ADAPTIVE APPROXIMATE FILTERING OF STATE-SPACE MODELS

Kamil Dedecius

Institute of Information Theory and Automation
Czech Academy of Sciences
Pod Vodárenskou věží 1143/4, 182 08 Prague, Czech Republic

ABSTRACT

Approximate Bayesian computation (ABC) filtration of state-space models replaces popular particle filters in cases where the observation models (i.e. likelihoods) are either computationally too demanding or completely intractable, but it is still possible to simulate from them. These sequential Monte Carlo methods evaluate importance weights based on the distance between the true observation and the simulated pseudo-observations. The paper proposes a new adaptive method consisting of probability kernel-based evaluation of importance weights with online determination of kernel scale. It is shown that the resulting algorithm achieves performance close to particle filters in the case of well-specified models, and outperforms generic particle filters and state-of-art ABC filters under heavy-tailed noise and model misspecification.

Index Terms— Approximate Bayesian computation, ABC, filtration, adaptive kernels

1. INTRODUCTION

In many applications we are interested in a sequential estimation of a discrete-time state-space model with hidden states $\{X_n\}_{n=1,2,\dots}$ and observations $\{Y_n\}_{n=1,2,\dots}$ given by

$$X_n|(X_{n-1}=x_{n-1}) \sim f(x_n|x_{n-1}) \quad (1)$$

$$Y_n|(X_n=x_n) \sim g(y_n|x_n), \quad (2)$$

where f and g are known nonlinear scalar or multivariate functions, and the prior $X_0 \sim \pi(x_0)$. Such models abound in econometrics, target tracking, computer vision, computational biology and many other fields, see, e.g. [1, 2].

The optimal Bayesian filtering aims to sequentially infer the distribution $\pi(x_{1:n}|y_{1:n})$ from observations $y_1, \dots, y_n$ by virtue of the Bayes' theorem

$$\pi(x_{0:n}|y_{1:n}) \propto \pi(x_0) \prod_{k=1}^{n} g(y_k|x_k) f(x_k|x_{k-1}). \quad (3)$$

However, except for a limited number of rather special cases, the posterior distribution is analytically intractable and forces one to resort to approximate inference, mostly based on Monte Carlo methods [1–3]. Their specific branch termed *particle filters* (PFs) approximate the target density $\pi(x_{0:n}|y_{1:n})$ by drawing samples $x_{0:n}^{(i)}$ from convenient proposal densities $q_n(x_n|y_n, x_{1:n-1})$ and representing the target density by $\{W_n^{(i)}, X_n^{(i)}\}_{i=1,\dots,I}$, where the importance weights take values in the unit I-simplex. The Bayesian update (3) then recursively incorporates new observations y_n,

$$W_n^{(i)} \propto W_{n-1}^{(i)} \frac{g\left(y_n|x_n^{(i)}\right) f\left(x_n^{(i)}|x_{n-1}^{(i)}\right)}{q\left(x_n^{(i)}|y_n, x_{n-1}^{(i)}\right)}. \quad (4)$$

A subsequent resampling follows to prevent particle depletion [3]. There remains to choose $q_n(x_n|y_n, x_{n-1})$. Often the density $f(x_n|x_{n-1})$ is a reasonable choice, resulting in the so-called bootstrap filter [3, 4].

In certain cases the observation model (2) is either too complex to be evaluated analytically or numerically (but it is still possible to sample from it by plugging the state), or it is a rather rough approximation of the true model. In such situations, the particle filter can be superseded by the so-called approximate Bayesian computation (ABC) filter [5]. The ABC methods avoid evaluation of the observation model likelihood (2) by matching observations with simulated pseudo-observations (see [6, 7] for recent reviews). The resulting approximation of posterior distribution (3) is based upon defining a probability distribution on an extended state space with the pseudo-observations lying on the data-space and (usually) distributed according to the true model [8]. The first SMC filter replacing (4) by its ABC counterpart was proposed quite recently (2012) by Jasra *et al.* [5]. Later, Calvet and Czellar [9] proposed its modification, replacing the uniform kernel by nonuniform kernels with variable kernel scales.

Inspired by both the original [5] and the improved filter [9], a new method is proposed in this contribution. It also exploits probability kernels, but alleviates some restrictions imposed in [9]. The approximation tolerance is driven by the required number of pseudo-observations to be covered by a preset credibility region of the kernel. The resulting filter has a good performance and stability under heavy-tailed noise.

