Statistics and Computing (2022) 32:7
https://doi.org/10.1007/s11222-021-10070-2

Stochastic approximation cut algorithm for inference in modularized Bayesian models

Yang Liu[1] · **Robert J. B. Goudie**[1]

Received: 22 April 2021 / Accepted: 6 November 2021

Abstract

Bayesian modelling enables us to accommodate complex forms of data and make a comprehensive inference, but the effect of partial misspecification of the model is a concern. One approach in this setting is to modularize the model and prevent feedback from suspect modules, using a cut model. After observing data, this leads to the cut distribution which normally does not have a closed form. Previous studies have proposed algorithms to sample from this distribution, but these algorithms have unclear theoretical convergence properties. To address this, we propose a new algorithm called the stochastic approximation cut (SACut) algorithm as an alternative. The algorithm is divided into two parallel chains. The main chain targets an approximation to the cut distribution; the auxiliary chain is used to form an adaptive proposal distribution for the main chain. We prove convergence of the samples drawn by the proposed algorithm and present the exact limit. Although SACut is biased, since the main chain does not target the exact cut distribution, we prove this bias can be reduced geometrically by increasing a user-chosen tuning parameter. In addition, parallel computing can be easily adopted for SACut, which greatly reduces computation time.

Keywords Cutting feedback · Stochastic approximation Monte Carlo · Intractable normalizing functions · Discretization

1 Introduction

Bayesian models mathematically formulate our beliefs about the data and parameter. Such models are often highly structured models that represent strong assumptions. Many of the desirable properties of Bayesian inference require the model to be correctly specified. We say a set of models $f(x|\theta)$, where $\theta \in \Theta$, are misspecified if there is no $\theta_0 \in \Theta$ such that data X is independently and identically generated from $f(x|\theta_0)$ (Walker 2013). In practice, models will inevitably fall short of covering every nuance of the truth. One popular approach when a model is misspecified is fractional (or power) likelihood. This can be used in both classical (e.g. Nakaya et al. 2005; Huang et al. 2010; Liu et al. 2018) and Bayesian (e.g. Miller and Dunson 2019; Bhattacharya et al. 2019) frameworks. However, this method treats all of the models as equally misspecified.

We consider the situation when the assumptions of the model are thought to partially hold: specifically, we assume that one distinct component (or module in the terminology of Liu et al. 2009) is thought to be incorrectly specified, whereas the other component is correctly specified. In standard Bayesian inference, these distinct modules are linked by Bayes' theorem. Unfortunately, this means the reliability of the whole model may be affected even if only one component is incorrectly specified. To address this, in this paper we adopt the idea of "cutting feedback" (Lunn et al. 2009b; Liu et al. 2009; Plummer 2015; Jacob et al. 2017, 2020) which modifies the links between modules so that estimation of non-suspect modules is unaffected by information from suspect modules. This idea has been used in a broad range of applications including the study of population pharmacokinetic/pharmacodynamic (PK/PD) models (Lunn et al. 2009a), analysis of computer models (Liu et al. 2009),

Yang Liu was supported by a Cambridge International Scholarship from the Cambridge Commonwealth, European and International Trust. Robert J.B. Goudie was funded by the UK Medical Research Council [programme code MC_UU_00002/2]. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

✉ Yang Liu
yang.liu@mrc-bsu.cam.ac.uk

Robert J. B. Goudie
robert.goudie@mrc-bsu.cam.ac.uk

[1] MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Published online: 06 December 2021

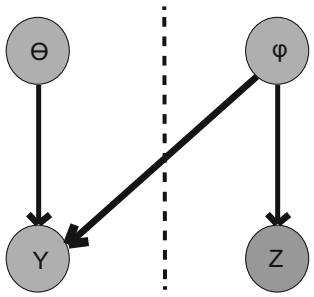

Fig. 1 DAG representation of a generic two-module model. The two modules are separated by a dashed line

Bayesian estimation of causal effects with propensity scores (McCandless et al. 2010; Zigler 2016) and Bayesian analysis of health effect of air pollution (Blangiardo et al. 2011).

Consider the generic two-module model with observable quantities (data) Y and Z and parameters θ and φ, shown in the directed acyclic graph (DAG) in Fig. 1. The joint distribution is

$$p(Y, Z, \theta, \varphi) = p(Y|\theta, \varphi)p(Z|\varphi)p(\theta)p(\varphi),$$

and the standard Bayesian posterior, given observations of Y and Z, is

$$\begin{aligned} p(\theta, \varphi|Y, Z) &= p(\theta|Y, \varphi)p(\varphi|Y, Z) \\ &= \frac{p(Y|\theta, \varphi)p(\theta)}{p(Y|\varphi)} \frac{p(Y|\varphi)p(Z|\varphi)p(\varphi)}{p(Y, Z)}. \end{aligned}$$

Suppose we are confident that the relationship between φ and Z is correctly specified but not confident about the relationship between φ and Y. To prevent this possible misspecification affecting estimation of φ, we can "cut" feedback by replacing $p(\varphi|Y, Z)$ in the standard posterior with $p(\varphi|Z)$, making the assumption that φ should be solely estimated by Z,

$$\begin{aligned} p_{\text{cut}}(\theta, \varphi) &:= p(\theta|Y, \varphi)p(\varphi|Z) \\ &= \frac{p(Y|\theta, \varphi)p(\theta)}{p(Y|\varphi)} \frac{p(Z|\varphi)p(\varphi)}{p(Z)}. \end{aligned} \tag{1}$$

We call (1) the "cut distribution". The basic idea of cutting feedback is to allow information to "flow" in the direction of the directed edge, but not in the reverse direction (i.e. a "valve" is added to the directed edge).

Sampling directly from $p_{\text{cut}}(\theta, \varphi)$ is difficult because the marginal likelihood $p(Y|\varphi) = \int p(Y|\theta, \varphi)p(\theta)\mathrm{d}\theta$ depends on a parameter of interest φ and is not usually analytically tractable, except in the simple case when $p(\theta)$ is conditionally conjugate to $p(Y|\theta, \varphi)$, which we do not wish to assume. This intractable marginal likelihood is a conditional posterior normalizing constant: it is the normalizing function for the posterior distribution $p(\theta|Y, \varphi)$, conditional on φ, of a parameter θ of interest:

$$p(\theta|Y, \varphi) = \frac{p(Y, \theta|\varphi)}{p(Y|\varphi)}. \tag{2}$$

This differs importantly to intractable likelihood normalizing constants, as discussed in the doubly intractable literature (e.g. Park and Haran 2018), in which the normalizing function $H(\varphi) = \int h(Y|\varphi)dY$ for the likelihood is intractable.

$$p(Y|\varphi) = \frac{h(Y|\varphi)}{H(\varphi)}.$$

The normalizing function $H(\varphi)$ is obtained by marginalizing the likelihood, with respect to the observable quantity Y, in contrast to the normalizing function $p(Y|\varphi)$, which is obtained by marginalizing likelihood $p(Y, \theta|\varphi)$ with respect to a parameter θ of interest. This difference means that standard methods for doubly intractable problems (e.g. Møller et al. 2006; Murray et al. 2006; Liang 2010; Liang et al. 2016), which introduce an auxiliary variable, with the same distribution (or proposal distribution) as the distribution of the *a posteriori* observed and fixed Y to cancel the intractable normalizing function shared by them, do not directly apply to (2).

A simple algorithm that aims to draw samples from $p_{\text{cut}}(\theta, \varphi)$ is implemented in WinBUGS (Lunn et al. 2009b). It is a Gibbs-style sampler that involves updating θ and φ with a pair of transition kernels $q(\theta'|\theta, \varphi')$ and $q(\varphi'|\varphi)$ that satisfy detailed balance with $p(\theta|Y, \varphi')$ and $p(\varphi|Z)$, respectively. However, the chain constructed by the WinBUGS algorithm may not have the cut distribution as its stationary distribution (Plummer 2015) since

$$\int p_{\text{cut}}(\theta, \varphi)q(\theta'|\theta, \varphi')q(\varphi'|\varphi)\mathrm{d}\theta\mathrm{d}\varphi = w(\theta', \varphi')p_{\text{cut}}(\theta', \varphi'),$$

where the weight function w is

$$w(\theta', \varphi') = \int \frac{p(\theta|Y, \varphi)}{p(\theta|Y, \varphi')} q(\varphi|\varphi')q(\theta|\theta', \varphi')\mathrm{d}\theta\mathrm{d}\varphi.$$

The WinBUGS algorithm is inexact since $w(\theta', \varphi') \neq 1$, except in the simple case (conditionally conjugate) when it is possible to draw exact Gibbs updates from $p(\theta'|Y, \varphi')$. Plummer (2015) proposed two algorithms that address this problem by satisfying $w(\theta', \varphi') = 1$ approximately. One is a nested MCMC algorithm, which updates θ from $p(\theta'|Y, \varphi')$ by running a separate internal Markov chain with transition kernel $q^*(\theta'|\theta, \varphi')$ satisfying detailed balance with the target distribution $p(\theta|Y, \varphi')$. The other is a linear path algorithm, which decomposes the complete MCMC move from (θ, φ) to (θ', φ') into a series of substeps along a linear path from

φ to φ' and drawing a new θ at each substep. However, these methods require either the length of the internal chain or the number of substeps to go to infinity, meaning that in practice, these algorithms will not necessarily converge to p_{cut}.

In this article, we propose a new sampling algorithm for $p_{\text{cut}}(\theta, \varphi)$, called the stochastic approximation cut (SACut) algorithm. Since φ can be easily sampled from tractable part $p(\varphi|Z)$, our algorithm aims to sample θ from the intractable part $p(\theta|Y, \varphi)$. Our algorithm is divided into two chains that are run in parallel: the main chain that approximately targets $p_{\text{cut}}(\theta, \varphi)$ and an auxiliary chain that is used to form a proposal distribution for $\theta|\varphi$ in the main chain (Fig. 2b). The auxiliary chain uses stochastic approximation Monte Carlo (SAMC) (Liang et al. 2007) to approximate the intractable marginal likelihood $p(Y|\varphi)$ and subsequently form a discrete set of distribution $p(\theta|Y, \varphi)$ for each $\varphi \in \Phi_0 = \{\varphi_0^{(1)}, \ldots, \varphi_0^{(m)}\}$, a set of pre-selected auxiliary parameters (Fig. 2a).

The basic "naive" form of our algorithm has convergence in distribution, but stronger convergence properties can be obtained by building a proposal distribution $p_n^{(\kappa)}(\theta|Y, \varphi)$ to target an approximation $p^{(\kappa)}(\theta|Y, \varphi)$ instead of the true distribution $p(\theta|Y, \varphi)$ (Fig. 2c, d). We prove a weak law of large numbers for the samples $\{(\theta_n, \varphi_n)\}_{n=1}^N$ drawn from the main chain. We also prove that the bias due to targeting $p^{(\kappa)}(\theta|Y, \varphi)$ can be controlled by the precision parameter κ, and that the bias decreases geometrically as κ increases. Our algorithm is inspired by the adaptive exchange algorithm (Liang et al. 2016), but replaces the exchange step with a direct proposal distribution for θ given φ in the main chain.

2 Main result

Let the product space $\Theta \times \Phi$ be the supports of θ and φ under p_{cut}. We assume the following throughout for simplicity.

Assumption 1 (a) Θ and Φ are compact and (b) p_{cut} is continuous with respect to θ and φ over $\Theta \times \Phi$.

