Transl Clin Pharmacol. 2023 Jun;31(2):69-84
https://doi.org/10.12793/tcp.2023.31.e9
pISSN 2289-0882·eISSN 2383-5427

Translational and Clinical Pharmacology TCP

Review Article

Check for updates

A review of the Bayesian approach with the MCMC and the HMC as a competitor of classical likelihood statistics for pharmacometricians

Kyungmee Choi *

College of Science and Technology, Hongik University, Sejong 30016, Korea

OPEN ACCESS

Received: May 25, 2023
Revised: Jun 14, 2023
Accepted: Jun 18, 2023
Published online: Jun 26, 2023

***Correspondence to**
Kyungmee Choi
College of Science and Technology, Hongik University, 2639 Sejong-ro, Jochiwon-eup, Sejong 30016, Korea.
Email: kmchoi@hongik.ac.kr

ORCID iDs
Kyungmee Choi
https://orcid.org/0000-0001-9977-8362

Funding
This research was funded by 2023 Hongik University Research Funds.

Conflict of Interest
- Authors: Nothing to declare
- Reviewers: Nothing to declare
- Editors: Nothing to declare

Reviewer
This article was reviewed by peer experts who are not TCP editors.

ABSTRACT

This article reviews the Bayesian inference with the Monte Carlo Markov Chain (MCMC) and the Hamiltonian Monte Carlo (HMC) samplers as a competitor of the classical likelihood statistical inference for pharmacometricians. The MCMC and the HMC samplers have greatly contributed to realization of the Bayesian methods with minimal requirement of mathematical theory. They do not require any closed form of the posterior density nor linear approximation of complex nonlinear models in high dimension even with non-conjugate priors. The HMC even weakens the dependency of the chain and improves computational efficiency. Pharmacometrics is one of great beneficiaries since they use complex multivariate multilevel nonlinear mixed effects models based on the restricted maximum likelihood estimation. Comprehension of the Bayesian approach will help pharmacometricians to access the data analysis more conveniently.

Keywords: Likelihood Estimates; Bayesian Inference; Monte Carlo Markov Chain; Hamiltonian Monte Carlo

INTRODUCTION

The main objective of this article is to provide a review of Bayesian methods to pharmacometricians. Statistical inference on the clinical trial data collected from subjects among vast populations is a routine task in pharmacometrics. Even a simple 'PK analysis' with only one compartment involves a complicated nonlinear mixed effects model with multiple parameters expressing the human body in probability [1-3]. To fit the non-linear mixed effects models, the classical statistical inference linearizes the non-linear models based on the first-order or second-order Taylor approximations and looks for the estimators based on the restricted likelihood estimation (REML) which involves complicated high dimensional integration [2-5]. The traditional Bayesian inference needs calculation of the posterior density, which requires high dimensional integration. Meanwhile, Bayesian sampling methods like the Monte Carlo Markov Chain (MCMC) and the Hamiltonian Monte Carlo (HMC) do not seek for an analytical solution, do not linearize non-linear models, do avoid high dimensional and analytical integrations, and thus have quickly evolved with their computational efficiency [3,6-10]. As a target function, they require only the likelihood

function and prior densities instead of an explicit posterior density. They have become daily necessities in Pharmacometrics as an appealing alternative of the classical likelihood statistics and the traditional Bayesian methods. This article will review principles of the classical inference as well to provide an outline of Bayesian data analysis.

The major difference between frequentists and Bayesians lies in whether the target parameter θ is an unknown fixed constant or a random variable following a prior distribution carrying its information. [11-14] There have been already innumerable literatures on the Bayesian approach. A recent article by Lee [3] provides an excellent mathematical review on the Bayesian approach on the nonlinear mixed effects models commonly used in Pharmacometrics. This article starts with the traditional statistical inference to position the Bayesian approach within its outline. The best way not to get lost in the overflow of statistical techniques is to keep the map in mind and keep referring back to global and local goals.

Section 2 will review classical likelihood statistics with the goodness-of-fit criteria including the Akaike information criterion (AIC). Section 3 will review basic Bayesian inference with the Bayesian information criterion (BIC) and the posterior odds (PO) and the Bayes factor (BF) [15-18]. Section 4 will review the Monte Carlo Method and the Markov Chain to be combined as the Bayesian samplers like the MCMC and the HMC.

CLASSICAL LIKELIHOOD INFERENCE

Estimation and test

For a set of data $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, we assume a linear model such as

$$y_i = x_i^T\beta + \varepsilon_i, \qquad (1)$$

where x_i in R^p and the errors ε_i are independent and identically distributed with a probability density or mass function f which is often a normal distribution with the mean 0 and variance σ^2. If x_i=1, then the model includes only the intercept which is the grand mean μ. The overall goal of the data analysis is to estimate the unknown parameter $\theta=\{\beta,\sigma^2\}$ and test hypotheses in order to select a good model [11-14,17,18]. Let $\hat{\theta}$ be a point estimator of θ. The bias is defined as the difference between the expected value of the estimator and the true parameter,

$$\text{Bias}(\hat{\theta})=E[\hat{\theta}]-\theta.$$

The estimator is said to be unbiased if the bias is zero. The unbiasedness is called consistency for large samples and ergodicity for stochastic processes with dependency. At the significance level α, a two-sided hypothesis is

$$H_0: \theta=\theta_0 \quad H_1: \theta\neq\theta_0.$$

For p-value less than α, H_0 is rejected. The 100(1−α)% confidence interval (CI) satisfies

$$P(\theta\in CI)=1-\alpha.$$

If the 95% CI does not include θ_0, then H_0 is rejected at the significance level 0.05. A common goodness-of-fit criterion of the point estimator $\hat{\theta}$ is the mean squared error (MSE) which is [17]

$$MSE=E[(\theta-\hat{\theta})^2]=(Bias(\hat{\theta}))^2+Var(\hat{\theta}).$$

The estimator with smaller MSE is better. There is a bias-variance tradeoff as the bias gets big when the variance gets small.

https://doi.org/10.12793/tcp.2023.31.e9

The most popular point estimator has been the maximum likelihood estimator (MLE) which maximizes the likelihood function $L(\theta)$ and the log-likelihood function $\ell(\theta)$ [11-14,17,18]

$$L(\theta) = \prod_{i=1}^{n} f_Y(y_i; x_i\theta),\ \ell(\theta) = \log L(\theta).$$

