Approximate Inference for Observation-Driven Time Series Models with Intractable Likelihoods

AJAY JASRA, National University of Singapore
NIKOLAS KANTAS and ELENA EHRLICH, Imperial College London

In this article, we consider approximate Bayesian parameter inference for observation-driven time series models. Such statistical models appear in a wide variety of applications, including econometrics and applied mathematics. This article considers the scenario where the likelihood function cannot be evaluated pointwise; in such cases, one cannot perform exact statistical inference, including parameter estimation, which often requires advanced computational algorithms, such as Markov Chain Monte Carlo (MCMC). We introduce a new approximation based upon Approximate Bayesian Computation (ABC). Under some conditions, we show that as $n \to \infty$, with n the length of the time series, the ABC posterior has, almost surely, a Maximum A Posteriori (MAP) estimator of the parameters that is often different from the true parameter. However, a noisy ABC MAP, which perturbs the original data, asymptotically converges to the true parameter, almost surely. In order to draw statistical inference, for the ABC approximation adopted, standard MCMC algorithms can have acceptance probabilities that fall at an exponential rate in n and slightly more advanced algorithms can mix poorly. We develop a new and improved MCMC kernel, which is based upon an exact approximation of a marginal algorithm, whose cost per iteration is random, but the expected cost, for good performance, is shown to be $\mathcal{O}(n^2)$ per iteration. We implement our new MCMC kernel for parameter inference from models in econometrics.

Categories and Subject Descriptors: I.6 [**Simulation and Modeling**]; I.6.0: General

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Observation-driven time series models, approximate Bayesian computation, asymptotic consistency, Markov chain Monte Carlo

ACM Reference Format:
Ajay Jasra, Nikolas Kantas, and Elena Ehrlich. 2014. Approximate inference for observation-driven time series models with intractable likelihoods. ACM Trans. Model. Comput. Simul. 24, 3, Article 13 (April 2014), 25 pages.
DOI: http://dx.doi.org/10.1145/2592254

1. INTRODUCTION

Observation-driven time-series models, introduced by Cox [1981], have a wide variety of real applications, including econometrics (GARCH models) and applied mathematics (inferring initial conditions and parameters of ordinary differential equations). The model can be described as follows. We observe $\{Y_k\}_{k\in\mathbb{N}_0}$, $Y_k \in \mathsf{Y}$, which are associated to an unobserved process $\{X_k\}_{k\in\mathbb{N}_0}$, $X_k \in \mathsf{X}$ that is potentially unknown. Define the process $\{Y_k, X_k\}_{k\in\mathbb{N}_0}$ (with y_0 some arbitrary point on Y) on a probability space $(\Omega, \mathscr{F}, \mathbb{P}_\theta)$, where

This work is supported by the MOE Singapore.

Authors' addresses: A. Jasra, Department of Statistics and Applied Probability, Faculty of Science, Block S16, 6 Science Drive 2, National University of Singapore, Singapore, 117546, Republic of Singapore; email: staja@nus.edu.sg; N. Kantas and E. Ehrlich, Department of Mathematics, Imperial College London, London, SW7 2AZ, UK.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2014 ACM 1049-3301/2014/04-ART13 $15.00

DOI: http://dx.doi.org/10.1145/2592254

for every $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$, $\mathbb{P}_\theta$ is a probability measure. Denote by $\mathscr{F}_k = \sigma(\{Y_n, X_n\}_{0\leq n\leq k})$. The model is defined as, for $k \in \mathbb{N}_0$,

$$\begin{aligned}\mathbb{P}_\theta(Y_{k+1} \in A|\mathscr{F}_k) &= \int_A H^\theta(x_k, dy) \quad A \times \mathsf{X} \in \mathscr{F} \\ X_{k+1} &= \Phi^\theta(X_k, Y_{k+1}) \\ \mathbb{P}_\theta(X_0 \in B) &= \int_B \Pi_\theta(x_0)dx_0 \quad \mathsf{Y} \times B \in \mathscr{F},\end{aligned}$$

where $H : \Theta \times \mathsf{X} \times \sigma(\mathsf{Y}) \to [0, 1]$, $\Phi : \Theta \times \mathsf{X} \times \mathsf{Y} \to \mathsf{X}$, $\Pi_\theta(x_0)$ is a probability density on X for every $\theta \in \Theta$, and dx_0 is Lebesgue measure. Throughout, we assume that for any $(x, \theta) \in \mathsf{X} \times \Theta$ $H^\theta(x, \cdot)$ admits a density with respect to some $\sigma-$finite measure μ, which we denote as $h^\theta(x, y)$. Next, we define a prior probability distribution $\nu(\theta)d\theta$, $d\theta$ is Lebesgue measure on $(\Theta, \mathcal{B}(\Theta))$, and write $\xi(x_0, \theta) = \Pi_\theta(x_0)\nu(\theta)$ assumed to be a proper joint probability density on $\mathsf{X} \times \Theta$. Thus, given n observations $y_{1:n} := (y_1, \ldots, y_n)$, the object of inference is the posterior distribution on $\Theta \times \mathsf{X}$:

$$\Pi_n(d(\theta, x_0)|y_{1:n}) \propto \left(\prod_{k=1}^{n} h^\theta\left(\Phi^\theta_{k-1}(y_{0:k-1}, x_0), y_k\right)\right)\xi(x_0, \theta)dx_0d\theta, \tag{1}$$

where we have used the notation for $k > 1$, $\Phi^\theta_{k-1}(y_{0:k-1}, x_0) = \Phi^\theta \circ \cdots \circ \Phi^\theta(x_0, y_1)$, $\Phi^\theta_1(y_0, x_0) := \Phi^\theta(x_0, y_0)\Phi^\theta_0(x_0, y_0) := x_0$. In most applications of practical interest, the posterior cannot be computed pointwise, and one has to resort to numerical methods such as Markov Chain Monte Carlo (MCMC) to draw inference on θ and/or x_0.

In this article, we are not only interested in inferring the posterior distribution, but the scenario for which $h^\theta(x, y)$ cannot be evaluated pointwise, nor do we have access to a positive unbiased estimate of it (it is assumed that we can simulate from the associated distribution). In such a case, it is not possible to draw inference from the true posterior, even using numerical techniques. The common response in Bayesian statistics is now to adopt an approximation of the posterior using the notion of Approximate Bayesian Computation (ABC); see Marin et al. [2012] for a recent overview. ABC approximations of posteriors are based upon defining a probability distribution on an extended state-space, with the additional random variables lying on the data-space and usually distributed according the true likelihood. The closeness of the ABC posterior distribution is controlled by a tolerance parameter $\epsilon > 0$, and for some ABC approximations (but not all) the approximation is exact as $\epsilon \to 0$; the approximation introduced in this article will be exact when $\epsilon = 0$.

In this work, we introduce a new ABC approximation of observation-driven time-series models that is closely associated to that developed in Jasra et al. [2012] for Hidden Markov Models (HMMs) and later for static parameter inference from HMMs [Dean et al. 2011]. This latter ABC approximation is particularly well behaved and a noisy variant (which perturbs the data; (e.g., see Dean et al. [2011] and Fearnhead and Prangle [2012]) is shown under some assumptions to provide Maximum-Likelihood Estimators (MLEs) that asympotically in n (with n the length of the time series) are the true parameters. The new ABC approximation that we develop is studied from a theoretical perspective. Relying on the recent work of Douc et al. [2012], we show that, under some conditions, as $n \to \infty$, the ABC posterior has, almost surely, a MAP estimator of θ that converges to a point (or collection of points) that is typically different from the true parameter θ^* say. However, a noisy ABC MAP of θ asymptotically converges to the true parameter, almost surely. These results establish that the particular approximation adopted is reasonably sensible.

The other main contribution of this article is a development of a new MCMC algorithm designed to sample from the ABC approximation of the posterior. Due to the nature of the ABC approximation, it is easily seen that standard MCMC algorithms (e.g., Majoram et al. [2003]) will have an acceptance probability that will fall at an exponential rate in n. In addition, more advanced ideas, such as those based upon the pseudomarginal [Beaumont 2003], have recently been shown to perform rather poorly in theory; see Lee and Latuszynski [2012]. These latter algorithms are based upon exact approximations of marginal algorithms [Andrieu et al. 2010; Andrieu and Vihola 2012], which in our context is just sampling θ, x_0. We develop an MCMC kernel, related to recent work in Lee [2012], which is designed to have a random running time per iteration, with the idea of improving the exploration ability of the Markov chain. We show that the expected cost per iteration of the algorithm, under some assumptions and for reasonable performance, is $\mathcal{O}(n^2)$, which compares favourably with competing algorithms. We also show, empirically, that this new MCMC method out-performs standard pseudomarginal algorithms.

This article is structured as follows. In Section 2, we introduce our ABC approximation and give our theoretical results on the MAP estimator. In Section 3, we give our new MCMC algorithm, along with some theoretical discussion about its computational cost and stability. In Section 4, our approximation and MCMC algorithm are illustrated on toy and real examples. In Section 5, we conclude the article with some discussion of future work. The proofs of our theoretical results are given in the Appendix.

2. APPROXIMATE POSTERIORS USING ABC APPROXIMATIONS

2.1. ABC Approximations and Noisy ABC

As it was emphasised in Section 1, we are interested in performing inference when $h^\theta(x, y)$ cannot be evaluated pointwise and we do not have access to a positive unbiased estimate of it. We will instead assume that it is possible to sample from h^θ. In such scenarios, one cannot use standard simulation-based methods. For example, in an MCMC approach, the Metropolis-Hastings (M-H) acceptance ratio cannot be evaluated even though it may be well defined.

Following the work in Dean et al. [2011] and Jasra et al. [2012] for HMMs, we introduce an ABC approximation for the density of the posterior in (1) as follows:

$$\pi_n^\epsilon(\theta, x_0|y_{1:n}) \propto \xi(x_0, \theta) \prod_{k=1}^{n} h^{\theta,\epsilon}\big(\Phi_{k-1}^\theta(y_{0:k-1}, x_0), y_k\big), \tag{2}$$

with $\epsilon > 0$ and

$$h^{\theta,\epsilon}\big(\Phi_{k-1}^\theta(y_{0:k-1}, x_0), y_k\big) = \frac{\int_{B_\epsilon(y_k)} h^\theta\big(\Phi_{k-1}^\theta(y_{0:k-1}, x_0), y\big)\mu(dy)}{\mu(B_\epsilon(y_k))}, \tag{3}$$

where we denote $B_\epsilon(y)$ as the open ball centred at y with radius ϵ and write $\mu(B_\epsilon(y)) = \int_{B_\epsilon(y)} \mu(dx)$. Note that whilst $h^{\theta,\epsilon}$ is not available in closed form, we will be able to design algorithms that can sample from $\pi_n^\epsilon(\theta, x_0|y_{1:n})$ by sampling on an extended state-space; this is discussed in Section 3. A similar approximation can be found in Jasra et al. [2012] and Barthelmé and Chopin [2011] but for different models. As noted in the aforementioned articles, approximations of the form (2) are particularly sensible in that they not only retain the probabilistic structure of the original statistical model but, in addition, facilitate simple implementation of statistical computational methods. The former point is particularly useful in that one can study the properties (and accuracy) of the ABC approximation using similar mathematical tools to the original model; this is illustrated in Section 2.2.

In general, we will refer to ABC as the procedure of performing inference for the posterior in (2). In addition, we will call *noisy ABC* the inference procedure that uses a perturbed sequence of data instead of the observed one. This sequence is $\{\hat{Y}_k\}_{k\geq 0}$, where, conditional upon the observations and independently, $\hat{Y}_k|Y_k = y_k \sim \mathcal{U}_{B_\epsilon(y_k)}$ (uniformly distributed on $B_\epsilon(Y_k)$). We remark that noisy ABC had been developed elsewhere in Dean et al. [2011] and Fearnhead and Prangle [2012]. In particular, the theoretical results presented in Section 2.2 have been proved for ABC approximations of HMMs in Dean et al. [2011]; indeed, the results of this article are less precise than in Dean et al. [2011] with a similar deduction—that noisy ABC can remove bias in parameter estimation as the number of data grow. Dean et al. [2011] is based upon a particular identity for ABC approximations of HMMs and is preceded by a more general (and related) identity in Wilkinson et al. [2013].