Assumption 1(a) is restrictive, but is commonly assumed in the study of adaptive Markov chains (Haario et al. 2001). In most applied settings, it is reasonable to choose a prior for θ and φ with support on only a compact domain, making the domain of the cut distribution compact without any alteration to the likelihood. Note that Assumption 1 implies that p_{cut} is bounded over $\Theta \times \Phi$. From now on, define a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Denote Lebesgue measure μ on Θ and Φ and let P_{cut} be the measure on $\Theta \times \Phi$ defined by its density p_{cut}.

In following sections, we describe the construction of the algorithm. The naive version of our algorithm builds a discrete proposal distribution for θ, based on Liang et al. (2016). Note that, in Liang et al. (2016), this proposal distribution only draws auxiliary variables that are discarded once the parameter of interest is drawn, and so strong convergence results for samples drawn by this probability distribution are not needed by Liang et al. (2016). This does not always apply to our problem since θ is the parameter of interest and its parameter space can be continuous. The naive version does not allow us to prove stronger convergence and theoretical properties, so we apply a simple function approximation technique with a specially designed partition of the parameter space that enables a straightforward implementation. Although this approximation leads to bias, we show that it can be controlled.

2.1 Naive stochastic approximation cut algorithm

To introduce ideas that we will use in Sect. 2.3, we first describe a naive version of the stochastic approximation cut algorithm. The overall naive algorithm (Algorithm 1) is divided into two chains that are run in parallel.

The auxiliary chain $h_n = (\tilde{\theta}_n, \tilde{\varphi}_n)$, $n = 0, 1, 2, \ldots$, uses stochastic approximation Monte Carlo (Liang et al. 2007) to estimate $p(Y|\varphi)$ at a set of m pre-selected auxiliary parameter values $\Phi_0 = \{\varphi_0^{(1)}, \ldots, \varphi_0^{(m)}\}$. As we detail in supplementary materials, the set Φ_0 is chosen from a set of MCMC samples drawn from $p(\varphi|Z)$, chosen using the Max–Min procedure (Liang et al. 2016) that repeatedly adds the sample that has the largest Euclidean distance to the hitherto selected Φ_0. This ensures that Φ_0 covers the major part of the support of $p(\varphi|Z)$. A reasonably large m ensures the distribution $p(\theta|Y, \varphi)$ overlaps each other for neighbouring $\varphi_0^{(a)}$ and $\varphi_0^{(b)}$. The target density for $(\tilde{\theta}, \tilde{\varphi}) \in \Theta \times \Phi_0$, which is proportional to $p(\theta|Y, \varphi)$ in (1) at the values Φ_0, is

$$p(\tilde{\theta}, \tilde{\varphi}) = \frac{1}{m} \sum_{i=1}^{m} \frac{p(Y|\tilde{\theta}, \varphi_0^{(i)}) p(\tilde{\theta})}{p(Y|\varphi_0^{(i)})} \mathbb{1}_{\{\tilde{\varphi}=\varphi_0^{(i)}\}}. \tag{3}$$

Given proposal distributions $q_1(\tilde{\theta}'|\tilde{\theta})$ (e.g. symmetric random walk proposal) and $q_2(\tilde{\varphi}'|\tilde{\varphi})$ (e.g. uniformly drawing $\tilde{\varphi}'$ from a neighbouring set of $\tilde{\varphi}$) for $\tilde{\theta}$ and $\tilde{\varphi}$ individually, at each iteration n, proposals $\tilde{\theta}'$ and $\tilde{\varphi}'$ are drawn from a mixture proposal distribution, with a fixed mixing probability p_{mix},

$$q(\tilde{\theta}', \tilde{\varphi}'|\tilde{\theta}_{n-1}, \tilde{\varphi}_{n-1}) = \begin{cases} p_{\text{mix}} q_1(\tilde{\theta}'|\tilde{\theta}_{n-1}), & \text{for } \tilde{\theta}' \neq \tilde{\theta}_{n-1} \\ (1 - p_{\text{mix}}) q_2(\tilde{\varphi}'|\tilde{\varphi}_{n-1}), & \text{for } \tilde{\varphi}' \neq \tilde{\varphi}_{n-1} \\ 0, & \text{otherwise} \end{cases}$$

and accepted according to the Metropolis–Hastings acceptance probability with an iteration-specific target

$$p_n(\tilde{\theta}, \tilde{\varphi}) \propto \sum_{i=1}^{m} \frac{p(Y|\tilde{\theta}, \varphi_0^{(i)}) p(\tilde{\theta})}{\tilde{w}_{n-1}^{(i)}} \mathbb{1}_{\{\tilde{\varphi}=\varphi_0^{(i)}\}}, \ \tilde{\theta} \in \Theta, \ \tilde{\varphi} \in \Phi_0.$$

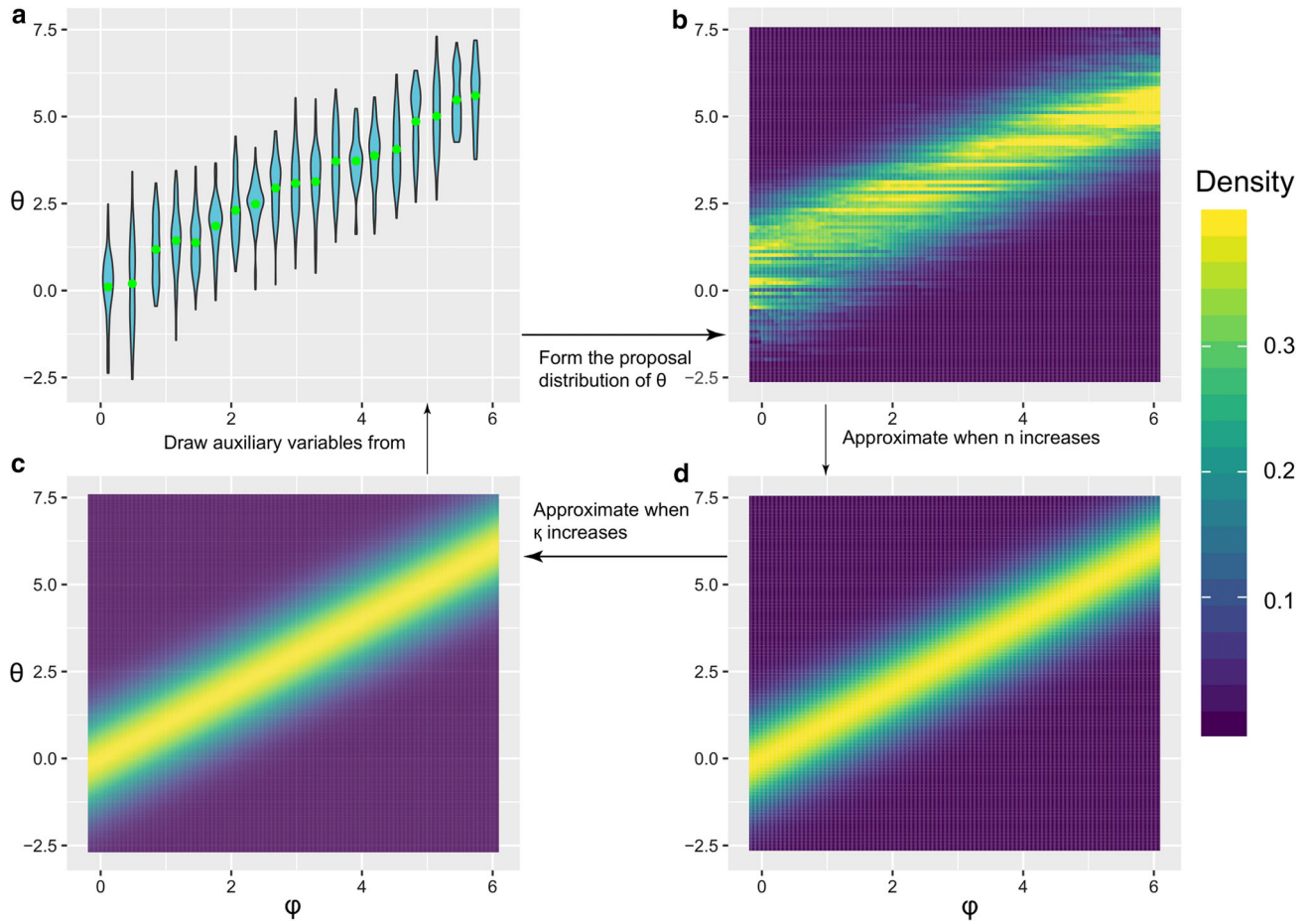

Fig. 2 Relationship between $p(\theta|Y,\varphi_0^{(i)})$, $p(\theta|Y,\varphi)$, $p^{(\kappa)}(\theta|Y,\varphi)$ and $p_n^{(\kappa)}(\theta|Y,\varphi)$. This is a toy example when the conditional distribution of θ, given $Y=1$ and φ, is $\mathrm{N}(\varphi, Y^2)$. Samples of the auxiliary variable $\tilde{\theta}$ are drawn from a mixture of discretized densities $p(\theta|Y,\varphi_0^{(i)})$, $i=1,\ldots,m$, shown in the violin plot in **a**, with the green dots showing the median of each component (see Sect. 2.1). Then, $p_n^{(\kappa)}(\theta|Y,\varphi)$, shown in **b**, is formed by using these auxiliary variables (see Sect. 2.2). Lemma 1 (Sect. 2.3) shows that $p_n^{(\kappa)}(\theta|Y,\varphi)$ converges to $p^{(\kappa)}(\theta|Y,\varphi)$, which is shown in **d**, while (8) shows that $p^{(\kappa)}(\theta|Y,\varphi)$ converges to the original density $p(\theta|Y,\varphi)$, shown in **c**

Here, $\tilde{w}_n^{(i)}$ is the estimate of $p(Y|\varphi_0^{(i)})$, $i=1,\ldots,m$, up to a constant, and $\tilde{w}_n=(\tilde{w}_n^{(1)},\ldots,\tilde{w}_n^{(m)})$ is a vector of these estimates at each of the pre-selected auxiliary parameter values Φ_0. We set $\tilde{w}_0^{(i)}=1$, $i=1,\ldots,m$ at the start. As described in Liang et al. (2007) and Liang et al. (2016), the estimates are updated by

$$\log(\tilde{w}_n^{(i)})=\log(\tilde{w}_{n-1}^{(i)})+\xi_n(e_{n,i}-m^{-1}),\ i=1,\ldots,m, \tag{4}$$

where $e_{n,i}=1$ if $\tilde{\varphi}_n=\varphi_0^{(i)}$ and $e_{n,i}=0$ otherwise, and $\xi_n=n_0/\max(n_0,n)$ decreases to 0 when n goes to infinity (the shrink magnitude n_0 is a user-chosen fixed constant). Note that in this auxiliary chain, when the number of iterations is sufficiently large, we are drawing (θ,φ) from (3). Hence, by checking whether the empirical sampling frequency of each $\varphi_0^{(i)}\in\Phi_0$ strongly deviates from m^{-1}, we can detect potential non-convergence of the auxiliary chain.