This work is supported by the Czech Science Foundation, grant 14-06678P.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. ABC FILTERS

The ABC filters approximate the target density $\pi(x_{0:n}|y_{1:n})$ with samples $x_n^{(i)}$, obtained again from a suitable proposal distribution $q(x_n|y_n, x_{n-1})$, that give rise to *pseudo-observations* $u_n^{(i)} \sim g(y_n|x_n^{(i)})$ in some sense close to the true observed y_n. This closeness is determined by a kernel probability density function $\tilde{g}_{\varepsilon_n}(y_n, u_n^{(i)})$. The approximate analogue of the posterior distribution (3) has the form [5]

$$\begin{aligned}\tilde{\pi}(x_{0:n}|y_{1:n}) &= \pi(x_0) \int \tilde{\pi}(x_{1:n}, u_{1:n}|y_{1:n}) du_{1:n} \\ &\propto \pi(x_0) \prod_{k=1}^{n} \left[\int \tilde{g}_{\varepsilon_n}(y_k, u_k) g(y_k|x_k) du_k \right] f(x_k|x_{k-1}).\end{aligned} \tag{5}$$

The resulting ABC filtering algorithm is a sequential importance resampling (SIR) type Monte Carlo algorithm similar to the particle filter presented above. The ABC importance weights update has the form

$$W_n^{(i)} \propto W_{n-1}^{(i)} \tilde{g}_{\varepsilon_n}\left(y_n, u_n^{(i)}\right). \tag{6}$$

Obviously, this basic setting leads to a bootstrap-type filter with the proposal distribution identified with the state evolution function.

The original ABC filter of Jasra et al. [5] uses (similarly to traditional nonsequential ABC methods) the uniform kernel

$$\tilde{g}_{\varepsilon_n}\left(y_n, u_n^{(i)}\right) = \mathbb{1}_{A_{\varepsilon_n, y_n}}\left(u_n^{(i)}\right), \tag{7}$$

where

$$A_{\varepsilon_n, y_n} = \{u : \rho(u, y_n) \leq \varepsilon_n\}, \qquad \varepsilon_n > 0. \tag{8}$$

Here, the metric $\rho(u, y_n)$ measures the distance of the pseudo-observation u from the true observation y_n. For instance, it can be the Manhattan L_1 or the Euclidean L_2 norm. It is shown in [5] that under fixed ε the presented ABC filter converges to a biased estimator as the number of particles N tends to infinity and that the bias itself tends to zero as ε_n goes to zero.

The uniform kernel (7) suffers two major drawbacks. First, its sequential adaptation is mandatory, otherwise with fixed ε, the filter may abruptly fail if the true observation y_n is an outlier, making the set A_{ε_n, y_n} empty. Second, the resulting importance weights are either zero or proportional to one: the pseudo-observations are either in A_{ε, y_n} or not. The remedy for the first issue consists in presetting a number $\alpha \in \{1, \ldots, I\}$ of particles closest to y_n to be always accepted. Consequently, $\varepsilon_n = \rho(u_n^{([\alpha])}, y_n)$ where $u_n^{([\alpha])}$ denotes the pseudo-observation having the αth least distance from y_n. Inspired by nonparametric kernel density estimation, Calvet and Czellar [9] propose to resolve the second issue and the problem of double convergence in I and ε_n by means of probability kernels with scale dependent on I. They develop a plug-in rule for the choice of kernel scale. Their approach follows the traditional kernel theory, e.g. [10, Chap. 3], where the merits and drawbacks of presented methods (including the one adopted in [9]) are discussed.

2.1. Contribution

A new method for determination of particle weights is proposed in this contribution. It is based on probability kernels, that is (possibly non-normalized) *symmetric probability density functions centered at zero and with a scale parameter* ε_n. This definition of kernels is much less restrictive than the one assumed in the kernel density estimation framework adopted by Calvet and Czellar [9, Assumption 1], imposing additional restrictions on the existence of the first two moments, and thus ruling out some popular and computationally attractive kernels like the Cauchy one. The optimal kernel scale is determined each time step based on the preset filter tolerance. Under stable behavior of observations (symmetric noise centered at zero), the scale shrinks, reducing the estimator bias. If outliers occur, the scale immediately grows, suppressing the influence of such observations. Examples demonstrate, that particularly the computationally appealing heavy-tailed Cauchy kernel leads to good filtering performance.