2.2. Consistency Results for the MAP Estimator

In this section, we will investigate some interesting properties of the ABC posterior in (2). In particular, we will look at the asymptotic behaviour with n of the resulting MAP estimators for θ. The properties of the MAP estimator reveal information about the mode of the posterior distribution as we obtain increasingly more data. Throughout the section, we will extend the process to have doubly infinite time indices (i.e., on $\mathbb{Z}$) and use notations such as $y_{-\infty:k}$, $k > -\infty$ to denote sequences from the infinite past. Throughout, $\epsilon > 0$ is fixed. To simplify the analysis in this section, we will assume that

(A1) —x_0 is fixed and known.
—$\nu(\theta)$ is bounded and positive everywhere in Θ.
—H and h do not depend upon θ. Thus, we have the following model recursions for the true model, for some $\theta^* \in \Theta$:

$$\begin{aligned} \mathbb{P}_{\theta^*}(Y_{k+1} \in A|\mathscr{F}_k) &= \int_A H(x_k, dy), \quad A \times \mathsf{X} \in \mathscr{F}, \\ X_{k+1} &= \Phi^{\theta^*}(X_k, Y_{k+1}), \end{aligned} \tag{4}$$

where we will denote associated expectations to $\mathbb{P}_{\theta^*}$ as $\mathbb{E}_{\theta^*}$.

In addition, for this section we will introduce some extra notations: (X, d) is a compact, complete and separable metric space and (Θ, d) is a compact metric space, with $\Theta \subset \mathbb{R}^{d_\theta}$. Let $\mathbb{Q}_\epsilon$ be the conditional probability law associated to the random sequence $\{\hat{Y}_k\}_{k\in\mathbb{Z}}$, defined above.

We proceed with some additional technical assumptions:

(A2) $\{X_k, Y_k\}_{k\in\mathbb{Z}}$ is a stationary stochastic process, with $\{Y_k\}_{k\in\mathbb{Z}}$ strict sense stationary and ergodic, following (4). See Fan and Yao [2005, Definition 2.2] for a definition of strict sense stationary.

(A3) For every $(x, y) \in \mathsf{X} \times \mathsf{Y}$, $\theta \mapsto \Phi^\theta(x, y)$ is continuous. In addition, there exist $0 < C < \infty$ such that for any $(x, x') \in \mathsf{X}$, $\sup_{y\in\mathsf{Y}} |h(x, y) - h(x', y)| \leq Cd(x, x')$. Finally, $0 < \underline{h} \leq h(x, y) \leq \overline{h} < \infty$, for every $(x, y) \in \mathsf{X} \times \mathsf{Y}$.

(A4) There exist a measurable function $\varrho : \mathsf{Y} \to (0, \overline{\varrho})$, $0 < \overline{\varrho} < 1$, such that for every $(\theta, y, x, x') \in \Theta \times \mathsf{Y} \times \mathsf{X}^2$

$$d(\Phi^\theta(x, y), \Phi^\theta(x', y)) \leq \varrho(y)d(x, x').$$

Under the assumptions thus far, for any $x \in \mathsf{X}$, $\lim_{m\to\infty} \Phi^\theta_{m+1}(Y_{-m:0}, x)$ exists (resp. $\lim_{m\to\infty} \Phi^\theta_{m+1}(\hat{Y}_{-m:0}, x)$) and is independent of x, $\mathbb{P}_{\theta^*}$ a.s. (resp. $\mathbb{P}_{\theta^*} \otimes \mathbb{Q}_\epsilon$ (product measure) a.s.); write the limit $\Phi^\theta_\infty(Y_{-\infty:0})$ (resp. $\Phi^\theta_\infty(\hat{Y}_{-\infty:0})$). See the proof of Lemma 20 of Douc et al. [2012] for details.

(A5) The following statements hold:
   (a) If $H(x, \cdot) = H(x', \cdot)$, then $x = x'$.
   (b) If $\Phi^{\theta}_{\infty}(\hat{Y}_{-\infty:0}) = \Phi^{\theta^*}_{\infty}(\hat{Y}_{-\infty:0})$ holds $\mathbb{P}_{\theta^*} \otimes \mathbb{Q}_{\epsilon}$−a.s., then $\theta = \theta^*$.

Assumptions (A2–A5) and the compactness of Θ are standard assumptions for maximum-likelihood estimation, and they can be used to show the uniqueness of the MLE; see Douc et al. [2012] for more details (see also Appendix A, where the assumptions of Douc et al. [2012] are given, and Cappé et al. [2005, Chapter 12] in the context of HMMs). Note that $0 < \underline{h} \leq h(x, y) \leq \overline{h} < \infty$, for every $(x, y) \in \mathsf{X} \times \mathsf{Y}$ will typically hold if $\mathsf{X} \times \mathsf{Y}$ is compact and, again, is a typical assumption used in the analysis of HMMs for the observation density (although weaker assumptions have been adopted). If the prior $\nu(\theta)$ is bounded and positive everywhere on Θ, it is a simple corollary that the MAP estimator will correspond to the MLE. In the remaining part of this section, we will adapt the analysis in Douc et al. [2012] for MLE to the ABC setup. We remark that the assumptions are similar to Douc et al. [2012], as the asymptotic analysis of such models is in its infancy; the objective of this article is not to make advances in the theory but more to consider the approximation introduced in this article. We note also that some of the assumptions are verified in Douc et al. [2012] for some examples, and we direct the reader to that article for further discussion.

In particular, we are to estimate θ using the log-likelihood function:

$$l_{\theta,x}(y_{1:n}) := \frac{1}{n}\sum_{k=1}^{n} \log\left(h^{\epsilon}\left(\Phi^{\theta}_{k-1}(y_{0:k-1}, x), y_k\right)\right).$$

We define the ABC MLE for n observations as

$$\theta_{n,x,\epsilon} = \arg\max_{\theta\in\Theta} l_{\theta,x}(y_{1:n}).$$

We proceed with the following proposition, whose proof is in Appendix A:

PROPOSITION 2.1. *Assume (A1–A4). Then for every $x \in \mathsf{X}$ and fixed $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathsf{d}(\theta_{n,x,\epsilon}, \Theta^*_{\epsilon}) = 0 \quad \mathbb{P}_{\theta*} - a.s.,$$

where $\Theta^*_{\epsilon} = \arg\max_{\theta\in\Theta}\mathbb{E}_{\theta^*}[\log(h^{\epsilon}(\Phi^{\theta}_{\infty}(Y_{-\infty:0}), Y_1))]$.

The result establishes that the estimate will converge to a point (or collection of points), which is typically different to the true parameter. That is to say, it is not always the case (without additional assumptions) that $\Theta_{\epsilon} = \{\theta^*\}$, which is shown next. Hence, there is often an intrinsic asymptotic bias for the plain ABC procedure. To correct this bias, we consider the noisy ABC procedure; this replaces the observations by $\hat{Y}_k$. The noisy ABC MLE estimator is then

$$\hat{\theta}_{n,x,\epsilon} = \arg\max_{\theta\in\Theta}\frac{1}{n}\sum_{k=1}^{n} \log\left(h^{\epsilon}\left(\Phi^{\theta}_{k-1}(\hat{y}_{0:k-1}, x), \hat{y}_k\right)\right).$$

We have the following result, whose proof is also in Appendix A:

PROPOSITION 2.2. *Assume (A1–A5) and that $X_0 = \Phi^{\theta}_{\infty}(\hat{Y}_{-\infty:0})$. Then for every $x \in \mathsf{X}$ and fixed $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathsf{d}(\hat{\theta}_{n,x,\epsilon}, \theta^*) = 0 \quad \mathbb{P}_{\theta*} \otimes \mathbb{Q}_{\epsilon} - a.s..$$

The result shows that the noisy ABC MLE estimator is asymptotically unbiased. Therefore, given that in our setup the ABC MAP estimator corresponds to the ABC MLE, we can conclude that the mode of the posterior distribution as we obtain increasingly more data is converging towards the true parameter.

3. COMPUTATIONAL METHODOLOGY

Recall that we formulated the ABC posterior in (2), which can also be written as

$$\pi_n^{\epsilon}(\theta, x_0|y_{1:n}) = \frac{p^{\epsilon}_{\theta,x_0}(y_{1:n})\xi(x_0,\theta)}{\int p^{\epsilon}_{\theta,x_0}(y_{1:n})\xi(x_0,\theta)dx_0 d\theta},$$

with

$$p^{\epsilon}_{\theta,x_0}(y_{1:n}) = \int \prod_{k=1}^{n} \frac{\mathbb{I}_{B_\epsilon(y_k)}(u_k)}{\mu(B_\epsilon(y_k))} h^{\theta}\left(\Phi^{\theta}_{k-1}(y_{0:k-1}, x_0), u_k\right) du_{1:n}.$$

Note that we have just used Fubini's theorem to rewrite the likehood $p^{\epsilon}_{\theta,x_0}(y_{1:n})$ as an integral of a product instead of a product of integrals of $\prod_{k=1}^{n} h^{\theta,\epsilon}(\Phi^{\theta}_{k-1}(y_{0:k-1}, x_0), y_k)$. In this article, we will focus only on MCMC algorithms and in particular on the M-H approach; in Section 3.2.5, we discuss alternative methodologies and our contribution in this context. In order to sample from the posterior π_n^{ϵ} one runs an ergodic Markov chain with the invariant density being π_n^{ϵ}. Then, after a few iterations when the chain has reached stationarity, one can treat the samples from the chain as approximate samples from π_n^{ϵ}. This is shown in Algorithm 1, where for convenience we denote $\gamma = (\theta, x_0)$. The one-step transition kernel of the MCMC chain is usually described as the *M-H kernel* and follows from Step 2 in Algorithm 1.

ALGORITHM 1: A marginal M-H algorithm for $\pi^{\epsilon}(\gamma|y_{1:n})$

```
(1) (Initialisation) At $t = 0$, sample $\gamma_0 \sim \xi$.
(2) (M-H kernel) For $t \geq 1$:
      —Sample $\gamma'|\gamma_{t-1}$ from a proposal $Q(\gamma_{t-1}, \cdot)$ with density $q(\gamma_{t-1}, \cdot)$.
      —Accept the proposed state and set $\gamma_t = \gamma'$ with probability
          $1 \wedge \frac{p^{\epsilon}_{\gamma'}(y_{1:n})}{p^{\epsilon}_{\gamma_{t-1}}(y_{1:n})} \times \frac{\xi(\gamma')q(\gamma', \gamma_{t-1})}{\xi(\gamma_{t-1})q(\gamma_{t-1}, \gamma')}$,
       otherwise set $\gamma_t = \gamma_{t-1}$. Set $t = t + 1$ and return to the start of 2.
```

Unfortunately, $p^{\epsilon}_{\theta,x_0}(y_{1:n})$ is not available analytically and cannot be evaluated, so this rules out the possibility of using traditional MCMC approaches such as Algorithm 1. However, one can resort to the so-called pseudomarginal approach whereby positive unbiased estimates of $p^{\epsilon}_{\theta,x_0}(y_{1:n})$ are used instead within an MCMC algorithm; e.g., see Andrieu et al. [2010] and Andrieu and Vihola [2012]. We will refer to this algorithm as ABC-MCMC (despite the fact that this terminology has been used in other contexts in the literature). The resulting algorithm can be posed as one targeting a posterior defined on an extended state-space so that its marginal coincides with $\pi_n^{\epsilon}(\theta, x_0|y_{1:n})$. We will use these ideas to present ABC-MCMC as an M-H algorithm that is an exact approximation to an appropriate marginal algorithm.

To illustrate an example of these ideas, we proceed by writing a posterior on an extended state-space $\Theta \times \mathsf{X} \times \mathsf{Y}^n$ as follows:

$$\pi_n^{\epsilon}(\theta, x_0, u_{1:n}|y_{1:n}) \propto \xi(x_0,\theta) \prod_{k=1}^{n} \mathbb{I}_{B_\epsilon(y_k)}(u_k) h^{\theta}\left(\Phi^{\theta}_{k-1}(y_{0:k-1}, x_0), u_k\right). \tag{5}$$

It is clear that (2) is the marginal of (5), and hence the similarity in the notation. As we will show later in this section, extending the target space in the posterior as in (5) is not the only choice. We emphasise that the only essential requirement for each choice is that the marginal of the extended target is $\pi_n^\epsilon(\theta, x_0|y_{1:n})$, but one should be cautious because the particular choice will affect the mixing properties and the efficiency of the MCMC scheme that will be used to sample from $\pi_n^\epsilon(\theta, x_0, u_{1:n}|y_{1:n})$ in (5) or another variant.

3.1. Standard Approaches for ABC-MCMC

We will now look at two basic different choices for extending the ABC posterior while keeping the marginal fixed to $\pi^\epsilon(\theta, x_0|y_{1:n})$. In the remainder of the article, we will denote $\gamma = (\theta, x_0)$ as we did in Algorithm 1.

