The University of
Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

White, S.R. and Kypraios, Theodore and Preston, Simon P. (2015) Piecewise Approximate Bayesian Computation: fast inference for discretely observed Markov models using a factorised posterior distribution. Statistics and Computing, 25 (2). pp. 289-301. ISSN 1573-1375

Piecewise Approximate Bayesian Computation: fast inference for discretely observed Markov models using a factorised posterior distribution

S.R. White · T. Kypraios · S.P. Preston

Abstract Many modern statistical applications involve inference for complicated stochastic models for which the likelihood function is difficult or even impossible to calculate, and hence conventional likelihood-based inferential techniques cannot be used. In such settings, Bayesian inference can be performed using Approximate Bayesian Computation (ABC). However, in spite of many recent developments to ABC methodology, in many applications the computational cost of ABC necessitates the choice of summary statistics and tolerances that can potentially severely bias the estimate of the posterior.

We propose a new "piecewise" ABC approach suitable for discretely observed Markov models that involves writing the posterior density of the parameters as a product of factors, each a function of only a subset of the data, and then using ABC within each factor. The approach has the advantage of side-stepping the need to choose a summary statistic and it enables a stringent tolerance to be set, making the posterior "less approximate". We investigate two methods for estimating the posterior density based on ABC samples for each of the factors: the first is to use a Gaussian approximation for each factor, and the second is to use a kernel density estimate. Both methods have their merits. The Gaussian approximation is simple, fast, and probably adequate for many

applications. On the other hand, using instead a kernel density estimate has the benefit of consistently estimating the true piecewise ABC posterior as the number of ABC samples tends to infinity. We illustrate the piecewise ABC approach with four examples; in each case, the approach offers fast and accurate inference.

Keywords Approximate Bayesian Computation · Simulation · Stochastic Lotka–Volterra

1 Introduction

Stochastic models are commonly used to model processes in the physical sciences (Wilkinson 2011a; Van Kampen 2007). For many such models the likelihood is difficult or costly to compute making it infeasible to use conventional inference techniques such as maximum likelihood estimation. However, provided it is possible to simulate from a model, then "implicit" methods such as Approximate Bayesian Computation (ABC) methods enable inference without having to calculate the likelihood. These methods were originally developed for applications in population genetics (Pritchard et al. 1999) and human demographics (Beaumont et al. 2002), but are now being used in a wide range of fields including epidemiology (McKinley et al. 2009), evolution of species (Toni et al. 2009), finance (Dean et al. 2011), and evolution of pathogens (Gabriel et al. 2010), to name a few.

Intuitively, ABC methods involve simulating data from the model using various parameter values and making inference based on which parameter values produced realisations that are "close" to the observed data. Let the data $x = (x_1, \ldots, x_n) \equiv (x(t_1), \ldots, x(t_n))$ be a vector comprising observations of a possibly vector state variable $X(t)$ at

S.R. White (✉)
MRC Biostatistics Unit, Cambridge CB2 0SR, UK
e-mail: simon.white@mrc-bsu.cam.ac.uk

T. Kypraios · S.P. Preston
School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, UK

T. Kypraios
e-mail: Theodore.Kypraios@nottingham.ac.uk

S.P. Preston
e-mail: Simon.Preston@nottingham.ac.uk

Algorithm 1 Exact Bayesian Computation (EBC)

1: Sample θ^* from $\pi(\theta)$.
2: Simulate dataset x^* from the model using parameters θ^*.
3: Accept θ^* if $x^* = x$, otherwise reject.
4: Repeat.

Algorithm 2 Approximate Bayesian Computation (ABC)

As Algorithm 1, but with step 3 replaced by:
3': Accept θ^* if $d(s(x), s(x^*)) \leq \varepsilon$, otherwise reject.

---

time points $t_1, \ldots, t_n$. We assume that the data arise from a Markov stochastic model (which encompasses IID data as a special case) parameterised by the vector θ, which is the target of inference, and we denote by $\pi(x \mid \theta)$ the probability density of the data given a specific value of θ. Prior beliefs about θ are expressed via a density denoted $\pi(\theta)$. Algorithm 1 generates *exact* samples from the Bayesian posterior density $\pi(\theta \mid x)$ which is proportional to $\pi(x \mid \theta)\pi(\theta)$.

This algorithm is only of practical use if $X(t)$ is discrete, else the acceptance probability in Step 3 is zero. For continuous distributions, or discrete ones in which the acceptance probability in step 3 is unacceptably low, Pritchard et al. (1999) suggested Algorithm 2, where $d(\cdot, \cdot)$ is a distance function, usually taken to be the L^2-norm of the difference between its arguments; $s(\cdot)$ is a function of the data; and ε is a tolerance. Note that $s(\cdot)$ can be the identity function but in practice, to give a tolerable acceptance rate, it is usually taken to be a lower-dimensional vector comprising summary statistics that characterise key aspects of the data.

The output of the ABC algorithm is a sample from the ABC posterior density $\tilde{\pi}(\theta \mid x) = \pi(\theta \mid d(s(x), s(x^*)) \leq \varepsilon)$. Provided $s(\cdot)$ is sufficient for θ, then the ABC posterior density converges to $\pi(\theta \mid x)$ as $\varepsilon \to 0$ (Marin et al. 2012). However, in practice it is rarely possible to use an $s(\cdot)$ which is sufficient, or to take ε especially small (or zero). Hence ABC requires a careful choice of $s(\cdot)$ and ε to make the acceptance rate tolerably large, at the same time as trying not to make the ABC posterior too different from the true posterior, $\pi(\theta \mid x)$. In other words, there is a balance which involves trading off Monte Carlo error with "ABC error" owing to the choice of $s(\cdot)$ and tolerance ε.

Over the last decade, a wide range of extensions to the original ABC algorithm have been developed, including Markov Chain Monte Carlo (MCMC) (Marjoram et al. 2003) and sequential (Toni et al. 2009; Dean and Singh 2011) implementations, the incorporation of auxiliary regression models (Beaumont et al. 2002; Blum and François 2010), and (semi-)automatic choice of summary statistics (Fearnhead and Prangle 2012); see Marin et al. (2012) for a review. In all of these ABC variants computational cost is still a central issue, since it is always the computational cost

that determines the balance that can be made between controlling Monte Carlo error and controlling bias arising from using summary statistics and/or non-zero tolerance.

