Criterio para determinar el tamaño de muestra en procesos de simulación estocástica [a]

Criterion to Determine the Sample Size in Stochastic Simulation Processes

Recibido: Diciembre 04, 2020 | Aceptado: Agosto 10, 2021 | Publicado: Julio 29, 2022

Juan Daniel Molina-Muñoz*
Centro de Investigación en Matemáticas, CIMAT, Guanajuato, México
ORCID: https://orcid.org/0000-0001-8583-8889

José Andrés Christen
Centro de Investigación en Matemáticas, CIMAT, Guanajuato, México
ORCID: https://orcid.org/0000-0002-5795-4345

[a] Artículo de investigación

* Autor de correspondencia. Correo: juan.molina@cimat.mx

DOI: https://doi.org/10.11144/javeriana.iued26.cdss

Como citar este artículo
J.D. Molina Muñoz, J. Christen, "Criterio para determinar el tamaño de muestra en procesos de simulación estocástica" Ing. Univ. vol. 26, 2022. https://doi.org/10.11144/javeriana.iued26.cdss

Resumen

Objetivo: Proponer un criterio para determinar el tamaño de muestra en simulaciones estocásticas de MC (Monte Carlo) y MCMC (Markov chain Monte Carlo), garantizando una determinada precisión en la estimación de parámetros. Se busca que la precisión se garantice de forma adimensional. ***Materiales y métodos:*** El presente artículo propone un criterio buscando cumplir con el objetivo planteado. Además, de una metodología para la aplicación del mismo. ***Resultados y discusión:*** Se presenta la aplicación de la metodología en 3 contextos diferentes: Simulación de MC en que la muestra de interés presenta variabilidad moderada, simulación de MC en que la muestra de interés presenta variabilidad excesiva y simulación de MCMC. En todos los casos se obtienen adecuadas estimaciones del número de corridas MC y MCMC a partir de muestras relativamente pequeñas. Además, la aplicación de la metodología representa únicamente un costo computacional adicional marginal. ***Conclusiones:*** El criterio presentado en este artículo permite determinar el tamaño de muestra en simulaciones estocásticas, garantizando precisión adimensional en la estimación de parámetros.

Palabras clave: Simulación estocástica, tamaño de muestra, Monte Carlo, MCMC, coeficiente de variación.

Abstract

Objective: To propose a criterion to determine the sample size in stochastic simulations of MC (Monte Carlo) and MCMC (Markov chain Monte Carlo), guaranteeing certain precision in estimating parameters. It is intended that the accuracy is guaranteed in a dimensionless way. ***Materials and methods:*** This paper proposes a criterion that seeks to meet the stated objective. In addition, a methodology for its application. ***Results and discussion:*** The application of the methodology is presented in 3 different contexts: MC simulation in which the sample of interest presents moderate variability, MC simulation in which the sample of interest presents excessive variability, and MCMC simulation. In all cases, adequate estimates of the number of MC and MCMC runs are obtained from relatively small samples. Furthermore, the application of the methodology represents only a marginal additional computational cost. ***Conclusions:*** The criterion presented in this paper allows for determining the sample size in stochastic simulations, guaranteeing dimensionless precision in estimating parameters.

Keywords: Stochastic simulation, sample size, Monte Carlo, MCMC, coefficient of variation.

Introducción

El método de simulación de MC (Monte Carlo) se basa en el principio de muestreo aleatorio, principalmente permite: Generar valores de una distribución de probabilidad de interés, realizar integración numérica (comúnmente usada para la estimación de parámetros), o desarrollar procesos de optimización [1], [2], [3], [4]. El método de simulación de MCMC (Markov chain Monte Carlo) hace parte de la familia de métodos de MC. Este es especialmente útil cuando se desea simular valores de una distribución de probabilidad de la que no se conoce explícitamente su constante de normalización. Esta situación es habitual en el contexto de la estadística bayesiana, sin ser el área exclusiva de aplicación del método [5], [6], [1]. En general, las herramientas que ofrecen los métodos de MC y MCMC son útiles para modelar fenómenos físicos, económicos, biológicos, médicos, sociales, lo cual hace que sean ampliamente utilizados en múltiples áreas del conocimiento [7], [8], [9], [10], [11].

En la ejecución de simulaciones de MC, o simulación estocástica (con o sin MCMC), una decisión fundamental es determinar del tamaño de muestra (o el número de simulaciones a realizar). La importancia de esta decisión está dada en que el tamaño de muestra influye sobre el error de estimación de cantidades de interés (QoI, por sus siglas en inglés) y sobre la precisión de las estimaciones que se realicen. Podría pensarse como solución tener un tamaño de muestra tan grande como sea posible, sin embargo, cada nueva simulación tiene un costo computacional que dependerá de la complejidad del sistema a simular. Por esto, se tiene interés en construir un criterio que permita encontrar el tamaño de muestra necesario para garantizar una precisión dada en la estimación de parámetros.

En el desarrollo de procesos de simulación estocástica una práctica común es que se fije el tamaño de muestra con valores estándar (100 ó 1000 ó 10,000 ó $1 \cdot 10^6$ u otros). Estos se convierten en "números mágicos" que, aunque son aceptados en muchos entornos académicos, no garantizan la precisión en la estimación de parámetros ni en el control del error de simulación. Dicha práctica sigue estando vigente, como puede verse por ejemplo en estas publicaciones muy recientes: [11], [12], [10], [13], [9].

No existen muchas alternativas para determinar el tamaño de muestra en una simulación de MC de forma cuantitativa. Por ejemplo, en [4], que es un libro ampliamente conocido en el área de simulación estocástica, se plantea el siguiente procedimiento para fijar el tamaño de muestra de la simulación:

1. Escoger un valor aceptable d para la desviación estándar del estimador de un parámetro de interés.
2. Generar una muestra de la variable de interés de tamaño al menos 100.

3. Continuar generando valores adicionales, detenerse cuando se haya generado k valores y $\frac{S}{\sqrt{k}} < d$, donde S representa la desviación estándar muestral de los valores generados.

El método propuesto por [4] puede considerarse arbitrario en cuanto a la fijación de la muestra inicial de tamaño en al menos 100. Además, que resulta ambigua la elección del valor d.

Algunos libros populares de simulación estocástica no presentan un criterio explícito para determinar el tamaño de muestra en simulaciones de MC, sino que se limitan a proponer el uso del teorema de limite central y los intervalos de confianza que de este se desprenden, como una herramienta para evaluar la convergencia de los estimadores [1], [14].