In the main Markov chain (θ_n,φ_n), $n=1,2,\ldots$, we draw φ' from a proposal distribution $q(\varphi'|\varphi)$ and then draw θ' according to a random measure

$$P_n^*(\theta\in\mathcal{B}|Y,\varphi')=\frac{\sum_{j=1}^n\sum_{i=1}^m\tilde{w}_{j-1}^{(i)}\frac{p(Y|\tilde{\theta}_j,\varphi')}{p(Y|\tilde{\theta}_j,\varphi_0^{(i)})}\mathbb{1}_{\{\tilde{\theta}_j\in\mathcal{B},\varphi_0^{(i)}=\tilde{\varphi}_j\}}}{\sum_{j=1}^n\sum_{i=1}^m\tilde{w}_{j-1}^{(i)}\frac{p(Y|\tilde{\theta}_j,\varphi')}{p(Y|\tilde{\theta}_j,\varphi_0^{(i)})}\mathbb{1}_{\{\varphi_0^{(i)}=\tilde{\varphi}_j\}}}, \tag{5}$$

where $\mathcal{B}\subset\Theta$ is any Borel set. Given a φ, the random measure (5) is formed via a dynamic importance sampling procedure proposed in Liang (2002) with intention to approximate the unknown distribution $p(\theta|Y,\varphi)$ (see supplementary material for a detailed explanation). For any Borel set $\mathcal{B}\subset\Theta$, we have

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m}\tilde{w}_{j-1}^{(i)}\frac{p(Y|\tilde{\theta}_j,\varphi)p(\tilde{\theta}_j)}{p(Y|\tilde{\theta}_j,\varphi_0^{(i)})p(\tilde{\theta}_j)}\mathbb{1}_{\{\tilde{\theta}_j\in\mathcal{B},\varphi_0^{(i)}=\tilde{\varphi}_j\}}$$
$$\to\sum_{i=1}^{m}\int_{\mathcal{B}}mp(Y|\varphi_0^{(i)})\frac{p(Y|\theta,\varphi)p(\theta)}{p(Y|\theta,\varphi_0^{(i)})p(\theta)}\frac{1}{m}\frac{p(Y|\theta,\varphi_0^{(i)})p(\theta)}{p(Y|\varphi_0^{(i)})}\mathrm{d}\theta$$
$$=m\int_{\mathcal{B}}p(Y|\theta,\varphi)p(\theta)\mathrm{d}\theta,$$

and similarly, the denominator of (5) converges to the $mp(Y|\varphi)$. Hence, by Lemma 3.1 of Liang et al. (2016), since $\Theta\times\Phi$ is compact, for any Borel set $\mathcal{B}\subset\Theta$ and on any outcome ω of probability space Ω, we have:

$$\lim_{n\to\infty}\sup_{\varphi\in\Phi}\left|P_n^*(\theta\in\mathcal{B}|Y,\varphi)-\int_{\mathcal{B}}p(\theta|Y,\varphi)\mathrm{d}\theta\right|=0. \tag{6}$$

This implies that the distribution of $\{\theta_n\}$, drawn from (5), converges in distribution to $p(\theta|Y,\varphi)$, and this convergence occurs uniformly over Φ. Note that the probability measure $P_n^*(\theta\in\mathcal{B}|Y,\varphi')$ is adapted to filtration $\mathcal{G}_n=\sigma(\cup_{j=1}^n(\tilde{\theta}_j,\tilde{\varphi}_j,\tilde{w}_j))$ on $(\Omega,\mathcal{F},\mathbb{P})$ and has a Radon–Nikodym derivative with respect to a mixture of Dirac measures determined by $\tilde{\Theta}_n=\cup_{j=1}^n\{\tilde{\theta}_j\}$ (Gottardo and Raftery 2008), because it is the law of a discrete random variable defined on $\tilde{\Theta}_n$. To achieve stronger convergence results, we will build a continuous probability distribution based on (5).

Algorithm 1: Naive stochastic approximation cut algorithm

Initialize at starting points $h_0=(\tilde{\theta}_0,\tilde{\varphi}_0)$, $\tilde{w}_0$ and (θ_0,φ_0);
For $n=1,\ldots,N$;

(a) Auxiliary chain:

(1) Draw a proposal $(\tilde{\theta}',\tilde{\varphi}')$ according to $q(\tilde{\theta}',\tilde{\varphi}'|\tilde{\theta}_{n-1},\tilde{\varphi}_{n-1})$.
(2) Accept the proposal, and set $(\tilde{\theta}_n,\tilde{\varphi}_n)=(\tilde{\theta}',\tilde{\varphi}')$ according to the iteration-specific acceptance probability.
(3) Calculate $\tilde{w}_n^{(i)}$ according to (4), $i=1,\ldots,m$.

(b) Main chain:

(1) Draw a proposal φ' according to $q(\varphi'|\varphi_{n-1})$.
(2) Set $\varphi_n=\varphi'$ with probability:

$$\alpha(\varphi'|\varphi_{n-1})=\min\left\{1,\frac{p(\theta'|Y,\varphi')p(\varphi'|Z)q(\varphi_{n-1}|\varphi')p(\theta_{n-1}|Y,\varphi_{n-1})}{p(\theta_{n-1}|Y,\varphi_{n-1})p(\varphi_{n-1}|Z)q(\varphi'|\varphi_{n-1})p(\theta'|Y,\varphi')}\right\}$$
$$=\min\left\{1,\frac{p(\varphi'|Z)q(\varphi_{n-1}|\varphi')}{p(\varphi_{n-1}|Z)q(\varphi'|\varphi_{n-1})}\right\}.$$

(3) If φ' is accepted, draw θ' according to $P_n^*(\theta'|Y,\varphi')$ defined in (5) and set $\theta_n=\theta'$.
(4) Otherwise if φ' is rejected, set $(\theta_n,\varphi_n)=(\theta_{n-1},\varphi_{n-1})$.

End For;

2.2 Simple function approximation cut distribution

The convergence in distribution (6) presented in the naive stochastic approximation cut algorithm is not sufficiently strong to infer a law of large numbers or ergodicity of the drawn samples. We will show that these properties can be satisfied by targeting an approximation of the density function $p(\theta|Y,\varphi)$.

We adopt a density function approximation technique which uses a simple function as the basis. The use of a simple function to approximate a density function has been discussed previously (Fu and Wang 2002; Malefaki and Iliopoulos 2009), but here we use a different partition of the support of the function, determined by rounding to a user-specified number of decimal places. The general theory is presented in supplementary material.

Given the d-dimensional compact set Θ and user-specified number of decimal places κ, we partition Θ in terms of (partial) hypercubes Θ_r whose centres θ_r are the rounded elements of Θ to κ decimal places,

$$\Theta_r=\Theta\cap\{\theta:\|\theta-\theta_r\|_\infty\le 5\times10^{-\kappa-1}\},\ r=1,\ldots,R_\kappa, \tag{7}$$

where R_κ is the total number of rounded elements. The boundary set $\bar{\Theta}_\kappa$, which has Lebesgue measure 0, is:

$$\bar{\Theta}_\kappa=\Theta\cap\left(\bigcup_{r=1}^{R_\kappa}\{\theta:\|\theta-\theta_r\|_\infty=5\times10^{-\kappa-1}\}\right).$$

Using this partition, we are able to build a simple function density that approximates $p(\theta|Y,\varphi)$:

$$p^{(\kappa)}(\theta|Y,\varphi)=\sum_{r=1}^{R_\kappa}\frac{1}{\mu(\Theta_r)}\int_{\Theta_r}p(\theta'|Y,\varphi)\mathrm{d}\theta'\mathbb{1}_{\{\theta\in\Theta_r\}},$$

and let $P^{(\kappa)}$ be the corresponding probability measure on Θ. The simple function approximation cut distribution is then formed by replacing the exact conditional distribution with this approximation

$$p_{\mathrm{cut}}^{(\kappa)}(\theta,\varphi)=p^{(\kappa)}(\theta|Y,\varphi)p(\varphi|Z).$$

Let $P_{\mathrm{cut}}^{(\kappa)}$ be the corresponding probability measure on $\Theta\times\Phi$.

Given the general theory presented in supplementary material, we have

$$p^{(\kappa)}(\theta|Y,\varphi)\xrightarrow{\text{a.s.}}p(\theta|Y,\varphi),\quad\text{as }\kappa\to\infty. \tag{8}$$

The rate of convergence is tractable if we further assume density $p(\theta|Y,\varphi)$ is continuously differentiable.

Corollary 1 *If density function $p(\theta|Y,\varphi)$ is continuously differentiable, there exists a set $\mathscr{E} \subset \Theta$ with $\mu(\mathscr{E}) = \mu(\Theta)$ such that the local convergence holds:*

$$|p^{(\kappa)}(\theta|Y,\varphi) - p(\theta|Y,\varphi)| \leq (\varepsilon(\theta,\kappa) + \|\nabla p(\theta|Y,\varphi)\|_2)\frac{\sqrt{d}}{10^{\kappa}}, \ \forall\theta \in \mathscr{E},$$

where $\varepsilon(\theta,\kappa) \to 0$ as $\kappa \to \infty$.

In addition, the global convergence holds:

$$\sup_{\theta\in\mathscr{E}} |p^{(\kappa)}(\theta|Y,\varphi) - p(\theta|Y,\varphi)| \leq \sup_{\theta\in\Theta} \|\nabla p(\theta|Y,\varphi)\|_2 \frac{\sqrt{d}}{10^{\kappa}}.$$

Proof See the general theory in supplementary material. □

2.3 Stochastic approximation cut algorithm

We now refine the naive stochastic approximation cut algorithm by replacing in the main chain the proposal distribution P_n^*, which concentrates on the discrete set $\tilde{\Theta}_n$, by a distribution, with support on the compact set Θ, that we will show converges almost surely to $P^{(\kappa)}$.

Let $\mathscr{W}_n(\varphi) = (W_n(\Theta_1|Y,\varphi), \ldots, W_n(\Theta_{R_\kappa}|Y,\varphi))$ be a random weight process based on the probability of the original proposal distribution P_n^* taking a value in each partition component Θ_r as

$$W_n(\Theta_r|Y,\varphi) = \frac{P_n^*(\theta \in \Theta_r|Y,\varphi) + (nR_\kappa)^{-1}}{1 + n^{-1}}, \tag{9}$$

where $r = 1, \ldots, R_\kappa$. Note that $W_n(\Theta_r|Y,\varphi)$ is adapted to the auxiliary filtration $\mathcal{G}_n$. By adding a $(nR_\kappa)^{-1}$, each $W_n(\Theta_r|Y,\varphi)$, $r = 1, \ldots, R_\kappa$, is strictly positive and yet this modification does not affect the limit since $(nR_\kappa)^{-1} \to 0$. That is, on any outcome ω of probability space Ω, we have

$$\lim_{n\to\infty} \sup_{\varphi\in\Phi;\, 1\leq r\leq R_\kappa} \left| W_n(\Theta_r|Y,\varphi) - \int_{\Theta_r} p(\theta|Y,\varphi)\mathrm{d}\theta \right| = 0. \tag{10}$$

We now define the random measure process $P_n^{(\kappa)}$ that replaces P_n^* used in the naive stochastic approximation cut algorithm. For any Borel set $\mathcal{B}$,

$$P_n^{(\kappa)}(\theta \in \mathcal{B}|Y,\varphi) = \int_{\mathcal{B}} \sum_{r=1}^{R_\kappa} \frac{1}{\mu(\Theta_r)} W_n(\Theta_r|Y,\varphi)\mathbb{1}_{\{\theta\in\Theta_r\}}\mathrm{d}\theta. \tag{11}$$

Clearly, $P_n^{(\kappa)}(\theta \in \Theta|Y,\varphi) = 1$ so $P_n^{(\kappa)}$ is a valid probability measure on Θ. Additionally, since $\mathscr{W}_n(\varphi)$ is adapted to filtration $\mathcal{G}_n$, $P_n^{(\kappa)}$ is adapted to filtration $\mathcal{G}_n$. The Radon–Nikodym derivative of $P_n^{(\kappa)}$ with respect to the Lebesgue measure μ on Θ is

$$p_n^{(\kappa)}(\theta|Y,\varphi) = \sum_{r=1}^{R_\kappa} \frac{1}{\mu(\Theta_r)} W_n(\Theta_r|Y,\varphi)\mathbb{1}_{\{\theta\in\Theta_r\}}. \tag{12}$$

This density is not continuous, but it is bounded on Θ. In addition, since Θ is the support of $p(\theta|Y,\varphi)$ and $\mathscr{W}_n(\varphi)$ is strictly positive, the support of $p_n^{(\kappa)}$ is Θ for all $\varphi \in \Phi$ as well.