The MLE $\hat{\theta}_{MLE}$ is defined as

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta)$$

and it is obtained by solving the equation $\frac{\partial \ell}{\partial \theta} = 0$. At $\hat{\theta}_{MLE}$, $\ell(\theta)$ takes its maximum, its derivative is zero, and $\hat{\theta}_{MLE}$ is the most likely value of the parameter θ for given data. For a large n, the MLE follows an asymptotic normal distribution,

$$\hat{\theta}_{MLE} \approx N\left(\theta, \frac{1}{n} I^{-1}(\hat{\theta}_{MLE})\right).$$

The Fisher Information $I(\theta)$ is said to be

$$I(\theta) = -E_\theta\left[\sum_{i=1}^{n} \frac{\partial^2 \ell(\theta)}{\partial\theta\, \partial\theta^T}\right].$$

The MLE is consistent, which means it is asymptotically unbiased for a large sample or $E[\hat{\theta}_{MLE}] \approx \theta$ [14-17]. Under the normality condition, the MLEs and the maximum log-likelihood (MLL) are

$$\hat{\beta}_{MLE} = (X^T X)^{-1} X^T X,\ \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (2)$$

$$\ell(\hat{\theta}_{MLE}) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2_{MLE}) - \frac{n}{2}.$$

In pharmacometrics, multilevel non-linear mixed effects models are of the form [3,5,19,20]

$$y = f(t,\theta) + \varepsilon$$

$$\theta = F(\beta, x, \eta)$$

$$\varepsilon \sim N(0, \sigma^2, I),\ \eta \sim N(0, \omega^2 I).$$

In the individual-level model, f is a differentiable real-valued function to describe the non-linear model, t is time, the parameter θ is subject-specific, and ε is the within-subject error with a variance $\sigma^2 I$. In the population model, F is often a linear model, β is a population-level fixed effect, and x is a covariate, and η is the random effect. The simplest population model is $\theta = x^T\beta + \eta$.

The REML estimation method is often used and its likelihood function [3] is complicated as

$$L(\beta, \sigma^2, \Omega | y) = \int f_Y(y | \beta, \eta, \sigma^2) f_\eta(\eta | \Omega)\, d\eta.$$

Compared to ordinary linear models, it has an additional random effect η with its variance Ω as a kind of penalty. Often, the non-linear $f(t,\theta)$ is linearized based on a first-order or second-order Taylor approximations and then the Newton-Raphson method or the expectation-maximization (EM) algorithm are applied to iteratively compute the MLE [5,21]. The EM-algorithm alternates two steps, calculates the expected log-likelihood function in the expectation step, and estimates and updates the parameter in the maximization step. The

https://doi.org/10.12793/tcp.2023.31.e9

algorithm repeats the two steps until the estimate converges. The software NONMEM [17] fits the pharmacokinetics/pharmacodynamics (PK/PD) models in pharmacometrics. See Kim et al. [20] for details.

Model selection criteria

The Wilks' test can be used as a model selection criterion [5,17,18,22] when the MLE is used. For a large sample, the MLL follows an approximate chi-square distribution with degrees of freedom d which is the dimension of the parameter space or the number of parameters to be estimated in the model,

$$-2MLL=-2\ \ell(\hat{\theta}_{MLE})\rightarrow\chi^2(d).$$

The greater MLL provides the better evidence for the model. Suppose we compare the two nested models, a reduced model versus a general model. Then the hypotheses are

H_0: reduced model

H_1: general model,

where $H_0 \subseteq H_1$. Let $\hat{\theta}_r$ be the MLE under H_0 and $\hat{\theta}_g$ be the MLE under H_1. The test statistic Δ is defined and its asymptotic distribution is as follows:

$$\Delta=2\ (\ell(\hat{\theta}_g)-\ell(\hat{\theta}_r))\rightarrow\chi^2(d_g-d_r).$$

The degrees of freedom (d_g-d_r) is the dimension difference between the two parameter spaces. At the significance level 0.05, the decision is to reject H_0 if $\Delta > \chi^2_{0.05}(d_g - d_r)$ [5,17]. The AIC [15,16] assesses goodness-of-fit,

$$AIC=-2\ \ell(\hat{\theta}_{MLE})+2\ d.$$

The AIC also includes penalty on the dimension d of the parameter space to prevent including too many explanatory variables in the model. We select a model with the smallest AIC. If the Fisher information matrix $I(\theta)$ is not positive definite for all θ, the generalized Watanabe-AIC (WAIC) can be used [23]. As a general criterion of goodness-of-fit of the model, the prediction error is assessed by the empirical MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

The model with smaller MSE is preferred. As mentioned in section 2.1, there is the bias-variance tradeoff and note that it has the same form as eq. (2).

BAYESIAN INFERENCE

As computing power is no longer a barrier, the Bayesian method is more competitive than ever, replacing the classical likelihood statistics [11-18]. For a target parameter θ in eq. (1), let us first define the prior density function (pdf) $p(\theta)$ such as

$$\int p(\theta)d\theta = 1.$$

The joint pdf of Y and θ is defined as product of the conditional density function $f_{Y|\Theta}(y|\theta)$ and the prior density function

$$f_{Y,\Theta}(y,\theta)=f_{Y|\Theta}(y|\theta)p(\theta).$$

https://doi.org/10.12793/tcp.2023.31.e9

Note that $f_{Y,\Theta}(y,\theta)$ is the likelihood function $L(\theta)$. The marginal pdf of Y is obtained by integrating the joint pdf with respect to θ,

$$f_Y(y)=\int f_{Y,\Theta}(y,\theta)d\theta=\int f_{Y,\Theta}(y|\theta)p(\theta)d\theta$$

The posterior pdf $p(\theta|x,y)$ is then obtained as

$$p(\theta|x,y)=\frac{f_{Y,\Theta}(y,\theta)}{f_Y(y)}=\frac{f_{Y|\Theta}(y|\theta)p(\theta)}{f_Y(y)}\propto f_{Y|\Theta}(y|\theta)p(\theta).$$