In this paper we propose a novel algorithm called *piece-wise ABC* (PW-ABC), the aim of which is to substantially reduce the computational cost of ABC. The algorithm is applicable to a particular (but fairly broad) class of models, namely those with the Markov property and for which the state variable is observable at discrete time points. The algorithm is based on a factorisation of the posterior density such that each factor corresponds to only a subset of the data. The idea is to apply Algorithm 2 for each factor (a task which is computationally very cheap), to compute the density estimates for each factor, and then to estimate the full posterior density as the product of these factors. Taking advantage of the factorisation lowers the computational burden of ABC such that the choice of summary statistic and tolerance—and the accompanying biases—can potentially be avoided completely.

In the following section we describe PW-ABC in more detail. The main practical issue of the method is how to use the ABC samples from each posterior factor to estimate the full posterior density. We discuss two approaches to estimating the relevant densities and products of densities, then we apply PW-ABC, using both approaches, to four examples: a toy illustrative example of inferring the probability of success in a binomial experiment, a stochastic-differential-equation model, an autoregressive time-series model, and a dynamical predator–prey model. We conclude with a discussion of the strengths and limitations of PW-ABC, and of potential further generalisations.

2 Piece-wise ABC (PW-ABC)

Our starting point is to use the Markov property to write the likelihood as

$$\pi(x \mid \theta) = \left(\prod_{i=2}^{n} \pi(x_i \mid x_{i-1}, \ldots, x_1, \theta)\right) \pi(x_1 \mid \theta)$$

$$= \left(\prod_{i=2}^{n} \pi(x_i \mid x_{i-1}, \theta)\right) \pi(x_1 \mid \theta). \tag{1}$$

The likelihood contribution of the first data point x_1 can be included in inference, but this contribution is asymptotically irrelevant as the number of observations, n, increases, and we henceforth follow the common practice to ignore the factor $\pi(x_1 \mid \theta)$ in (1). Accounting for this, and by using multiple applications of Bayes' theorem, the posterior density can be written in the following factorised form,

Algorithm 3 Piece-Wise Approximate Bayesian Computation (PW-ABC)

for $i = 2$ to n **do**

　　a: Apply the ABC Algorithm to draw m approximate (or exact, if $s(\cdot) = \text{Identity}(\cdot)$ and $\varepsilon = 0$) samples, $\theta^*_{i(1)}, \ldots, \theta^*_{i(m)}$, from $\tilde{\varphi}_i(\theta)$, the implied ABC density;

　　b: Using the samples $\theta^*_{i(1)}, \ldots, \theta^*_{i(m)}$ and either (6) or (12), calculate a density estimate, $\hat{\varphi}_i(\theta)$, of $\tilde{\varphi}_i(\theta)$.

end for

Substitute the density estimates $\hat{\varphi}_i(\theta)$ into (3) to calculate an estimate, $\hat{\pi}(\theta \mid x)$, of $\pi(\theta \mid x)$.

$$\pi(\theta \mid x) \propto \pi(x \mid \theta)\pi(\theta)$$

$$= \left(\prod_{i=2}^n \frac{\pi(x_i \mid x_{i-1}, \theta)\pi(\theta)}{\pi(\theta)} \right) \pi(\theta)$$

$$\propto \pi(\theta)^{(2-n)} \left(\prod_{i=2}^n \varphi_i(\theta) \right), \tag{2}$$

where

$$\varphi_i(\theta) = c_i^{-1} \pi(x_i \mid x_{i-1}, \theta)\pi(\theta)$$

$$c_i = \int \pi(x_i \mid x_{i-1}, \theta)\pi(\theta)\mathrm{d}\theta.$$

Essentially, in (2) the posterior density, $\pi(\theta \mid x)$, of θ given the full data x has been decomposed into a product involving densities $\varphi_i(\theta)$, each of which depends only on a pair of data points, $\{x_{i-1}, x_i\}$.

The key idea now is to use ABC to draw approximate samples from each of the densities $\varphi_i(\theta)$. Applying Algorithm 2 involves (i) drawing θ^* from $\pi(\theta)$, (ii) simulating $x_i^* \mid x_{i-1}, \theta^*$, and (iii) accepting θ^* if $d(s(x_i), s(x_i^*)) \leq \varepsilon$. We use $\tilde{\varphi}_i(\theta)$ to denote the implied ABC density from which these samples are drawn (with $\tilde{\varphi}_i(\theta) = \varphi_i(\theta)$ if $s(\cdot) = \text{Identity}(\cdot)$ and $\varepsilon = 0$). By repeating (i)—(iii) we generate samples of, say, m draws, $\theta^*_{i(1)}, \ldots, \theta^*_{i(m)}$, from each $\tilde{\varphi}_i(\theta)$. Now, suppose that $\hat{\varphi}_i(\theta)$ is an estimate, based on the sample $\theta^*_{i(1)}, \ldots, \theta^*_{i(m)}$, of the density $\tilde{\varphi}_i(\theta)$ (and hence of the density $\varphi_i(\theta)$). Then the posterior density (2) can be estimated by

$$\hat{\pi}(\theta \mid x) = g(\theta) \Big/ \int g(\theta)\mathrm{d}\theta, \tag{3}$$

where

$$g(\theta) = \pi(\theta)^{(2-n)} \left(\prod_{i=2}^n \hat{\varphi}_i(\theta) \right). \tag{4}$$

The steps of PW-ABC are summarised in Algorithm 3.

The rationale of the piecewise approach is to reduce the dimension for ABC, replacing a high-dimensional problem with multiple low-dimensional ones. In standard ABC the summary statistic, $s(\cdot)$, is the tool used to reduce the dimension, but in PW-ABC, with dimension already reduced by the factorisation in (2), we can take $s(\cdot) = \text{Identity}(\cdot)$ and typically use a much smaller ε.

The question remains of how to calculate the density estimates, $\hat{\varphi}_i(\theta)$. Below we discuss two approaches: (i) using a Gaussian approximation, and (ii) using a kernel density estimate. Henceforth, quantities based on (i) are denoted by superscript g, and those based on (ii) are denoted by superscript k. In both cases we discuss the behaviour of the estimators in the asymptotic regime in which the number of observations, n, is kept fixed while the size of each ABC sample increases, $m \to \infty$.