3. PROPOSED ADAPTIVE ABC FILTER

Suppose that a functional form of a probability kernel (a symmetric probability density function with some scale ε_n) is chosen and time n is fixed. The core idea of the proposed method is to adjust its scale parameter ε_n so that the kernel p-credibility region (i.e. the p-highest probability density region) covers exactly $\alpha \in \{1, \ldots, N\}$ pseudo-observations $u_n^{(i)}$ generated by particles $x_n^{(i)}$. This means to find ε_n such that the αth least distant pseudo-observation $u_n^{([\alpha])}$ is the $(p+1)/2$ quantile of the kernel. In statistics, one typically chooses $p \geq 0.95$.

After the user sets the tuning parameters p and α, sequential computation of the kernel scale involves (i) computation of distances $||u_n^{(i)}, y_n||$, (ii) finding $u_n^{([\alpha])}$, the αth least distant pseudo-observation, (iii) computation of ε_n. The resulting scale is then directly plugged in the kernel, and the update of importance weights (6) is performed. Algorithm 1 summarizes the proposed adaptive filter.

The determination of the scale ε_n is in many cases of standard statistical distributions quite easy and computationally non-intensive. Below, two particularly appealing examples are given: the popular Gaussian kernel and the Cauchy kernel. It will be demonstrated later in simulated examples that the latter brings additional filter robustness due to its very heavy tails.

Gaussian kernel

$$\tilde{g}_{\varepsilon_n}\left(y_n, u_n^{(i)}\right) \propto \exp\left(\frac{\left|u_n^{(i)} - y_n\right|^2}{\varepsilon_n^2}\right) \tag{9}$$

$$\text{with} \qquad \varepsilon_n = \frac{\left|u_n^{([\alpha])} - y_n\right|}{\Phi^{-1}\left(\frac{p+1}{2}\right)}, \tag{10}$$

where Φ^{-1} is the quantile (inverse cumulative distribution) function of $\mathcal{N}(0,1)$.

Cauchy kernel

$$\tilde{g}_{n,\varepsilon_n}\left(y_n, u_n^{(i)}\right) \propto \left(1 + \frac{\left|u_n^{(i)} - y_n\right|^2}{\varepsilon_n^2}\right)^{-1} \tag{11}$$

$$\text{with} \qquad \varepsilon_n = \frac{\left|u_n^{([\alpha])} - y_n\right|}{\tan(\pi p)}. \tag{12}$$

3.1. Properties

Thorough analysis of the proposed ABC filter with kernel adaptation is beyond the scope and available extent of this contribution. However, it is immediately possible to recognize certain important properties:

Stability under model misspecificaton (outliers, heavy-tailed noise) – Under stable conditions, the pseudo-observations concentrate and the kernel scale shrinks. However, if an outlier occurs, the resulting scale becomes high in order to cover the required number of particles, which makes the associated weights to follow a flatter distribution (Fig. 1). This prevents the particle weights from degeneration, which would usually occur in the generic particle filter. On the other hand, informative outliers can be taken into account with a suitable kernel shape.

Filtration with intractable/computationally demanding observation likelihoods – a design property of ABC methods [6].

Straightforward application to higher dimensions – often, it is easy to evaluate credibility regions of standard multivariate statistical distributions. This task is not so straightforward in KDE-based methods [10].

Relatively cheap computations with popular kernels – e.g., in the case of the Cauchy kernel it is necessary to (i) find the αth least distant particle (search in $N \times 1$ array), (ii) calculate the scale ε_n (evaluation of tangent function) and (iii) evaluation of particles weights as Cauchy kernel values (no special functions). Thus the complexity can be lower than in particle filters and it is comparable to Jasra's filter (which is a special case of the proposed filter).