En el caso de simulaciones de MCMC en la mayoría de los libros sobre este tema se presentan criterios para garantizar la convergencia de la cadena conformada por los valores simulados. Sin embargo, no se presentan criterios que garanticen precisión en las estimaciones calculadas bajo condiciones de convergencia [5], [1]. Una de las pocas alternativas es el criterio planteado en [15] para determinar el burn-in en simulaciones de MCMC, además del tamaño de muestra necesario para garantizar precisión en la estimación de un cuantil de una función de los parámetros del modelo.

La mayoría de los artículos sobre este tema representan aplicaciones en áreas del conocimiento específicas. En muchos de estos artículos, los métodos que emplean para calcular el tamaño de muestra se basan en la fórmula de intervalo de confianza para la media que se desprende del teorema de limite central, y de la expresión del ancho del intervalo se despeja el tamaño de muestra. La variación de cada método está dada en qué conjunto de los siguientes elementos se asumen como fijos o conocidos: el nivel de confianza, el ancho del intervalo, el error de estimación admisible y el coeficiente de variación [16], [17], [18], [19], [20], [21], [22]. Sin embargo, en la mayoría de los casos no resulta intuitivo determinar un adecuado valor para el ancho del intervalo o el error de estimación admisible. Respecto al coeficiente de variación, se tiene que son casi nulas las situaciones en que este se conoce de antemano y, además, si el cálculo del tamaño de muestra depende de este valor, se incurre en un argumento circular en caso de querer estimar la media.

En este artículo se presenta un criterio para determinar el tamaño de muestra de una simulación estocástica (con o sin MCMC), que garantice una determinada precisión (establecida por el usuario) en la estimación de los parámetros. Con el importante atenuante que dicha precisión se garantiza de forma adimensional basado en el número de "cifras significativas" que el estimador de MC tiene. Es decir, sea $a \in \mathbb{R}^+$, expresando este número en notación científica tenemos el resultado presentado en la ecuación (1).

$$a = m_a\, 10^q = a_1.a_2a_3 \ldots 10^q \text{ with } a_1 \neq 0, q \in \mathbb{Z}. \quad (1)$$

Se conoce como la mantisa de a a m_a, $1 \leq m_a < 10$ y $a_0.a_1a_2$ ... es la expansión decimal de m_a. El usuario establece cuantas cifras p necesita/quiere que estén correctas con probabilidad cercana a 1 (> 0.9999) en un estimador de MC de a. En ese caso, q y de a_0 a a_p estarían correctas. Se plantea un algoritmo para la implementación del criterio, simple y de bajo costo computacional, mediante una estimación preliminar del coeficiente de variación (C_V, la desviación estándar dividida entre el valor esperado de un funcional de interés). Es de mencionar que el aporte matemático de la metodología propuesta en este artículo no es grande. Su valor está dado en su utilidad práctica y en su sencillez computacional.

Primero, el presente artículo expone el criterio planteado para determinar el tamaño de muestra, se propone una metodología para implementar el criterio, y se presentan consideraciones para la determinación del tamaño de muestra en el caso de una simulación MCMC. Después se desarrollan múltiples ejemplos del uso de la metodología propuesta. Y finalmente, se presentan las conclusiones.

Materiales y métodos

Notación y definición del criterio

Sea $X \in \mathbb{R}^n$ una variable aleatoria con función de densidad de probabilidad (o de probabilidad de masa) $f_X(\cdot)$, sea además $g: \mathbb{R}^n \longrightarrow \mathbb{R}^+$ un funcional. Consideramos los valores presentados en la ecuación (2).

$$\mu = \mathbb{E}[g(X)] \text{ and } \sigma^2 = \mathbb{V}[g(X)] \quad (2)$$

Sea $X_1, X_2, \ldots, X_T$ una muestra de variables independientes e idénticamente distribuidas de $f_X(\cdot)$ obtenida a partir de simulación de MC. Considerando el estimador para μ, utilizamos la media simple presentada en la ecuación (3).

$$h_T = \frac{1}{T}\sum_{i=1}^{T} g\,(X_i). \quad (3)$$

Por el teorema central del límite [23] tenemos el resultado presentado en la ecuación (4).

$$\sqrt{T}\frac{(h_T-\mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1). \tag{4}$$

Expresando a μ en notación científica como en (1), tenemos que μ es igual a lo que se muestra en la ecuación (5).

$$\mu = m_\mu\ 10^q. \tag{5}$$

Luego, tomando el exponente de μ, se considera a h_T y a σ expresados como se muestra en la ecuación (6).

$$h_T = m_T\ 10^q \text{ and } \sigma = m_\sigma\ 10^q. \tag{6}$$

Nótese que m_T y m_σ no necesariamente son las mantisas de la convención de notación científica descritas en (1). Luego, (2) puede re-escribirse como se presenta en la ecuación (7).

$$\sqrt{T}\frac{(m_T\ 10^q - m_\mu\ 10^q)}{m_\sigma\ 10^q} = \sqrt{T}\frac{(m_T-m_\mu)}{m_\sigma} \xrightarrow{d} \mathcal{N}(0,1). \tag{7}$$

Ahora, sea $p \in \mathbb{N}$ un valor establecido, el interés está dado en garantizar con alta probabilidad (cercana a 1) una precisión de p cifras significativas en la estimación de la mantisa de m_μ, es decir, que $m_\mu = m_T$ si ambas cantidades se redondean a p cifras significativas. Lo cual ocurre si se satisface la desigualdad presentada en la ecuación (8).

$$\left|m_T - m_\mu\right| < 0.5 \cdot 10^{-(p-1)}. \tag{8}$$

Para garantizar el resultado (4) se consideran los cálculos presentados en la ecuación (9).

$$\begin{aligned} &\mathbb{P}\left(\left|m_T - m_\mu\right| < 0.5 \cdot 10^{-(p-1)}\right) = \mathbb{P}\left(-0.5 \cdot 10^{-(p-1)} < m_T - m_\mu < 0.5 \cdot 10^{-(p-1)}\right) \\ =\ &\mathbb{P}\left(-\frac{\sqrt{T}}{2m_\sigma}10^{-(p-1)} < \sqrt{T}\frac{(m_T-m_\mu)}{m_\sigma} < \frac{\sqrt{T}}{2m_\sigma}10^{-(p-1)}\right) \\ \approx\ &\mathbb{P}\left(-\frac{\sqrt{T}}{2m_\sigma}10^{-(p-1)} < \mathrm{Z} < \frac{\sqrt{T}}{2m_\sigma}10^{-(p-1)}\right), \end{aligned} \tag{9}$$

Esto último considerando el resultado (3), con Z una variable aleatoria tal que $\mathrm{Z}\sim\mathcal{N}(0,1)$. De esta forma, se tiene que $\mathbb{P}\left(\left|m_T - m_\mu\right| < 0.5 \cdot 10^{-(p-1)}\right) \approx \mathbb{P}(-z < Z < z)$ con $z = \frac{\sqrt{T}}{2m_\sigma}10^{-(p-1)}$ un cuantil de la distribución de Z tal que $\mathbb{P}(-z < Z < z) \approx 1$. Tomando $z =$

4, se tiene $\mathbb{P}(-4 < Z < 4) > 1 - 1 \cdot 10^{-4}$. Entonces $z = 4 = \frac{\sqrt{T}}{2m_\sigma} 10^{-(p-1)}$ y $\sqrt{T} = 8\, m_\sigma\, 10^{(p-1)}$.

