N_{eff}/nm para la simulación (calibrada) de la posterior los cantantes de ópera.

FIGURA 9. Perfil de correlación para la simulación calibrada.

3. EL ESTADO DEL ARTE

El método de Metropolis-Hastings es muy flexible y existe una colección numerable de versiones que pueden ser empleadas en contextos muy particulares. Una buena referencia que incluye métodos de simulación por medio de cadenas de Markov se encuentra en [9], donde incluso se pueden encontrar generalizaciones con **transiciones Markovianas asimétricas** y extensiones a **problemas de dimensión variable**. El libro [1] presenta el estado del arte al 2010.

El cómputo Bayesiano se popularizó con el muestreador de Gibbs. En particular, el avance en teoría de grafos para representar una distribución conjunta como un Grafo Acíclico Dirigido (DAG) que se implementó en software como BUGS o WinBUGS. Pueden consultar el libro de Kruschke [8] para su explicación.

La desventaja del muestreador de Gibbs es que tiende a ser muy lento en problemas de tamaño grande. Ha habido estrategias que aceleran la simulación aunque al **costo de utilizar aproximaciones**. Estas estrategias han sido materializadas en lenguajes de programación mas generales como Infer.NET.

JAGS (Just Another Gibbs Sampler), es una generalización donde se implementan métodos MCMC para generar simulaciones de distribuciones posteriores. Los paquetes `rjags` y `R2jags` permiten ajustar modelos en JAGS desde R [7]. Es muy fácil utilizar estos programas pues uno simplemente debe especificar las distribuciones iniciales, la verosimilitud y los datos observados. Igual el libro de Kruschke [8].

Al depender de gradientes para construir propuestas para las cadenas de Markov ha sido natural el desarrollo de herramientas de muestreo basadas en diferenciadores automáticos. Por ejemplo, Pyro utiliza PyTorch. Tenemos también Tensorflow Probability que utiliza Tensorflow. Pymc (antes Pymc3) utiliza Theano (ahora llamado Aesara). NumPyro utiliza numpy y JAX como *backend*.

Pymc es un muestreador híbrido que permite utilizar Metropolis-Hastings, Gibbs y HMC para la simulación de la posterior [10]. También es mucho más flexible y brinda muestreadores más modernos basados en partículas e información de primer orden (gradientes).

Además, hay herramientas que utilizan las librerías de muestreo para análisis específicos. Por ejemplo, tenemos **cmdstanarm** ajusta **modelos de regresión** utilizando **Stan** como *backend*. La herramienta de Facebook, **Prophet**, utiliza **Stan** (ver [aquí](#)) como *backend* y se especializa en **series de tiempo**. **fenics-pymc3** se especializa en soluciones de **ecuaciones diferenciales** escritas en FEniCS. También tenemos **beat** para **análisis probabilístico de terremotos** y **exoplanet** para series de tiempo en **astronomía**. Por supuesto, no podía faltar una integración **scikit** que se llama **Pymc-Learn**.

Existen otras alternativas para construir cadenas de Markov. Por ejemplo, hay algoritmos que buscan evolucionar una colección de muestras de θ como un enjambre que se comunican entre sí para generar una caminata aleatoria en el espacio del soporte de la distribución. Ejemplos de éstos son el **t-walk** [2] o un ensamble de cadenas linealmente relacionadas como en la herramienta de **emcee** [4],

Finalmente, hay muchos más mecanismos que tienen como objetivo aproximar la distribución posterior. En problemas donde la verosimilitud es **computacionalmente costosa** existen alternativas para crear aproximaciones. El artículo [5] provee de una alternativa utilizando una combinación de técnicas bien establecidas (difusiones Langevin, ensamble de partículas interactivas y filtros de Kalman).

REFERENCIAS

- [1] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011. [9](#)
- [2] J. A. Christen and C. Fox. A general purpose sampling algorithm for continuous distributions (the t-walk). *Bayesian Analysis*, 5(2):263–281, jun 2010. ISSN 1936-0975, 1931-6690. . [10](#)
- [3] N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2015. [1](#)
- [4] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. Emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306–312, mar 2013. ISSN 00046280, 15383873. . [10](#)
- [5] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin Diffusions: Gradient Structure And Ensemble Kalman Sampler. *arXiv:1903.08866 [math]*, oct 2019. [10](#)
- [6] 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](#)
- [7] K. Hornik, F. Leisch, and A. Zeileis. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of DSC*, volume 2, pages 1–1, 2003. [9](#)
- [8] J. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2014. [9](#)
- [9] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer New York, New York, NY, 2004. ISBN 978-0-387-76369-9 978-0-387-76371-2. . [9](#)
- [10] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016. [10](#)