Ignoring the normalizing marginal function $f_Y(y)$, the posterior pdf is proportional to product of the likelihood function and the prior density,

$$\text{posterior}\propto\text{likelihood}\times\text{prior}.$$

The most popular loss function is the squared error loss function

$$L(\theta,\hat{\theta}_B)=(\theta-\hat{\theta}_B)^2$$

and its expectation is called the risk function $R(\theta,\hat{\theta}_B)$. The Bayesian estimator $\hat{\theta}_B$ is defined to minimize the risk function and minimize the conditional risk function for given Y,

$$\hat{\theta}_B(Y)=\underset{\theta}{\operatorname{argmin}}\,E\left[L(\theta,\hat{\theta}_B)\right]=\underset{\theta}{\operatorname{argmin}}\,E\left[E\left[L\left(\theta,\hat{\theta}_B(Y)\right)|Y\right]\right]=\underset{\theta}{arg\min}\,E\left[L\left(\theta,\hat{\theta}_B(Y)\right)|Y\right].$$

The Bayesian estimator $\hat{\theta}_B(Y)$ is obtained at its minimum value where its derivative is zero,

$$\frac{\partial}{\partial\hat{\theta}_B}E\left[L(\theta,\hat{\theta}_B)|Y\right]=\frac{\partial}{\partial\hat{\theta}_B}\int\left(\theta-\hat{\theta}_B\right)^2p(\theta|x,y)d\theta=-2\int\left(\theta-\hat{\theta}_B\right)p(\theta|x,y)d\theta=0.$$

Solving the last equality, we can get the Bayesian estimator $\hat{\theta}_B(Y)$ such as

$$\hat{\theta}_B(Y)=E[\theta|Y]=\int\theta p_{\Theta|Y}(\theta|y)d\theta.$$

The Bayes estimator for the squared error loss function is the posterior mean.

There are some issues on practical use of the Bayesian method because calculating the posterior density and posterior mean requires intricate mathematics when models are complex in high dimension. In addition, selecting a prior is an important issue to avoid subjectivity argument. When there is no prior information, we can try a noninformative symmetric pdf such as the uniform or normal or Cauchy distributions. The transformation invariant Jeffrey's prior is another choice, which is proportional to the square root of the determinant of the Fisher information $\sqrt{|I(\theta)|}$ [24]. If the prior and the posterior densities belong to the same distribution family, then the prior is called the conjugate prior. Well-known conjugate priors are Normal-Normal, Normal-Inverse Gamma, Poisson-Gamma, Geometric-Gamma, Multinomial-Dirichlet, Uniform-Pareto, Exponential-Gamma, and so on.

For large samples, the Bayesian estimator is very close to the MLE [14]. Under appropriate regularity conditions, the asymptotic distribution of the Bayesian estimator $\hat{\theta}_B$ for large n is

$$\hat{\theta}_B\approx N\left(\hat{\theta}_{MLE},\ \frac{1}{n}I^{-1}\left(\hat{\theta}_{MLE}\right)\right),$$

where $\hat{\theta}_{MLE}$ is the MLE and $I(\theta)$ is the Fisher Information. The Bayesian estimator is consistent like the MLE. The $100(1-\alpha)\%$ Bayesian credible interval is defined as

https://doi.org/10.12793/tcp.2023.31.e9

$$P(\theta_L^*(Y) < \theta < \theta_U^*(Y)|Y) = 1 - \alpha.$$

The $\theta_L^*(Y)$ and $\theta_U^*(Y)$ are $\alpha/2$ and $(1-\alpha/2)$ posterior quantiles, respectively. A smart numerical method using the Laplace integration method [25] was applied to earn the Bayesian estimators. Nevertheless, its practical calculation, for example an approximate posterior mean $E[\theta|Y]$, can be still an agony if it has to be done with complex models in high dimension.

The BIC is defined as

$$BIC = -2\ell(\hat{\theta}_{MLE}) + d \ln n.$$

The BIC also include penalty on the dimension d of the parameter space to prevent including too many explanatory variables in the model. If the Fisher Information is not positive definite, the Watanabe-BIC can be used [23]. Let us call the general data set as $Z=X$ or $Z=Y$ or $Z=(X,Y)$ in eq. (1). For models M_1, M_2, and M, the PO is defined as

$$PO = \frac{p(M_1|Z)}{p(M_2|Z)} = \frac{p(Z|M_1)}{p(Z|M_2)} \frac{p(M_1)}{p(M_2)}.$$

Here, $p(Z|M)$ corresponds to the likelihood, $p(M|Z)$ to the posterior density, and $p(M)$ to the prior, and the first term on the righthand side is called the BF. If PO >1, then the model M_1 is selected. In terms of the BIC, the posterior probability of the model M_i among m models approximates to

$$p(M_i|Z) \approx \frac{\exp\left(-\frac{1}{2}BIC(M_i)\right)}{\exp\left(-\frac{1}{2}BIC(M_1)\right) + \exp\left(-\frac{1}{2}BIC(M_2)\right) + \cdots + \exp\left(-\frac{1}{2}BIC(M_m)\right)}.$$

The model with the minimum BIC can be selected with the largest posterior probability [17].

SAMPLING METHODS FOR BAYESIAN INFERENCE

Monte Carlo method

From this section, let us generalize the notation as $Z=X$ or $Z=Y$ or $Z=(X,Y)$. If $Z=(X,Y)$, Y is a random variable for fixed X in eq. (1). The Monte Carlo is a widespread numerical integration method which was originally developed by John von Neumann and Stanislaw Ulam in the mid-1940s [26,27]. Adopting the Monte Carlo method, let us simulate $\{\theta^{(1)},\theta^{(2)},\ldots,\theta^{(n)}\}$ from $p(\theta|Z)$ and take the average as an estimate of the posterior mean of θ. By the law of large numbers, the sample mean converges to the posterior mean of θ

$$E[\theta|Z] \approx \frac{1}{n}\sum_{i=1}^{n} \theta^{(i)}$$

as $n \to \infty$. For the 95% credible interval $(\theta_L^*(z), \theta_U^*(z))$, the 0.025 quantile and the 0.975 quantile of the sample can be used for $\theta_L^*(z)$ and $\theta_U^*(z)$. As an example, we generate a sample of size 1000 from the posterior density $p(\theta|z)$. With their order statistics $\theta_{(1)} \le \theta_{(2)} \le \theta_{(3)} \le \ldots \le \theta_{(1000)}$, the equal-tailed 95% Bayesian credible interval can be calculated as $(\theta_{(25)}, \theta_{(975)})$, where the 0.025 quantile is $\theta_L^*(z) = \theta_{(25)}$ and the 0.975 quantile is $\theta_U^*(z) = \theta_{(975)}$.