Por lo tanto, el mínimo tamaño de muestra que garantiza una precisión de p cifras significativas en la estimación de m_μ es la expresión presentada en la ecuación (10).

$$T^* = 64\, m_\sigma^2\, 10^{2(p-1)}. \tag{10}$$

Ahora, considerando que $\mathrm{C_V} = \frac{\sigma}{\mu} = \frac{m_\sigma\, 10^q}{m_\mu\, 10^q} = \frac{m_\sigma}{m_\mu}$, es decir $m_\sigma = \mathrm{C_V}\, m_\mu$, tenemos que $T^* = 64\, \mathrm{C_V}^2\, m_\mu^2\, 10^{2(p-1)}$. Pero $m_\mu < 10$, por ser una mantisa, entonces $T^* < 64\, \mathrm{C_V}^2\, 10^2\, 10^{2(p-1)}$, con lo cual se obtiene el resultado presentado en la ecuación (11).

$$T^* < 64\, \mathrm{C_V}^2\, 10^{2p}. \tag{11}$$

Un caso estándar es cuando $\mathrm{C_V} < \frac{1}{4}$, en este caso la expresión (5) es equivalente a la expresión presentada en la ecuación (12).

$$T^* < 64 \left(\frac{1}{4}\right)^2 10^{2p} = 4 \cdot 10^{2p}. \tag{12}$$

Ahora, suponiendo que se tienen T simulaciones independientes de un funcional de interés. Para determinar la precisión que con este número de corridas se garantiza en la estimación de m_μ del funcional, se puede partir de la expresión (5), hacer $T^* = 64\, \mathrm{C_V}^2\, 10^{2p}$ y de esta igualdad despejar p , la expresión resultante se muestra en la ecuación (13).

$$p = 0.5 \log_{10}(T) - \log_{10}(8\mathrm{C_V}). \tag{13}$$

Así, el p que se calcula por medio de la expresión (6) representa una cota superior para la precisión que puede garantizarse en la estimación de m_μ del funcional con T simulaciones. Así, podemos fijar una precisión p y calcular el T^* requerido o con un número actual de muestras T dado calcular el número real de cifras significativas p en un estimador.

En general, no se puede estimar directamente el $\mathrm{C_V} = \frac{\sigma}{\mu}$ pues de esta forma se incurre en un argumento circular respecto al objetivo principal de garantizar precisión en la estimación de m_μ. Se plantea entonces realizar una estimación preliminar del C_V en la que no se utilice la

totalidad de la información muestral. Esta estimación preliminar se propone sea realizada por medio del método de Bland [24], el cual se presenta en el Apéndice A.

Corrección para MCMC

Sea $X_1, X_2, \ldots$ una cadena de Markov reversible. Sea como se muestra en la ecuación (14).

$$\gamma_t = \gamma_{-t} = \text{Cov}(\, g(X_i), g(X_{i+t})), \tag{14}$$

la autocovarianza en el resago t de la serie de tiempo estacionaria $g(X_1), g(X_2), \ldots$, en [25] demuestran que para una cadena de Markov estacionaria, irreducible y reversible tenemos el resultado presentado en la ecuación (15).

$$T\ \mathbb{V}(h_T)\ \overrightarrow{C.S.}\ \ \sigma^2 = \sum_{t=-\infty}^{\infty} \gamma_t. \tag{15}$$

Además, si $\sigma^2 < \infty$, tenemos la ecuación resultante (16).

$$\sqrt{T}\frac{(h_T - \mu)}{\sigma}\ \overrightarrow{d}\ \mathcal{N}(0,1). \tag{16}$$

Es decir, la expresión (7) representa el teorema central del límite para el caso de muestras dependientes construidas a partir de una cadena de Markov estacionaria, irreducible y reversible. En [26] se plantea un método para estimar σ^2, el cual se presenta en el apéndice B.

Así, dado el resultado presentado en (7) los resultados propuestos en la sección anterior son igualmente válidos para muestras construidas vía MCMC. Debe notarse que la definición e interpretación de σ^2 cambia entre el caso de la simulación de una muestra independiente vía MC al caso de la simulación de una muestra dependiente vía MCMC.

En el caso MCMC, para evitar incurrir en un argumento circular al momento de estimar el C_V, existen 2 alternativas: Primero, estimar el tiempo de autocorrelación integrado (IAT por sus siglas en inglés) de la muestra y, a partir de este, extraer una muestra pseudo-independiente de la muestra dependiente, y sobre esta última aplicar el procedimiento planteado en la siguiente sección. La segunda opción es, dado que el argumento circular recae únicamente sobre μ y que este parámetro se estima de la misma forma para muestras independientes y dependientes, se plantea estimar μ por medio del método de Bland [24] y σ usando el método de Geyer [26]. Para esta última alternativa la aplicación del procedimiento

planteado en la siguiente sección para estimar T^*es válido, salvo el cambio en la estimación de σ.

Método heurístico para calcular T^*

Se planteada una metodología iterativa para calcular el T^* que garantice una precisión de p cifras significativas:

1. Sean $\tau = 2^{-2}$, $l = 0$ y $T_l = 64\,\tau^2\,10^2$. Generar vía simulación Monte Carlo una muestra de variables independientes del funcional de interés de tamaño T_l, y de esta calcular $\widehat{C_{VB}}$ que representa la estimación del C_V por medio del método de Bland.
2. Mientras $\widehat{C_{VB}} > \tau$ y $\tau \leq 4$, hacer $\tau = 2\tau$ y $T_l = 64\,\tau^2\,10^2$. Generar una muestra de variables independientes del funcional de interés de tamaño T_l, y de esta calcular $\widehat{C_{VB}}$.
3. Si $\tau > 4$ imprimir el mensaje "El procedimiento se detiene pues la muestra de interés posee una dispersión excesiva" y detener el algoritmo. Sino, continuar con el siguiente paso.
4. Hacer $T_l = 64\,\widehat{C_{VB}}^2 10^2$ (con $\widehat{C_{VB}}$ el último estimador calculado por medio del método de Bland). Generar una muestra de variables independientes del funcional de interés de tamaño T_l, y de esta calcular $\widehat{C_V} = \frac{S}{\bar{X}}$ (estimador tradicional). Hacer $l = l + 1$ y $T_l = 64\,\widehat{C_V}^2 10^{2l}$.
5. Mientras $l < p$, generar una muestra de variables independientes del funcional de interés de tamaño T_l, y de esta calcular $\widehat{C_V} = \frac{S}{\bar{X}}$. Hacer $l = l + 1$ y $T_l = 64\,\widehat{C_V}^2 10^{2l}$.
6. Retornar $T^* = T_p$.