Initially consider the ABC approximation when extended to the space $\Theta \times \mathsf{X} \times \mathsf{Y}^n$:

$$\pi_n^\epsilon(\gamma, u_{1:n}|y_{1:n}) = \frac{\xi(\gamma)p_\gamma^\epsilon(y_{1:n})}{\int \xi(\gamma)p_\gamma^\epsilon(y_{1:n})d\gamma} \frac{\prod_{k=1}^n \frac{\mathbb{I}_{B_\epsilon(y_k)}(u_k)}{\mu(B_\epsilon(y_k))}}{p_\gamma^\epsilon(y_{1:n})} \prod_{k=1}^n h^\theta\left(\Phi_{k-1}^\theta(y_{0:k-1}, x_0), u_k\right).$$

Recall that one cannot evaluate $h^\theta(\Phi_{k-1}^\theta(y_{0:k-1}, x_0), u_k)$ and is only able to simulate from it. In Algorithm 2, we present a natural M-H proposal that could be used to sample from $\pi_n^\epsilon(\gamma, u_{1:n}|y_{1:n})$ instead of the one shown in Step 2 of Algorithm 1. Note that this time, the state of the MCMC chain is composed of $(\gamma, u_{1:n})$. Here, each u_k assumes the role of an auxiliary variable to be eventually integrated out at the end of the MCMC procedure.

ALGORITHM 2: M-H proposal for basic ABC-MCMC

```
—Sample $\gamma'|\gamma$ from a proposal $Q(\gamma, \cdot)$ withdensity $q(\gamma, \cdot)$.
—Sample $u_{1:n}^{'}$ from a distribution with joint density $\prod_{k=1}^n h^{\theta'}\left(\Phi_{k-1}^{\theta'}(y_{0:k-1}, x_0), u_k\right)$.
—Accept the proposed state $\left(\gamma', u_{1:n}'\right)$ with probability:
    $$1 \wedge \frac{\prod_{k=1}^n \mathbb{I}_{B_\epsilon(y_k)}(u_k^{'})}{\prod_{k=1}^n \mathbb{I}_{B_\epsilon(y_k)}(u_k)} \times \frac{\xi(\gamma')q(\gamma', \gamma)}{\xi(\gamma)q(\gamma, \gamma')}.$$
```

As n increases, the M-H kernel in Algorithm 2 will have an acceptance probability that falls quickly with n. In particular, for any fixed γ, the probability of obtaining a sample in $B_\epsilon(y_1) \times \cdots \times B_\epsilon(y_n)$ will fall at an exponential rate in n. This means that this basic ABC-MCMC approach will be inefficient for moderate values of n.

This issue can be dealt with by using N multiple trials so that at each k, some auxiliary variables (or pseudoobservations) are in the ball $B_\epsilon(y_k)$. This idea originates from Beaumont [2003] and Majoram et al. [2003] and in fact augments the posterior to a larger state-space, $\Theta \times \mathsf{X} \times \mathsf{Y}^{nN}$, in order to target the following density:

$$\widetilde{\pi}_n^\epsilon\left(\gamma, u_{1:n}^{1:N}|y_{1:n}\right) = \frac{\xi(\gamma)p_\gamma^\epsilon(y_{1:n})}{\int \xi(\gamma)p_\gamma^\epsilon(y_{1:n})d\gamma} \frac{\prod_{k=1}^n \frac{\sum_{j=1}^N \mathbb{I}_{B_\epsilon(y_k)}(u_k^j)}{N\mu(B_\epsilon(y_k))}}{p_\gamma^\epsilon(y_{1:n})} \prod_{k=1}^n \prod_{j=1}^N h^\theta\left(\Phi_{k-1}^\theta(y_{0:k-1}, x_0), u_k^j\right).$$

Again, one can show that the marginal of interest $\pi^\epsilon(\gamma|y_{1:n})$ is preserved—that is,

$$\pi_n^\epsilon(\gamma|y_{1:n}) = \int_{\mathsf{Y}^{nN}} \widetilde{\pi}_n^\epsilon\left(\gamma, u_{1:n}^{1:N}|y_{1:n}\right)du_{1:n}^{1:N} = \int_{\mathsf{Y}^n} \pi_n^\epsilon(\gamma, u_{1:n}|y_{1:n})du_{1:n}.$$

In Algorithm 3, we present an M-H kernel with invariant density $\widetilde{\pi}_n^{\epsilon}$. The state of the MCMC chain now is $(\gamma, u_{1:n}^{1:N})$. We remark that as N grows, one expects to recover the properties of the ideal M-H algorithm in Algorithm 1. Nevertheless, it has been shown in Lee and Latuszynski [2012] that even the M-H kernel in Algorithm 3 does not always perform well. It can happen that the chain often gets stuck in regions of the state-space $\Theta \times \mathsf{X}$ where

$$\alpha_k(y_{1:k}, \epsilon, \gamma) := \int_{B_\epsilon(y_k)} h^{\theta}\big(\Phi_{k-1}^{\theta}(y_{0:k-1}, x_0), u\big)du$$

is small. Given this notation, we remark that

$$p_\gamma^\epsilon(y_{1:n}) = \prod_{k=1}^{n} \frac{\alpha_k(y_{1:k}, \epsilon, \gamma)}{\mu(B_\epsilon(y_k))},$$

which is useful to note from here on.

ALGORITHM 3: M-H proposal for ABC with N trials

```
—Sample $\gamma'|\gamma$ from a proposal $Q(\gamma, \cdot)$ withdensity $q(\gamma, \cdot)$.
—Sample $u'^{1:N}_{1:n}$ from a distribution with jointdensity
  $\prod_{k=1}^{n}\prod_{j=1}^{N} h^{\theta'}\big(\Phi_{k-1}^{\theta'}(y_{0:k-1}, x_0), u'^{j}_{k}\big)$.
—Accept the proposed state $\big(\gamma', u'^{1:N}_{1:n}\big)$ with probability:
  $1 \wedge \frac{\prod_{k=1}^{n}(\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}_{B_\epsilon(y_k)}(u'^{j}_{k}))}{\prod_{k=1}^{n}(\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}_{B_\epsilon(y_k)}(u^{j}_{k}))} \times \frac{\xi(\gamma')q(\gamma', \gamma)}{\xi(\gamma)q(\gamma, \gamma')}$.
```

3.2. An M-H Kernel for ABC with a Random Number of Trials

We will address this shortfall detailed previously by proposing an alternative augmented target and corresponding M-H kernel. Intrinsically, on inspection of Algorithm 3, one would like to ensure that many of the $N > 1$ samples, $u_k^{'j}$, will lie in $B_\epsilon(y_k)$, whilst not necessarily being more clever about the proposal mechanism. The basic idea is that one will use the same simulation mechanism but *ensure* that we will have all N samples, $u_k^{'j}$, in $B_\epsilon(y_k)$. The by-product of the strategy we adopt, so that we sample from $\pi_n^\epsilon(\gamma|y_{1:n})$, will be that the simulation time per iteration of the MCMC kernel will be a random variable; in words, this idea is as follows. At a given iteration of the MCMC, we will sample the timesteps of the model sequentially after first proposing a new γ, call this γ'. At each timestep k (associated to the model), we keep on sampling the $u_k^{'j}$ until there are exactly N in $B_\epsilon(y_k)$. The number of samples to achieve this is a random variable, and conditional on γ' its distribution is known. This is a negative binomial random variable with success probability $\alpha_k(y_{1:k}, \epsilon, \gamma')$. The contribution here will be to formulate a particular extended target distribution on γ and the number of simulated samples at each timestep so that one can use the proposal mechanism hinted at previously and still sample from $\pi_n^\epsilon(\gamma|y_{1:n})$.

Consider an alternative extended target, for $N \geq 2$, $m_k \in \{N, N+1, \ldots, \}$, $1 \leq k \leq n$:

$$\hat{\pi}_n^{\epsilon}(\gamma, m_{1:n}|y_{1:n}) = \frac{\xi(\gamma)p_\gamma^\epsilon(y_{1:n})}{\int \xi(\gamma)p_\gamma^\epsilon(y_{1:n})d\gamma} \frac{\prod_{k=1}^{n}\frac{N-1}{\mu(B_\epsilon(y_k))(m_k-1)}}{p_\gamma^\epsilon(y_{1:n})} \times \prod_{k=1}^{n}\binom{m_k-1}{N-1}\alpha_k(y_{1:k}, \gamma, \epsilon)^N(1-\alpha_k(y_{1:k}, \gamma, \epsilon))^{m_k-N}.$$

Standard results for negative binomial distributions (see Neuts and Zacks [1967] and Zacks [1980] for more details, as well as Appendix B.1) imply that

$$\sum_{m_k=N}^{\infty} \frac{1}{m_k - 1}\binom{m_k - 1}{N-1}\alpha_k(y_{1:k}, \epsilon, \gamma)^N(1 - \alpha_k(y_{1:k}, \epsilon, \gamma))^{m_k - N} = \frac{\alpha_k(y_{1:k}, \epsilon, \gamma)}{N-1}. \qquad (6)$$

It then follows that from (6), that $p^{\epsilon}_{\gamma}(y_{1:n})$ is equal to

$$\sum_{m_{1:n}\in\{N,N+1,\dots\}^n} \prod_{k=1}^{n}\left[\frac{N-1}{\mu(B_\epsilon(y_k))(m_k-1)}\binom{m_k-1}{N-1}\alpha_k(y_{1:k}, \gamma, \epsilon)^N(1-\alpha_k(y_{1:k}, \gamma, \epsilon))^{m_k-N}\right]$$

and thus that the marginal with respect to γ is the one of interest:

$$\pi^{\epsilon}_n(\gamma|y_{1:n}) = \sum_{m_{1:n}\in\{N,N+1,\dots\}^n} \widehat{\pi}^{\epsilon}_n(\gamma, m_{1:n}|y_{1:n}).$$

In Algorithm 4, we present an M-H kernel with invariant density $\widehat{\pi}^{\epsilon}_n$. The state of the MCMC chain this time is $(\gamma, m_{1:n},)$ and the proposal mechanism is as described at the start of this section.

ALGORITHM 4: M-H proposal with a random number of trials

```
—Sample $\gamma'|\gamma$ from a proposal $Q(\gamma, \cdot)$ withdensity $q(\gamma, \cdot)$.
—For $k = 1, \dots, n$ repeat the following: sample $u^1_k, u^2_k, \dots$ with probability density
  $h^{\theta'}(\Phi^{\theta'}_{k-1}(y_{0:k-1}, x'_0), u_k)$ until there are $N$ samples lying in $B_\epsilon(y_k)$; thenumber of samples to
  achieve this (including the successful trial) is $m'_k$.
—Accept $(\gamma', m'_{1:n})$ with probability:
      $1 \wedge \frac{\prod_{k=1}^{n}\frac{1}{m'_k-1}}{\prod_{k=1}^{n}\frac{1}{m_k-1}} \times \frac{\pi(\gamma')q(\gamma', \gamma)}{\pi(\gamma)q(\gamma, \gamma')}.$
```

The potential benefit of the kernel associated to Algorithm 4 is that one expects the probability of accepting a proposal is higher than the previous M-H kernel associated with Algorithm 3 (for a given N). This comes at a computational cost that is both increased and random; this may not be a negative, in the sense that the associated mixing time of the MCMC kernel may fall relative to the proposal considered in Algorithm 3 whose computational cost per iteration is deterministic. The proposed kernel is based on the N*−hit* kernel of Lee [2012], which has been adapted here to account for the data being a sequence of observations resulting from a time series.

3.2.1. On the choice of N. To implement Algorithm 4, one needs to select N. We now present a theoretical result that can provide some intuition on choosing a sensible value of N. Let $\mathbb{E}_{\gamma,N}[\cdot]$ denote expectation with respect to $\prod_{k=1}^{n}\binom{m_k-1}{N-1}\alpha_k(y_{1:k}, \epsilon, \gamma)^N(1 - \alpha_k(y_{1:k}, \epsilon, \gamma))^{m_k-N}$ given γ, N; we use the capital symbols $M_{1:n}$ in the expectation operator. One sensible way to select N as function of n, in Algorithm 4, is so that the relative variance associated with (c.f. the acceptance probability in Algorithm 4)

$$\prod_{k=1}^{n}\frac{1}{m_k-1}$$

(conditional on γ, $m_{1:n}$ are generated from $\prod_{k=1}^{n}\binom{m_k-1}{N-1}\alpha_k(y_{1:k}, \epsilon, \gamma)^N(1-\alpha_k(y_{1:k}, \epsilon, \gamma))^{m_k-N}$) will not grow with n; in general, one might expect the algorithm to get worse as n grows.

In other words, if N can be chosen to control the relative variance described previously, with respect to n, then one might hope that a major contributor to the instability of the M-H algorithm, that is growing n, is controlled and the resulting M-H algorithm can converge quickly. We will assume that the observations are fixed and known and will adopt the additional assumption:

(A6) For any fixed $\epsilon > 0$, $\gamma \in \Theta \times \mathsf{X}$, we have $\alpha_k(y_{1:k}, \epsilon, \gamma) > 0$.