2.1 Gaussian approximation for $\hat{\varphi}_i(\theta)$

Denote the d-dimensional multivariate Gaussian density with mean, μ, and covariance, Σ, by

$$K(\theta; \mu, \Sigma) = (2\pi)^{-d/2}(\det \Sigma)^{-1/2}$$

$$\times \exp\left(-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1}(\theta - \mu) \right). \tag{5}$$

A Gaussian approximation for $\hat{\varphi}_i(\theta)$ is

$$\hat{\varphi}_i^{\text{g}}(\theta) = K\big(\theta; \bar{\theta}_i^*, Q_i\big), \tag{6}$$

where

$$\bar{\theta}_i^* = \frac{1}{m} \sum_{j=1}^m \theta^*_{i(j)},$$

$$Q_i = \frac{1}{m-1} \sum_{j=1}^m \big(\theta^*_{i(j)} - \bar{\theta}_i^*\big)\big(\theta^*_{i(j)} - \bar{\theta}_i^*\big)^T,$$

are the sample mean and sample covariance of the ABC posterior sample $\theta^*_{i(1)}, \ldots, \theta^*_{i(m)}$. A consequence of using (6) is that the product of the density approximations is also Gaussian (though in general unnormalised):

$$\prod_{i=2}^n \hat{\varphi}_i^{\text{g}}(\theta) = w \cdot K(\theta; a, B), \tag{7}$$

where

$$B = \left(\sum_{i=2}^n Q_i^{-1} \right)^{-1}, \tag{8}$$

$$a = B\left(\sum_{i=2}^n Q_i^{-1}\bar{\theta}_i^* \right), \tag{9}$$

$$w = \det(2\pi B)^{1/2} \prod_{i=2}^n \det(2\pi Q_i)^{-1/2}$$

$$\times \prod_{s=2}^{n} \prod_{t>s}^{n} \exp\left(-\frac{1}{2}\left(\bar{\theta}_s^* - \bar{\theta}_t^*\right)^T R_{st}\left(\bar{\theta}_s^* - \bar{\theta}_t^*\right)\right), \qquad (10)$$

$$R_{st} = Q_s^{-1} B Q_t^{-1}. \qquad (11)$$

We note the following properties of approximation (6) (see, for example, Mardia et al. 1979). If the densities $\tilde{\varphi}_i(\theta)$ from which the $\theta_{i(1)}^*, \ldots, \theta_{i(m)}^*$ are drawn are Gaussian, i.e., $\tilde{\varphi}_i(\theta) = K(\theta; \mu_i, \Sigma_i)$, then $\bar{\theta}_i^*$ and Q_i are unbiased and consistent estimators of μ_i and Σ_i, respectively, and hence a and B are consistent estimators of the true mean and covariance of $\prod \tilde{\varphi}_i(\theta)$. More generally, for $\tilde{\varphi}_i(\theta)$ which is not necessarily Gaussian, $\bar{\theta}_i^*$ and Q_i are consistent estimators of the mean and the variance of the Gaussian density, $\hat{\varphi}_i^{\mathrm{g}}(\theta)$, which minimises the Kullback–Leibler divergence,

$$\mathrm{KL}\left(\tilde{\varphi}_i(\theta)\|\hat{\varphi}_i^{\mathrm{g}}(\theta)\right) = \int \tilde{\varphi}_i(\theta) \log\left(\tilde{\varphi}_i(\theta)/\hat{\varphi}_i^{\mathrm{g}}(\theta)\right)\mathrm{d}\theta;$$

i.e., for each i, $\hat{\varphi}_i^{\mathrm{g}}(\theta)$ is asymptotically the "optimal" Gaussian approximation to $\tilde{\varphi}_i(\theta)$. No such relevant optimality holds for the product of densities, however: the (normalised) product of Gaussians, each of which is closest in the KL sense to $\tilde{\varphi}_i(\theta)$, is in general not the Gaussian closest to (the normalised version of) $\prod \tilde{\varphi}_i(\theta)$; and indeed it may be very substantially different. In other words, as $m \to \infty$, a and B do *not* in general minimise

$$\mathrm{KL}\left(\left\{\prod \tilde{\varphi}_i(\theta) / \int \left(\prod \tilde{\varphi}_i(\theta)\right)\right\} \Big\| K(\theta, a, B)\right).$$

2.2 Kernel density estimate for $\hat{\varphi}_i(\theta)$

A second method we consider is to estimate each density $\tilde{\varphi}_i(\theta)$ using a kernel density estimate (see for instance Silverman 1986 and Wand and Jones 1995). A kernel density estimate based on Gaussian kernel functions (5) is

$$\hat{\varphi}_i^{\mathrm{k}}(\theta) = \frac{1}{m} \sum_{j=1}^{m} K\left(\theta; \theta_{i(j)}^*, H_i\right), \qquad (12)$$

where H_i is a bandwidth matrix. We follow the approach of Fukunaga (1972) in choosing the bandwidth matrix such that the shape of the kernel mimics the shape of the sample, in particular by taking H_i to be proportional to the sample covariance matrix, Q_i. Using bandwidth matrix

$$H_i = q \cdot m^{-2/(d+4)} Q_i, \qquad (13)$$

where $q > 0$ is a constant not dependent on m, ensures desirable behaviour as the sample size $m \to \infty$. In particular, in terms of the little-o notation ($a_m = o(b_m)$ as $m \to \infty$ denotes $\lim_{m\to\infty} |a_m/b_m| = 0$) and with E denoting expectation, using choice of bandwidth (13), subject to mild regularity conditions on $\tilde{\varphi}_i(\theta)$ (Wand and Jones 1995),

$$E\left\{\hat{\varphi}_i^{\mathrm{k}}(\theta)\right\} = \tilde{\varphi}_i(\theta) + o(1), \qquad (14)$$

$$E\left\{\hat{\varphi}_i^{\mathrm{k}}(\theta)^2\right\} = \tilde{\varphi}_i(\theta)^2 + o(1). \qquad (15)$$

From (14)–(15), the bias, $\mathrm{b}\{\hat{\varphi}_i^{\mathrm{k}}(\theta)\} = E\{\hat{\varphi}_i^{\mathrm{k}}(\theta)\} - \tilde{\varphi}_i(\theta)$, the variance, $\mathrm{var}\{\hat{\varphi}_i^{\mathrm{k}}(\theta)\} = E\{\hat{\varphi}_i^{\mathrm{k}}(\theta)^2\} - E\{\hat{\varphi}_i^{\mathrm{k}}(\theta)\}^2$, and the mean integrated squared error,

$$\mathrm{MISE}\{\hat{\varphi}_i^{\mathrm{k}}\} = E \int \left(\hat{\varphi}_i^{\mathrm{k}}(\theta) - \tilde{\varphi}_i(\theta)\right)^2 \mathrm{d}\theta, \qquad (16)$$

are all $o(1)$. These results generalise routinely to the case of a product of n kernel density estimates, that is, in which $\prod \hat{\varphi}_i^{\mathrm{k}}(\theta)$ is used as an estimator for $\prod \tilde{\varphi}_i(\theta)$. It follows that since the $\theta_{i(j)}^*$ are independent for all i, j, then, using (14)–(15),

$$\mathrm{b}\left\{\prod \hat{\varphi}_i^{\mathrm{k}}(\theta)\right\} = \left\{\prod E\hat{\varphi}_i^{\mathrm{k}}(\theta)\right\} - \prod \tilde{\varphi}_i(\theta) = o(1),$$

$$\mathrm{var}\left\{\prod \hat{\varphi}_i^{\mathrm{k}}(\theta)\right\} = \prod E\left\{\hat{\varphi}_i^{\mathrm{k}}(\theta)^2\right\} - \prod \left\{E\hat{\varphi}_i^{\mathrm{k}}(\theta)\right\}^2 = o(1),$$

$$\mathrm{MISE}\left\{\prod \hat{\varphi}_i^{\mathrm{k}}\right\} = E \int \left(\prod \hat{\varphi}_i^{\mathrm{k}}(\theta) - \prod \tilde{\varphi}_i(\theta)\right)^2 \mathrm{d}\theta = o(1).$$