Using $P_n^{(\kappa)}$ as the proposal distribution has the advantage that $p_n^{(\kappa)}$ converges almost surely to $p^{(\kappa)}$, in contrast to the convergence in distribution for the naive algorithm in (6).

Lemma 1 *Given Assumption 1, on any outcome ω of probability space Ω, we have:*

$$p_n^{(\kappa)}(\theta|Y,\varphi) \xrightarrow{a.s.} p^{(\kappa)}(\theta|Y,\varphi),$$

and this convergence is uniform over $(\Theta \setminus \tilde{\Theta}_\kappa) \times \Phi$.

Note that the convergence is to $p^{(\kappa)}(\theta|Y,\varphi)$ rather than $p(\theta|Y,\varphi)$, but we will show in Corollary 2 that this bias reduces geometrically as the precision parameter κ increases.

The complete stochastic approximation cut algorithm (SACut) is shown in Algorithm 2. The key idea is that we propose samples for θ from a density $p_n^{(\kappa)}(\theta|Y,\varphi)$, which approximates $p(\theta|Y,\varphi)$ and from which we can draw samples, but we accept these proposals according to $p^{(\kappa)}(\theta|Y,\varphi)$, which then cancels. This results in the acceptance probability being determined only by the proposal distribution for φ; the proposal distribution for θ is not involved. Indeed, the acceptance probability is the same as the partial Gibbs sampler that we will discuss in Sect. 3.1.1.

2.4 Parallelization and simplification of computation

The main computational bottleneck of the stochastic approximation cut algorithm is the updating and storage of the cumulative set of auxiliary variable values $\tilde{\Theta}_n = \cup_{j=1}^n \{\tilde{\theta}_j\}$. Since we draw a new φ' at each iteration, in order to calculate all possible probabilities defined by (5) and (9), the density $p(Y|\tilde{\theta},\varphi')$ must be calculated $|\tilde{\Theta}_n|$ times. This is equivalent to running $|\tilde{\Theta}_n|$ internal iterations at each step of external iteration for the existing approximate approaches proposed in Plummer (2015). Note that $\tilde{\Theta}_n$ is solely generated from the auxiliary chain so $|\tilde{\Theta}_n|$ is not affected by the precision parameter κ. If the calculation of this density is computationally expensive, the time to perform each update of the chain will become prohibitive when $|\tilde{\Theta}_n|$ is large. However, the calculation of $p(Y|\tilde{\theta},\varphi')$ for different values of $\tilde{\theta}$ is embarrassingly parallel so can be evaluated in parallel whenever multiple computer cores are available, enabling a considerable speed up.

The speed of the computation can be further improved by reducing the size of $\tilde{\Theta}_n$. Given the precision parameter κ, we

Algorithm 2: Stochastic approximation cut (SACut) algorithm

Initialize at starting points $h_0 = (\tilde{\theta}_0, \tilde{\varphi}_0)$, $\tilde{w}_0$ and (θ_0, φ_0);
For $n = 1, \ldots, N$;

(a) Auxiliary chain:

(1) Draw a proposal $(\tilde{\theta}', \tilde{\varphi}')$ according to $q(\tilde{\theta}', \tilde{\varphi}'|\tilde{\theta}_{n-1}, \tilde{\varphi}_{n-1})$.
(2) Accept the proposal, and set $(\tilde{\theta}_n, \tilde{\varphi}_n) = (\tilde{\theta}', \tilde{\varphi}')$ according to the iteration-specific acceptance probability.
(3) Calculate $\tilde{w}_n^{(i)}$ according to (4), $i = 1, \ldots, m$.

(b) Main chain:

(1) Draw a proposal φ' according to $q(\varphi'|\varphi_{n-1})$.
(2) Set $\varphi_n = \varphi'$ with probability:

$$\alpha(\varphi'|\varphi_{n-1}) = \min\left\{1, \frac{p^{(\kappa)}(\theta'|Y,\varphi')p(\varphi'|Z)q(\varphi_{n-1}|\varphi')p^{(\kappa)}(\theta_{n-1}|Y,\varphi_{n-1})}{p^{(\kappa)}(\theta_{n-1}|Y,\varphi_{n-1})p(\varphi_{n-1}|Z)q(\varphi'|\varphi_{n-1})p^{(\kappa)}(\theta'|Y,\varphi')}\right\}$$
$$= \min\left\{1, \frac{p(\varphi'|Z)q(\varphi_{n-1}|\varphi')}{p(\varphi_{n-1}|Z)q(\varphi'|\varphi_{n-1})}\right\}.$$

(3) If φ' is accepted, calculate $W_n(\Theta_r|Y,\varphi')$ defined in (9), $r = 1, \ldots, R_\kappa$. Draw a proposal θ' according to $p_n^{(\kappa)}(\theta'|Y,\varphi')$ defined in (11) and set $\theta_n = \theta'$.
(4) Otherwise if φ' is rejected, set $(\theta_n, \varphi_n) = (\theta_{n-1}, \varphi_{n-1})$.

End For;

round all elements from $\tilde{\Theta}_n$ to their κ decimal places and let $\tilde{\Theta}_n^{(\kappa)}$ be the set that contains these rounded elements. At each iteration, the number of calculations of density $p(Y|\tilde{\theta}, \varphi')$ is equal to the number of d-orthotopes that auxiliary chain $\{\tilde{\theta}_j\}_{j=1}^n$ has visited up to iteration n and by (6) we know that the distribution of auxiliary samples of $\tilde{\theta}$ converges to the true distribution $p(\theta|Y, \varphi)$. Hence, the computational speed is mainly determined by the precision parameter κ and the target distribution $p(\theta|Y, \varphi)$. In particular, for any fixed κ and a sufficiently long auxiliary chain the computational cost is upper bounded by the case of uniform distribution since it equally distributes over the space Θ.

Theorem 1 *Given an arbitrary d-dimensional compact parameter space Θ and a precision parameter κ and suppose that the auxiliary chain has converged before we start collecting auxiliary variable $\tilde{\theta}$, for any fixed number of iteration n. Then, the expected number of d-orthotopes visited $\mathbb{E}\left(|\tilde{\Theta}_n^{(\kappa)}|\right)$ is maximized when the target distribution is uniform distribution.*

Proof See supplementary material (Online Resource 1). □

For example, given a d-dimensional parameter space $\Theta = [0 - 5 \times 10^{-\kappa-1}, 1 + 5 \times 10^{-\kappa-1}]^d$ and its partition Θ_r, $r = 1, \ldots, 11^{d\kappa}$, we consider the uniform distribution as the target distribution. Assuming the auxiliary chain has converged, the expectation of $|\tilde{\Theta}_n^{(\kappa)}|$ is

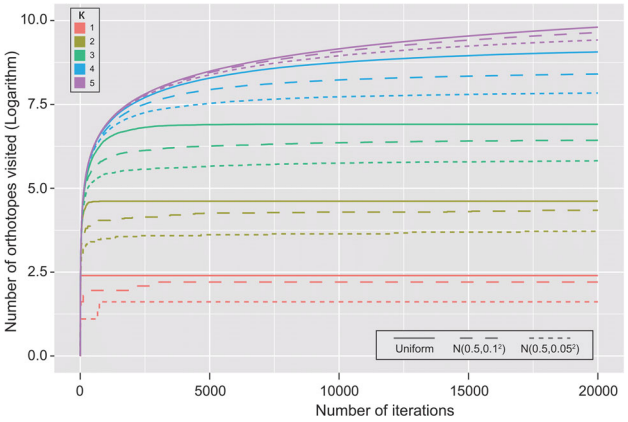

Fig. 3 Relationship between the number of orthotopes visited and the number of iterations when precision parameter $\kappa = 1, 2, 3, 4, 5$. Separate Monte Carlo simulations were conducted for uniform distribution and truncated normal distribution with standard deviation 0.1 and 0.05

$$\mathbb{E}\left(|\tilde{\Theta}_n^{(\kappa)}|\right) = 11^{d\kappa} - \frac{\left(11^{d\kappa} - 1\right)^n}{11^{d\kappa(n-1)}}.$$

In the case of $d = 1$, Fig. 3 compares the number of orthotopes visited between the uniform distribution and truncated normal distribution when the standard deviation is 0.1 and 0.05. It shows that larger precision parameter κ means more evaluations of $p(Y|\tilde{\theta}, \varphi')$ are required. Hence, a wise choice of a small κ can significantly reduce computation time.

While small κ means a loss of precision since local variations of original target distribution are smoothed by rounding the value of its samples, in most applied settings only a small number of significant figures are meaningful, and so the ability to trade-off the precision and computational speed is appealing. Comparing short preliminary run of chains for different candidates of κ may be useful when a suitable choice of κ is unclear. We will discuss this in Sect 4.1.

3 Convergence properties

In this section, we study the convergence properties of samples drawn by the stochastic approximation cut algorithm. We establish a weak law of large numbers with respect to the simple function approximation cut distribution $P_{\text{cut}}^{(\kappa)}$, under some regularity conditions, by proving that the conditions required by Theorem 3.2 in Liang et al. (2016) are satisfied. We then prove that the bias with respect to P_{cut} can be reduced geometrically by increasing the precision parameter κ. To aid exposition of the convergence properties, it is necessary to first introduce two simpler but infeasible alternative algorithms. Then, we prove the convergence of the algorithm.

The framework of our proofs follow Liang et al. (2016). However, adjustments are made for two key differences. Firstly, the parameter of interest here has two components,

instead of just one, and we require completely different proposal distributions to those in Liang et al. (2016): the proposal distribution of θ involves an auxiliary chain and simple function approximation, and the proposal distribution of φ is a standard MCMC algorithm. Secondly, the parameter drawn by (12) here is retained, rather than being discarded as in Liang et al. (2016). This means the distributions involved here are different and more complicated.