The following result holds, whose proof can be found in Appendix B.2:

Proposition 3.1. *Assume (A6) and let* $\beta \in (0, 1)$, $n \geq 1$, *and* $N \geq \frac{2n}{1-\beta} \vee 3$. *Then for fixed* $(\gamma, \epsilon) \in \Theta \times \mathsf{X} \times \mathbb{R}^+$, *we have*

$$\mathbb{E}_{\gamma,N}\left[\left(\frac{\prod_{k=1}^{n}\frac{1}{\mu(B_\epsilon(y_k))(M_k-1)}}{p_\gamma^\epsilon(y_{1:n})} - 1\right)^2\right] \leq \frac{Cn}{N},$$

where $C = 1/\beta$.

The result shows that one should set $N = \mathcal{O}(n)$ for the relative variance not to grow with n, which is unsurprising, given the conditional independence structure of the $m_{1:n}$. To get a better handle on the variance, suppose that $n = 1$, then for γ fixed and taking expectations with respect to $\prod_{j=1}^{N} h^\theta(x_0, u_1^j)$ (i.e., considering the proposal in Algorithm 3)

$$\mathbb{V}\mathrm{ar}_{\gamma,N}\left[\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}_{B_\epsilon(y_1)}(u_1^j)\right] = \frac{\alpha_1(y_1, \epsilon, \gamma)(1 - \alpha_1(y_1, \epsilon, \gamma))}{N}. \tag{7}$$

In the context of Algorithm 4, one can show that when taking expectations with respect to $\binom{m_1-1}{N-1}\alpha_1(y_1, \epsilon, \gamma)^N(1 - \alpha_k(y_1, \epsilon, \gamma))^{m_1-N}$ (see the Remarks in Appendix B.1),

$$\mathbb{V}\mathrm{ar}_{\gamma,N}\left[\frac{N-1}{M_1-1}\right] \leq \frac{\alpha_1(y_1, \epsilon, \gamma)^2}{(N-2)}. \tag{8}$$

The variance in (8) is less than or equal to that in (7) if

$$\frac{N}{N-2} \leq \frac{1 - \alpha_1(y_1, \epsilon, \gamma)}{\alpha_1(y_1, \epsilon, \gamma)},$$

which is likely to occur if $\alpha_1(y_1, \epsilon, \gamma)$ is not too large (recall that we want ϵ to be small so that we have a good approximation of the true posterior) and N is moderate—this is precisely the scenario in practice. Note that the issue of computational cost, which is not taken into account, is very important. This reduces the possible impact of the previous discussion, but now we have some information on when (and if ever) the new proposal in Algorithm 4 could perform better than that in Algorithm 3. This at least suggests that one might want to try Algorithm 4 in practice.

Remark 3.1. In the context of Algorithm 3, it can be shown, when taking expectations with respect to $\prod_{k=1}^{n}\prod_{j=1}^{N} h^\theta(\Phi_{k-1}^\theta(y_{0:k-1}\ , x_0), u_k^j)$ and fixing γ, N (writing this again as $\mathbb{E}_{\gamma,N}$), that

$$\mathbb{E}_{\gamma,N}\left[\left(\prod_{k=1}^{n}\left[\left(\frac{1}{N\mu(B_\epsilon(y_k))}\sum_{j=1}^{N}\mathbb{I}_{B_\epsilon(y_k)}(u_k^j)\right)\right]\Big/ p_\gamma^\epsilon(y_{1:n}) - 1\right)^2\right] = \prod_{k=1}^{n}\left[\frac{1}{\alpha_k(y_{1:k}, \epsilon, \gamma)N} + \frac{N-1}{N}\right] - 1 \tag{9}$$

compare to the acceptance probability in Algorithm 3 to see the relevance of this. Note that the preceding quantity is not uniformly upper bounded in γ unless $\inf_{k,\gamma} \alpha_k(y_{1:k}, \epsilon, \gamma) \geq C > 0$, which may not occur. Conversely, Proposition 3.1 shows that the relative variance associated with the proposal in Algorithm 4 is uniformly upper bounded in γ under minimal conditions. We suspect that this means in practice that the kernel with random number of trials may mix faster than an MCMC kernel using the proposal in Algorithm 3.

3.2.2. Computational Considerations. As the cost per iteration of Algorithm 4 is random, we will investigate this further. We denote the proposal of $\gamma, m_{1:n}$ as $\tilde{Q}$ (i.e., as in Algorithm 4). Let ζ be the initial distribution of the MCMC chain and ζK^t the distribution of the state at time t. In addition, denote by m_k^t the proposed state for m_k at iteration t. We will write the expectation with respect to $\zeta K^t \otimes \tilde{Q}$ as $\mathbb{E}_{\zeta K^t \otimes \tilde{Q}}$. We will assume that the observations are fixed and known. Then, we have the following result:

PROPOSITION 3.2. *Let $\epsilon > 0$, and suppose that there exists a constant $C > 0$ such that for any $n \geq 1$, we have $\inf_{k,\gamma} \alpha_k(y_{1:k}, \gamma, \epsilon) \geq C$, μ–a.e.. Then, it holds for any $N \geq 2$, $t \geq 1$, that*

$$\mathbb{E}_{\zeta K^t \otimes \tilde{Q}}\left[\sum_{k=1}^{n} M_k^t\right] \leq \frac{nN}{C}.$$

The expected value of $\sum_{k=1}^{n} m_k^t$ grows at most linearly with n when taking expectations with respect to $\zeta K^t \otimes \tilde{Q}$. By Proposition 3.1, we should scale N linearly with n to control the relative variance of the proposed $\prod_{k-1}^{n} 1/(m_k^t - 1)$ (uniformly in γ and irrespective of t), and on average, we expect that we will need to wait $\mathcal{O}(n)$ timesteps to generate all of the $\{m_k^t\}_{k=1}^{n}$ so that approximately the computational cost is $\mathcal{O}(n^2)$ per iteration. This approximate cost of $\mathcal{O}(n^2)$ per iteration is comparable to many exact approximations of MCMC algorithms (e.g., Andrieu et al. [2010]), albeit in a much simpler situation.

Note also that the kernel in Algorithm 3 is expected to require a cost of $\mathcal{O}(n^2)$ per iteration for reasonable performance (i.e., controlling a relative variance, as described in Remark 3.1), although this cost here is deterministic. This can be shown by assuming that $\inf_{k,\gamma} \alpha_k(y_{1:k}, \epsilon, \gamma) \geq C > 0$ and yielding the upper bound of $(1 + 1/(CN))^n$ (on the term on the L.H.S. of (9)), and then one should set $N = \mathcal{O}(n)$ for the upper bound to go to a limit as n grows; this is done n times. As mentioned previously, one expects the approach with random number of trials to work better with regards to the mixing time, especially when the values of $\alpha_k(y_{1:k}, \epsilon, \gamma)$ are not large. We attribute this to Algorithm 4, providing a more targetted way to use the simulated auxiliary variables. This will be illustrated numerically in Section 4.

3.2.3. Relating the Variance of the Estimator or $p_\gamma^\epsilon(y_{1:n})$ with the Efficiency of ABC-MCMC. A comparison of our results with the interesting work in Doucet et al. [2012] seems relevant. There the authors deal with a more general context that includes the proposals in both Algorithms 3 and 4 as special cases. Doucet et al. [2012] show that we should choose N as a particular asymptotic (in N) variance; the main point is that the (asymptotic) variance of the estimate of $p_\gamma^\epsilon(y_{1:n})$ should be the same for each γ. We conjecture that in the context of Algorithm 4, a finite sample version of the work of Doucet et al. [2012] would be to choose N such that the actual variance (variance in the sense of Proposition 3.1) of the estimate of $p_\gamma^\epsilon(y_{1:n})$ is constant with respect to γ. In this scenario, on inspection of the proof of Proposition 3.1, for a given γ, one should set N to be the solution of

$$\left(\prod_{k=1}^{n}\alpha_k(y_{1:k},\epsilon,\gamma)^2\right)\left(\frac{1}{(N-1)^n(N-2)^n}-\frac{1}{(N-1)^{2n}}\right)=C$$

for some desired (upper bound on the) variance C (whose optimal value would need to be obtained). This would lead to N changing at each iteration, in addition, but does not change the simulation mechanism. Unfortunately, one cannot do this in practice, as the $\alpha_k(y_{1:k},\epsilon,\gamma)$ are unknown.

3.2.4. On the Ergodicity of the Sampler. We now give a comment regarding the ergodicity of the MCMC kernel associated with Algorithm 4. If there exists a constant $C<\infty$ independent of γ such that

$$\frac{1}{\prod_{k=1}^{n}\alpha_k(y_{1:k},\epsilon,\gamma)}\leq C \tag{10}$$

and the marginal MCMC kernel in Algorithm 1 is geometrically ergodic, then by Andrieu and Vihola [2012, Propositions 7 and 9], the MCMC kernel of Algorithm 4 is also geometrically ergodic. This result follows because the weight w_x in Andrieu and Vihola [2012] is

$$\frac{\prod_{k=1}^{n}\frac{N-1}{\mu(B_\epsilon(y_k))(m_k-1)}}{p_\gamma^\epsilon(y_{1:n})},$$

which is upper bounded uniformly in γ under (10), which allows one to apply Propositions 7 and 9 of Andrieu and Vihola [2012] [Andrieu and Vihola 2012, Propositions 7, 9] (along with the geometric ergodicity of the marginal MCMC kernel).

3.2.5. Some Comments on Alternative Simulation Schemes. We have introduced a new MCMC kernel for our ABC approximation. However, there are many contributions to the literature on simulation-based methods for ABC approximations, particularly those based on sequential Monte Carlo methods; see, for instance, Beaumont et al. [2009] and Del Moral et al. [2012], which might arguably be considered the gold-standard approaches. In terms of the approach of Del Moral et al. [2012], this generally performs well when the underlying MCMC kernels have a fast rate of convergence, and as such, this is the idea of the method introduced here; the MCMC kernel associated to the proposal in Algorithm 4 could be used within the SMC approach of Del Moral et al. [2012] (although one would need to modify the procedure, as it cannot be used as presented in Del Moral et al. [2012]), potentially enhancing it. The approach in Beaumont et al. [2009] is possibly unsuitable for this particular model structure (at least as described in Beaumont et al. [2009, Section 3]), as the acceptance probability per sample is likely to fall at an exponential rate in n.

4. EXAMPLES

Two examples are now presented. It is remarked that the assumptions in Section 2.2 do not hold in these examples. However, it is found that some of the results derived there seem to hold; it is conjectured that our results in Section 2.2 can be proved under weaker hypotheses than we have adopted.

4.1. Scalar Normal Means Model

4.1.1. Model. For this example, let each of Y_k, X_k, θ be a scalar real random variable and consider the model

$$Y_{k+1}=\theta X_k+\kappa_k,\quad X_{k+1}=X_k$$

with $X_0 = 1$ and $\kappa_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, where we denote $\mathcal{N}(0, \sigma^2)$ the zero mean normal distribution with variance σ^2. The prior on θ is $\mathcal{N}(0, \phi)$. This model is usually referred to as the standard normal means model in one dimension, and the posterior is given by

$$\theta | y_{1:n} \sim \mathcal{N}\left(\frac{\sigma_n^2}{\sigma^2}\sum_{k=1}^{n} y_k, \sigma_n^2\right),$$

where $\sigma_n^2 = (\frac{1}{\phi} + \frac{n}{\sigma^2})^{-1}$. Note that if $Y_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta^*, \sigma^2)$, then the posterior on θ is consistent and concentrates around θ^* as $n \to \infty$.