Con los pasos 1 y 2 del algoritmo se estructura la estimación preliminar del C_V. La función de τ es definir una adecuada cota superior para C_V. Inicialmente se asume $C_V < \frac{1}{4}$ que representa el caso estándar o regular, si este supuesto no es factible ($\widehat{C_{VB}} > \frac{1}{4}$) se asume $C_V < \frac{1}{2}$ y así sucesivamente, multiplicando la cota por 2. El último escenario que se considera en la estimación preliminar es $C_V < 4$. Con los pasos 4 y 5 del algoritmo se refina la estimación del C_V, usando toda la información muestral disponible.

Para mayor eficiencia se recomienda que en cada paso del algoritmo que se requiera generar una muestra de variables independientes de $f_X(\cdot)$, no descartar los valores que previamente se hayan generado. Teniendo en cuenta que de esta forma en ningún momento se viola el supuesto de independencia entre los elementos de la muestra.

Si se tiene una muestra independiente del funcional de interés de tamaño T y sobre está quiere calcularse la precisión p que se garantiza en la estimación de m_μ, el algoritmo anterior sigue siendo válido. Únicamente es necesario que, en cada uno de los pasos, cuando se habla de generar una muestra de tamaño T_l, tomar esta de la muestra disponible. En este caso el algoritmo se detiene en el momento que $T_l > T$ y la precisión que se garantiza es $p = l - 1$.

Implementación en Python

El algoritmo planteado en la sección anterior se implementó en Python. Tanto para el caso de simulaciones de MC como para simulaciones de MCMC se definieron dos funciones:

- Una que calcula el tamaño de muestra T^* necesario para garantizar una precisión p en la estimación de la mantisa de un funcional de interés; los argumentos que recibe está función son la precisión deseada, el mecanismo de generación de la muestra y el funcional; esta función retorna el valor de T^*, una muestra de tamaño T^* del funcional de interés, estimaciones en la etapa inicial y de refinamiento del coeficiente de variación y del tamaño de muestra, y una estimación de μ (calculada sobre la muestra de tamaño T^*) en notación científica redondeando su mantisa con p cifras significativas.
- La otra permite calcular la precisión que se garantiza para una determinada muestra; el argumento de esta función es la muestra a evaluar; esta función retorna el valor del número p de cifras significativas que se pueden garantizar con la muestra, y una estimación de μ en notación científica redondeando su mantisa con p cifras significativas.

Los códigos en Python con la implementación del algoritmo junto con los ejemplos que se presentan en este artículo pueden encontrarse en la plataforma GitHub, en el enlace: https://github.com/jdmolinam/Sample_Size_Criterion

Resultados

Se desarrollaron tres ejemplos para presentar la aplicación del algoritmo planteado en diferentes condiciones. En el ejemplo 1 se presenta el caso en que el coeficiente de variación del funcional de interés es menor a $\frac{1}{4}$. Este puede considerarse un caso estándar de dispersión moderada. En el ejemplo 2 el funcional tiene una mayor dispersión pues se tiene un coeficiente de variación mayor a $\frac{1}{4}$. Finalmente, el ejemplo 3 presenta la aplicación del

algoritmo en el contexto de una simulación de MCMC en la cual no se conoce de antemano el verdadero valor del coeficiente de variación del funcional de interés.

Simulación de Monte Carlo con $C_V < \frac{1}{4}$

Sea $X = (X_1, X_2, X_3) \sim \mathcal{N}_3(\mu_X, \Sigma)$, con $\mu_X = (3, 3, 3)$ y $\Sigma = \mathbb{I}_3$. Bajo estas condiciones se tiene que las X_i son independientes entre si y $X_i \sim \mathcal{N}(\mu_i, 1)$, para $i = 1,2,3$. Por lo tanto, la función de densidad de probabilidad de X es como se presenta en la ecuación (17).

$$\begin{aligned} f(X) &= \prod_{i=1}^{3}\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x_i-\mu_i)^2\right)\right] \\ &= (2\pi)^{-\frac{3}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{3}(x_i-\mu_i)^2\right). \end{aligned} \tag{17}$$

Considerando el funcional g: $\mathbb{R}^3 \longrightarrow \mathbb{R}$, tal que $g(X) = \sum_{i=1}^{3} X_i$, tenemos para μ la expresión (18),

$$\mu = \mathbb{E}[g(X)] = \mathbb{E}\left[\sum_{i=1}^{3} X_i\right] = \sum_{i=1}^{3}\mathbb{E}(X_i) = 9 \tag{18}$$

Y por σ^2 la expresión (19).

$$\sigma^2 = \mathbb{V}[g(X)] = \mathbb{V}\left[\sum_{i=1}^{3} X_i\right] = \sum_{i=1}^{3}\mathbb{V}(X_i) = 3. \tag{19}$$

Por lo tanto, $\mu = 9 \cdot 10^0 = m_\mu\, 10^0$ es decir, $m_\mu = 9$ y $q = 0$. Además, $\mathrm{C_V} = \frac{\sigma}{\mu} = \frac{\sqrt{3}}{9} \approx 0.1925$.

Ahora, se procede a calcular una cota superior para el tamaño de muestra T^* que garantice una precisión en la estimación de m_μ de $p = 3$ cifras significativas. Desarrollando el proceso iterativo planteado anteriormente se encontró $T^* = 2{,}335{,}380$, en la Figura 1 se presentan los diferentes tamaños de muestra y estimaciones del $\mathrm{C_V}$ consideradas en el proceso de estimación de T^*. Se generó una muestra aleatoria por medio de simulación de MC de tamaño T^*, y de esta se obtuvo $h_T = 9.0003 \cdot 10^0$, es decir, $m_T = 9.0003$. Así, recordando que

$m_\mu = 9$ se tiene que $|m_T - m_\mu| = 0.0003 < 0.5 \cdot 10^{-2} = 0.005$. Además, si se redondea tanto a m_μ como a m_T a 3 cifras significativas se tiene $m_\mu = m_T \approx 9.00$. Por lo tanto, se confirma la precisión de $p = 3$ cifras significativas en la estimación de m_μ.