Hence, in the sense defined by the latter equation, the density estimator $\prod \hat{\varphi}_i^{\mathrm{k}}(\theta)$ converges to the true density $\prod \tilde{\varphi}_i(\theta)$ as $m \to \infty$.

Regarding the choice of q in (13), in certain settings it is possible to determine an optimal value. Suppose that the true density $\tilde{\varphi}_i(\theta)$ is Gaussian and let $\hat{\varphi}_i^{\mathrm{k}}(\theta)$ in (12) be a kernel density estimate of $\tilde{\varphi}_i(\theta)$. Then

$$q = \left\{(d+2)/4\right\}^{-2/(d+4)} \qquad (17)$$

is optimal in the sense that (13) is then an unbiased and consistent estimator of the bandwidth that minimises the leading term of the large-m asymptotic expansion of (16); see Wand and Jones (1995, p. 111). Analogous calculations are rather more involved in the product case, however: even with the assumption that each $\tilde{\varphi}_i(\theta)$ is Gaussian, no closed expression for q is possible. Hence, in the examples in the following section, Sect. 4, we opted to tune q in the heuristic way described by Wand and Jones (1995), starting with a large q (ten times that in (17)) then reducing it manually until "random" fluctuations begin to appear in the density estimates.

A consequence of using Gaussian kernel functions (5) in (12) is that the product of the density approximations is then itself a weighted mixture of $(n-1)^m$ Gaussians,

$$\prod_{i=2}^{n} \hat{\varphi}_i^{\mathrm{k}}(\theta) = m^{(1-n)} \prod_{i=2}^{n} \sum_{j=1}^{m} K\left(\theta; \theta_{i(j)}^*, H_i\right)$$

$$= m^{(1-n)} \sum_{j_2,\ldots,j_n} \prod_{i=2}^{n} K\left(\theta; \theta_{i(j_i)}^*, H_i\right)$$

$$= \sum_{j_2,\ldots,j_n}^{m} w_{j_2,\ldots,j_n} K\left(\theta; a_{j_2,\ldots,j_n}, B_{j_2,\ldots,j_n}\right), \qquad (18)$$

where expressions for the covariances $B_{j_2,...,j_n}$, means $a_{j_2,...,j_n}$, and weights $w_{j_2,...,j_n}$, analogous to those in (8)–(10), are given in Appendix 1.

2.3 Estimating the posterior density

Sections 2.1 and 2.2 describe methods for computing the factor $\prod \hat{\varphi}_i(\theta)$ in (3). For calculating an estimate of the full posterior, $\hat{\pi}(\theta \mid x)$ in (3), we must multiply $\prod \hat{\varphi}_i(\theta)$ by $\pi(\theta)^{(2-n)}$ and normalise. Let us suppose that the prior is Gaussian, $\pi(\theta) = K(\theta; \mu_{\mathrm{pri}}, \Sigma_{\mathrm{pri}})$. For the case where we are using the Gaussian approximation, $\hat{\varphi}_i^{\mathrm{g}}(\theta)$ from (6), for each $\hat{\varphi}_i(\theta)$, then the posterior is

$$\hat{\pi}^{\mathrm{g}}(\theta \mid x) = K(\theta; \mu_{\mathrm{post}}, \Sigma_{\mathrm{post}}), \tag{19}$$

where

$$\Sigma_{\mathrm{post}} = \left((2-n)\Sigma_{\mathrm{pri}}^{-1} + B^{-1}\right)^{-1}, \tag{20}$$

$$\mu_{\mathrm{post}} = \Sigma_{\mathrm{post}}\left((2-n)\Sigma_{\mathrm{pri}}^{-1}\mu_{\mathrm{pri}} + B^{-1}a\right), \tag{21}$$

and a and B are as defined in (7).

If instead we use the kernel approximation, $\hat{\varphi}_i^{\mathrm{k}}(\theta)$ from (12), for each $\hat{\varphi}_i(\theta)$, then the posterior density is

$$\hat{\pi}^{\mathrm{k}}(\theta \mid x) = \sum_{j_2,...,j_n}^{m} w'_{j_2,...,j_n} K\left(\theta; a'_{j_2,...,j_n}, B'_{j_2,...,j_n}\right) \Big/ \sum_{j_2,...,j_n}^{m} w'_{j_2,...,j_n}, \tag{22}$$

where expressions for $B'_{j_2,...,j_n}$, $a'_{j_2,...,j_n}$ and $w'_{j_2,...,j_n}$ are in Appendix 1.

2.4 An expression for the posterior density

In the preceding sections we considered how to sample from the $\varphi_i(\theta)$ and then use the samples to estimate the posterior density $\pi(\theta \mid x)$. Here we consider in more detail the implied posterior density which is targeted by PW-ABC. For either of PW-ABC and ABC, the posterior can be written as

$$\tilde{\pi}(\theta \mid x) \propto \tilde{\pi}(x \mid \theta)\pi(\theta), \tag{23}$$

where $\tilde{\pi}(x \mid \theta)$ is, respectively, either the implied PW-ABC or ABC approximation to the likelihood. First, we define the function

$$K_{\varepsilon,p}(z) = V^{-1}\mathbb{1}\{\|z\|_p \leq \varepsilon\}, \tag{24}$$

where argument z is of dimension, say, u, and either continuous- or discrete-valued in accord with the support of the data; $\|\cdot\|_p$ is the L^p-norm; $\mathbb{1}\{\cdot\}$ is an indicator function; and V, which depends on u, ε, and p, is such that