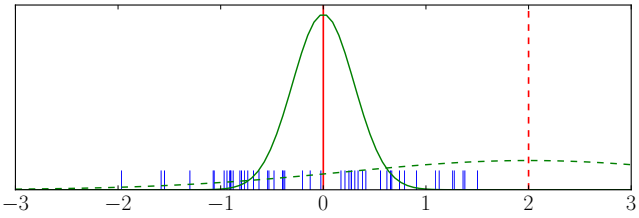

Fig. 1. Kernel adaptation. If the pseudo-observations $u_n^{(i)}$ (blue ticks) are close to the true observation $y_n = 0$ (solid red line), a narrow kernel (solid green line) covers the required number of them. If the true observation is an outlier $y_n = 2$ (dashed red line), the kernel must be very flat (dashed green).

Algorithm 1 ABC FILTER WITH ADAPTIVE KERNEL

Sample initial particles $x_0^{(i)}, i = 1, \ldots, I$ from a suitable prior distribution $\pi(x_0)$, assign uniform initial importance weights $W_0^{(i)} = 1/I$. Choose a kernel function $\tilde{g}_{\varepsilon_n}$, e.g. (9) or (11), set the credibility region level p and the associated number of particles α to be covered by it.

For $n = 1, 2, \ldots$ **do:**

1. Obtain observation y_n.
2. Propagate particles $x_n^{(i)} \sim q_n(x_n | y_n, x_{n-1})$.
3. Simulate pseudo-observation $u_n^{(i)} \sim g_n\left(y_n \middle| x_n^{(i)}\right)$.
4. Kernel adaptation:
   (a) Calculate distances $\left|\left|u_n^{(i)} - y_n\right|\right|$.
   (b) Find $u_n^{([\alpha])}$, the αth least distant pseudo-observation.
   (c) Calculate kernel scale ε_n, e.g. (10) or (12).
5. Update weights $W_n^{(n)} \propto W_{n-1}^{(i)} \tilde{g}_{\varepsilon_n}\left(y_n, u_n^{(i)}\right)$.
6. Resample if the effective sample size drops below a specified threshold.

4. EXAMPLES

Two following two examples demonstrate performance of the proposed method. First, a state-space model with nonlinear state and linear observation equations with known normal noise terms is adopted. This setting, ideal for the particle filters, demonstrates that the proposed adaptive ABC filter performance is only very slightly worse. The second example considers a completely nonlinear multimodal state-space model with a heavy-tailed observation noise and model misspecification. The performance measure is the mean squared error (MSE),

$$\begin{aligned} MSE &= \frac{1}{N}\sum_{n=1}^{N}(\hat{x}_n - x_n)^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{i=1}^{I} W_n^{(i)} x_n^{(i)} - x_n\right)^2. \end{aligned} \tag{13}$$

Both examples exploit a one-dimensional state-space model popular in particle filtering literature [11–13], which we simplify in the first example. Three ABC filters are employed: the proposed kernel-based filters with the Gaussian and Cauchy kernel, and the original filter of Jasra *et al.* [5]. The ABC filter of Calvet and Czellar [9] with the Gaussian and quasi-Cauchy kernels were assessed as well; for their lower performance only MSE values are reported. A particle filter is used for comparison with the ABC filters. All methods are "off-the-shelf", i.e. without any additional tuning. They are all bootstrap filters with 1000 particles. The multinomial resampling is performed each n, followed by a random walk with variance 0.5. The absolute value $|u_n^{(i)} - y_n|$ serves as the distance function. The ABC credibility intervals with $p = 0.95$ cover 300 pseudo-observations. All filters are initialized with identical set of particles sampled from the uniform distribution $\mathcal{U}(-100, 100)$.

The commented source code is freely available on the website `http://diffest.utia.cas.cz`.

Example 1: Linear observation model with normal noise

We consider a state-space model of the form

$$x_n = \frac{x_{n-1}}{2} + \frac{25x_{n-1}}{1 + x_{n-1}^2} + 8\cos(1.2n) + v_n,$$
$$y_n = x_n + w_n,$$

initialized from $x_0 = 0$. The series have 100 samples ($n = 1, \ldots, 100$). v_n and w_n are independent identically distributed zero-mean normal noise variables with standard deviations 1 and 10, respectively; the realizations of the observation noise are depicted in Figure 2.