Simulación de Monte Carlo con $C_V > \frac{1}{4}$

Sea $X = (X_1, X_2, X_3, X_4)$ con X_1, X_2, X_3, X_4 independientes e idénticamente distribuidas tal que $X_i \sim \text{Exp}(\lambda = 1)$, para $i = 1, \cdots, 4$. Por lo tanto, la función de densidad de probabilidad de X es como se presenta en la ecuación (20).

$$f_X(X) = \prod_{i=1}^{4} [\exp(-x_i)] = \exp\left(-\sum_{i=1}^{4} x_i\right). \tag{20}$$

Considerando el funcional $g: \mathbb{R}^4 \longrightarrow \mathbb{R}^+$, tal que $g(X) = \frac{1}{4}\sum_{i=1}^{4} X_i$, tenemos para μ la expresión (21),

$$\mu = \mathbb{E}[g(X)] = \mathbb{E}\left[\frac{1}{4}\sum_{i=1}^{4} X_i\right] = \frac{1}{4}\sum_{i=1}^{4} \mathbb{E}(X_i) = \frac{1}{4} 4\lambda^{-1} = 1 \tag{21}$$

Y para σ^2 la expresión (22).

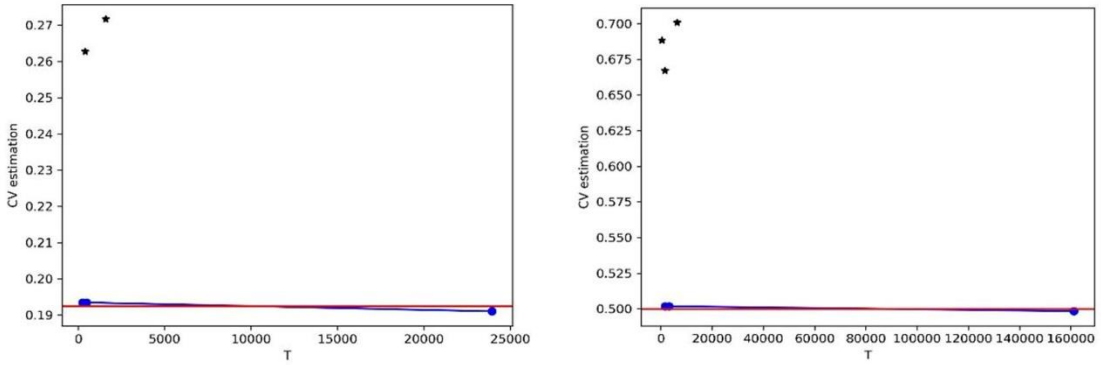
$$\sigma^2 = \mathbb{V}[g(X)] = \mathbb{V}\left[\frac{1}{4}\sum_{i=1}^{4} X_i\right] = \frac{1}{16}\sum_{i=1}^{4} \mathbb{V}(X_i) = \frac{1}{16} 4\lambda^{-2} = 0.25. \tag{22}$$

Por lo tanto, $\mu = 1 \cdot 10^0 = m_\mu \, 10^0$ es decir, $m_\mu = 1$ y $q = 0$. Además, $\mathrm{C_V} = \frac{\sigma}{\mu} = \frac{\sqrt{0.25}}{1} = 0.5$.

Ahora, se procede a calcular una cota superior para el tamaño de muestra T^* que garantice una precisión en la estimación de m_μ de $p = 3$ cifras significativas. Desarrollando el proceso iterativo planteado anteriormente se encontró $T^* = 15{,}965{,}508$, en la figura 1 se presentan los diferentes tamaños de muestra y estimaciones del $\mathrm{C_V}$ consideradas en el proceso de estimación de T^*. Se generó una muestra aleatoria por medio de simulación de MC de tamaño

T^*, y de esta se obtuvo $h_T = 1.0001 \cdot 10^0$, es decir, $m_T = 1.0001$. Así, recordando que $m_\mu = 1$ se tiene que $|m_T - m_\mu| = 0.0001 < 0.5 \cdot 10^{-2} = 0.005$. Además, si se redondea tanto a m_μ como a m_T a 3 cifras significativas se tiene $m_\mu = m_T \approx 1.00$. Por lo tanto, se confirma la precisión de $p = 3$ cifras significativas en la estimación de m_μ.

Figura 1. Valores de T y estimaciones del C_V considerados en la estimación de T^*, para los ejemplos 1 y 2. A la izquierda los resultados del ejemplo 1 a la derecha los del ejemplo 2.



* Los asteriscos representan las estimaciones del C_V en la etapa inicial y los puntos las estimaciones en la etapa de refinamiento.

Fuente: Elaboración propia

Simulación de MCMC: Estimación bayesiana en el modelo Lotka–Volterra

Este ejemplo se desarrolló alrededor del sistema de ecuaciones de Lotka–Volterra, el cual describe la dinámica de dos poblaciones de animales: una depredadora y otra presa. Se consideró el sistema bajo las condiciones presentadas en la ecuación (23).

$$\begin{aligned} \frac{du_1}{dt} &= u_1(1 - u_2), \\ \frac{du_2}{dt} &= u_2(u_1 - 1). \end{aligned} \tag{23}$$

Donde $u_1(t)$ y $u_2(t)$ representan la población (miles de especímenes) en el tiempo t de la especie presa y depredador respectivamente. $u_1(0) = u_1^0$ y $u_2(0) = u_2^0$ son desconocidos. Los parámetros que caracterizan el modelo son $\theta = (u_1^0, u_2^0)$ y de estos, el parámetro de interés es u_1^0. Así, el funcional a considerar es $g(\theta) = u_1^0$.

Se plantea un problema inverso bayesiano [27] en el cual los datos disponibles tienen la siguiente estructura presentada en la ecuación (24)

$$y_i = u_1(t_i, \theta) + \varepsilon_i, \quad i = 1, \dots, n, \tag{24}$$

con $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Así, $y_i|\theta \sim \mathcal{N}(\, u_1(t_i, \theta)\,, \sigma^2)$. Para la simulación de los datos se fijó $\theta = (0.5\,,\, 2)$, $\sigma = \max_{\{i=1,\cdots,n\}}(\, u_1(t_i, \theta)\,) \cdot 0.1$, $n = 5$ con las observaciones tomadas sobre los tiempos $\{0.75, 1.5, 2.25, 3.0, 3.75\}$. Se consideró la distribución previa presentada en la ecuación (25).

$$\theta \sim U([0.5 - e\,,\, 0.5 + e] \times [2 - e\,,\, 2 + e]), \tag{25}$$

Con $e = 0.2$. De esta forma, lo tenemos como se muestra en la ecuación (26).

$$f(\theta) = f(u_1^0, u_2^0) = \frac{1}{0.16}\, \mathbb{I}_{[0.3\,,0.7]}(u_1^0)\mathbb{I}_{[1.8\,,2.2]}(u_2^0). \tag{26}$$

Sea $Y = (y_1, \dots, y_n)$ el vector de observaciones, la función de verosimilitud está caracterizada por la expresión (27).

$$f(Y|\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\, y_i - u_1(t_i, \theta))^2 \right). \tag{27}$$

Así, la distribución a posteriori está determinada por el teorema de Bayes, es decir: $f(\theta|Y) \propto f(Y|\theta) f(\theta)$.