$\int K_{\varepsilon,p}(z)\mathrm{d}z = 1$, with this integral interpreted as a sum in the discrete case. For ABC with distance $d(\cdot, \cdot)$ taken to be the L^p-norm of the difference between its arguments, the implied ABC approximation to the likelihood (Wilkinson 2013) is the convolution

$$\tilde{\pi}_{\mathrm{ABC}}(x \mid \theta) = \int \pi(y \mid \theta) K_{\varepsilon,p}(y - x)\mathrm{d}y. \tag{25}$$

Hence ABC replaces the true likelihood with an approximate version averaged over an L^p-ball of radius ε centred on the data vector, x. In PW-ABC, we target each $\varphi_i(\theta)$ by an ABC approximation $\tilde{\varphi}_i(\theta) \propto \tilde{\pi}_{\mathrm{ABC}}(x_i \mid x_{i-1}, \theta)\pi(\theta)$, with

$$\tilde{\pi}_{\mathrm{ABC}}(x_i \mid x_{i-1}, \theta) = \int \pi(y_i \mid x_{i-1}, \theta) K_{\varepsilon,p}(y_i - x_i)\mathrm{d}y_i,$$

and the implied PW-ABC likelihood is the product

$$\tilde{\pi}_{\mathrm{PW\text{-}ABC}}(x \mid \theta) = \prod \tilde{\pi}_{\mathrm{ABC}}(x_i \mid x_{i-1}, \theta). \tag{26}$$

Now, to compare directly the implied ABC and PW-ABC likelihood approximations, we neglect as before the likelihood contribution from the first observation x_1, then denote by x' the vector x with x_1 removed (and similar for y); hence we can write (25) and (26), respectively, as

$$\int \pi(y_2 \mid x_1, \theta)\left[\prod_{i=3}^{n} \pi(y_i \mid y_{i-1}, \theta)\right] K_{\varepsilon,p}(y' - x')\mathrm{d}y', \tag{27}$$

and

$$\int \pi(y_2 \mid x_1, \theta)\left[\prod_{i=3}^{n} \pi(y_i \mid x_{i-1}, \theta)\right] K^*_{\varepsilon,p}(y' - x')\mathrm{d}y', \tag{28}$$

where

$$K^*_{\varepsilon,p}(z') = \prod_{i=2}^{n} K_{\varepsilon,p}(z'_i). \tag{29}$$

Two differences between ABC and PW-ABC are clear: first, in ABC the conditioning is on the simulated trajectory, whereas in PW-ABC the conditioning is on the data; and second, in PW-ABC the convolution is with respect to a different kernel (29). This implied kernel seems intuitively reasonable; for example, if the x_i are scalar then the convolution in (28) amounts to an averaging over a hypercube of side length 2ε centred on x'. The difference in the shapes of the regions defined by $K_{\varepsilon,p}(\cdot)$ and $K^*_{\varepsilon,p}(\cdot)$ is of secondary importance, however, since PW-ABC enables use of a much smaller ε than ABC, so the averaging will be over a much smaller region around x', and the approximate likelihood will typically be much closer to the true.

3 Some other considerations

3.1 Practical issues in drawing samples

The independence of the samples $\theta_{i(j)}^*$ for all i, j means that drawing samples for PW-ABC is "embarrassingly parallel", i.e., the task can be divided easily between multiple cores. For example, one approach is use all available cores simultaneously to sample from $\tilde{\varphi}_2(\theta)$ until m draws $\theta_{2(1)}^* \cdots \theta_{2(m)}^*$ are accepted, and then do likewise for $\tilde{\varphi}_3(\theta)$, and so on. Another possibility, which requires less coordination between the cores, is to have different cores sampling from different $\tilde{\varphi}_i(\theta)$, then reassign cores as appropriate whenever any of the $\tilde{\varphi}_i(\theta)$ reaches m accepted samples.

Another benefit of the $\theta_{i(j)}^*$ being independent is that samples can be reused in the event of deciding retrospectively to perform PW-ABC with a smaller ε: the subset of original samples acceptable with the new smaller ε can be retained, leaving only the need, for each $\tilde{\varphi}_i(\theta)$, to "top-up" the number of samples to m. Similarly, samples can obviously be retained given a retrospective decision to use a larger m.

3.2 Estimating the marginal likelihood

In some applications, especially when model comparison is of interest, it is useful to compute the marginal likelihood of the data given the model. The marginal likelihood is

$$\pi(x) = \int \pi(x \mid \theta)\pi(\theta)\mathrm{d}\theta \qquad (30)$$

$$= \left(\prod_{i=2}^{n} c_i\right) \int \left(\prod_{i=2}^{n} \varphi_i(\theta)\right) \pi(\theta)^{2-n}\mathrm{d}\theta. \qquad (31)$$

The unknown c_i can be estimated by $\hat{c}_i = m/(V M_i)$, where M_i equals the number of ABC draws necessary in the ith interval to achieve m acceptances, and V is defined in (24); see Appendix 2. For the integral in (31), using the Gaussian approximation (7) leads to

$$\int \left(\prod_{i=2}^{n} \hat{\varphi}_i^{\mathrm{g}}(\theta)\right) \pi(\theta)^{2-n}\mathrm{d}\theta$$

$$= w \cdot (\det B)^{-1/2} \cdot (\det \Sigma_{\mathrm{post}})^{1/2} \cdot \left(\det(2\pi \Sigma_{\mathrm{pri}})\right)^{(n/2-1)}$$

$$\times \exp\left\{-\frac{1}{2}(a - \mu_{\mathrm{pri}})^T \left((2-n)^{-1} \Sigma_{\mathrm{pri}} + B\right)^{-1}\right.$$

$$\left. \times (a - \mu_{\mathrm{pri}})\right\}, \qquad (32)$$

whereas using the kernel approximation (12) gives

$$\int \left(\prod_{i=2}^{n} \hat{\varphi}_i^{\mathrm{k}}(\theta)\right) \pi(\theta)^{2-n}\mathrm{d}\theta = \sum_{j_2,\ldots,j_n}^{m} w_{j_2,\ldots,j_n}'. \qquad (33)$$

3.3 Practical numerical calculations for the kernel approximation

Since expressions (18), (22), (33) for the kernel case involve sums with $(n - 1)^m$ terms, these expressions are largely of academic interest and are typically not suitable for practical calculations. For the examples in this paper we used a more direct numerical approach, first writing (4) as

$$g(\theta) = \exp\left(\sum_{i=2}^{n} h_i(\theta)\right)\pi(\theta),$$

where $h_i(\theta) = \log(\varphi_i^{\mathrm{k}}(\theta)/\pi(\theta))$, and then evaluating $h_i(\theta)$, $\pi(\theta)$ and hence $g(\theta)$ pointwise on a fine lattice. Performing calculations in this way on the log scale avoids underflow errors and improves numerical stability compared with trying to evaluate (4) directly. As a further check for robustness, we varied the lattice position and resolution to make sure the results were insensitive to the particular choices.