Figure 3 depicts the evolutions of estimation residues $\hat{x}_n - x_n$ and the associated final MSE. The results indicate that the proposed adaptive ABC filter outperforms the original ABC filter of Jasra *et al.* and, moreover, that it attains MSE performance close to the bootstrap particle filter with both kernels. Both kernels exhibit very similar behavior, the Cauchy kernel is slightly less sensitive to the noise. The ABC filter of Calvet and Czellar [9] with the Gaussian and quasi-Cauchy kernels attains MSEs 36.7 and 110.6, respectively.

Finally, Figure 4 shows evolution of the kernel scales ε_n. Apparently, the Cauchy kernel is more stable due to its heavy tails. A fast adaptation from a very flat prior information is apparent in both cases.

Example 2: Nonlinear model with Cauchy noise

The second example deals with the popular completely nonlinear multimodal state-space model of the form

$$x_n = \frac{x_{n-1}}{2} + \frac{25x_{n-1}}{1 + x_{n-1}^2} + 8\cos(1.2n) + v_n,$$
$$y_n = \frac{x_n^2}{20} + w_n,$$

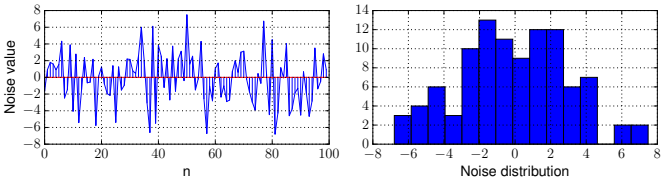

Fig. 2. Example 1: Evolution of noise realizations and their histogram.

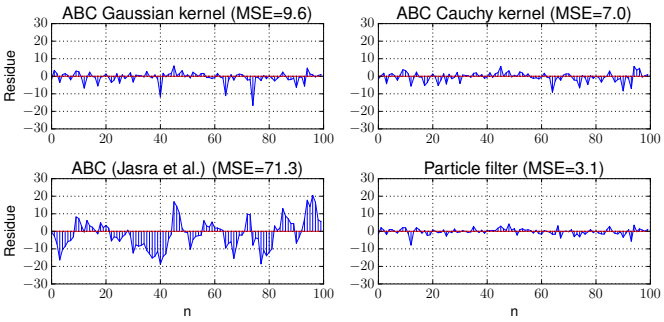

Fig. 3. Example 1: Evolution of estimation residues $\hat{x}_n - x_n$.

initialized in the same way as in the previous example with the exception that w_n has a Cauchy distribution with zero mode and a unit scale. Figure 5 depicts evolution of this heavy-tailed noise: clearly, there are several outliers in the series.

Figure 6 shows the evolution of estimation residues of all four filters. The Cauchy-kernel ABC filter reaches the best results with reasonably stable estimates and fast recovery from outliers, its MSE=29.9. The Gaussian kernel is more sensitive, but it also reaches reasonable results. On the other hand, the original ABC filter exhibits degraded performance due to its weighting strategy. As expected, the off-the-shelf bootstrap particle filter suffers from model misspecification and outliers. Finally, the filter of Calvet and Czellar [9] with the Gaussian and quasi-Cauchy kernels has MSE values 73.3 and 122.7, respectively.

Figure 7 shows fast adaptation of kernel scales ε_n in both kernels. The spikes correlate well with the departures from the expected observation values. The filters are stabilized during these events due higher scale leading to lower difference in weights.

Conclusion

A novel adaptive approximate Bayesian computation method for filtration of nonlinear state-space models with computationally demanding or intractable likelihoods was proposed. Its performance is close to the particle filters in settings where the latter can be used (and dominate). However, in difficult scenarios where the particle filter struggles, the ABC method provides a reasonable estimation accuracy and stability.

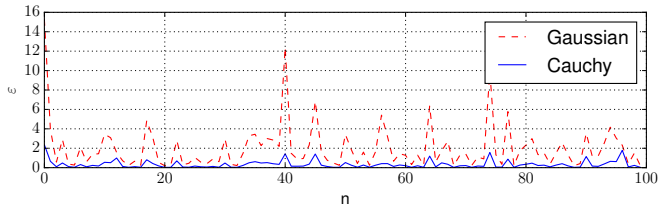

Fig. 4. Example 1: Evolution of kernel scales ε_n of the Gaussian and Cauchy kernels, respectively.