3.1 Infeasible alternative algorithms

Definition 1 Given a signed measure $\mathcal{M}$ defined on a set E, and a Borel set $\mathcal{B} \subset E$, define the total variation norm of $\mathcal{M}$ as

$$\|\mathcal{M}(\cdot)\|_{\mathrm{TV}} = \sup_{\mathcal{B}\subset E} |\mathcal{M}(\mathcal{B})|.$$

3.1.1 A partial Gibbs sampler

The most straightforward algorithm that draws samples from $p_{\mathrm{cut}}^{(\kappa)}(\theta, \varphi)$ is a standard partial Gibbs sampler, which draws proposals θ' from $p^{(\kappa)}(\theta'|Y, \varphi')$, given a φ' drawn from a proposal distribution $q(\varphi'|\varphi_{n-1})$. The transition kernel is

$$\begin{aligned}
&\mathbf{u}^{(1)}((\theta_n, \varphi_n)|(\theta_{n-1}, \varphi_{n-1})) \\
&= \alpha(\varphi_n|\varphi_{n-1}) p^{(\kappa)}(\theta_n|Y, \varphi_n) q(\varphi_n|\varphi_{n-1}) \\
&+ \left(1 - \int_{\Theta\times\Phi} \alpha(\varphi|\varphi_{n-1}) p^{(\kappa)}(\theta|Y, \varphi) q(\varphi|\varphi_{n-1}) \mathrm{d}\theta \mathrm{d}\varphi\right) \\
&\quad \delta((\theta_n, \varphi_n) - (\theta_{n-1}, \varphi_{n-1})) \\
&= \alpha(\varphi_n|\varphi_{n-1}) p^{(\kappa)}(\theta_n|Y, \varphi_n) q(\varphi_n|\varphi_{n-1}) \\
&+ \left(1 - \int_{\Phi} \alpha(\varphi|\varphi_{n-1}) q(\varphi|\varphi_{n-1}) \mathrm{d}\varphi\right) \\
&\quad \delta((\theta_n, \varphi_n) - (\theta_{n-1}, \varphi_{n-1})),
\end{aligned}$$

where δ is the multivariate Dirac delta function and

$$\alpha(\varphi_n|\varphi_{n-1}) = \min\left\{1, \frac{p(\varphi_n|Z) q(\varphi_{n-1}|\varphi_n)}{p(\varphi_{n-1}|Z) q(\varphi_n|\varphi_{n-1})}\right\}.$$

This transition kernel is Markovian and admits $p_{\mathrm{cut}}^{(\kappa)}$ as its stationary distribution, provided a proper proposal distribution $q(\varphi_n|\varphi_{n-1})$ is used. We write $\mathbf{U}^{(1)}$ for the corresponding probability measure.

Let $\mathbf{u}^{(s)}$ denote the s-step transition kernel and write $\mathbf{U}^{(s)}$ for the corresponding probability measure. By Meyn et al. (2009), we have ergodicity on $\Theta \times \Phi$,

$$\lim_{s\to\infty} \left\|\mathbf{U}^{(s)}(\cdot) - P_{\mathrm{cut}}^{(\kappa)}(\cdot)\right\|_{\mathrm{TV}} = 0,$$

and for any bounded function f defined on $\Theta \times \Phi$, we have a strong law of large numbers

$$\frac{1}{N}\sum_{n=1}^{N} f(\theta_n, \varphi_n) \xrightarrow{\text{a.s.}} \int_{\Theta\times\Phi} f(\theta, \varphi) P_{\mathrm{cut}}^{(\kappa)}(\mathrm{d}\theta, \mathrm{d}\varphi).$$

Note, however, that this algorithm is infeasible because $p^{(\kappa)}(\theta|Y, \varphi)$ is intractable, since $p(\theta|Y, \varphi)$ is intractable, and so we cannot directly draw proposals for θ.

3.1.2 An adaptive Metropolis–Hastings sampler

An adaptive Metropolis–Hastings sampler can be built by replacing $p^{(\kappa)}$ in the calculation of acceptance probability of the stochastic approximation cut algorithm by its approximation $p_n^{(\kappa)}$, which is the exact proposal distribution for θ at the nth step. The acceptance probability is determined by both θ and φ,

$$\begin{aligned}
&\alpha_n((\theta', \varphi')|(\theta_{n-1}, \varphi_{n-1})) \\
&= \min\left\{1, \frac{p^{(\kappa)}(\theta'|Y, \varphi') p(\varphi'|Z) q(\varphi_{n-1}|\varphi') p_n^{(\kappa)}(\theta_{n-1}|Y, \varphi_{n-1})}{p^{(\kappa)}(\theta_{n-1}|Y, \varphi_{n-1}) p(\varphi_{n-1}|Z) q(\varphi'|\varphi_{n-1}) p_n^{(\kappa)}(\theta'|Y, \varphi')}\right\},
\end{aligned}$$

and we can write the transition kernel,

$$\begin{aligned}
&\mathbf{v}_n^{(1)}((\theta_n, \varphi_n)|(\theta_{n-1}, \varphi_{n-1}), \mathcal{G}_n) \\
&\quad = \alpha_n((\theta_n, \varphi_n)|(\theta_{n-1}, \varphi_{n-1})) p_n^{(\kappa)}(\theta_n|Y, \varphi_n) q(\varphi_n|\varphi_{n-1}) \\
&\quad + \left(1 - \int_{\Theta\times\Phi} \alpha_n((\theta, \varphi)|(\theta_{n-1}, \varphi_{n-1})) p_n^{(\kappa)}(\theta|Y, \varphi)\right. \\
&\quad \left. q(\varphi|\varphi_{n-1}) \mathrm{d}\theta \mathrm{d}\varphi\right) \delta((\theta_n, \varphi_n) - (\theta_{n-1}, \varphi_{n-1})),
\end{aligned}$$

where δ is the multivariate Dirac delta function. Conditional on the filtration $\mathcal{G}_n$, $\mathbf{v}_n^{(1)}$ is Markovian. We write $\mathbf{V}_n^{(1)}$ for the corresponding probability measure. Note that this sampler is not a standard Metropolis–Hastings algorithm since the transition kernel is not constant. Instead, it is an *external* adaptive MCMC algorithm (Atchadé et al. 2011).

Given information up to $\mathcal{G}_n$, if we stop updating auxiliary process, then $P_n^{(\kappa)}$ is fixed and not random, and this sampler reduces to a standard Metropolis–Hastings sampler. The transition kernel $\mathbf{V}_n^{(1)}$ admits $p_{\mathrm{cut}}^{(\kappa)}$ as its stationary distribution provided a proper proposal distribution is used. That is, define

$$\begin{aligned}
\mathbf{v}_n^{(s)} &= \int_{\Theta^{s-1}\times\Phi^{s-1}} \prod_{k=1}^{s} \mathbf{v}_n^{(1)}((\theta_k, \varphi_k)|(\theta_{k-1}, \varphi_{k-1}), \mathcal{G}_n) \\
&\quad \mathrm{d}\theta_{1:s-1} \mathrm{d}\varphi_{1:s-1},
\end{aligned}$$

and $\mathbf{V}_n^{(s)}$ as the corresponding probability measure. Then, on $\Theta \times \Phi$ we have

$$\lim_{s\to\infty} \left\|\mathbf{V}_n^{(s)}(\cdot) - P_{\mathrm{cut}}^{(\kappa)}(\cdot)\right\|_{\mathrm{TV}} = 0.$$

Note, however, that this algorithm is also infeasible because, while we can draw proposals for θ, since $p_n^{(\kappa)}$ is known up to $\mathcal{G}_n$, $p^{(\kappa)}(\theta|Y,\varphi)$ remains intractable so we cannot calculate the acceptance probability.

3.2 Convergence of the stochastic approximation cut algorithm

The infeasibility of the partial Gibbs sampler and the adaptive Metropolis–Hastings sampler motivates the development of the stochastic approximation cut algorithm, which replaces the proposal distribution $p_n^{(\kappa)}$ by its target $p^{(\kappa)}$ in the accept–reject step of the adaptive Metropolis–Hastings sampler. This leads to the same acceptance probability as is used by the partial Gibbs sampler, so the proposed algorithm can be viewed as combining the advantages of both the partial Gibbs sampler and the adaptive Metropolis–Hastings sampler. The transition kernel of the stochastic approximation cut algorithm is

$$\begin{aligned}
&\mathbf{t}_n^{(1)}((\theta_n,\varphi_n)|(\theta_{n-1},\varphi_{n-1}),\mathcal{G}_n)\\
&=\alpha(\varphi_n|\varphi_{n-1})p_n^{(\kappa)}(\theta_n|Y,\varphi_n)q(\varphi_n|\varphi_{n-1})\\
&+\left(1-\int_{\Theta\times\Phi}\alpha(\varphi|\varphi_{n-1})p_n^{(\kappa)}(\theta|Y,\varphi)q(\varphi|\varphi_{n-1})\mathrm{d}\theta\mathrm{d}\varphi\right)\\
&\delta((\theta_n,\varphi_n)-(\theta_{n-1},\varphi_{n-1}))\\
&=\alpha(\varphi_n|\varphi_{n-1})p_n^{(\kappa)}(\theta_n|Y,\varphi_n)q(\varphi_n|\varphi_{n-1})\\
&+\left(1-\int_{\Phi}\alpha(\varphi|\varphi_{n-1})q(\varphi|\varphi_{n-1})\mathrm{d}\varphi\right)\\
&\delta((\theta_n,\varphi_n)-(\theta_{n-1},\varphi_{n-1}))\,,
\end{aligned}$$

where δ is the multivariate Dirac delta function. Conditionally to $\mathcal{G}_n$, the transition kernel $\mathbf{t}_n^{(1)}$ is Markovian. We write $\mathbf{T}_n^{(1)}$ for the corresponding probability measure. Given information up to $\mathcal{G}_n$ and stopping updating the auxiliary process, $P_n^{(\kappa)}$ is fixed and not random, and we define the s-step transition kernel as

$$\mathbf{t}_n^{(s)}=\int_{\Theta^{s-1}\times\Phi^{s-1}}\prod_{k=1}^{s}\mathbf{t}_n^{(1)}((\theta_k,\varphi_k)|(\theta_{k-1},\varphi_{k-1}),\mathcal{G}_n)\,\mathrm{d}\theta_{1:s-1}\mathrm{d}\varphi_{1:s-1},$$

and write $\mathbf{T}_n^{(s)}$ for the corresponding probability measure.

We now present several lemmas required to prove a weak law of large numbers for this algorithm (proofs in supplementary material (Online Resource 1)), appropriately modifying the reasoning of Meyn and Tweedie (1994), Roberts and Tweedie (1996) and Liang et al. (2016) for this setting.

Assumption 2 The posterior density $p(\varphi|Z)$ is continuous on Φ and the proposal distribution $q(\varphi'|\varphi)$ is continuous with respect to (φ',φ) on $\Phi\times\Phi$.

Lemma 2 (Diminishing adaptation) *Given Assumptions 1 and 2, then*

$$\lim_{n\to\infty}\sup_{\theta\in\Theta,\varphi\in\Phi}\left\|\boldsymbol{T}_{n+1}^{(1)}(\cdot|(\theta,\varphi),\mathcal{G}_{n+1})-\boldsymbol{T}_n^{(1)}(\cdot|(\theta,\varphi),\mathcal{G}_n)\right\|_{\mathrm{TV}}=0.$$

Before presenting the next lemma, we introduce the concept of *local positivity*.

Definition 2 A proposal distribution $q(\psi'|\psi)$ satisfies local positivity if there exists $\delta>0$ and $\varepsilon>0$ such that for every $\psi\in\Psi$, $|\psi'-\psi|\leq\delta$ implies that $q(\psi'|\psi)>\varepsilon$.