3.4 Sampling from the posterior distribution

In some circumstances it may be desirable to draw samples from the approximate posterior density. In the Gaussian case, drawing from (19) is straightforward. For the kernel case, (22), in principle sampling can be achieved by normalising the weights, randomly choosing a component with probability equal to these normalised weights, then sampling from the selected Gaussian component. But in practice, again, the large number of terms in (22) will typically preclude this approach. Other possibilities include using a Gibbs sampler, or sampling approximately using Gaussian mixtures with fewer components; see Sudderth et al. (2003).

4 Examples

In this section we test PW-ABC on synthetic data from four models. The first, as a toy illustrative example, involves inferring from IID data the probability of success in a binomial experiment. Second is the Cox–Ingersoll–Ross model, a stochastic-differential-equation model for which the continuous state variable has known transition density, which we use to investigate PW-ABC with $\varepsilon > 0$. Third, we consider an integer-valued time series model called INAR(1), a model for which the likelihood is available (albeit awkward to compute) and enables comparison of our approach with a "gold standard" MCMC approach. Finally, we consider a stochastic Lotka–Volterra model, a simple example from a common class of models (which occur, for instance, in modelling stochastic chemical kinetics) in which the likelihood, and therefore many standard methods of inference, are unavailable.
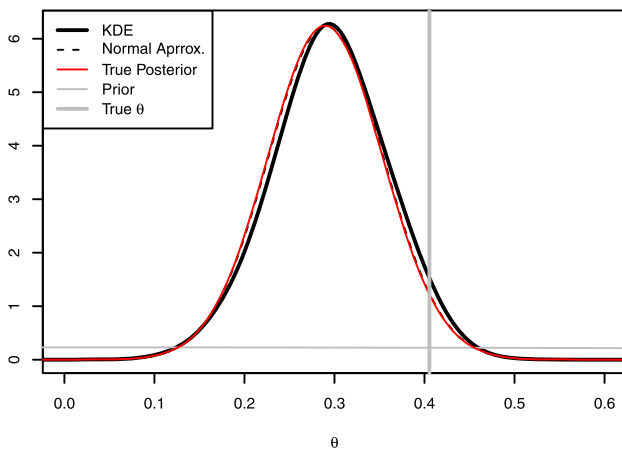
Fig. 1 Results for the binomial model in Sect. 4.1. Shown are the true posterior density, $\pi(\theta \mid x)$, the posterior density approximations $\hat{\pi}^g(\theta \mid x)$ and $\hat{\pi}^k(\theta \mid x)$, the prior, and the true θ

4.1 Binomial model

For this toy example we suppose the data is the set $x = \{x_1, \ldots, x_{10}\}$ of $n = 10$ observations from the model $X_i \sim$ Binom($k_i = 100, p = 0.6$). We work in terms of the transformed parameter $\theta = \text{logit}(p)$, using a prior $\pi(\theta) \sim N(0, 3^2)$. For this model the data are IID, so that $\pi(x_i \mid x_{i-1}, \theta) = \pi(x_i \mid \theta)$. Exact samples from $\varphi_i(\theta)$ can be obtained by sampling θ^* from the prior, sampling $X_i^* \sim \text{Binom}(100, \theta^*)$, and then accepting θ^* if and only if $X_i^* = x_i$. We follow the PW-ABC approach described in Sect. 2, drawing $m = 5000$ samples from each $\varphi_i(\theta)$, using these samples to construct Gaussian $\hat{\varphi}_i^g(\theta)$ and kernel density $\hat{\varphi}_i^k(\theta)$ approximations, then using these density approximations to construct approximate posterior densities, $\hat{\pi}^g(\theta \mid x)$ and $\hat{\pi}^k(\theta \mid x)$. Figure 1 shows that the approximate posterior densities are very close to the true posterior density for this example. The true log marginal likelihood, $\log \pi(x)$, computed by direct numerical integration of (30), is -31.39; using approximation $\hat{\varphi}_i^g(\theta)$ and (32) gives -31.44; and using approximation $\hat{\varphi}_i^k(\theta)$ and numerical integration of the left-hand side of (33) gives -31.48.

4.2 Cox–Ingersoll–Ross Model

The Cox–Ingersoll–Ross (CIR) model (Cox et al. 1985) is a stochastic differential equation (SDE) describing evolution of an interest rate, $X(t)$. The model is

$$dX(t) = a\big(b - X(t)\big)dt + \sigma\sqrt{X(t)}dW(t),$$

where a, b and σ respectively determine the reversion speed, long-run value and volatility, and where $W(t)$ denotes a standard Brownian motion. The density of $X(t_i) \mid X(t_j)$, a, b, σ ($t_i > t_j$) is a non-central chi-square (Eq. (18), Cox et al.

1985), and hence the likelihood is known in closed form. Since the likelihood is known, (PW-)ABC is unnecessary (indeed, in general for SDEs with unknown likelihoods, approaches that exploit the SDE structure—e.g., the likelihood approximations of Aït-Sahalia (2002), or the Monte Carlo methods developed by Durham and Gallant (2002)—are likely to be better choices for inference than (PW-)ABC; however, we include the CIR model here as a simple example of PW-ABC applied to a problem with a continuous state variable, where non-zero choice of ε is necessary, and where the true posterior distribution is available for comparison.

We generated $n = 10$ equally spaced observations from a CIR process with parameters $(a, b, \sigma) = (0.5, 1, 0.15)$ and $X(0) = 1$ on the interval $t \in [0, 4.5]$. Treating a and σ as known, we performed inference on the transformed parameter $\theta = \log(b)$ with a Uniform prior on the interval $(-5, 2)$. Using $\varepsilon = 10^{-2}$ we drew samples of size $m = 10,000$ for each $\varphi_i(\theta)$, $i = 2, \ldots, 10$, achieving acceptance rates around 1.5 % on average.

Figure 2(a) shows the true posterior density, $\pi(\theta \mid \mathcal{X})$, together with the Gaussian- and kernel-based PW-ABC approximations, $\hat{\pi}^g(\theta \mid x)$ and $\hat{\pi}^k(\theta \mid x)$. The figure shows that for sufficiently large m the kernel approximation $\hat{\pi}^k(\theta \mid x)$ agrees very well with the true posterior. The Gaussian approximation $\hat{\pi}^g(\theta \mid x)$, even for large m, does badly here, which is due to skewness of the densities $\varphi_i(\theta)$. Figure 2(b) shows how the posterior density targeted by PW-ABC (see Sect. 2.4) depends on ε, and in particular how it converges to the true posterior density as $\varepsilon \to 0$.