The ABC approximation after marginalizing out the auxiliary variables has a likelihood given by

$$p_\theta^\epsilon(y_{1:n}) = \frac{1}{\epsilon^n}\prod_{k=1}^{n}\left[F\left(\frac{y_k + \epsilon - \theta}{\sigma}\right) - F\left(\frac{y_k - \epsilon - \theta}{\sigma}\right)\right],$$

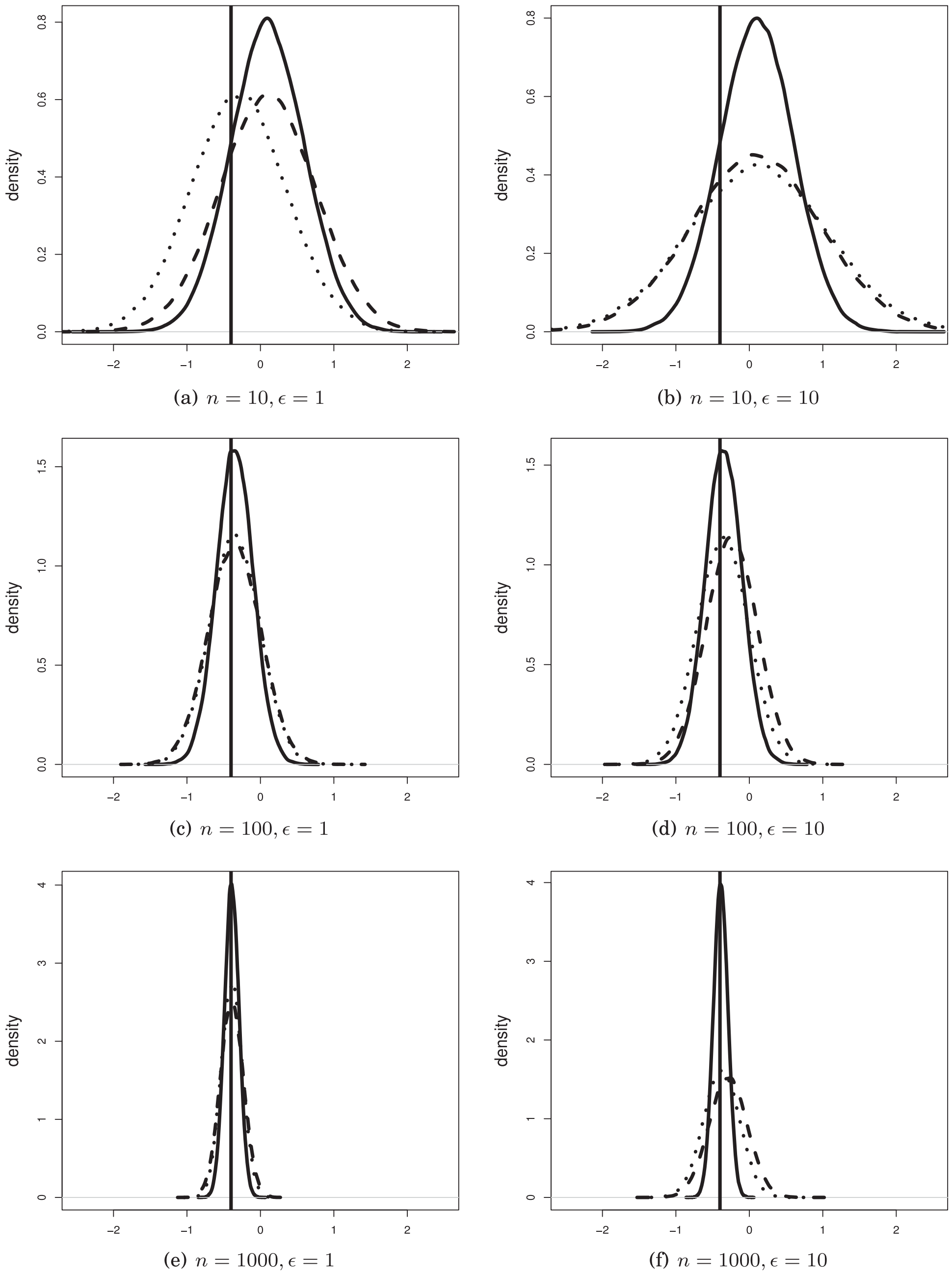
where F is the standard normal cumulative density function. Thus, this is a scenario where we can perform the marginal MCMC.

4.1.2. Simulation Results. Three datasets are generated from the model with $n \in \{10, 100, 1,000\}$ and $\sigma^2 = 1$. In addition, for $\epsilon = 1$, we perturb the datasets in order to use them for noisy ABC. For the sake of comparison, we also generate a noisy ABC dataset for $\epsilon = 10$. We will also use a prior with $\phi = 1$.

We run the proposal in Algorithm 4 (we will frequently use the expression *Algorithm* to mean an MCMC kernel with the given proposal mechanism of the Algorithm), Algorithm 3, and a marginal MCMC algorithm that just samples on the parameter space $\mathbb{R}$ (i.e., the invariant density is proportional to $p_\theta^\epsilon(y_{1:n})\nu(\theta)$). Each algorithm is run with a normal random walk proposal on the parameter space with the same scaling. The scaling chosen yields an acceptance rate of around 0.25 for each run of the marginal MCMC algorithm. Algorithm 4 is run with $N = n$ and Algorithm 3 with a slightly higher value of N so that the computational times are about the same (thus, the running time of Algorithm 4 is not a problem in this example). The algorithms are run for 10,000 iterations, and the results can be found in Figures 1, 2, and 3.

In Figure 1, the density plots for the posterior samples on θ, from the marginal MCMC, can be seen for $\epsilon \in \{1, 10\}$ and each value of n. When $\epsilon = 1$, we can observe that both ABC and noisy ABC get closer to the true posterior as n grows. For noisy ABC, this is the behavior that is predicted in Section 2.2. In particular, Proposition 2.1 suggests that the ABC can have some asymptotic bias, whereas this should not be the case for noisy ABC in Proposition 2.2; this is seen for finite n, especially in Figure 1(a). For the ABC approximation, following the proof of Theorem 1 in Jasra et al. [2012] (if one can adopt the same assumptions for h^θ for g [of that paper], the proof [and hence result] in Jasra et al. [2012] can be used, as it does not make any assumption on the hidden Markov chain), one can see that the bias falls with ϵ; hence, in this scenario, there is not a substantial bias for the standard ABC approximation. When we make ϵ larger, a more pronounced difference between ABC and noisy ABC can be seen, and it appears as n grows that the noisy ABC approximation is slightly more accurate (relative to ABC).

We now consider the similarity of Algorithms 3 and 4 to the marginal algorithm (i.e., the kernel that both procedures attempt to approximate); the results are in Figures 2 and 3, and $\epsilon = 1$ throughout. With regards to both the density plots (Figure 2) and autocorrelations (Figure 3—only for noisy ABC; we found similar results when considering plain ABC), we can see that both MCMC kernels appear to be quite similar to



(a) $n = 10, \epsilon = 1$

(b) $n = 10, \epsilon = 10$

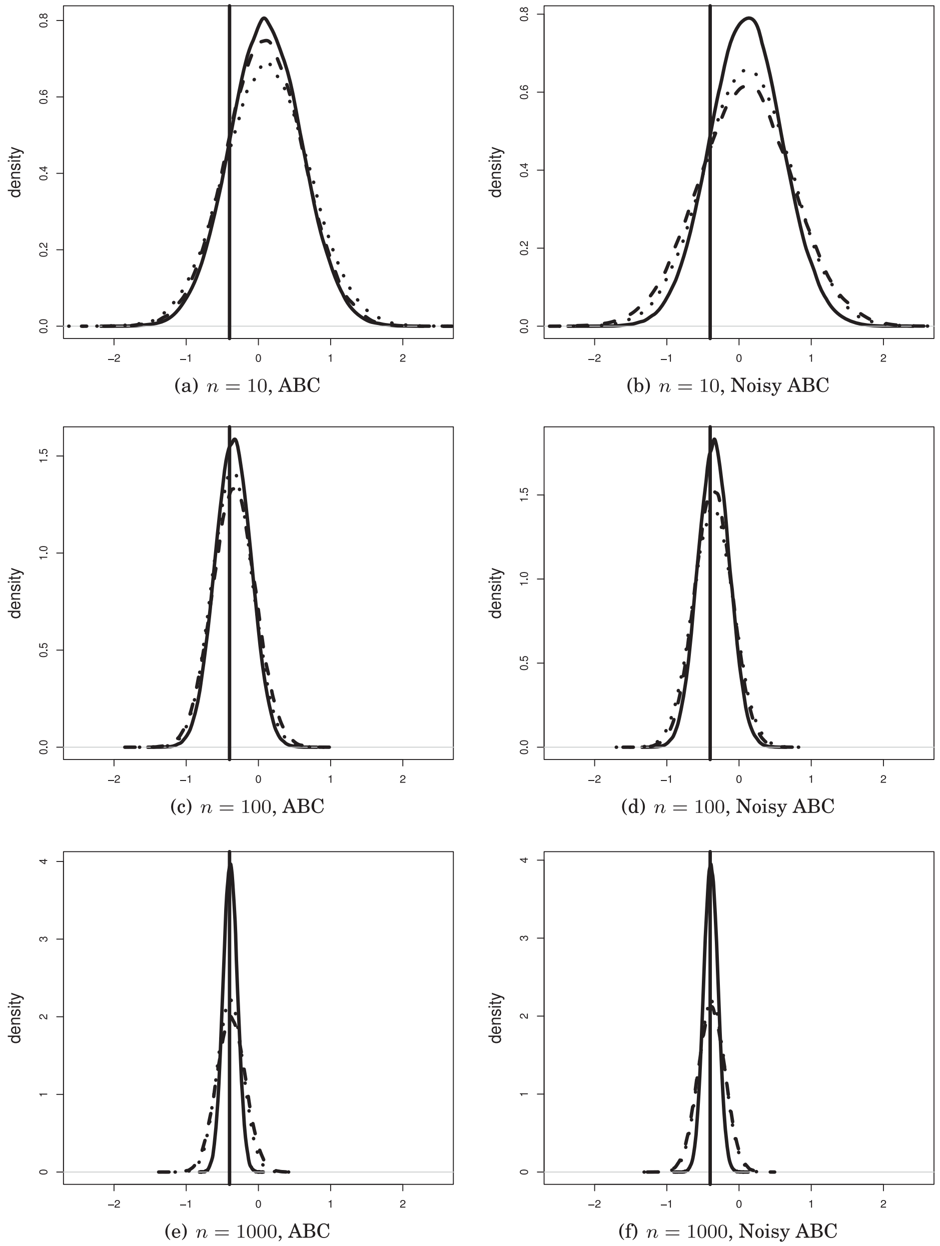
(c) $n = 100, \epsilon = 1$

(d) $n = 100, \epsilon = 10$

(e) $n = 1000, \epsilon = 1$

(f) $n = 1000, \epsilon = 10$

Fig. 1. Marginal MCMC density plots for normal means example. In each plot, the true posterior (full), noisy ABC (dot), ABC (dash) densities of θ are plotted for different values of n (10, first row; 100, second row; 1,000, third row) and ϵ (1, first column; 100, second column). The vertical line is the value of θ that generated the data.



(a) $n = 10$, ABC

(b) $n = 10$, Noisy ABC

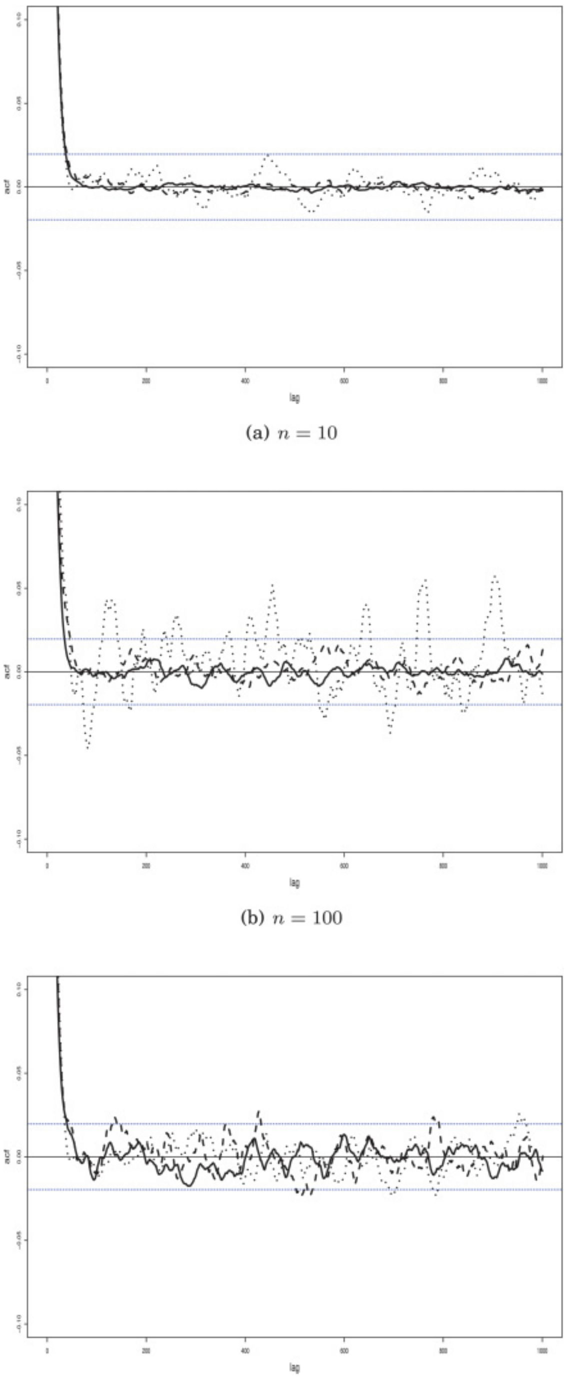
(c) $n = 100$, ABC

(d) $n = 100$, Noisy ABC

(e) $n = 1000$, ABC

(f) $n = 1000$, Noisy ABC

Fig. 2. MCMC density plots for normal means example. In each plot, the true posterior (black), ABC (first column) or noisy ABC (second column) densities of θ are plotted for different values of n (10, first row; 100, second row; 1,000, third row). In addition, the plots are for Algorithm 3 (dot) and Algorithm 4 (dash). The vertical line is the value of θ that generated the data. Throughout, $\epsilon = 1$.