Bajo las condiciones de este ejemplo, el parámetro sobre el que desea garantizarse precisión en su estimación es:

$$\mu = \mathbb{E}(u_1^0 \mid Y). \tag{28}$$

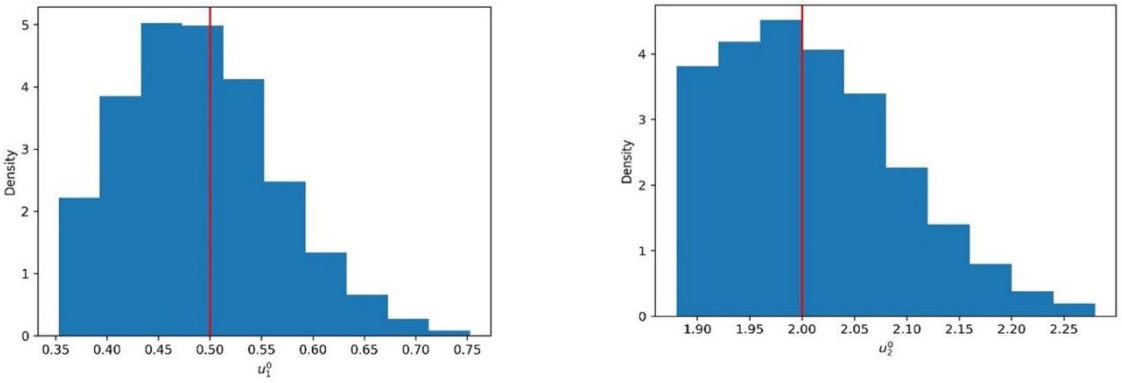
Para generar valores de la distribución a posteriori se usó el algoritmo MCMC, t-walk [28]. Inicialmente se corrieron 10.000 iteraciones, a partir de las cuales se plantea que con un *burn-in* de 500 iteraciones, la cadena conformada por los valores simulados alcanza el estado estacionario.

Ahora, se procede a calcular una cota superior para el tamaño de muestra T^* que garantice una precisión en la estimación de m_μ de $p = 2$ cifras significativas. Se desarrolló el procedimiento planteado con la salvedad de realizar la estimación preliminar del coeficiente de variación estimando la media con el método de Bland y la desviación estándar de la muestra dependiente con el método de Geyer. Se obtuvo $T^* = 942{,}270$, en la Figura 2 se

presentan las distribuciones a posteriori obtenidas a partir de una muestra MCMC de tamaño T^*.

Adicionalmente, en la Tabla 1 presenta la precisión que se garantiza para cada uno de los ejemplos, en el caso en que se cuenta con una muestra de tamaño T (independiente en los ejemplos 1 y 2, obtenida vía MCMC en el 3) del funcional de interés. El cálculo de la precisión se hizo utilizando el algoritmo heurístico planteado, asumiendo el C_V desconocido.

Figura 2. Histogramas de las distribuciones a posteriori de u_1^0 y u_2^0



* A la izquierda la distribución a posteriori de u_1^0, a la derecha la de u_2^0. Las líneas rojas en u_1^0=0.5 y u_2^0=2 representan los valores de los parámetros usados para generar los datos sintéticos.

Fuente: Elaboración propia

Tabla 1. Precisión p que se garantiza con diferentes tamaños de muestra en cada uno de los ejemplos*

Sample Size			
$C_V < \frac{1}{4}$	$C_V > \frac{1}{4}$	MCMC (IAT = 50, $C_V = 1.2$)	p
250	1,700	9,500	1
25,000	170,000	950,000	2
2,500,000	17,000,000	95,000,000	3

* En el caso $C_V < \frac{1}{4}$ las diferentes precisiones se garantizan con tamaños de muestra más pequeños, para el caso $C_V > \frac{1}{4}$ se requieren muestras más grandes y en el caso MCMC mucho más. Por lo tanto, elegir arbitrariamente el tamaño de la muestra con "números mágicos" en general es un error, que se agudiza si se tiene un $C_V > \frac{1}{4}$ o habitualmente para simulaciones de MCMC. Obsérvese el conocido, y comúnmente descuidado, aumento de 100 veces el tamaño de la muestra con un aumento de una cifra significativa en la precisión, resultante de la tasa de convergencia de $\frac{1}{2}$ en el CLT.

Fuente: Elaboración propia

Conclusiones

En muchos casos, el problema de determinación del tamaño de muestra en simulaciones estocásticas se aborda de manera ligera por medio de "números mágicos", o de técnicas que no ofrecen garantías importantes, como el control del error de simulación o precisión en la estimación de parámetros. Esto representa un serio problema, especialmente si se tiene un $C_V > \frac{1}{4}$ ó si se realiza una simulación de MCMC donde el IAT sea mayor a 1, que es lo más habitual en casos no triviales.

El criterio que se presenta en este artículo permite determinar el tamaño de muestra en simulaciones estocásticas, lo que garantiza la precisión adimensional en la estimación de parámetros. La importancia de este resultado no radica en su aporte matemático, sino en el valor práctico del mismo.

La metodología de tipo heurística presentada en este artículo para aplicar el criterio en cada uno de los ejemplos de aplicación mostró ser eficiente al no requerir de grandes muestras para realizar una buena estimación del C_V y de T^*. Además, que esta metodología no agrega un gran costo computacional al proceso global de simulación. Los autores consideran que la implementación de esta metodología en softwares estadísticos de uso común como R y Python se convertiría en una herramienta de gran utilidad.

Referencias

[1] C. Robert and G. Casella, *Monte Carlo statistical methods*, Springer Science & Business Media, 2004, https://doi.org/10.1007/978-1-4757-4145-2

[2] G. Fishman, *Monte Carlo: concepts, algorithms, and applications*, Springer Science & Business Media, 2013.

[3] J. S. Liu, *Monte Carlo strategies in scientific computing*, Springer Science & Business Media, 2008, https://doi.org/10.1007/978-0-387-76371-2

[4] S. Ross, *Simulation*, 5th ed., Elsevier Science, 2012.

[5] D. Gamerman and H. F. Lopes, Markov chain Monte Carlo: stochastic simulation for Bayesian inference, Chapman and Hall/CRC, 2006.

[6] B. A. Berg and A. Billoire, Markov chain Monte Carlo simulations, Wiley Encyclopedia of Computer Science and Engineering. Wiley Online Library, 2007.