Lemma 3 *Given Assumption 1, the proposal distributions with densities $p_n^{(\kappa)}:\Theta\to\mathbb{R}$ and $p^{(\kappa)}:\Theta\to\mathbb{R}$ are both uniformly lower bounded away from 0 and satisfy local positivity uniformly for all values $\varphi\in\Phi$.*

Lemma 4 (Stationarity) *Given Assumptions 1 and 2, and the filtration $\mathcal{G}_n$ (i.e. $P_n^{(\kappa)}$ is not random), then if the transition kernel measures $\boldsymbol{U}^{(1)}$ and $\boldsymbol{V}_n^{(1)}$ both admit an irreducible and aperiodic Markov chain, then the transition kernel measure $\boldsymbol{T}_n^{(1)}$ admits an irreducible and aperiodic chain. Moreover, if the proposal distribution $q(\varphi'|\varphi)$ satisfies local positivity, then there exists a probability measure Π_n on $\Theta\times\Phi$ such that for any $(\theta_0,\varphi_0)\in\Theta\times\Phi$,*

$$\lim_{s\to\infty}\left\|\boldsymbol{T}_n^{(s)}(\cdot)-\Pi_n(\cdot)\right\|_{\mathrm{TV}}=0,$$

and this convergence is uniform over $\Theta\times\Phi$.

Lemma 5 (Asymptotic Simultaneous Uniform Ergodicity) *Given Assumptions 1 and 2 and the assumptions in Lemma 4, for any initial value $(\theta_0,\varphi_0)\in\Theta\times\Phi$, and any $\varepsilon>0$ and $e>0$, there exist constants $S(\varepsilon)>0$ and $N(\varepsilon)>0$ such that*

$$\mathbb{P}\left(\left\{P_n^{(\kappa)}:\left\|\boldsymbol{T}_n^{(s)}(\cdot)-P_{\mathrm{cut}}^{(\kappa)}(\cdot)\right\|_{\mathrm{TV}}\leq\varepsilon\right\}\right)>1-e,$$

for all $s>S(\varepsilon)$ and $n>N(\varepsilon)$.

Lemma 2 leads to condition (c) (diminishing adaptation), Lemma 4 leads to condition (a) (stationarity) and Lemma 5 leads to condition (b) (asymptotic simultaneous uniform ergodicity) in Theorem 3.2 of Liang et al. (2016). Hence, we have the following weak law of large numbers.

Theorem 2 (WLLN) *Suppose that the conditions of Lemma 5 hold. Let f be any measurable bounded function on $\Theta\times\Phi$. Then, for samples (θ_n,φ_n), $n=1,2,\ldots$ drawn using the stochastic approximation cut algorithm, we have that*

$$\frac{1}{N}\sum_{n=1}^{N}f(\theta_n,\varphi_n)\to\int_{\Theta\times\Phi}f(\theta,\varphi)P_{\mathrm{cut}}^{(\kappa)}(\mathrm{d}\theta,\mathrm{d}\varphi),$$

in probability.

Proof This follows by Theorem 3.2 in Liang et al. (2016). □

Given further conditions and combining Corollary 1 with Theorem 2, we have the following corollary.

Corollary 2 *Given the conditions in Corollary 1 hold for the cut distribution p_{cut} and conditions in Theorem 2 hold. Then, given a measurable and bounded function $f : \Theta \times \Phi \to \mathbb{R}$, there exists, for any $\varepsilon > 0$ and $e > 0$, a precision parameter κ and iteration number N, such that for samples (θ_n, φ_n), $n = 1, 2, \ldots$ drawn using the stochastic approximation cut algorithm, we have that*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} f(\theta_n, \varphi_n) - \int_{\Theta\times\Phi} f(\theta,\varphi) P_{\text{cut}}(\mathrm{d}\theta, \mathrm{d}\varphi)\right| \le \varepsilon\right) > 1 - e.$$

More specifically, the bias

$$\left|\int_{\Theta\times\Phi} f(\theta,\varphi) P_{\text{cut}}(\mathrm{d}\theta,\mathrm{d}\varphi) - \int_{\Theta\times\Phi} f(\theta,\varphi) P_{\text{cut}}^{(\kappa)}(\mathrm{d}\theta,\mathrm{d}\varphi)\right|$$

can be controlled by

$$\sup_{\theta\in\Theta,\varphi\in\Phi} \|\nabla_\theta p(\theta|Y,\varphi)\|_2 \frac{\sqrt{d}}{10^\kappa}\left(\int_{\Theta\times\Phi} f(\theta,\varphi) p(\varphi|Z)\mathrm{d}\theta\mathrm{d}\varphi\right).$$

Corollary 2 shows that, although the convergence established by Theorem 2 is biased with respect to the true cut distribution P_{cut}, the bias can be geometrically reduced by selecting a large precision parameter κ.

4 Illustrative examples

We demonstrate the proposed algorithm in this section. First, we use a simulation example to introduce a simple method for choosing the precision parameter κ and demonstrate that the proposed algorithm can eliminate the feedback from a suspect module. We then examine a simulated case designed to highlight when existing algorithms will perform poorly. We finally apply our algorithm to an epidemiological example and compare results with existing studies. The R package *SACut* and code to replicate these examples can be downloaded from GitHub.[1]

4.1 Simulated random effects example

In this example, we discuss a simple method for selecting the precision parameter κ and show that the proposed algorithm can effectively cut the feedback from a suspect module.

[1] https://github.com/MathBilibili/Stochastic-approximation-cut-algorithm.

We consider a simple normal–normal random effect example previously discussed by Liu et al. (2009), with groups $i = 1, \ldots, 100 = N$, observations $Y_{ij} \sim \mathrm{N}(\beta_i, \varphi_i^2)$, $j = 1, \ldots, 20$ in each group, and random effects distribution $\beta_i \sim \mathrm{N}(0, \theta^2)$. Our aim is to estimate the random effects standard deviation θ and the residual standard deviation $\varphi = (\varphi_1, \ldots, \varphi_N)$. By sufficiency, the likelihood can be equivalently represented in terms of the group-specific means $\bar{Y}_i = \frac{1}{20}\sum_{j=1}^{20} Y_{ij}$ and the sum of squared deviations $s_i^2 = \sum_{j=1}^{20}(Y_{ij} - \bar{Y}_i)^2$ as

$$\bar{Y}_i \sim \mathrm{N}(\beta_i, \frac{\varphi_i^2}{20}),$$
$$s_i^2 \sim \text{Gamma}\left(\frac{20-1}{2}, \frac{1}{2\varphi_i^2}\right).$$

Given the sufficient statistics $\bar{Y} = (\bar{Y}_1, \ldots, \bar{Y}_N)$ and $s^2 = (s_1^2, \ldots, s_N^2)$, the model consists of two modules: module 1 involving (s^2, φ) and module 2 involving $(\bar{Y}, \beta, \varphi)$, where $\beta = (\beta_1, \ldots, \beta_N)$.

We consider the situation when an outlier group is observed, meaning that module 2 is misspecified, and compare the standard Bayesian posterior distribution with the cut distribution. Specifically, we simulate data from the model with $\theta^2 = 2$, and φ_i^2 drawn from a Unif(0.5, 1.5) distribution ($\varphi_1^2 = 1.60$), but we artificially set $\beta_1 = 10$, making the first group an outlier and thus our normal assumption for the random effects misspecified. Given priors $p(\varphi_i^2) \propto (\varphi_i^2)^{-1}$ and $p(\theta^2|\varphi^2) \propto (\theta^2 + \bar{\varphi}^2/20)^{-1}$, Liu et al. (2009) showed the standard Bayesian marginal posterior distribution for the parameters of interest is:

$$\begin{aligned} p(\theta,\varphi|\bar{Y},s^2) &= p(\theta|\bar{Y},\varphi)p(\varphi|\bar{Y},s^2) \\ &\propto \frac{1}{\theta^2+\bar{\varphi}^2/20}\prod_{i=1}^{100}(\varphi_i^2)^{-\frac{21}{2}}\exp\left(-\frac{s_i^2}{2\varphi_i^2}\right) \\ &\quad \frac{1}{(\theta^2+\varphi_i^2/20)^{1/2}}\exp\left(-\frac{\bar{Y}_i^2}{2(\theta^2+\varphi_i^2/20)}\right). \end{aligned}$$

Since we are confident about our assumption of normality of Y_{ij} but not confident about our distributional assumption for the random effects β_i, following Liu et al. (2009), we consider the cut distribution in which we remove the influence of $\bar{Y}$ on φ, so that possible misspecification of the first module does not affect φ:

$$p_{\text{cut}}(\theta,\varphi) := p(\theta|\bar{Y},\varphi)p(\varphi|s^2),$$

where

$$p(\varphi|s^2) \propto \prod_{i=1}^{100}\varphi_i^{-21}\exp\left(-\frac{s_i^2}{2\varphi_i^2}\right).$$

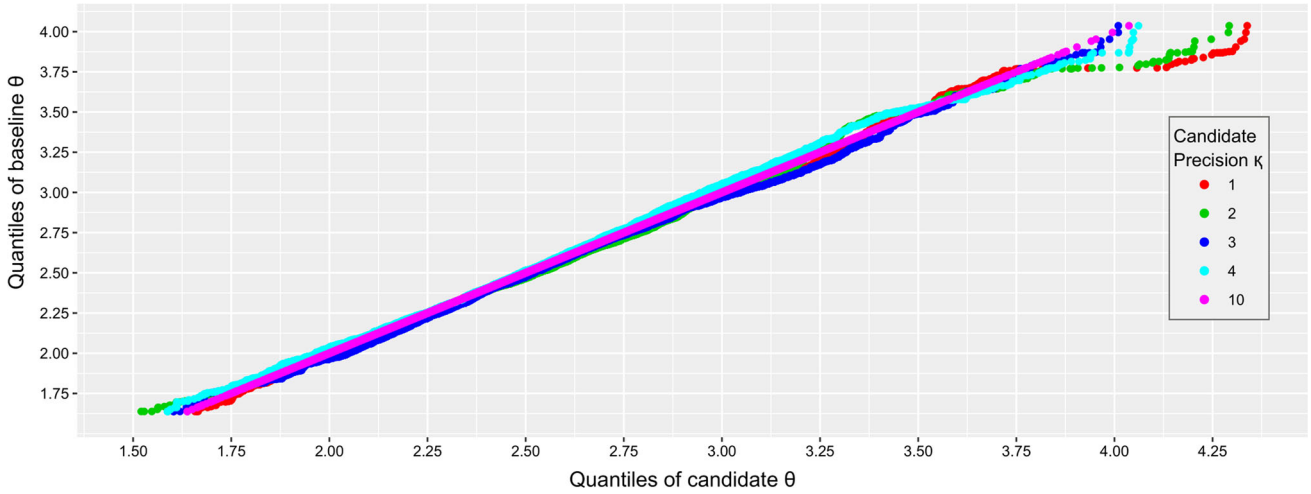

Fig. 4 Quantile–quantile plot for θ drawn from (13) with precision parameter $\kappa = 1, 2, 3, 4, 10$. The x-axis of the quantile–quantile plot is the quantile of samples under different κ, and the y-axis is the quantile of samples under the gold standard $\kappa = 10$

To apply the proposed algorithm we first construct the auxiliary parameter set for the parameter φ by selecting 70 samples selected from posterior samples of $p(\varphi|s^2)$ by the Max-Min procedure (Liang et al. 2016). We set the shrink magnitude $n_0 = 1000$ and run only the auxiliary chain for 10^4 iterations before starting to store the auxiliary variable h_n, as suggested by Liang et al. (2016).