(a) $n = 10$

(b) $n = 100$

(c) $n = 1000$

Fig. 3. Autocorrelation plots for normal means example. In each plot, the autocorrelation for every fifth iteration is plotted, all for noisy ABC, for marginal MCMC (full), Algorithm 3 (dot), and Algorithm 4 (dash). Three different values of n are presented, and $\epsilon = 1$ throughout. The dotted horizontal lines are a default confidence interval generated by the R package.

the marginal MCMC. It is also noted that the acceptance rates of these latter kernels are also not far from that of the marginal algorithm (results not shown). These results are unsurprising given the simplicity of the density that we target, but still reassuring; a more comprehensive comparison is given in the next example. Encouragingly, Algorithms 3 and 4 do not seem to noticeably worsen as n grows; this shows that, at least for this example, the recommendation of $N = \mathcal{O}(n)$ is quite useful. We remark that whilst these results are for a single batch of data, the results with regards to the performance of the MCMC are consistent with other datasets.

4.2. Real Data Example

4.2.1. Model. Set, for $(Y_k, X_k) \in \mathbb{R} \times \mathbb{R}^+$

$$\begin{aligned} Y_{k+1} &= \kappa_k \quad k \in \mathbb{N}_0 \\ X_{k+1} &= \beta_0 + \beta_1 X_k + \beta_2 Y_{k+1}^2 \quad k \in \mathbb{N}_0, \end{aligned}$$

where $\kappa_k | x_k \overset{\text{ind}}{\sim} \mathcal{S}(0, x_k, \varphi_1, \varphi_2)$ (i.e., a stable distribution, with location 0, scale X_k and asymmetry and skewness parameters φ_1, φ_2; see Chambers et al. [1976] for more information). We set

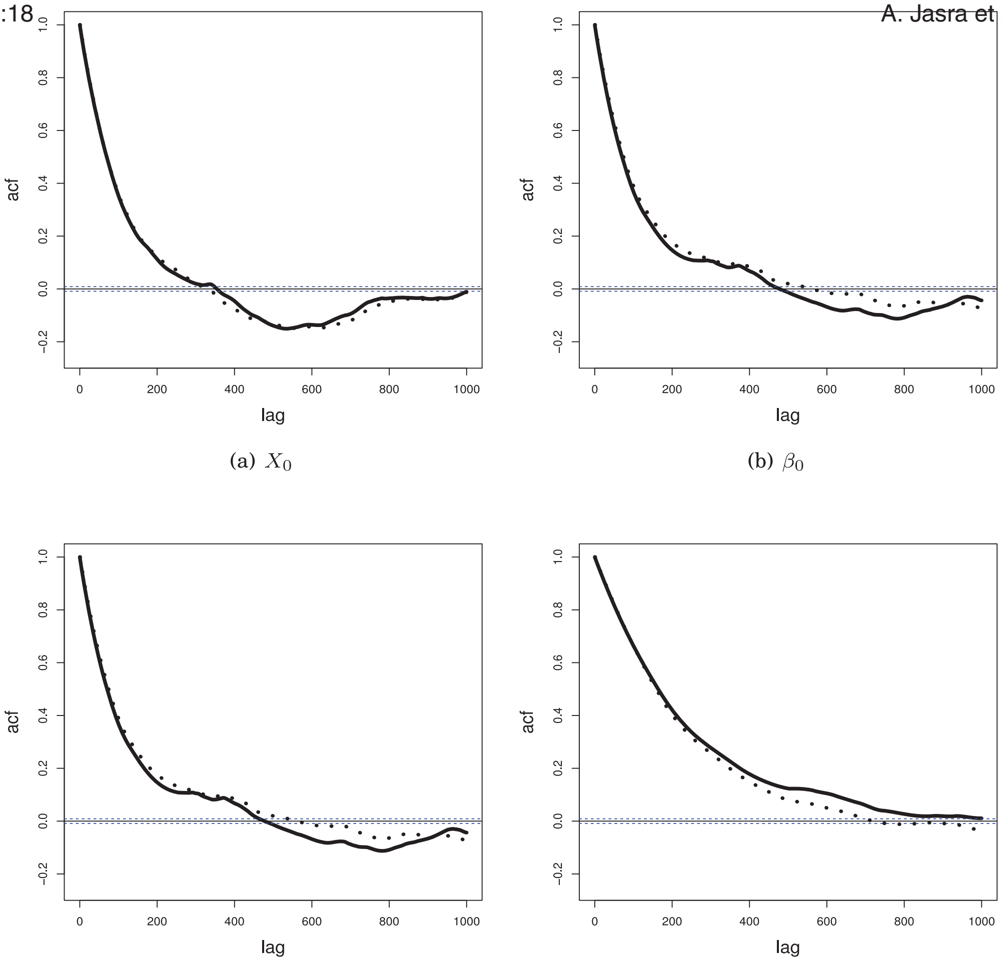
$$X_0 \sim \mathcal{G}a(a, b), \quad \beta_0, \beta_1, \beta_2 \sim \mathcal{G}a(c, d),$$

where $\mathcal{G}a(a, b)$ is a Gamma distribution with mean a/b and $\theta = (\beta_{0:2}) \in (\mathbb{R}^+)^3$. This is a GARCH(1,1) model with an *intractable likelihood*—that is, one cannot perform exact parameter inference and has to resort to approximations.

4.2.2. Simulation Results. We consider daily log-returns data from the S&P 500 index from 03/1/11 to 14/02/13, which constitutes 533 data points. In the priors, we set $a = c = 2$ and $b = d = 1/8$, which are not overly informative. In addition, $\varphi_1 = 1.5$ and $\varphi_2 = 0$. The values of $\varphi_1 = 1.5$ means that the observation density has very heavy tails (characteristic of financial data) and $\varphi_2 = 0$ that the distribution of the log-returns is symmetric about zero; in general, this may not occur for all log-returns data but is a reasonable assumption in an initial data analysis. We consider $\epsilon \in \{0.01, 0.5\}$ and only a noisy ABC approximation of the model. The values of ϵ are chosen so as to illustrate two scenarios: one where the proposal in Algorithm 3 seems to mix very well, with little efforts—that is, constructing q—and one where it does not seem to mix well, even with considerable effort. Algorithms 3 and 4 are to be compared. The MCMC proposals on the parameters are normal random walks on the log-scale, and for both algorithms, we set $N = 250$. It should be noted that our results are fairly robust to changes in $N \in [100, 500]$, which are the values with which we tested the algorithm.

In Figure 4, we present the autocorrelation plot of 50,000 iterations of both MCMC kernels when $\epsilon = 0.5$. Algorithm 3 took about 0.30 seconds per iteration, and Algorithm 4 took about 1.12 seconds per iteration During preliminary runs for the case $\epsilon = 0.5$, we modified the proposal variances to yield an acceptance rate of around 0.3. The plot shows that both algorithms appear to mix across the state-space in a very reasonable way. The MCMC procedure associated with Algorithm 4 takes much longer and in this situation does not appear to be required. This run is one of many that we performed, and we observed this behaviour in many of our runs.

In Figure 5, we can observe the autocorrelation plots from a particular (typical) run when $\epsilon = 0.01$. In this case, both algorithms are run for 200,000 iterations. Algorithm 3 took about 0.28 seconds per iteration, and Algorithm 4 took about 2.06 seconds per iteration; this issue is discussed next. During preliminary runs for the case $\epsilon = 0.01$, we attempted to modify the proposal variances to yield an acceptance rate of around 0.3; this was not achieved for either algorithm as we now report. In this scenario, considerable effort was expended for Algorithm 3 to yield an acceptance rate around 0.3,



(a) X_0

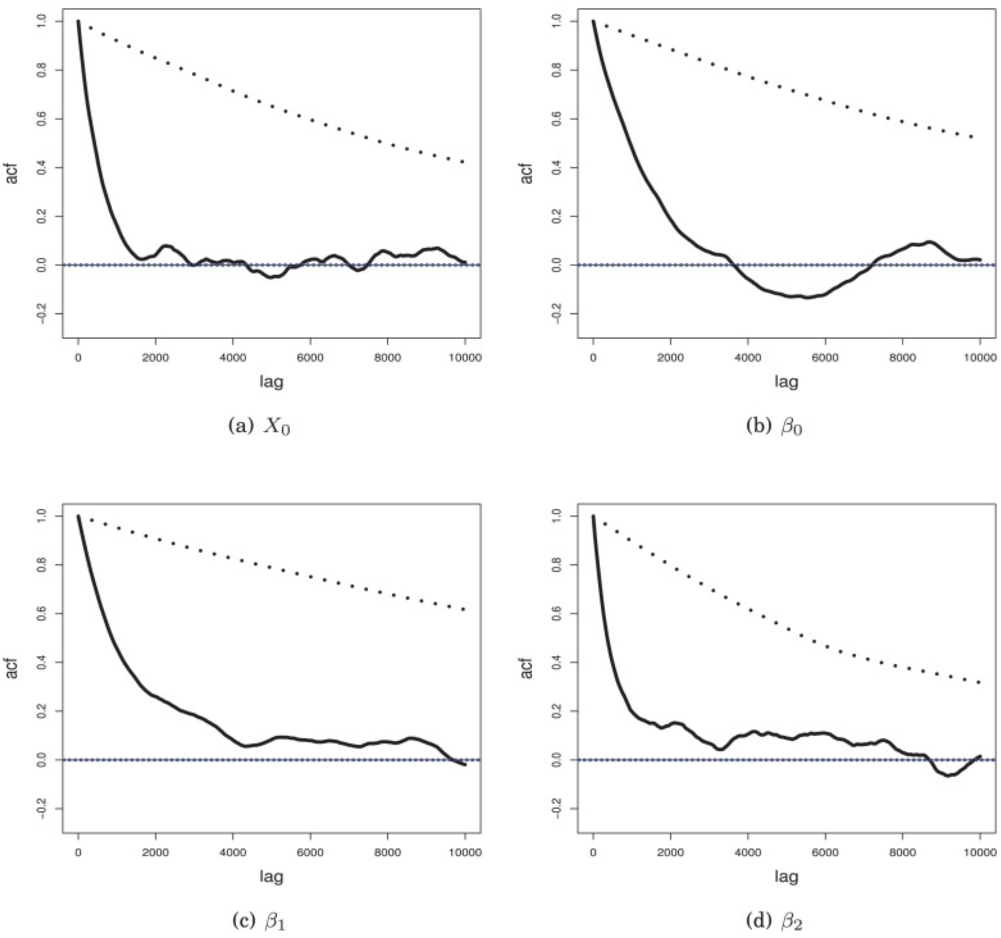
(b) β_0

(c) β_1

(d) β_2

Fig. 4. Autocorrelation plots for the sampled parameters of the example in Section 4.2. We run Algorithm 3 (dot) and 4 (full) for 50,000 iterations (both with $N = 250$) on the S & P 500 data, associated to noisy ABC and $\epsilon = 0.5$. The dotted horizontal lines are a default confidence interval generated by the R package.

but despite this, we were unable to make the algorithm traverse the state-space. In contrast, with less effort, Algorithm 4 appears to perform quite well and move around the parameter space (the acceptance rate was around 0.15 vs. 0.01 for Algorithm 3). Whilst the computational time for Algorithm 4 is considerably more than Algorithm 3, in the same amount of computation time, it still moves more around the state-space as suggested by Figure 5; algorithm runs of the same length are provided for presentational purposes. To support this point, we computed the ratio of the effective sample size from Algorithm 4 to that of Algorithm 3 when standardizing for computational time; this value is 2.04, indicating (very roughly) that Algorithm 4 is twice as efficient as Algorithm 3 for this example. We remark that whilst we do not claim that it is impossible to make Algorithm 3 mix well in this example, we were unable to do so, and alternatively for Algorithm 4, we expended considerably less effort for very reasonable performance. This example is typical of many runs of the algorithm and examples that we have investigated and is consistent with the discussion in Section 3.2.2, where we stated that Algorithm 4 is likely to outperform Algorithm 3 when the $\alpha_k(y_{1:k}, \epsilon, \gamma)$ are not large, which is exactly the scenario in this example.



(a) X_0

(b) β_0

(c) β_1

(d) β_2

Fig. 5. Autocorrelation plots for the sampled parameters of the example in Section 4.2. We run Algorithm 3 (dot) and 4 (full) for 200,000 iterations (both with $N = 250$) on the S&P 500 data, associated to noisy ABC and $\epsilon = 0.01$. The dotted horizontal lines are a default confidence interval generated by the R package.