https://doi.org/10.12793/tcp.2023.31.e9

Translational and Clinical Pharmacology TCP

Sampling methods

For the Monte Carlo method to efficiently work for the Bayesian approach, sampling from the posterior density should be affordable. Sampling methods have started with the very basic inversion method [12] and evolved to rejection sampling methods [13,28,29], the importance sampling method [17], and the Gibbs sampling method which simulates iteratively $p(x|y,z), p(y|x,z), p(z|x,y)$ for random variables X, Y, and Z in turn [17].

As an integrating concept of all these various sampling methods, in 1953 the Metropolis algorithm was first published [6] and in 1970, Hastings extended it to more general cases [7]. In the 1990s, the name "Metropolis-Hastings Algorithm" was mentioned by Chib and Greenberg [8]. The HMC sampler employed the very classical Newton's law in the stochastic process and efficiency has dramatically improved [3,6-10]. Both the MCMC and the HMC samplers have introduced a different paradigm from both classical likelihood statistics and traditional Bayesian statistics. Above all, the Bayesian methods with the two samplers do require only the shape of the posterior density as product of the likelihood and the prior [7,28-32]. Since then, the MCMC and HMC have played the major role among Bayesian methods. The Gibbs sampling method is applied as a part of the MCMC algorithm [17]. The HMC sampling method adopting differentials turned out to work better with high computational efficiency. Since Markov chain is not independent, we need stationarity for its convergence and ergodicity for its consistency.

Markov chain

As computing power is no longer worry, the Markov chain [33] was adopted to the Bayesian method. There are states which communicate one another with transition probabilities. A Markov chain describes a memoryless process of transition events, where any future state depends only on the present state and does not remember the past states [34,35]. The stochastic process $\{Z_1, Z_2, \ldots, Z_n, \ldots\}$ is a Markov chain if it satisfies

$$P(Z_{n+1}|Z_0, Z_1, Z_2, \ldots Z_n) = P(Z_{n+1}|Z_n).$$

A process moves from state to state according to the transition probability.

For example, let us consider a Markov chain with two states S_1 and S_2 as in **Fig. 1**. For transition probabilities $P_{11}=P(Z_{n+1}=1|Z_n=1)=0.3$, $P_{12}=P(Z_{n+1}=2|Z_n=1)=0.7$, $P_{21}=P(Z_{n+1}=1|Z_n=2)=0.4$, $P_{22}=P(Z_{n+1}=2|Z_n=2)=0.6$, the corresponding transition probability matrix is given by

$$\mathrm{P} = \begin{pmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{pmatrix}.$$

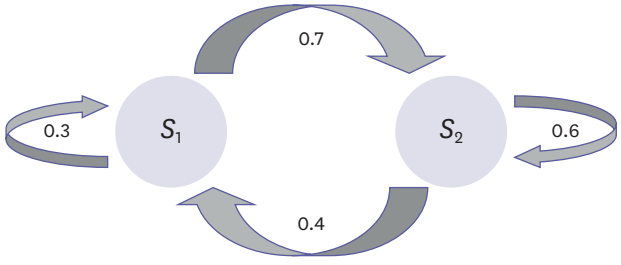

Figure 1. A Markov Chain with two states.

https://doi.org/10.12793/tcp.2023.31.e9

The rows represent the current state and the columns represent the future state. Let us start at an initial state π_0=(1,0) which means that the chain is located at S_1 with probability 1. Then, when n=1, it will be found at S_1 with probability 0.3 and at S_2 with probability 0.7, which can be calculated by multiplying π_0 and P as

$$\pi_1=\pi_0 P=(0.3,0.7)$$

$$\pi_2=\pi_1 P=(\pi_0 P)P=\pi_0 P^2=(0.37,0.63)$$

Let transition of the chain continue. Then, the probability π_n of being in states S_1 and S_2 at step n converges to the equilibrium distribution π as $n\rightarrow\infty$.

$$\pi_n=\pi_0 P^n\rightarrow\pi=(0.3636364,0.6363636)$$

It means that eventually the probability of chain being found at S_1 is 0.3636364 and the probability of chain being found at S_2 is 0.6363636. Regardless of the starting states, after an adequate number of transitions, the limiting probability of the Markov chain being found at any of the two states is π. Then, the general balance equation is as follows:

$$\pi P=\pi.$$

Samples from the Markov chains are not independent, which violates the well-known independence assumption of the classical likelihood statistics and traditional Bayesian methods. The Markov chain is said to be stationary if its statistical properties do not change according to time. Stationarity assures existence of the limiting distribution π, mean, and variance. A Markov chain is said to be ergodic if its time average approximates to the statistical mean, which implies its asymptotic unbiasedness. If a Markov chain is irreducible, aperiodic, and recurrent, it is ergodic. Stationarity and ergodicity of the Markov chain are required to assure existence of asymptotically unbiased estimator of the parameter [35].

Even though ignoring details does not hurt, let us mention brief definitions before moving to the next step. A Markov chain is irreducible if one state can be reached from every other state in finite time regardless of the initial sate, recurrent if all states are recurrent, aperiodic to ensure that the states of the whole chain are not partitioned into subsets with periodic recurrent states [36].

The MCMC algorithm

The major issue of the MCMC is stationarity and ergodicity so that the time average converges to a Bayes estimator which is a consistent estimator of θ [6,7,35,36]. The MCMC solves local or detailed balance equations instead of the big simultaneous general balance equation. For a pair of states S_i and S_j in the Markov chain and the target distribution π, let us consider the proposal probability of moving from the S_i to S_j as $P(S_i\rightarrow S_j)$. The Markov chain is said to be time reversible if The Markov chain satisfies detailed balance

$$\pi(S_i)P(S_i\rightarrow S_j)=\pi(S_j)P(S_j\rightarrow S_i), \qquad (3)$$

where both forward and the backward processes are the same. If detailed balance holds for every pair of states, then the Markov chain is known to be generally balanced and a stationary distribution θ exists.