[7] C. Forastero, L. Zamora, D. Guirado a A. Lallena, "A Monte Carlo tool to simulate breast cancer screening programmes," *Physics in Medicine & Biology*, vol. 55, no. 17, p. 5213, 2010, https://doi.org/10.1088/0031-9155/55/17/021

[8] H. MacGillivray, R. Dodd, B. McNally, J. Lightfoot, H. Corwin and S. Heathcote, "Monte-Carlo simulations of galaxy systems," *Astrophysics and Space Science*, vol. 81, no. 1-2, pp. 231-250, 1982, https://doi.org/10.1007/BF00683346

[9] T. Flouri, X. Jiao, B. Rannala and Z. Yang, "A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis," *Molecular Biology and Evolution*, vol. 37, nº 4, pp. 1211-1223, 2020, https://doi.org/10.1093/molbev/msz296

[10] C. L. Ritt, J. R. Werber, A. Deshmukh and M. Elimelech, "Monte Carlo simulations of framework defects in layered two-dimensional nanomaterial desalination membranes: implications for permeability and selectivity," Environmental Science & Technology, vol. 53, nº 11, pp. 6214-6224, 2019, https://doi.org/10.1021/acs.est.8b06880

[11] I. Ciufolini and A. Paolozzi, "Mathematical prediction of the time evolution of the COVID-19 pandemic in Italy by a Gauss error function and Monte Carlo simulations," *The European Physical Journal Plus*, vol. 135, nº 4, p. 355, 2020, https://doi.org/10.1140/epjp/s13360-020-00383-y

[12] R. Al, C. R. Behera, K. V. Gernaey and G. Sin, "Stochastic simulation-based superstructure optimization framework for process synthesis and design under uncertainty," *Computers & Chemical Engineering*, Vol. 143, pp. 107-118, 2020, https://doi.org/10.1016/j.compchemeng.2020.107118

[13] E. Spitoni, K. Verma, V. S. Aguirre and F. Calura, "Galactic archaeology with asteroseismic ages-II. Confirmation of a delayed gas infall using Bayesian analysis based on MCMC methods," *Astronomy & Astrophysics*, vol. 635, p. A58, 2020, https://doi.org/10.1051/0004-6361/201937275

[14] O. Jones, R. Maillardet and A. Robinson, *Introduction to scientific programming and simulation using R*, Chapman and Hall/CRC, 2014.

[15] A. E. Raftery and S. Lewis, How many iterations in the gibbs sampler? In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds., *Bayesian Statistics*, vol. 4, Oxford University Press, 1992.

[16] I. Lerche and B. S. Mudford, "How many Monte Carlo simulations does one need to do?," *Energy exploration & exploitation*, vol. 23, no. 6, pp. 405-427, 2005, https://www.jstor.org/stable/43754693

[17] F. E. Ritter, M. J. Schoelles, K. S. Quigley and L. C. Klein, "Determining the number of simulation runs: Treating simulations as theories by not sampling their behavior," in L. Rothrock and S. Narayanan (eds.), *Human in the loop simulations*, Springer, 2011, pp. 97-116.

[18] M. Liu, "Optimal Number of Trials for Monte Carlo Simulation," VRC-Valuation Research Report, 2017.

[19] L. T. Truong, M. Sarvi, G. Currie and T. M. Garoni, "How many simulation runs are required to achieve statistically confident results: a case study of simulation-based surrogate safety measures," *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, 2015, pp. 274-278, https://doi.org/10.1109/ITSC.2015.54

[20] G. Hahn, "Sample Sizes for Monte-Carlo Simultation," *IEEE Transactions on Systems Man and Cybernetics*, no. 5, p. 678, 1972. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4309200

[21] W. Oberle, Monte Carlo Simulations: Number of Iterations and Accuracy, Army Research Lab Aberdeen Proving Ground Md Weapons and Materials Research, 2015, https://apps.dtic.mil/sti/pdfs/ADA621501.pdf

[22] M. D. Byrne, "How many times should a stochastic model be run? An approach based on confidence intervals," *Proceedings of the 12th International conference on cognitive modeling*, Ottawa, 2013.

[23] R. J. Serfling, *Approximation theorems of mathematical statistics*, John Wiley & Sons, 2009.

[24] M. Bland, "Estimating mean and standard deviation from the sample size, three quartiles, minimum, and maximum," *International Journal of Statistics in Medical Research*, vol. 4, no. 1, pp. 57-64, 2014, https://doi.org/10.6000/1929-6029.2015.04.01.6

[25] C. Kipnis and S. S. Varadhan, "Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions," *Communications in Mathematical Physics*, vol. 104, no. 1, pp. 1-19, 1986, https://doi.org/10.1007/BF01210789

[26] C. J. Geyer, "Practical markov chain Monte Carlo," *Statistical Science*, vol. 7, no. 4, pp. 473-483, 1992, https://doi.org/10.1214/ss/1177011137

[27] M. A. Capistrán, J. A. Christen and S. Donnet, "Bayesian analysis of ODEs: solver optimal accuracy and Bayes factors," SIAM/ASA Journal on Uncertainty Quantification, vol. 4, no. 1, pp. 829--849, 2016, https://doi.org/10.1137/140976777

[28] J. A. Christen and C. Fox, "A general purpose sampling algorithm for continuous distributions (the t-walk)," *Bayesian Analysis*, vol. 5, no. 2, pp. 263-281, 2010, https://doi.org/10.1214/10-BA603

Apéndice A. Método para la estimación preliminar del C_V

Dado el interés en realizar una estimación preliminar del C_V, se considera el método de Bland [24], el cual permite estimar la media y varianza de una muestra de variables independientes e idénticamente distribuidas sin utilizar el total de la información muestral. A continuación, se explica el método.

Sea X una variable aleatoria en $\mathbb{R}^+$, con función de densidad de probabilidad (o probabilidad de masa) $f_X(\cdot)$, sea $X_1, X_2, \cdots, X_n$ una muestra de variables independientes e idénticamente distribuidas de $f_X(\cdot)$ y sean $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ los estadísticos de orden de dicha muestra. Se consideran las siguientes cantidades: $a = X_{(1)}$, q_1: el primer cuartil muestral, m: la mediana muestral, q_3: el tercer cuartil muestral y $b = X_{(n)}$. Considerando que la media y la varianza muestral están determinadas por las expresiones presentadas en la ecuación (29).