The precision parameter κ should be chosen large enough to obtain accurate results, while being small enough that computation is not prohibitively slow. To illustrate this, we compare results with $\kappa = 10$, which we regard as the gold standard, to results with $\kappa = 1, 2, 3, 4$. Different values of κ affect the sampling of θ only via (11), so we compare samples drawn from $p_n^{(\kappa)}(\theta|\bar{Y}, \varphi)$, averaged over the marginal cut distribution of φ:

$$p_n^{(\kappa)}(\theta|\bar{Y}, s^2) := \int p_n^{(\kappa)}(\theta|\bar{Y}, \varphi)\, p_{\text{cut}}(\varphi)\mathrm{d}\varphi, \tag{13}$$

where the marginal cut distribution $p_{\text{cut}}(\varphi)$ is

$$p_{\text{cut}}(\varphi) := \int p_{\text{cut}}(\theta, \varphi)\mathrm{d}\theta = p(\varphi|s^2) \propto p(s^2|\varphi)p(\varphi).$$

We draw 10^5 samples from (13) for each value of κ, after running the proposed algorithm with few iterations (10^4) as a preliminary trial. Figure 4 shows the quantile–quantile plot for 5 choices for κ. The fit appears good for all choices of κ, except in the tails, where $\kappa = 3$ and $\kappa = 4$ provide a closer match to the gold standard. Thus, we choose $\kappa = 3$ as it gives a sufficiently accurate approximation.

We apply both the standard Bayesian approach and the stochastic approximation cut algorithm ($\kappa = 3$), each with ten independent chains. All chains were run for 10^5 iterations, and we retain only every 100th value, after discarding the first 10% of the samples, and we summarize the results by the mean and credible interval (CrI). Pooling the ten chains for the cut distribution gave estimates of $\theta^2 = 2.54$ (95% CrI 1.93–3.44) and $\varphi_1^2 = 1.58$ (95% CrI 0.88–3.18), whereas the standard Bayesian approach gave estimates of $\theta^2 = 2.53$ (95% CrI 1.93–3.44) and $\varphi_1^2 = 1.69$ (95% CrI 0.91–3.76). Figure 5 presents the medians for the parameter of interest φ_1^2 under each of the ten independent runs for the cut distribution and the standard Bayesian posterior. Recalling the true value for $\varphi_1^2 = 1.60$, it is clear that when using the stochastic approximation cut algorithm the medians locate around its true value rather than deviating systematically towards one side. This indicates the proposed algorithm has successfully prevented the outlying observation from influencing the estimation of φ_1^2.

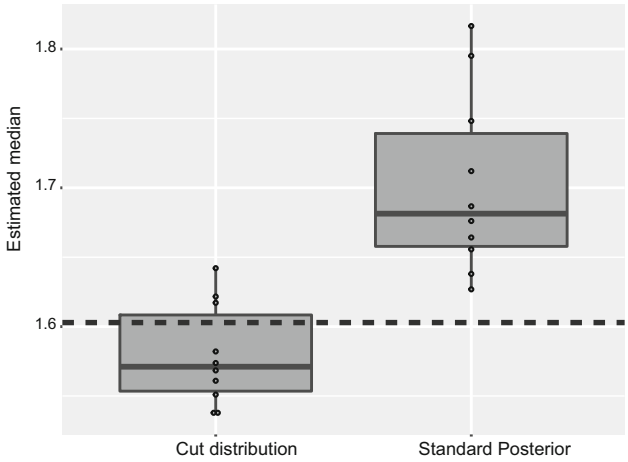

Fig. 5 Box plot of median estimates for φ_1^2 from each of ten independent runs, under the cut distribution and the standard Bayesian posterior. The dashed line indicates the true value of φ_1^2

4.2 Simulated strong dependence between θ and φ

In this section, we apply our algorithm in a simulated setting that illustrates when nested MCMC (Plummer 2015) can perform poorly. Consider the case when the distribution of θ is highly dependent on φ. In this case, if the distance between successive values φ' and φ is large in the external MCMC chain, the weight function may not be close to 1 and so the internal chain will typically require more iterations to reach convergence. This will be particularly problematic if the mixing time for the proposal distribution is large.

To simulate this scenario, we consider a linear regression for outcomes $Y_i, i = 1, \ldots, 50$, in which the coefficient vector $\theta = (\theta_1, \ldots, \theta_d)$ is closely related to the coefficient φ for covariate $X_i = (X_{\theta,i}, X_{\varphi,i})$. To assess the performance under small and moderate dimension of θ, we consider $d = 1$ and 20 in this illustration. In addition to observations of the outcome Y_i and the covariate X_i, we assume we have separate observations $Z_j, j = 1, \ldots, 100$ related to the coefficient φ.

$$\begin{aligned} Y_i &\sim \mathrm{N}(\theta^\intercal X_{\theta,i} + \varphi X_{\varphi,i}, 3), \ i = 1, \ldots, 50; \\ Z_j &\sim \mathrm{N}(\varphi, 1), \ j = 1, \ldots, 100. \end{aligned} \tag{14}$$

Suppose that we wish to estimate φ solely on the basis of $Z = (Z_1, \ldots, Z_{100})$, and so we cut the feedback from $Y = (Y_1, \ldots, Y_{50})$ to φ.

We generate Y and Z according to (14), with $\varphi = 1$ and $\theta_p = \sin(p)$, $p = 1, \ldots, d$, and compare the results of stochastic approximation cut (SACut) algorithm, naive SACut and nested MCMC with internal chain length $n_{\text{int}} = 1$, 10, 200, 500, 1000, 1500 and 2000. Notably, nested MCMC with $n_{\text{int}} = 1$ is the WinBUGS algorithm. The proposal distribution for each element of φ is a normal distribution, centred at the previous value and with standard deviation 0.25, and the proposal distribution for θ used in the nested MCMC is a normal distribution, centred at the previous value and with standard deviation 10^{-5}. The priors for both parameters are uniformly distributed within a compact domain. We set the shrink magnitude $n_0 = 2000$ and precision parameter $\kappa_p = 4$, $p = 1, \ldots, 20$. The SACut and naive SACut algorithms are processed in parallel on ten cores of Intel Xeon E7-8860 v3 CPU (2.2 GHz), and the (inherently serial) nested MCMC algorithm is processed on a single core. All algorithms were independently run 20 times, and the results are the averages across runs. Each run consists 5×10^4 iterations. We retain only every tenth value after discarding the first 40% samples as burn-in.

To assess the performance of these algorithms, we compare their estimation of $\mathbb{E}(\theta)$, lag-1 auto-correlation of samples, the Gelman–Rubin diagnostic statistic $\hat{R}$ (Gelman and Rubin 1992) and the average time needed for the whole run. The precision of the estimation of θ is measured by the mean square error (MSE) across its d (either 1 or 20) components. The convergence is evaluated by averaging the Gelman–Rubin diagnostic statistic of d components.

Results are shown in Table 1. The time required to run the nested MCMC algorithm increases as the length of the internal chain or dimension of θ increases, although the influence of dimension of θ is relatively small. In a low-dimensional case ($d = 1$), the time needed to run SACut and naive SACut is more than the time needed to run the WinBUGS algorithm and nested MCMC algorithm when the length of internal chain is less than 500, but both the MSE and the Gelman–Rubin statistic are lower when using the SACut algorithm. In particular, the bias of the WinBUGS algorithm is large. There is only trivial difference in bias between SACut and nested MCMC when $n_{\text{int}} \geq 1000$, but SACut is significantly faster than nested MCMC. In the higher-dimensional case ($d = 20$), both SACut and naive SACut significantly outperform the WinBUGS and nested MCMC algorithm in terms of MSE. Although the difference between SACut and nested MCMC with $n_{\text{int}} = 1000$ is small, the Gelman–Rubin statistic of the nested MCMC is still larger than the threshold 1.2 suggested by Brooks and Gelman (1998). The MCMC chains produced by the nested MCMC converge better and the bias is smaller when $n_{\text{int}} \geq 1500$, but the SACut algorithm still outperforms it according to MSE and Gelman–Rubin statistic, and takes less time. It is also clear that nested MCMC samples show very strong auto-correlation for both cases and thinning may not efficiently solve this issue (Link and Eaton 2012); both SACut and naive SACut do not show any auto-correlation. We also note that the estimates provided by SACut and naive SACut are almost identical in practice. However, since the time needed for both algorithm is almost the same, providing the full approach with a more solid theoretical foundation is a valuable contribution to the computational statistics literature for the cut distribution.

Jacob et al. (2020) recently proposed an unbiased coupling algorithm which can sample from the cut distribution. It requires running coupled Markov chains where samples from each chain marginally target the same true distribution. The estimator is completely unbiased when two chains meet. Drawing samples from the cut distribution using the unbiased coupling algorithm typically involves two stages. In general, the first stage involves running coupled chains for φ until they meet. For each sampled φ, the second stage involves running another set of coupled chains for θ until they meet. Although the algorithm is unbiased, as illustrated in Sects. 4.2, 4.3 and the discussion of Jacob et al. (2020), the number of iterations for coupled chains is determined by meeting times, which can be very large especially when the dimension of the parameter is high. As a comparison, we apply the unbiased coupled algorithm on this example by using the R package "unbiasedmcmc" provided by Jacob et al. (2020). To simplify the implementation and computation of the unbiased coupling algorithm, we consider a

Table 1 Mean squared error (MSE), lag-1 auto-correlation (in absolute value) |AC|, Gelman–Rubin statistic $\hat{R}$, and clock time for the stochastic approximation cut (SACut) algorithm, naive SACut algorithm, WinBUGS algorithm, the nested MCMC algorithm (with varying internal chain length n_{int}) and unbiased coupling algorithm

d	Algorithm	n_{int}	MSE $\times 10^3$	|AC|	$\hat{R}$	Time (min)
1	SACut	–	0.112	0.019	1.00	311
	Naive SACut	–	0.114	0.016	1.00	308
	WinBUGS	1	355.280	0.999	308.78	1
	Nested MCMC	10	217.907	0.999	29.87	10
	Nested MCMC	200	0.158	0.997	1.74	182
	Nested MCMC	500	0.138	0.993	1.25	454
	Nested MCMC	1000	0.109	0.990	1.07	910
	Nested MCMC	1500	0.113	0.986	1.08	1349
	Nested MCMC	2000	0.118	0.981	1.05	1771
	Unbiased Coupling	–	0.114	0.012	1.01	22
20	SACut	–	1.42	0.009	1.00	1239
	Naive SACut	–	1.47	0.002	1.01	1219
	WinBUGS	1	16387.69	0.999	209.55	2
	Nested MCMC	10	12490.25	0.999	22.73	11
	Nested MCMC	200	249.18	0.999	2.38	259
	Nested MCMC	500	10.76	0.997	1.33	517
	Nested MCMC	1000	1.86	0.994	1.22	1010
	Nested MCMC	1500	1.69	0.991	1.19	1515
	Nested MCMC	2000	1.60	0.988	1.11	2058
	Unbiased Coupling	–	1.36	0.013	1.00	2030

All values are means across 20 independent runs

simplified scenario with an informative conjugate prior for φ, meaning we can omit the first stage and instead directly draw 5×10^4 samples from $p(\varphi|Z)$. This prior is normal with mean equal to the true value of φ. We then ran preliminary coupled chains for θ that target $p(\theta|Y, \varphi)$ given these samples of φ so as to sample the meeting times. Over the 5×10^4 independent runs, the 95% and 99% quantiles of meeting times were 44 and 147, respectively, when $d = 1$. Although the majority of meeting times are, relatively, small, their 95% and 99% quantiles were 3525 and 5442, respectively, when $d = 20$. To ensure that the total number of iterations covers the majority of meeting times, following Jacob et al. (2020), we set the minimum number of iterations for each coupled chain to ten times the 95% quantile of meeting times. The algorithm was processed in parallel on the same ten cores as SACut, and the final result is shown in Table 1. Notably, unlike the nested MCMC algorithm, the computational time of the unbiased coupling algorithm increases significantly when the dimension of θ increases because it takes more time for coupled chains to couple in high-dimensional cases. In the low-dimensional case ($d = 1$), the unbiased coupling algorithm performs better according to all metrics. In the higher-dimensional case ($d = 20$), the unbiased coupling algorithm achieves similar MSE to the SACut algorithm, but it takes considerably more computation time than SACut, even though the unbiased coupling algorithm was been conducted under a simplified setting (i.e. no coupled chain for φ).