Let us construct a stochastic process $\{x^{(0)},x^{(1)},\ldots,x^{(n)},\ldots\}$ from the target function θ. Keeping the detailed balance eq. (3), the MCMC first proposes a candidate and accept only as much as the acceptance probability to keep the detailed balance of the chain. At step n, the algorithm

https://doi.org/10.12793/tcp.2023.31.e9

proposes the chain to move from the current state $x^{(n)}$ to a new state y with the proposal probability $q(y|x^{(n)})$. For implementation, we first sample y from a proposal distribution $q(y|x^{(n)})$ as a candidate of a new chain $x^{(n+1)}$. Secondly, at the current state $x^{(n)}$ the Hasting's acceptance probability of y

$$\alpha(x^{(n)}, y) = \min\left\{1, \frac{\pi(y)q(x^{(n)}|y)}{\pi(x^{(n)})q(y|x^{(n)})}\right\} \quad (4)$$

is calculated. For symmetric proposal distributions like the uniform distribution, the normal distribution, t or Cauchy distribution, the proposal rate is $q(x^{(n)}|y)/q(y|x^{(n)})=1$. Finally, sample u from $U(0,1)$ and accept y as $x^{(n+1)}$ if $u<\alpha(x^{(n)},y)$. The acceptance probability is maintained because $P(U<\alpha)=\alpha$.

Metropolis-Hastings Algorithm

STEP 1. Choose an initial $x^{(0)}$.
STEP 2. Sample a proposal $y\sim q(y|x^{(n)})$.
STEP 3. Calculate $\alpha(x^{(n)},y)$.
STEP 4. Sample $u\sim U[0,1]$.
STEP 5. If $u<\alpha(x^{(n)},y)$,
then accept the proposal $x^{(n+1)}=y$.
Otherwise reject the proposal y.
Repeat STEP 2-5.

The detailed balance holds for the Hasting's α as mentioned in the lecture note "MCMC and Bayesian Modeling" of Haugh in 2017

$$\begin{aligned}
&\alpha(x,y)\pi(x)q(y|x)\\
&= \min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\}\pi(x)q(y|x)\\
&= \min\{\pi(x)q(y|x),\ \pi(y)q(x|y)\}\\
&= \min\left\{1, \frac{\pi(x)q(y|x)}{\pi(y)q(x|y)}\right\}\pi(y)q(x|y)\\
&= \alpha(y,x)\pi(y)q(x|y)
\end{aligned}$$

The MCMC uses the product of likelihood and prior as a sampling target function avoiding the exact posterior density. For $Z=X$ or $Z=Y$ or $Z=(X,Y)$, let us apply the MCMC to the Bayesian methods to the posterior density $p(\theta|z)$ as a target function $\pi(\theta)$. Then, the Hasting's acceptance probability is as follows.

$$\begin{aligned}
&\alpha(\theta^{(n)},\theta^{new})\\
&= \min\left\{1, \frac{p(\theta^{new}|z)\ q(\theta^{(n)}|\theta^{new})}{p(\theta^{(n)}|z)q(\theta^{new}|\theta^{(n)})}\right\}\\
&= \min\left\{1, \frac{\frac{p(z|\theta^{new})\ p(\theta^{new})}{f_Z(z)}q(\theta^{(n)}|\theta^{new})}{\frac{p(z|\theta^{(n)})p(\theta^{(n)})}{f_Z(z)}\ q(\theta^{new}|\theta^{(n)})}\right\}\\
&= \min\left\{1, \frac{p(z|\theta^{new})p(\theta^{new})q(\theta^{(n)}|\theta^{new})}{p(z|\theta^{(n)})p(\theta^{(n)})\ q(\theta^{new}|\theta^{(n)})}\right\}
\end{aligned}$$

https://doi.org/10.12793/tcp.2023.31.e9

The target function changes from the posterior density to the product of likelihood and the prior,

$$\text{Target Function } \pi = \text{Likelihood} \times \text{Prior} = p(z|\theta)p(\theta). \quad (5)$$

Therefore, the Bayesian methods with the MCMC are free from analytical or numerical integrations to get posterior density. After a burn-in process, the MCMC can produce an ergodic Markov chain and the Bayesian estimator $\hat{\theta}_B$ is obtained as the simple average of its simulated process.

HMC and No-U-Turn samplers (NUTSs)

The MCMC still has limitations in its computation. First, its efficiency is low since acceptance rate is sometimes as less as about 25% [37]. Secondly, the dependency between successive states in the MCMC has been a barrier towards improvement of its approximation accuracy. The HMC sampler weakens the chain dependency between successive states by employing the very classical Newton's law and efficiency has dramatically improved [10,29-31]. The HMC is known to be the most efficient because of speedy mixing error and small discretization error [3]. The Stan [38,39] and the Tensorflow [40] use the HMC as the default sampler. The HMC works well with the data in high dimension. The HMC uses differentials instead of direct integrals.

In a vector field of the Newton's force, the Hamiltonian H is the total energy of the system as sum of the potential energy and the kinetic energy. By the law of energy conservation, the H is fixed while the particles move around the system over time. Assume that the position coordinate of the particles is $\theta=(\theta_1,\theta_2,\ldots,\theta_d)$ and an auxiliary variable $p=(p_1,p_2,\ldots,p_d)\sim N(0,M)$ is their momentum coordinates satisfying $p_i=m_iv_i$ or $v_i=p_i/m_i$ for mass m_i and velocity v_i. The kinetic energy is $p_i^2/(2m_i)$. For the target distribution $\pi(\theta)$, the joint density $f(p,\theta)$ defines a Hamiltonian system as follows.

$$A\ Total\ Energy\ H(p,\theta) = -\log f(p,\theta) = -\log(\pi(\theta)f(p|\theta)) = -\log\pi(\theta) - \log f(p|\theta)$$

The potential energy $U(\theta)$ and the kinetic energy $K(p)$ are given by

$$U(\theta) = -\log\pi(\theta).$$

$$K(p) = -\log f(p|\theta) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i} = \frac{1}{2}p^T M^{-1} p$$