We now turn to the cost of simulating Algorithm 4. For the case $\epsilon = 0.5$, we simulated the data an average of 148,000 times (per iteration), and for $\epsilon = 0.01$, this figure was 330,000. In this example, significant effort is expended in simulating the $m_{1:n}$. This shows, at least in this example, that one can run the algorithm without it failing to sample the $m_{1:n}$. The results here suggest that one should prefer Algorithm 4 only in challenging scenarios, as it can be very expensive in practice.

Finally, we remark that the MLE for a Gaussian GARCH model is $\beta_{0:2} = (4.1 \times 10^{-6}, 0.16, 0.82)$. This differs from the posterior means, so we consider the fit of the models. On inspection of the residuals, the ratio of Y_{k+1}/X_{k+1} under the estimated model, which are not presented, we did not find that either model fit the data well. This is in the sense that the residuals did not fit the hypothesized distribution of either model; it seems that perhaps this model is not appropriate for these data under either noise distribution.

5. CONCLUSIONS

In this article, we have considered approximate Bayesian inference from observation-driven time series models. We looked at some consistency properties of the corresponding MAP estimators and also proposed an efficient ABC-MCMC algorithm to sample from these approximate posteriors. The performance of the latter was illustrated using numerical examples.

There are several interesting extensions to this work. Firstly, the asymptotic analysis of the ABC posterior in Section 2.2 can be further extended. For example, one may consider Bayesian consistency or Bernstein Von-Mises theorems, which could provide further justification of the approximation that was introduced here. Alternatively, one could look at the the asymptotic bias of the ABC posterior with respect to ϵ or the asymptotic loss in efficiency of the noisy ABC posterior with respect to ϵ similar to the work in Dean et al. [2011] for HMMs. Secondly, the geometric ergodicity of the presented MCMC sampler can be further investigated in the spirit of Andrieu and Vihola [2012] and Lee and Latuszynski [2012]. Thirdly, an investigation to extend the ideas here for sequential Monte Carlo methods should be beneficial. This has been initiated in Jasra et al. [2013] in the context of particle filtering for Feynman-Kac models with indictors potentials (which includes the ABC approximation of HMMs); several results, in the context of Section 3.2, are derived.

A. PROOFS FOR SECTION 2

Before giving our proofs, we will remind the reader of the assumptions (B1–B3) used in Douc et al. [2012]. These are written in the context of a general observation-driven time series model. Define the process $\{Y_k, X_k\}_{k\in\mathbb{N}_0}$ (with y_0 some arbitrary point on Y) on a probability space $(\Omega, \mathscr{F}, \bar{\mathbb{P}}_\theta)$, where for every $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$, $\bar{\mathbb{P}}_\theta$ is a probability measure. Denote by $\mathscr{F}_k = \sigma(\{Y_n, X_n\}_{0\leq n\leq k})$. For $k \in \mathbb{N}_0$, $X_0 = x$,

$$\begin{aligned} \bar{\mathbb{P}}(Y_{k+1} \in A|\mathscr{F}_k) &= \int_A \bar{H}(x_k, dy) \quad A \times \mathsf{X} \in \mathscr{F} \\ X_{k+1} &= \bar{\Phi}^\theta(X_k, Y_{k+1}), \end{aligned}$$

where $\bar{H} : \mathsf{X} \times \sigma(\mathsf{Y}) \to [0, 1]$, $\bar{\Phi} : \Theta \times \mathsf{X} \times \mathsf{Y} \to \mathsf{X}$ and for every $\theta \in \Theta$. Throughout, we assume that for any $x \in \mathsf{X}$ $\bar{H}(x, \cdot)$ admits a density with respect to some $\sigma-$finite measure μ, which we denote as $\bar{h}(x, y)$. As in Section 2, we extend the definitions of the time index of the process to $\mathbb{Z}$. We denote $\max(v, 0) = (v)_+$ for some $v \in \mathbb{R}$.

(B1) $\{Y_k\}_{k\in\mathbb{Z}}$ a strict-sense stationary and ergodic stochastic process. Write the associated probability measure $\bar{\mathbb{P}}_\star$.
(B2) For all $(x, y) \in \mathsf{X} \times \mathsf{Y}$, the functions $\theta \mapsto \bar{\Phi}^\theta(x, y)$ and $v \mapsto \bar{h}(x, y)$ are continuous.
(B3) There exists a family of $\bar{\mathbb{P}}_\star$-a.s. finite random variables

$$\left\{\bar{\Phi}^\theta_\infty(Y_{-\infty:k}), (\theta, k) \in \Theta \times \mathbb{Z}\right\}$$

such that for each $x \in \mathsf{X}$
(i) $\lim_{m\to\infty} \sup_{\theta\in\Theta} d(\bar{\Phi}^\theta_\infty(Y_{-m:0}, x), \bar{\Phi}^\theta_\infty(Y_{-\infty:0})) = 0$, $\bar{\mathbb{P}}_\star$-a.s.
(ii) $\bar{\mathbb{P}}_\star$-a.s.

$$\lim_{k\to\infty} \sup_{\theta\in\Theta} |\log(\bar{h}\big(\bar{\Phi}^\theta_{k-1}(Y_{1:k-1}, x), Y_k)\big) - \log\big(\bar{h}\big(\bar{\Phi}^\theta_\infty(Y_{-\infty:k-1}), Y_k\big)\big)| = 0.$$

(iii)

$$\bar{\mathbb{E}}_\star\left[\sup_{\theta\in\Theta}\big(\log\big(\bar{h}\big(\bar{\Phi}^\theta_\infty(Y_{-\infty:k-1}), Y_k\big)\big)\big)_+\right] < +\infty,$$

with $\bar{\mathbb{E}}_\star$ denoting expectations with respect to $\bar{\mathbb{P}}_\star$.

The ideas of our proofs are essentially just to verify these assumptions for *our perturbed ABC model*, which uses the system (4), except that the observations (either the actual ones or perturbed ones for noisy ABC) are fitted with the density defined in (3).

Proof. [Proof of Proposition 2.1] The proof of $\lim_{n\to\infty} \mathsf{d}(\theta_{n,x,\epsilon}, \Theta_\epsilon^*) = 0 \quad \mathbb{P}_{\theta *} - a.s.$ follows from Douc et al. [2012, Theorem 19] if we can establish conditions (B1–B3) for our perturbed ABC model. Clearly, (B1) and part of (B2) hold. For part of (B2), we need to show that for any $y \in \mathsf{Y}$ that $x \mapsto h^\epsilon(x, y)$ is continuous. Consider

$$|h^\epsilon(x, y) - h^\epsilon(x', y)| = \frac{1}{\mu(B_\epsilon(y))}\left|\int_{B_\epsilon(y)} \big(h(x, y) - h(x', y)\big)\,\mu(dy)\right|.$$

Let $\varepsilon > 0$, then, by (A3), there exists a $\delta > 0$ such that for $d(x, x') < \delta$

$$\sup_{y\in\mathsf{Y}} |h(x, y) - h(x', y)| < \varepsilon$$

and hence for (x, x'), as shown previously,

$$|h^\epsilon(x, y) - h^\epsilon(x', y)| < \varepsilon,$$

which establishes (B2) of Douc et al. [2012].

(B3-i) holds via [Douc et al. 2012, Lemma 20] through (A4): by the proof of Douc et al. [2012, Lemma 20], $\lim_{m\to\infty} \Phi^\theta_{m+1}(Y_{-m:k}, x)$ exists (for any fixed $k \geq 0$, $x \in \mathsf{X}$) and is independent of x (call the limit $\Phi^\theta_\infty(Y_{-\infty:k})$). Now, for (B3-ii) of Douc et al. [2012], fix $m > 1, k > 1$, $x, x' \in \mathsf{X}$ we note that as $\underline{h} \leq h^\epsilon(x, y) \leq \overline{h} < \infty$ (see (A3)), $h \mapsto \log(h)$ is Lipschitz and

$$\begin{aligned}&\big|\log\big(h^\epsilon\big(\Phi^\theta_{k-1}(Y_{1:k-1}, x), Y_k\big)\big) - \log\big(h^\epsilon\big(\Phi^\theta_{m+k}(Y_{-m:k-1}, x'), Y_k\big)\big)\big|\\ &\quad\leq C\big|h^\epsilon\big(\Phi^\theta_{k-1}(Y_{1:k-1}), x\big), Y_k\big) - h^\epsilon\big(\Phi^\theta_{m+k}(Y_{-m:k-1}, x'), Y_k\big)\big|\end{aligned}$$

for some $C < \infty$ that does not depend upon $Y_{-m:k-1}, Y_k, x, x', \epsilon$. Now

$$\begin{aligned}&\big|h^\epsilon\big(\Phi^\theta_{k-1}(Y_{1:k-1}), x), Y_k\big) - h^\epsilon\big(\Phi^\theta_{m+k}(Y_{-m:k-1}, x'\big), Y_k\big)\big|\\ &\quad= (\mu(B_\epsilon(Y_k)))^{-1}\left|\int_{B_\epsilon(Y_k)} \big[h\big(\Phi^\theta_{k-1}(Y_{1:k-1}, x), y\big) - h\big(\Phi^\theta_{m+k}(Y_{-m:k-1}, x'), y\big)\big]\mu(dy)\right|\\ &\quad\leq (\mu(B_\epsilon(Y_k)))^{-1} \times \mu(B_\epsilon(Y_k)) \sup_{y\in\mathsf{Y}}\big|h\big(\Phi^\theta_{k-1}(Y_{1:k-1}, x), y\big) - h\big(\Phi^\theta_{m+k}(Y_{-m:k-1}, x'), y\big)\big]\big|.\end{aligned}$$

Thus, by (A3) and the preceding calculations, we have that

$$\begin{aligned}&\big|\log\big(h^\epsilon\big(\Phi^\theta_{k-1}(Y_{1:k-1}, x), Y_k\big)\big) - \log\big(h^\epsilon\big(\Phi^\theta_{m+k}(Y_{-m:k-1}, x'), Y_k\big)\big)\big|\\ &\quad\leq Cd\big(\Phi^\theta_{k-1}(Y_{1:k-1}, x), \Phi^\theta_{m+k}(Y_{-m:k-1}, x')\big)\end{aligned}$$

for some $C < \infty$ that does not depend upon $Y_{-m:k-1}, Y_k, x, \epsilon, \theta$. Then, by (A4),

$$\begin{aligned}&\big|\log\big(h^\epsilon\big(\Phi^\theta_{k-1}(Y_{1:k-1}, x), Y_k\big)\big) - \log\big(h^\epsilon\big(\Phi^\theta_{m+k}(Y_{-m:k-1}, x'), Y_k\big)\big)\big|\\ &\quad\leq Cd\big(x, \Phi^\theta_{m+1}(Y_{-m:0}, x')\big)\prod_{j=1}^{k-1}\varrho(Y_k)\\ &\quad\leq Cd\big(x, \Phi^\theta_{m+1}(Y_{-m:0}, x')\big)\overline{\varrho}^{k-1}.\end{aligned}$$

Taking suprema over θ and as X is compact, we have

$$\sup_{\theta\in\Theta}\big|\log\big(h^\epsilon(\Phi^\theta_{k-1}(Y_{1:k-1}, x), Y_k)\big) - \log\big(h^\epsilon\big(\Phi^\theta_{m+k}(Y_{-m:k-1}, x'), Y_k\big)\big)\big| \leq C'\overline{\varrho}^{k-1},$$

where $C' < \infty$ and does not depend $Y_{-m:k-1}, Y_k, x, \epsilon, \theta, m$. Taking limits as $m \to \infty$ in the preceding inequality, we have $\mathbb{P}_{\theta^*}$−a.s.

$$\sup_{\theta\in\Theta}\big|\log\big(h^\epsilon\big(\Phi^\theta_{k-1}(Y_{1:k-1}, x), Y_k\big)\big) - \log\big(h^\epsilon\big(\Phi^\theta_\infty(Y_{-\infty:k-1}), Y_k\big)\big)\big| \leq C'\overline{\varrho}^{k-1}.$$

Now we can conclude that $\mathbb{P}_{\theta^*}$−a.s.

$$\lim_{k\to\infty}\sup_{\theta\in\Theta}\big|\log\big(h^{\epsilon}(\Phi^{\theta}_{k-1}(Y_{1:k-1},x),Y_k)\big)-\log\big(h^{\epsilon}\big(\Phi^{\theta}_{\infty}(Y_{-\infty:k-1}),Y_k\big)\big)\big|=0,$$

which proves (B3-ii) of Douc et al. [2012]. Note, finally that (B3-iii) trivially follows by $h^{\epsilon}(x,y)\leq\overline{h}<\infty$. Hence, we have proved that

$$\lim_{n\to\infty}\mathsf{d}(\theta_{n,x,\epsilon},\Theta^*_{\epsilon})=0\quad\mathbb{P}_{\theta*}-a.s..$$ □