For this example, estimates of the log marginal likelihood, $\log \pi(\mathcal{X})$ are as follows: by direct numerical integration of (30), 8.14; using approximation $\hat{\varphi}_i^g(\theta)$, 2.78; and by using $\hat{\varphi}_i^k(\theta)$ in conjunction with numerical integration of the left-hand side of (33), 7.93.

4.3 An integer-valued autoregressive model

Integer-valued time series arise in contexts such as modelling monthly traffic fatalities (Neal and Subba Rao 2007) or the number of patients in a hospital at a sequence of time points (Moriña et al. 2011). Consider the following integer-valued autoregressive model of order p, known as INAR(p):

$$X_t = \sum_{i=1}^p \alpha_i \circ X_{t-i} + Z_t, \quad t \in \mathbb{Z}, \tag{34}$$

where Z_t for $t > 1$ are independent and identically distributed integer-valued random variables with $E[Z_t^2] < \infty$, with the Z_t assumed to be independent of the X_t. Here we assume $Z_t \sim Po(\lambda)$. Each operator $\alpha_i \circ$ denotes binomial thinning defined by

$$\alpha_i \circ W = \begin{cases} \text{Binomial}(W, \alpha_i), & W > 0, \\ 0, & W = 0, \end{cases} \tag{35}$$
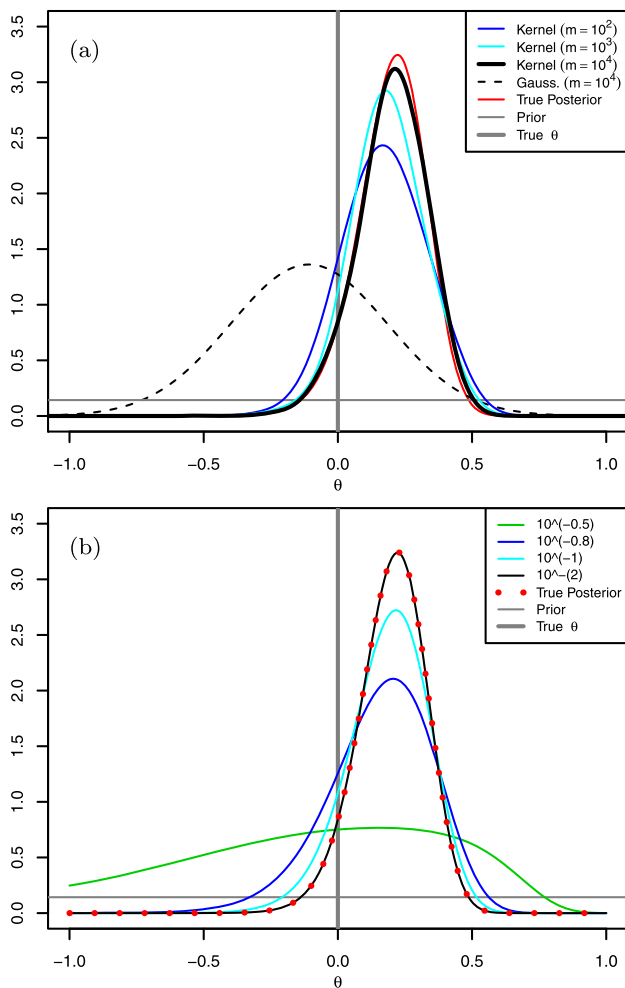
Fig. 2 Results for the CIR model of Sect. 4.2. (a) shows the true posterior density, $\pi(\theta \mid x)$; the PW-ABC posterior density approximations $\hat{\pi}^g(\theta \mid x)$ and $\hat{\pi}^k(\theta \mid x)$ using $\varepsilon = 10^{-2}$, with values of m indicated in the legend; the prior; and the true θ. (b) shows, for various values of ε, the true PW-ABC posterior (defined in Sect. 2.4)
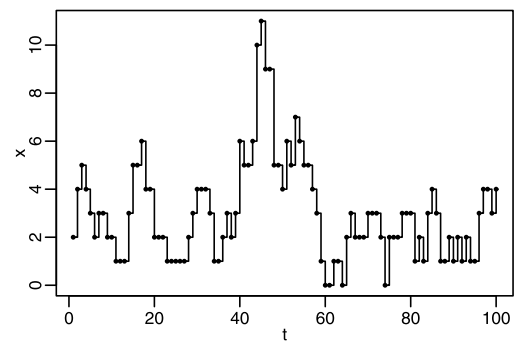

Fig. 3 The realisation of an INAR(1) process used in the example of Sect. 4.3, of length $n = 100$, generated using $\alpha = 0.7$ and $\lambda = 1.0$
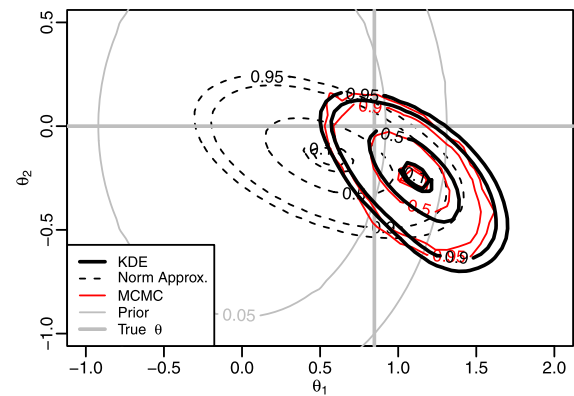

Fig. 4 Results for the INAR(1) example of Sect. 4.3. Shown are an MCMC approximation to the posterior density, $\pi(\theta \mid x)$, the posterior density approximations $\hat{\pi}^g(\theta \mid x)$ and $\hat{\pi}^k(\theta \mid x)$, the prior, and the true θ. The numbers on the contours denote the probability mass that they contain

for non-negative integer-valued random variable W. The operators $\alpha_i \circ$, $i = 1, \ldots p$, are assumed to be independent.

We consider the simplest example of this model, INAR(1) (see, for example, Al-Osh and Alzaid 1987), supposing that we have some observed data $x = \{x_1, \ldots, x_n\}$ from this model and wish to make inference for the parameters (α, λ). We generated $n = 100$ observations from an INAR(1) process using parameters $(\alpha, \lambda) = (0.7, 1)$ and $X(0) = 10$; the realisation is plotted in Fig. 3. Working in terms of the transformed parameter, $\theta = (\theta_1, \theta_2) = (\mathrm{logit}(\alpha), \log(\lambda))$, we used a prior of $\mathrm{Norm}(0, 3^2)$ for each of θ_1 and θ_2. For the EBC algorithm, the probability of acceptance is around 10^{-100} (as estimated from PW-ABC calculations described below), which is prohibitively small; even the ABC algorithm requires a value of ε so large that sequential approaches are needed.