Thus, the total energy is

$$H(p,\theta) = U(\theta) + K(p) = -\log\pi(\theta) + \frac{1}{2}p^T M^{-1} p. \quad (6)$$

There are two differential equations, one for velocity and the other for force. In the Hamiltonian dynamics, the velocity v is the derivative of the position, velocity is momentum divided by mass $v_i=p_i/m_i$, and the velocity is the derivative of the kinetic energy $K(p)$. So, the first differential equation holds

$$Velocity\ v = \frac{\partial\theta}{\partial t} = M^{-1}p$$

where $M=diag(m_1,m_2,\ldots,m_d)$. The force F is also the derivative of the momentum. The potential energy is defined as the negative integral of the force F along the path. In other words, the force F is the negative gradient of the potential energy. Thus, the second differential equation holds

https://doi.org/10.12793/tcp.2023.31.e9

$$Force\ F = \frac{\partial p}{\partial t} = \nabla \log \pi(\theta).$$

Simultaneously solving the two differential equations for θ, instead of seeking an analytic solution, the leapfrog integrator is adopted to implement Hamiltonian system. The path of the particle is discretized and the two derivatives are approximated according to a small step movement $\partial t \approx \varepsilon > 0$. From the two differential equations of v and F, we can get the small updates of θ and p,

$$\partial\theta \approx (\partial t)\ M^{-1}p,\ \ \partial p \approx (\partial t)\ \nabla \log \pi(\theta).$$

For the Bayesian methods, the target function $\pi(\theta)$ is the posterior density and then

$$\nabla \log \pi(\theta) = \nabla(\log f(y|\theta) + \log p(\theta) - \log f_Y(y)) = \nabla(\log f(y|\theta) + \log p(\theta))$$

which only needs the likelihood $f(y|\theta)$ and the prior $p(\theta)$. In the HMC, a particle changes half momentum, full position, and then another half momentum as a routine. Before and after the full-step update of θ, the momentum p updates twice only by half-steps $\varepsilon/2$ at each time. Then, for the current state $(p^{(n)}, \theta^{(n)})$ and the proposal state (p, θ), the acceptance rate is

$$\alpha(p^{(n)}, \theta^{(n)}, p, \theta) = \min(1, \exp\{H(p^{(n)}, \theta^{(n)}) - H(p, \theta)\}). \qquad (7)$$

If the proposal is accepted, then the current parameter is replaced by the proposal. If the proposal is rejected, the current parameter is kept. At step 1, the auxiliary momentum is simulated and updated from the proposal normal distribution. For the size of L, the natural rule could be $\varepsilon L = 1$. The brief sketch of the HMC is as follows.

Hamiltonian Monte Carlo Algorithm

Given $\theta^{(0)}$, $\varepsilon, L, M, \log \pi(\theta)$

STEP 1. Simulate a proposal momentum $p^{(n)} \sim N(0, M)$

STEP 2. For $i = 1, \ldots, L$

$$p \leftarrow p + \frac{1}{2}\varepsilon \nabla \log \pi(\theta)$$

$$\theta \leftarrow \theta + \varepsilon M^{-1}p$$

$$p \leftarrow p + \frac{1}{2}\varepsilon \nabla \log \pi(\theta)$$

STEP 3. Calculate $\alpha(p^{(n)}, \theta^{(n)}, p, \theta)$.

STEP 4. Sample $u \sim U[0,1]$.

if $u < \alpha(p^{(n)}, \theta^{(n)}, p, \theta)$

then accept the proposal $\theta^{(n)} = \theta$.

Otherwise reject θ.

Repeat STEP 1 to STEP 4.

The HMC algorithm has the three essential tuning parameters, the covariance matrix M of z, the discretization step size ε, and the number of leap frog steps L. A poor choice of those would lead to a low efficiency of the HMC.

https://doi.org/10.12793/tcp.2023.31.e9

As an extension of the HMC, the NUTS prevents trajectories from going back to the direction of the starting position [12]. During the burn-in period, the NUTS controls the discretization step size ε and the covariance matrix M of the momentum p. Once ε and M are adapted and then fixed, the NUTS controls the number of leapfrog steps L at each iteration. The NUTS moves in time back and forth to assure detailed balance of the chain for stationarity and ergodicity. For details, read the section of HMC of the stan Reference Manual in its homepage.

Software Stan

As a descendant of the BUGS (Bayesian inference Using Gibbs Sampling) released in 2007, the probabilistic program language Stan implemented the Bayesian HMC algorithms in C++ [38-42] including the MCMC with approximate numerical methods, the penalized maximum likelihood test. The Stan was named after Stanislaw Ulam (1909–1984) who invented the Monte Carlo Method. The Stan has a concrete structure of blocks such as data, transformed data, parameters, transformed parameters, model, and some generated quantities. See the official website https://mc-Stan.org/. The R package rstan is available and another R package brms has implemented the Stan in the syntax of lme4 [43,44]. In R [45], MCMCpack [46], hmclearn [31], and more packages are available. These Bayesian softwares fit real-world data to complex nonlinear mixed effects models using various non-conjugate priors even with high dimensional covariance matrices.

Bayesian regression with the MCMC and the HMC

Let us consider how the MCMC and the HMC can be applied to a simple linear regression model without the intercept and nonlinear mixed effects models. Thomas and Tu [31] presented an example of multiple regression, logistic regression, and Poisson regression with random subject effects with the HMC. We review the simple linear regression model without the intercept, which is for the given data $(x_i, y_i), i=1,\ldots,n$,

$$y_i=\beta x_i+\varepsilon_i,\ i=1,\ldots,n$$

$$\varepsilon_i \sim iid\ N(0,\sigma^2).$$

Let us consider conjugate priors which $\beta \sim N(0,\sigma_\beta^2)$ and $\sigma^2 \sim IG(a,b), (\sigma^2>0)$, where σ_β^2, a,b are constant. The target parameter is $\theta=(\beta,\sigma^2)$. The target function is proportional to product of the likelihood and the priors,

$$p(\beta,\sigma^2|y) \propto f(y|\beta,\sigma^2)p\left(\beta|\sigma_\beta^2\right)p(\sigma^2|a,b),$$

$$\pi(\beta,\sigma^2|y)=\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta x_i)^2\right)\exp\left(-\frac{\beta^2}{2\sigma_\beta^2}\right)(\sigma^2)^{-a-1}\exp\left(-\frac{b}{\sigma^2}\right).$$

We sample σ^2 from $IG(a,b)$, sample β from $N(0,\sigma_\beta^2)$, calculate the target function π, and the acceptance ratio α in eq. (4) assuming the proposal distribution as $U(0,1)$ or others. We repeat the iteration until enough (β,σ^2) are accepted after a burn-in process.