Proof. [Proof of Proposition 2.2] This result follows from Douc et al. [2012, Proposition 21]. One can establish assumptions (B1–B3) of Douc et al. [2012] using the proof of Proposition 2.1. Thus, we need only prove that

$$\text{if } H^{\epsilon}(x,\cdot)=H^{\epsilon}(x',\cdot),\quad\text{then } x=x'.$$

Now, for any $A\times\mathsf{X}\in\mathscr{F}$,

$$H^{\epsilon}(x,A)=\int_A\left[\frac{1}{\mu(B_{\epsilon}(y))}\int_{B_{\epsilon}(y)}H(x,du)\right]\mu(dy).$$

By (A5), $\int_{B_{\epsilon}(y)}H(x,du)=\int_{B_{\epsilon}(y)}H(x',du)$ means that $x=x'$, so

$$H^{\epsilon}(x,A)=H^{\epsilon}(x',A)$$

implies that $x=x'$, which completes the proof. □

B. REMARKS AND PROOFS FOR SECTION 3

B.1. Remarks

In order to deduce the result (6) (as well as a second inverse moment type identity in the proof of Proposition 3.1) from the work of Neuts and Zacks [1967] and Zacks [1980], some additional calculations are required. The notations in this section of the Appendix should be taken as independent of the rest of the article and are used to match those in Zacks [1980]. Using the results in Neuts and Zacks [1967], Zacks [1980] quotes the following. Zacks [1980, Eq. 1] gives a particular form for a negative binomial random variable X with probability mass function

$$\mathbb{P}(X=x)=\frac{\Gamma(\nu+x)}{\Gamma(x+1)\Gamma(\nu)}(1-\psi)^{\nu}\psi^{x}\quad x\in\{0,1,\ldots\},$$

with $\psi\in(0,1)$ and $\nu\in(0,\infty)$. Then, letting $\mathbb{E}$ denote expectations with respect to this given probability mass function, Zacks [1980, Eqs. 2 and 3] read:

$$\mathbb{E}\left[\frac{1}{\nu+X-1}\right]=\frac{1-\psi}{\nu-1}\quad\nu\geq2 \tag{11}$$

$$\mathbb{E}\left[\frac{1}{(\nu+X-1)(\nu+X-2)}\right]=\frac{(1-\psi)^2}{(\nu-1)(\nu-2)}\quad\nu\geq3. \tag{12}$$

To use these results in the context of the work in this article, we suppose that $\nu\in\mathbb{N}$ and make the change of variable $M=X+\nu$, which yields the probability mass function

$$\mathbb{P}(M=m)=\binom{m-1}{\nu-1}(1-\psi)^{\nu}\psi^{m-\nu}\quad m\in\{\nu,\nu+1,\ldots\},$$

which is a conventional negative binomial probability mass function associated to ν successes, with success probability $1-\psi$. Then, it follows from (11) and (12) that (using

$\mathbb{E}$ to denote expectations with respect to $\mathbb{P}(M = m)$)

$$\mathbb{E}\left[\frac{1}{M-1}\right] = \frac{1-\psi}{\nu-1} \quad \nu \geq 2$$

$$\mathbb{E}\left[\frac{1}{(M-1)(M-2)}\right] = \frac{(1-\psi)^2}{(\nu-1)(\nu-2)} \quad \nu \geq 3.$$

B.2. Proofs

Proof. [Proof of Proposition 3.1] We have

$$\mathbb{E}_{\gamma,N}\left[\left(\frac{\prod_{k=1}^n \frac{1}{M_k-1}}{\prod_{k=1}^n \frac{\alpha_k(y_{1:k},\epsilon,\gamma)}{N-1}} - 1\right)^2\right] = \frac{1}{(\prod_{k=1}^n \frac{\alpha_k(y_{1:k},\epsilon,\gamma)}{N-1})^2} \times \left(\prod_{k=1}^n \mathbb{E}_{\gamma,N}\left[\frac{1}{(M_k-1)^2}\right] - \left(\prod_{k=1}^n \frac{\alpha_k(y_{1:k},\epsilon,\gamma)}{N-1}\right)^2\right).$$

Now, by Neuts and Zacks [1967] and Zacks [1980], ($N \geq 3$) for any $k \geq 1$ (see also Appendix B.1)

$$\mathbb{E}_{\gamma,N}\left[\frac{1}{(M_k-1)(M_k-2)}\right] = \frac{\alpha_k(y_{1:k},\epsilon,\gamma)^2}{(N-1)(N-2)}$$

and thus clearly

$$\mathbb{E}_{\gamma,N}\left[\frac{1}{(M_k-1)^2}\right] \leq \frac{\alpha_k(y_{1:k},\epsilon,\gamma)^2}{(N-1)(N-2)}.$$

Hence,

$$\mathbb{E}_{\gamma,N}\left[\left(\frac{\prod_{k=1}^n \frac{1}{M_k-1}}{\prod_{k=1}^n \frac{\alpha_k(y_{1:k},\epsilon,\gamma)}{N-1}} - 1\right)^2\right] \leq (N-1)^{2n}\left(\frac{1}{(N-1)^n(N-2)^n} - \frac{1}{(N-1)^{2n}}\right). \tag{13}$$

Now, the R.H.S. of (13) is equal to

$$\frac{nN^{n-1} + \sum_{i=2}^n \binom{n}{i} N^{n-i}[(-1)^i - (-2)^i]}{N^n - 2nN^{n-1} + \sum_{i=2}^n \binom{n}{i} N^{n-i}(-2)^i}. \tag{14}$$

Now we will show

$$\sum_{i=2}^n \binom{n}{i} N^{n-i}[(-1)^i - (-2)^i] \leq 0. \tag{15}$$

The proof is given when n is odd. The case n even follows by the following proof as $n-1$ is odd and the additional term is nonpositive. Now we have for $k \in \{1, 2, 3, \ldots, (n-1)/2\}$ that the sum of consecutive even and odd terms is equal to

$$\frac{N^{n-2k}n!}{(n-2k-1)!(2k)!}\left[\frac{N(1-2^{2k})(2k+1) - (2^{2k+1}-1)(n-2k)}{(n-2k)(2k+1)N}\right],$$

which is nonpositive as

$$N \geq \frac{(2^{2k+1}-1)(n-2k)}{(1-2^{2k})(2k+1)}.$$

Thus, we have established (15). We will now show that

$$\sum_{i=2}^{n} \binom{n}{i} N^{n-i}(-2)^i \geq 0. \tag{16}$$

Following the same approach as shown previously (i.e., n is odd), the sum of consecutive even and odd terms is equal to

$$\frac{N^{n-2k}2^{2k}n!}{(n-2k-1)!(2k)!}\left[\frac{N(2k+1)-2(n-2k)}{(n-2k)(2k+1)N}\right].$$

This is nonnegative if

$$N \geq \frac{n-2k}{2k+1},$$

as $N \geq 2n/(1-\beta)$ and $6 \geq (1-\beta)$, it follows that $N \geq n/3 \geq (n-2k)/(2k+1)$; thus, one can establish (16).

Now, returning to (13) and noting (14), (15),and (16), we have

$$\mathbb{E}_{\gamma,N}\left[\left(\frac{\prod_{k=1}^{n}\frac{1}{M_k-1}}{\prod_{k=1}^{n}\frac{\alpha_k(y_{1:k},\epsilon,\gamma)}{N-1}} - 1\right)^2\right] \leq \frac{nN^{n-1}}{N^n - 2nN^{n-1}} = \frac{n}{N-2n},$$

as $N \geq 2n/(1-\beta)$, it follows that $n/(N-2n) \leq Cn/N$, and we conclude. □

Proof. [Proof of Proposition 3.2] We have (dropping the superscript t on M_k)

$$\begin{aligned}\mathbb{E}_{\zeta K^t \otimes \tilde{Q}}\left[\sum_{k=1}^{n} M_k\right] &= \int_{(\Theta\times\mathsf{X})^2} \sum_{\{N,N+1,\dots\}^n}\left(\sum_{k=1}^{n} m_k\right)\left\{\prod_{k=1}^{n}\binom{m_k-1}{N-1}\alpha_k(y_{1:k},\gamma',\epsilon)^N\right.\\ &\quad \left.\times(1-\alpha_k(y_{1:k},\gamma',\epsilon))^{m_k-N}\right\}q(\gamma,\gamma')\zeta K^t(d\gamma)d\gamma'\\ &= \int_{(\Theta\times\mathsf{X})^2}\left(\sum_{k=1}^{n}\frac{N}{\alpha_k(y_{1:k},\gamma,\epsilon)}\right)q(\gamma,\gamma')\zeta K^t(d\gamma)d\gamma' \leq \frac{nN}{C},\end{aligned}$$

where we have used the expectation of a negative binomial random variable and applied $\inf_{k,\gamma}\alpha_k(y_{1:k},\gamma,\epsilon) \geq C$ in the inequality. □

ACKNOWLEDGMENTS

A. Jasra acknowledges support from the MOE Singapore and funding from Imperial College London. N. Kantas was kindly funded by EPSRC under grant EP/J01365X/1. We thank two referees and the associate editor for extremely useful comments that have vastly improved the article.

REFERENCES

C. Andrieu, A. Doucet, and R. Holenstein. 2010. Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc. Ser. B* 72, 269–342.

C. Andrieu and M. Vihola. 2014. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann Appl. Probab*. Retrieved March 20, 2014, from arxiv.org/abs/1210.1484. (To appear).

S. Barthelmé and N. Chopin. 2014. Expectation-Propagation for Summary-Less, Likelihood-Free Inference. Technical Report. ENSAE. *J. Amer. Statist. Assoc*. (To appear).

M. A. Beaumont. 2003. Estimation of population growth or decline in genetically monitored populations. *Genetics* 164, 1139.

M. A. Beaumont, J. M. Cornuet, J. M. Marin, and C. P. Robert. 2009. Adaptive approximate Bayesian computation. *Biometrika* 86, 983–990.

O. Cappé, É. Moulines, and T. Ryden. 2005. *Inference in Hidden Markov Models*. Springer, New York, NY.

J. M. Chambers, C. L. Mallows, and B. W. Stuck. 1976. Method for simulating stable random variables. *J. Amer. Statist. Assoc.* 71, 340–344.

D. R. Cox. 1981. Statistical analysis of time-series: Some recent developments. *Scand. J. Statist.* 8, 93–115.

T. A. Dean, S. S. Singh, A. Jasra, and G. W. Peters. 2014. Parameter estimation for hidden Markov models with intractable likelihoods. *Scand. J. Statist*. Retrieved March 20, 2014, from arxiv.org/abs/1103.5399. (To appear).

P. Del Moral, A. Doucet, and A. Jasra. 2012. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statist. Comp.* 22, 1009–1020.

R. Douc, P. Doukhan, and E. Moulines. 2012. Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stoch. Proc. Appl.* 123, 2620–2647.

A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn. 2012. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. Retrieved March 20, 2014, from arxiv.org/abs/1210.1871.

J. Fan and Q. Yao. 2005. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York, NY.

P. Fearnhead and D. Prangle. 2012. Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. Roy. Statist. Soc. Ser. B* 74, 419–474.

A. Jasra, S. S. Singh, J. S. Martin, and E. McCoy. 2012. Filtering via approximate Bayesian computation. *Statist. Comp.* 22, 1223–1237.

A. Jasra, A. Lee, C. Yau, and X. Zhang. 2013. The alive particle filter. Retrieved March 20, 2014, from arxiv.org/abs/1304.0151.

A. Lee. 2012. On the choice of MCMC kernels for approximate Bayesian computation with SMC samplers. In *Proceedings of the Winter Simulation Conference (WSC'12)*. 1–12.

A. Lee and K. Latuszynski. 2012. Variance bounding and geometric ergodicity of Markov chain Monte Carlo for approximate Bayesian computation. Retrieved March 20, 2014, from arxiv.org/abs/1210.6703.

P. Majoram, J. Molitor, V. Plagnol, and S. Tavare. 2003. Markov chain Monte Carlo without likelihoods. *Proc. Nat. Acad. Sci.* 100, 15324–15328.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. Ryder. 2012. Approximate Bayesian computational methods. *Statist. Comp.* 22, 1167–1180.

M. F. Neuts and S. Zacks. 1967. On mixtures of χ^2 and $F-$ distributions which yield distributions of the same family. *Ann. Inst. Stat. Math.* 19, 527–536.

R. D. Wilkinson. 2013. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statist. Appl. Genetics Mole. Biol.* 12, 129–141.

S. Zacks. 1980. On some inverse moments of negative-binomial distributions and their application in estimation. *J. Stat. Comp. Sim.*, 10, 163–165.

Received June 2013; revised September 2013, November 2013; accepted November 2013