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2 \ , \tag{29}$$

Si por simplicidad se asume $n = 4Q + 1$, con $Q \in \mathbb{Z}^+$, es decir, $Q = (n-1)/4$, para la estimación de la media muestral, se tienen en cuenta las desigualdades presentadas en la ecuación (30).

$$\begin{array}{rcll}
a & \leq & X_{(1)} & \leq a \\
a & \leq & X_{(i)} & \leq q_1, \quad (i = 2, \cdots, Q) \\
q_1 & \leq & X_{(Q+1)} & \leq q_1 \\
q_1 & \leq & X_{(i)} & \leq m, \quad (i = Q+2, \cdots, 2Q) \\
m & \leq & X_{(2Q+1)} & \leq m \\
m & \leq & X_{(i)} & \leq q_3, \quad (i = 2Q+2, \cdots, 3Q) \\
q_3 & \leq & X_{(3Q+1)} & \leq q_3 \\
q_3 & \leq & X_{(i)} & \leq b, \quad (i = 3Q+2, \cdots, n-1) \\
b & \leq & X_{(n)} & \leq b.
\end{array} \tag{30}$$

Sumando todas las desigualdades anteriores y dividiendo por n se obtiene que $\alpha_p \leq \overline{X} \leq \beta_p$, donde α_p se define como se muestra en la ecuación (31),

$$\alpha_p = \frac{a + q_1 + m + q_3}{4} + \frac{4b - a - q_1 - m - q_3}{4n} \tag{31}$$

and β_p como se muestra en la ecuación (32).

$$\beta_p = \frac{q_1 + m + q_3 + b}{4} + \frac{4a - q_1 - m - q_3 - b}{4n}. \tag{32}$$

Entonces, tenemos el resultado presentado en la ecuación (33).

$$\overline{X} \approx \frac{\alpha_p + \beta_p}{2} = \overline{X}_B. \tag{33}$$

Para estimar la varianza muestral, se tienen en cuenta las desigualdades presentadas en la ecuación (34).

$$\begin{aligned}
aX_{(1)} &\leq X_{(1)}^2 \leq aX_{(1)} \\
aX_{(i)} &\leq X_{(i)}^2 \leq q_1X_{(i)}, \quad (i = 2, \cdots, Q) \\
q_1X_{(Q+1)} &\leq X_{(Q+1)}^2 \leq q_1X_{(Q+1)} \\
q_1X_{(i)} &\leq X_{(i)}^2 \leq mX_{(i)}, \quad (i = Q+2, \cdots, 2Q) \\
mX_{(2Q+1)} &\leq X_{(2Q+1)}^2 \leq mX_{(2Q+1)} \\
mX_{(i)} &\leq X_{(i)}^2 \leq q_3X_{(i)}, \quad (i = 2Q+2, \cdots, 3Q) \\
q_3X_{(3Q+1)} &\leq X_{(3Q+1)}^2 \leq q_3X_{(3Q+1)} \\
q_3X_{(i)} &\leq X_{(i)}^2 \leq bX_{(i)}, \quad (i = 3Q+2, \cdots, n-1) \\
bX_{(n)} &\leq X_{(n)}^2 \leq bX_{(n)}.
\end{aligned} \tag{34}$$

Sumando todas las desigualdades anteriores y con álgebra simple se obtiene que $\alpha_s \leq \sum_{i=1}^{n} X_i^2 \leq \beta_s$, donde α_s se define como se muestra en la ecuación (35),

$$\alpha_s = \frac{1}{8}[\,8b^2 + (n+3)(a^2 + q_1^2 + m^2q_3^2) + (n-5)(aq_1 + q_1m + mq_3 + q_3b)\,] \tag{35}$$

Y β_s como se muestra en la ecuación (36).

$$\beta_s = \frac{1}{8}[\,8a^2 + (n+3)(q_1^2 + m^2 + q_3^2 + b^2) + (n-5)(aq_1 + q_1m + mq_3 + q_3b)\,]. \tag{36}$$

Se plantea, $\sum_{i=1}^{n} X_i^2 \approx \frac{\alpha_s + \beta_s}{2} = \gamma_s$. Y teniendo en cuenta la expresión (37).

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 \; - n\overline{X}^2\right] \tag{37}$$

Entonces, tenemos el resultado (38).

$$S^2 \approx \frac{1}{n-1}\left(\gamma_s - n\overline{X}_B^2\right) = S_B^2. \tag{38}$$

Así, a partir del método de Bland se plantea la expresión presentada en la ecuación (39) como estimación preliminar de la C_V.

$$\widehat{C_{VB}} = \frac{S_B}{\overline{X}_B}. \tag{39}$$

Normalmente, se necesitará un tamaño de muestra muy pequeño para obtener una estimación inicial aproximada de C_V utilizando el estimador anterior. En el contexto del algoritmo propuesto en este trabajo, esto funciona bien para restringir el problema de estimación con un límite inicial para C_V utilizando una pequeña muestra de prueba y, a partir de ahí, establecer el tamaño de muestra necesario.

Apéndice B. Método para estimar la varianza de una muestra MCMC

En este apéndice se presenta el método de Geyer [26], para estimar la varianza de una muestra MCMC en estado estacionario. Considerando una cadena de Markov reversible $X_1, X_2, \ldots$, Sea como se muestra en la ecuación (40).

$$\gamma_t = \gamma_{-t} = \text{Cov}(\, g(X_i), g(X_{i+t})), \tag{40}$$

la autocovarianza en el rezago t de la serie de tiempo estacionaria $g(X_1), g(X_2), \cdots$. Esta cantidad puede estimarse a partir de la autocovarianza empírica presentada en la ecuación (41).

$$\hat{\gamma}_{n,t} = \hat{\gamma}_{n,-t} = \frac{1}{n}\sum_{i=1}^{n-t}(\, g(X_i) - h_n\,)(\, g(X_{i+t}) - h_n\,). \tag{41}$$

En [26] se demuestra que para una cadena de Markov estacionaria, irreducible y reversible, $\Gamma_m = \gamma_{2m} + \gamma_{2m+1}$ es una función de m estrictamente positiva, estrictamente decreciente y convexa. Además, $\widehat{\Gamma}_{n,m} = \hat{\gamma}_{n,2m} + \hat{\gamma}_{n,2m+1}$. Con base en estos resultados, en [26] se plantea el siguiente estimador para σ^2 presentado en la ecuación (42) se propone.

$$\widehat{\sigma}^2 = \hat{\gamma}_0 + 2\sum_{i=1}^{2m+1}\hat{\gamma}_{n,i} = -\hat{\gamma}_0 + 2\sum_{i=0}^{m}\widehat{\Gamma}_{n,m}\,, \tag{42}$$

donde m se escoge como el entero más grande de tal forma que se cumpla la expresión (43)

$$\widehat{\sigma}^2 = \hat{\gamma}_0 + 2\sum_{i=1}^{2m+1}\hat{\gamma}_{n,i} = -\hat{\gamma}_0 + 2\sum_{i=0}^{m}\widehat{\Gamma}_{n,m}\,, \tag{43}$$