For the HMC to be applied to the simple linear regression [31], $\pi(\theta)$ and $\nabla\log\pi(\theta)$ are necessary. The target function $\pi(\theta)$ is the product of the likelihood and the prior. For $\sigma^2>0$, we take a logarithmic transformation

$$\gamma=\log\sigma^2,\ \ d(\sigma^2)=e^\gamma d\gamma$$

to have $\gamma \in (-\infty,\infty)$. Then the log posterior density and the target function are

https://doi.org/10.12793/tcp.2023.31.e9

$$\log p(\beta,\gamma|y) = \log f(y|\beta,\gamma) + \log p\left(\beta\middle|\sigma_\beta^2\right) + \log p(\gamma|a,b) + Constant(y)$$

$$\log \pi(\beta,\gamma|y) = -\left(\frac{n}{2}+a\right)\gamma - \frac{e^{-\gamma}}{2}\sum_{i=1}^{n}(y_i-\beta x_i)^2 - be^{\gamma} - \frac{\beta^2}{2\sigma_\beta^2}.$$

We need its gradient which is going to be done as a part of the algorithm

$$\frac{\partial \log \pi(\beta,\gamma|y)}{\partial \beta} = e^{-\gamma}\sum_{i=1}^{n}(y_i-\beta x_i)x_i - \frac{\beta}{\sigma_\beta^2}$$

$$\frac{\partial \log \pi(\beta,\gamma|y)}{\partial \gamma} = -\left(\frac{n}{2}+a\right) + \frac{e^{-\gamma}}{2}\sum_{i=1}^{n}(y_i-\beta x_i)^2 + b\,e^{-\gamma}.$$

We get β and γ by solving the differential equations using the leapfrog integrator in the HMC. The HMC samples p from $N(0,M)$. Calculate H in eq. (6) and α in eq. (7).

Let us consider a multilevel non-linear mixed effects model as before [3],

$$y=f(t,\theta)+\varepsilon$$

$$\theta=F(\beta,x)+\eta$$

$$\varepsilon\sim N(0,\sigma^2 I),\ \eta\sim N(0,\omega^2 I)$$

The posterior density $p(\theta,\sigma^2,\omega^2|y)$ is proportional to product of conditional densities and priors which is the target function such as

$$p(\beta,\sigma^2,\omega^2|y) = \frac{f_{Y,\Theta}(y,\beta,\sigma^2,\omega^2)}{f_Y(y)} \propto \pi(\beta,\sigma^2,\omega^2|y) = f_Y(y|\beta,\sigma^2,\omega^2)\, f_\Theta(\theta|\,\omega^2)p(\beta)p(\sigma^2)p(\omega^2).$$

$f_Y(y|\theta,\sigma^2)$ is $N(f(t,\theta),\sigma^2 I)$ and $f_\Theta(\theta|\omega^2)$ is $N(F(\beta,x),\omega^2 I)$. Using the Gibbs sampling method again, we simulate β from $p(\beta)$, σ^2 from $p(\sigma^2)$, ω^2 from $p(\omega^2)$ [3]. Then we calculate $f_Y(y|\theta,\sigma^2)$, $f_\Theta(\theta|\omega^2)$, the target function π and the acceptance ratio α in eq. (4) assuming the proposal distribution as $U(0,1)$ or others.

Now let us look at the HMC [3]. We take gradient of log-likelihood and logarithm of priors as p_i from $N(0,m_i)$ for the given diagonal variance matrix $M=diag(m_i,\ldots,m_d)$,

$$-\log \pi(\theta|\sigma^2,\Omega,y) = \frac{1}{2\sigma^2}\|y-f(t,\theta)\|^2 + \frac{1}{2\omega^2}\|\theta-F(\beta,x)\|^2 + \log p(\beta) + \log p(\sigma^2) + \log p(\omega^2)$$

The HMC gets its gradient $\nabla\log\pi$, samples p from $N(0,M)$, and solves the differential equations using the leapfrog integrator. It calculates H in eq. (6) and α in eq. (7). Lee [3] reviewed details of the HMC applied to Bayesian nonlinear models for repeated measure data including PK/PD models.

Once the point estimators are acquired, the fitted value can be evaluated with the point estimators substituted in the model. Then, the residuals are evaluated as the difference between the observed and fitted values. For further model selection, the goodness-of-fit criteria like the MSE, the likelihood function, the log-likelihood function, AIC, BIC, WAIC, WBIC, BF, PO, and so on can be consequently calculated.

Both the MCMC and the HMC algorithms sample from only shape of the posterior density without knowing the exact posterior density and without linearization of non-linear

https://doi.org/10.12793/tcp.2023.31.e9

models. Since the resulting chain is supposed to be stationary and ergodic, a time average is an estimate of the posterior mean. They do not calculate integrals analytically, which is beneficiary to complex models in high dimension. Lee [3] has given more thorough details of application of the Bayes approach to pharmacometrics.

DISCUSSION

The recent Bayesian samplers like the MCMC and the HMC have opened a new era of the data analysis which can fit the complex real-world models without worrying about closed-from of the posterior density in high dimension. In addition, various informative non-conjugate priors can be applied for complex multidimensional covariance. Especially when the priors are the Uniform distribution, the target function is the likelihood function and the Bayesian samplers can be applied intactly to classical likelihood statistics. Pharmacometricians are great beneficiaries since even the simplest one-compartment model is the form of nonlinear mixed effects model using the complex restricted maximum likelihood method. As much as advantage of the Bayesian method is outstanding, its theoretical background is widened to classical statistical inference, stochastic processes, and computational algorithms. Therefore, a review can always help pharmacometricians understand the Bayesian inference. This article has provided a preliminary outline of statistical inference and the Bayesian samplers like the MCMC and HMC in view of their stationarity and ergodicity.