4.3 Epidemiological example

We now consider an epidemiological study of the relation between high-risk human papillomavirus (HPV) prevalence and cervical cancer incidence (Maucort-Boulch et al. 2008), which was previously discussed by Plummer (2015). In this study, age-stratified HPV prevalence data and cancer incidence data were collected from 13 cities. The model is divided into two modules. The first module concerns the number of people with HPV infection in city i, denoted as Z_i, out of a sample of N_i women:

$$Z_i \sim \text{Bin}(N_i, \varphi_i).$$

The second module describes the relation between the number of cancer cases Y_i from T_i person-years and incidence which is assumed to be linked with φ_i by a log linear relationship:

$$Y_i \sim \text{Poisson}\left(T_i \left(\exp(\theta_1 + \theta_2 \varphi_i)\right)\right).$$

The log-linear dose–response relationship is speculative, so we apply the cut algorithm to prevent the feedback from the second module to the estimation of φ_i (Plummer 2015).

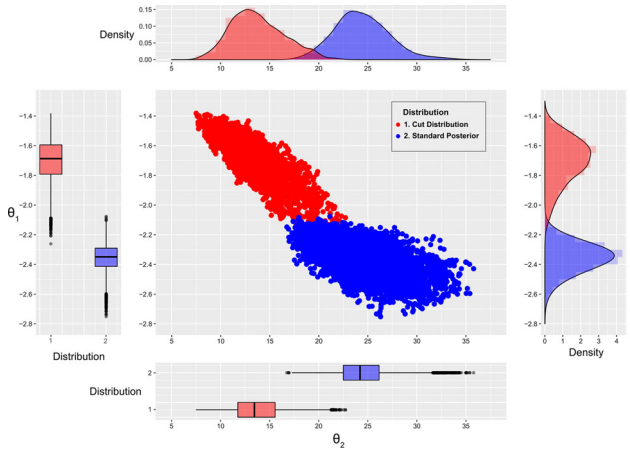

Fig. 6 Comparison of the distribution of θ_1 and θ_2 drawn from the cut distribution (red) and standard Bayesian posterior (blue)

We apply the stochastic approximation cut algorithm and compare results with the standard Bayesian approach (i.e. without a cut). Both algorithms were run ten times independently, each with 1.4×10^5 iterations. We set the shrink magnitude $n_0 = 20000$ and precision parameter $\kappa_1 = 3$ for θ_1 and $\kappa_2 = 2$ for θ_2. We retain only every 100th value after discarding the first 4×10^4 samples as burn-in. The pooled results of θ are shown in Fig. 6, highlighting the considerable effect of cutting feedback in this example. Our results are consistent with existing studies: specifically the scatter plot and density plot agree with Jacob et al. (2017) and Carmona and Nicholls (2020). Our results are also consistent with the results of nested MCMC algorithm when its internal chain length is largest (see Plummer (2015)). This again shows that the SACut algorithm provides similar estimates to the nested MCMC algorithm with a large internal chain length.

5 Conclusion

We have proposed a new algorithm for approximating the cut distribution that improves on the WinBUGS algorithm and approximate approaches in Plummer (2015). Our approach approximates the intractable marginal likelihood $p(Y|\varphi)$ using stochastic approximation Monte Carlo (Liang et al. 2007). The algorithm avoids the weakness of approximate approaches that insert an "internal limit" into each iteration of the main Markov chain. Obviously, one can argue that approximate approaches can be revised by setting the length of the internal chain to the number of iterations, i.e. $n_{\text{int}} = n$ so that the internal length diverges with n. However, since the sampling at each iteration is still not perfect and bias is inevitably introduced, the convergence of the main Markov chain remains unclear and the potential limit is not known. We proved convergence of the samples drawn by our algorithm and present the exact limit, though its convergence rate is not fully studied and needs further investigations. Although the bias is not completely removed by our algorithm, the degree of the bias is explicit in the sense that the shape of $p^{\kappa}(\theta|Y, \varphi)$ is known since the shape of $p(\theta|Y, \varphi)$ is normally obtainable given a fixed φ. Corollary 2 shows that the bias in our approach can be reduced by increasing the precision parameter κ. We proposed that κ be selected by comparing results across a range of choices; quantitative selection of this precision parameter still needs further study.

Existing approximate approaches (Plummer 2015) which need an infinitely long internal chain may be computationally slow, because the internal chain requires sequential calculation so parallelization is not possible. In contrast, thanks to the embarrassingly parallel calculation of (5), our algorithm can be more computationally efficient when multiple computer cores are available, although the per-iteration time of our algorithm decays as the Markov chain runs due to the increasing size of collection of auxiliary variables.

Lastly, while the adaptive exchange algorithm (Liang et al. 2016) is used for intractable normalizing problems when the normalizing function is an integral with respect to the observed data, it would be interesting to investigate the use of our algorithm for other problems involving a normalizing function that is an integral with respect to the unknown parameter. For example, our algorithm can be directly extended to sample from the recently developed semi-modular inference distribution (Carmona and Nicholls 2020) which generalizes the cut distribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-021-10070-2.

Acknowledgements The authors thank Daniela De Angelis and Simon R. White for helpful discussions and suggestions.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Atchadé, Y., Fort, G., Moulines, E., Priouret, P.: Adaptive Markov chain Monte Carlo: theory and methods. In: Barber, D., Cemgil, A.T., Chiappa, S. (Eds.) Bayesian Time Series Models, pp. 32–51. Cambridge University Press (2011)

Bhattacharya, A., Pati, D., Yang, Y.: Bayesian fractional posteriors. Ann. Stat. **47**(1), 39–66 (2019)

Blangiardo, M., Hansell, A., Richardson, S.: A Bayesian model of time activity data to investigate health effect of air pollution in time series studies. Atmos. Environ. **45**(2), 379–386 (2011)

Brooks, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. **7**(4), 434–455 (1998)

Carmona, C.U., Nicholls, G.K.: Semi-modular inference: enhanced learning in multi-modular models by tempering the influence of components. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, pp. 4226-4235. PMLR (2020)

Fu, J.C., Wang, L.: A random-discretization based Monte Carlo sampling method and its applications. Methodol. Comput. Appl. Probab. **4**(1), 5–25 (2002)

Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Stat. Sci. **7**(4), 457–472 (1992)

Gottardo, R., Raftery, A.E.: Markov chain Monte Carlo with mixtures of mutually singular distributions. J. Comput. Graph. Stat. **17**(4), 949–975 (2008)

Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli **7**(2), 223–242 (2001)

Huang, B., Wu, B., Barry, M.: Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. Int. J. Geogr. Inf. Sci. **24**(3), 383–401 (2010)

Jacob, P.E., Murray, L.M., Holmes, C.C., Robert, C.P.: Better together? Statistical learning in models made of modules. Preprint arXiv:1708.08719 (2017)

Jacob, P.E., O'Leary, J., Atchadé, Y.F.: Unbiased Markov chain Monte Carlo methods with couplings. J. R. Stat. Soc. B **6**, 66 (2020)

Liang, F.: Dynamically weighted importance sampling in Monte Carlo computation. J. Am. Stat. Assoc. **97**(459), 807–821 (2002)

Liang, F.: A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. J. Stat. Comput. Simul. **80**(9), 1007–1022 (2010)

Liang, F., Liu, C., Carroll, R.J.: Stochastic approximation in Monte Carlo computation. J. Am. Stat. Assoc. **102**(477), 305–320 (2007)

Liang, F., Jin, I.H., Song, Q., Liu, J.S.: An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants. J. Am. Stat. Assoc. **111**(513), 377–393 (2016)

Link, W.A., Eaton, M.J.: On thinning of chains in MCMC. Methods Ecol. Evol. **3**(1), 112–115 (2012)

Liu, F., Bayarri, M., Berger, J.: Modularization in Bayesian analysis, with emphasis on analysis of computer models. Bayesian Anal. **4**(1), 119–150 (2009)

Liu, Y., Lam, K.-F., Wu, J.T., Lam, T.T.-Y.: Geographically weighted temporally correlated logistic regression model. Sci. Rep. **8**(1), 1417 (2018)

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., Neuenschwander, B.: Combining MCMC with 'sequential' PKPD modelling. J. Pharmacokinet Phar. **36**(1), 19 (2009a)

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N.: The BUGS project: evolution, critique and future directions. Stat. Med. **28**(25), 3049–3067 (2009b)

Malefaki, S., Iliopoulos, G.: Simulation from a target distribution based on discretization and weighting. Commun. Stat. Simul. Comput. **38**(4), 829–845 (2009)

Maucort-Boulch, D., Franceschi, S., Plummer, M.: International correlation between human papillomavirus prevalence and cervical cancer incidence. Cancer. Epidem. Biomar. **17**(3), 717–720 (2008)

McCandless, L.C., Douglas, I.J., Evans, S.J., Smeeth, L.: Cutting feedback in Bayesian regression adjustment for the propensity score. Int. J. Biostat. **6**(2), 16 (2010)

Meyn, S.P., Tweedie, R.L.: Computable bounds for geometric convergence rates of Markov chains. Ann. Appl. Probab. **4**(4), 981–1011 (1994)

Meyn, S., Tweedie, R.L., Glynn, P.W.: Markov Chains and Stochastic Stability, 2nd edn. Cambridge Mathematical Library. Cambridge University Press (2009)

Miller, J.W., Dunson, D.B.: Robust Bayesian inference via coarsening. J. Am. Stat. Assoc. **114**(527), 1113–1125 (2019)

Møller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K.: An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika **93**(2), 451–458 (2006)

Murray, I., Ghahramani, Z., MacKay, D.J.C.: MCMC for doubly-intractable distributions. In: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, UAI'06, pp. 359–366. AUAI Press, Arlington, VA, USA (2006)

Nakaya, T., Fotheringham, A.S., Brunsdon, C., Charlton, M.: Geographically weighted Poisson regression for disease association mapping. Stat. Med. **24**(17), 2695–2717 (2005)

Park, J., Haran, M.: Bayesian inference in the presence of intractable normalizing functions. J. Am. Stat. Assoc. **113**(523), 1372–1390 (2018)

Plummer, M.: Cuts in Bayesian graphical models. Stat. Comput. **25**(1), 37–43 (2015)

Roberts, G.O., Tweedie, R.L.: Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika **83**(1), 95–110 (1996)

Walker, S.G.: Bayesian inference with misspecified models. J. Stat. Plan. Inference **143**(10), 1621–1633 (2013)

Zigler, C.M.: The central role of Bayes' theorem for joint estimation of causal effects and propensity scores. Am. Stat. **70**(1), 47–54 (2016)