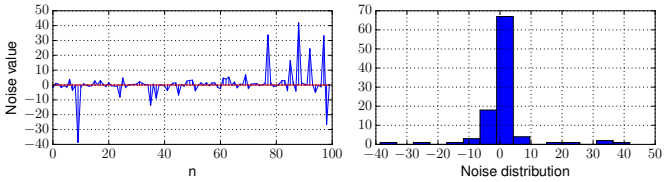

Fig. 5. Example 2: Evolution of noise realizations and their histogram.

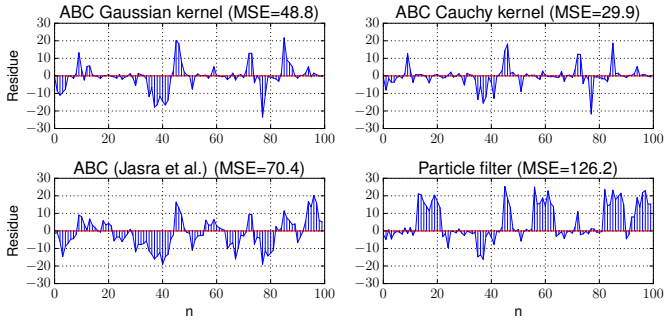

Fig. 6. Example 2: Evolution of estimation residues $\hat{x}_n - x_n$.

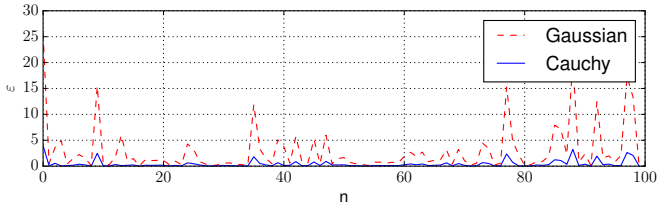

Fig. 7. Example 2: Evolution of kernel scales ε_n of the Gaussian and Cauchy kernels, respectively.

Future works include the topics of adaptive proposal distributions and thorough theoretical and experimental assessment of performance and convergence properties of the method. Also, recent results of Bornn et al. [14] indicating that multiple pseudo-observations do not necessarily improve efficiency of certain ABC algorithms deserve our focus.

REFERENCES

[1] O. Cappé, E. Moulines, and T. Rydén, *Inference In Hidden Markov Models*, Springer, New York, 2005.

[2] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*, Oxford University Press, 2012.

[3] A. Doucet and A. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," in *Oxford Handbook of Nonlinear Filtering*, 2011, pp. 656–704.

[4] J. Cornebise, E. Moulines, and J. Olsson, "Adaptive methods for sequential importance sampling with application to state space models," *Statistics and Computing*, vol. 18, no. 4, pp. 461–480, Aug. 2008.

[5] A. Jasra, S.S. Singh, J.S. Martin, and E. McCoy, "Filtering via approximate Bayesian computation," *Statistics and Computing*, vol. 22, no. 6, pp. 1223-1237, 2012.

[6] J. M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, "Approximate Bayesian computational methods," *Statistics and Computing*, vol. 22, no. 6, pp. 1167–1180, 2012.

[7] P. Green, K. Łatuszynski, M. Pereyra, and C. P. Robert, "Bayesian computation: A perspective on the current state, and sampling backwards and forwards," *Statistics and Computing*, to appear.

[8] A. Jasra, N. Kantas, and E. Ehrlich, "Approximate inference for observation-driven time series models with intractable likelihoods," *ACM Transactions on Modeling and Computer Simulation*, vol. 24, no. 3, pp. 1–25, 2014.

[9] L. Calvet and V. Czellar, "Accurate methods for approximate Bayesian computation filtering," *Journal of Financial Econometrics*, July 2014. *(to appear)*

[10] B. W. Silverman, "*Density Estimation for Statistics and Data Analysis*," Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1986.

[11] O. Cappé, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, May 2007.

[12] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 3, pp. 269–342, June 2010.

[13] J. S. Martin, A. Jasra, S. S. Singh, N. Whiteley, P. Del Moral, and E. McCoy, "Approximate Bayesian computation for smoothing," *Stochastic Analysis and Applications*, vol. 32, no. 3, pp. 397–420, Apr. 2014.

[14] L. Bornn, N. Pillai, A. Smith, and D. Woodard, "One pseudo-sample is enough in approximate Bayesian computation MCMC," *arXiv:1404.6298*, Apr. 2014.