ACKNOWLEDGEMENTS

The author appreciates Prof. Tae-Yoon Kim at Statistics, Keimyung University and Ph. D student Hane Lee at Statistics, Columbia University for their professional advices.

REFERENCES

1. Rowland M, Tozer TN. Clinical pharmacokinetics/pharmacodynamics. Philadelphia (PA): Lippincott Williams and Wilkins; 2005.
2. Gabrielsson J, Weiner D. Pharmacokinetic and Pharmacodynamic data analysis: concepts and applications. Boca Raton (FL): CRC Press; 2001.
3. Lee SY. Bayesian nonlinear models for repeated measurement data: an overview, implementation, and applications. Mathematics 2022;10:898.
CROSSREF
4. Price R. An essay towards solving a problem in the doctrine of chances. By the Late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a Letter to John Canton, A.M.F.R.S. Philos Trans R Soc Lond 1763;53:370-418.
CROSSREF
5. Pinheiro JC, Bates DM. Mixed-effects models in S and S-PLUS. New York (NY): Springer; 2004.
6. Metropolis NA, Rosenbluth M, Teller RM, Teller E. Equations of state calculations by fast computing machines. J Chem Phys 1953;21:1087-1092.
CROSSREF
7. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 1970;57:97-109.
CROSSREF
8. Chib S, Greenberg E. Understanding the metropolis-hastings algorithm. Am Stat 1995;49:327-335.
9. Brooks S, Gelman A, Jones G, Meng XL. Handbook of Markov chain Monte Carlo. Boca Raton (FL): CRC Press; 2011.

10. Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo. arXiv. July 16, 2018.
CROSSREF
11. Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics, 3rd ed. New York (NY): McGraw-Hill, Inc.; 1974.
12. Hogg R, Tanis E. Probability and statistical inference. 3rd ed. New York (NY): Macmillan Publishing Company; 1988.
13. Casella G, Berger RB. Statistical inference. 2nd ed. London: Duxbury Press; 2001.
14. Rice JA. Mathematical statistics and data analysis. Boston (MA): Cengage Learning; 2007.
15. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1974;19:716-723.
CROSSREF
16. Akaike H. Bayesian extension of the minimum AIC procedure. Biometrika 1979;66:237-242.
CROSSREF
17. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York (NY): Springer; 2017.
18. James G, Witten D, Hastie T, Tishrani R. An introduction to statistical learning with application in R, 2nd ed. New York (NY): Springer; 2021.
19. Beal SL, Sheiner LB, Boeckmann A, Bauer RJ. NONMEM users guides; NONMEM project group. San Francisco (CA): University of California; 1992.
20. Kim MG, Yim DS, Bae KS. R-based reproduction of the estimation process behind NONMEM Part 1: first-order approximation method. Transl Clin Pharmacol 2015;23:1-7.
CROSSREF
21. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 1977;39:1-38.
CROSSREF
22. Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 1938;9:60-62.
CROSSREF
23. Watanabe S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 2010;11:3571-3594.
24. Jeffreys H. An invariant form for the prior probability in estimation problems. Proc R Soc Lond A Math Phys Sci 1946;186:453-461.
PUBMED | CROSSREF
25. Tierney L, Kadane JB. Accurate approximations for posterior moments and marginal densities. J Am Stat Assoc 1986;81:82-86.
26. Metropolis N, Ulam S. The Monte Carlo method. J Am Stat Assoc 1949;44:335-341.
PUBMED | CROSSREF
27. Eckhardt R. Stan Ulam, John Von Neumann, and the Monte Carlo Method. Los Alamos Sci 1987;15:131-143.
28. Neal RM. Slice sampling. Ann Stat 2003;31:705-767.
CROSSREF
29. Neal RM. MCMC using Hamiltonian dynamics. Handbook of Markov chain Monte Carlo. New York (NY): Chapman and Hall/CRC; 2011.
30. Betancourt M, Girolami M. Hamiltonian Monte Carlo for hierarchical models. arXiv. December 3, 2013.
CROSSREF
31. Thomas S, Tu W. Learning Hamiltonian Monte Carlo in R. arXiv. December 18, 2020.
CROSSREF
32. Casella G, Robert CP, Wells MT. Generalized accept-reject sampling schemes. IMS Lecture Notes Monogr Ser 2004;45:342-347.
CROSSREF
33. Markov AA. Extension of the law of large numbers to dependent events. Bull Soc Phys Math 1906;2:155-156.
34. Ross SM. Stochastic processes. 2nd ed. New York (NY): John Wiley and Sons, Inc.; 1996.
35. Peebles PZ. Probability, random variables, random signal principles. 4th ed. New York (NY): McGraw Hill; 2001.
36. Aldous D, Fill JA. Reversible Markov Chains and random walks on graphs [Internet]. https://www.stat.berkeley.edu/~aldous/RWG/book.pdf. Accessed January 28, 2023.
37. Bedard M. Optimal acceptance rates for metropolis algorithms: moving beyond 0.234. Stoch Process Their Appl 2008;118:2198-2222.
CROSSREF

38. Carpenter B, Hoffman MD, Brubaker M, Lee D, Li P, Betancourt M. The Stan math library: reverse-mode automatic differentiation in C++'. arXiv. September 23, 2015.
CROSSREF

39. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. J Stat Softw 2017;76:1-32.
PUBMED | CROSSREF

40. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16); November 204, 2016; Savannah, GA. Berkeley (CA): USENIX Association; 2016, 265-283.

41. Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 2014;15:1593-1623.

42. Stan Development Team. Stan modeling language user's guide and reference manual, v.2.9.0 [Internet]. http://mc-stan.org/. Accessed January 28, 2023.

43. Burker PC. An R package for Bayesian multilevel models using Stan. J Stat Softw 2017;80:1-28.

44. Buerkner PC. Advanced Bayesian multilevel modeling with the R package brms. R J 2018;10:395-411.
CROSSREF

45. R Core Team. R: a language and environment for statistical computing [Internet]. https://www.R-project.org/. Accessed January 28, 2023.

46. Martin AD, Quinn KM, Park JH. MCMCpack: Markov chain Monte Carlo in R. J Stat Softw 2011;42:22.
CROSSREF