Using PW-ABC with $\varepsilon = 0$ we were able to draw exact samples from $\varphi_i(\theta)$ for all of the $i = 2, \ldots, 100$ factors, and achieve acceptance rates of around 9 %, on average. Figure 4 shows an estimate of the posterior density, $\pi(\theta \mid x)$ based on a gold-standard MCMC approach, together with Gaussian- and kernel-based PW-ABC approximations, $\hat{\pi}^g(\theta \mid x)$ and $\hat{\pi}^k(\theta \mid x)$, with $m = 10,000$ samples for each $\varphi_i(\theta)$. The figure shows good agreement between the MCMC posterior and the kernel approximation, $\hat{\pi}^k(\theta \mid x)$, but again somewhat poor agreement with the Gaussian approximation $\hat{\pi}^g(\theta \mid x)$. The poor performance of $\hat{\pi}^g(\theta \mid x)$ is caused by some of the densities $\varphi_i(\theta)$ being substantially different from Gaussian; see Fig. 5 which shows $\hat{\varphi}^g_{50}(\theta)$ and $\hat{\varphi}^k_{50}(\theta)$, for example. Using Gaussian approximations to non-Gaussian $\varphi_i(\theta)$ appears to have a strong impact on the accuracy of approximation $\hat{\pi}^g(\theta \mid x)$, even, as in the present case, where the true posterior $\pi(\theta \mid x)$, and most of individual $\varphi_i(\theta)$, are reasonably close to a Gaussian (cf. Fig. 4).

For this example, estimates of the log marginal likelihood, $\log \pi(x)$, are as follows: by direct numerical integra-
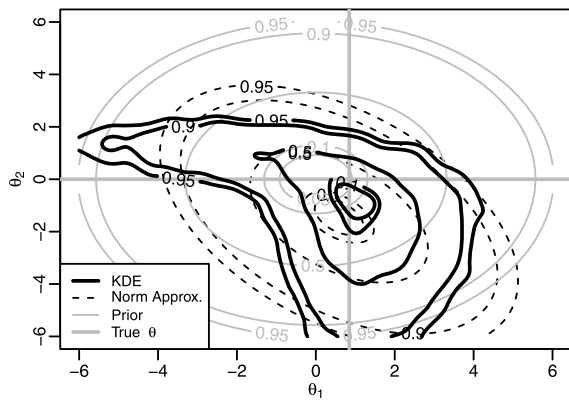
Fig. 5 For the INAR(1) example, an example of a factor with a "non–Gaussian" density: here $\hat{\varphi}_{50}^{g}(\theta)$ and $\hat{\varphi}_{50}^{k}(\theta)$ are substantially different from each other

tion of (30), -161.1; using approximation $\hat{\varphi}_{i}^{g}(\theta)$ and (32), -185.7; and by using $\hat{\varphi}_{i}^{k}(\theta)$ and numerical integration of the left-hand side of (33), -163.2.

We have used $p = 1$ for this example so that the likelihood is available, enabling comparison with MCMC and calculation of the true marginal likelihood. However, we stress that PW-ABC can be easily generalised for $p > 1$, a case for which the likelihood is essentially intractable and therefore one has to resort to either exact but less direct methods (such the Expectation–Maximization (EM) algorithm or data-augmented MCMC, both of which involve treating the terms $\alpha_i \circ X_{t-i}$ and Z_t as missing data) or methods of approximate inference, such as conditional least squares which involves minimizing $\sum_t (X_t - E[X_t \mid X_{t-1}])^2$; see, for example, McKenzie (2003) and references therein.

4.4 Stochastic Lotka–Volterra model

The stochastic Lotka–Volterra (LV) model is a model of predator–prey dynamics and an example of a stochastic discrete-state-space continuous-time Markov process (see, for example, Wilkinson 2011a). Predator–prey dynamics can be thought of in chemical kinetics terms: the predators and prey are two populations of "reactants" subject to three "reactions", namely prey birth, predation and predator death. Exact simulation of such models is straightforward, e.g., using the algorithm of Gillespie (1977). Inference is simple if the type and precise time of each reaction is observed. However, a more common setting is where the population sizes are only observed at discrete time points. In this case the number of reactions that have taken place is unknown and therefore the likelihood is not available and hence inference is much more difficult. Reversible-jump MCMC has been developed in this context (Boys et al. 2008) but it requires substantial expertise and input from the user to implement. Particle MCMC (pMCMC) methods (Andrieu et al.

2010), which provide an approximation to the likelihood via a Sequential Monte Carlo algorithm within an MCMC algorithm, have recently been proposed for stochastic chemical kinetics models (Golightly and Wilkinson 2011). Although being computationally intensive, such methods can work reliably provided the process is observed with measurement error. The R package smfsb, which accompanies (Wilkinson 2011a), contains a pMCMC implementation designed for stochastic chemical kinetics models, and we use this package to compare results for PW-ABC and pMCMC for the following example.

Let Y_1 and Y_2 denote the number of prey and predators respectively, and suppose Y_1 and Y_2 are subject to the following reactions

$$Y_1 \xrightarrow{r_1} 2Y_1, \qquad Y_1 + Y_2 \xrightarrow{r_2} 2Y_2, \qquad Y_2 \xrightarrow{r_3} \emptyset, \qquad (36)$$
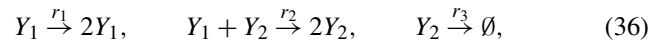
which respectively represent prey birth, predation and predator death. We consider the problem of making inference about the rates (r_1, r_2, r_3) based on observations of Y_1 and Y_2 made at fixed intervals.

We generated a realisation from the stochastic LV example of Wilkinson (Wilkinson, p. 208), that is, model (36) using $(r_1, r_2, r_3) = (1, 0.005, 0.6)$, $Y_1(0) = 50$ and $Y_2(0) = 100$. We performed inference in terms of transformed parameters, $\theta = (\theta_1, \theta_2, \theta_3) = (\log r_1, \log r_2, \log r_3)$, this time with priors $\pi(\theta_1) \sim \text{Norm}(\log(0.7), 0.5)$, $\pi(\theta_2) \sim \text{Norm}(\log(0.005), 0.5)$, and $\pi(\theta_3) \sim \text{Norm}(\log(0.3), 0.5)$. We again applied PW-ABC using $\varepsilon = 0$, in other words requiring an exact match between the observed and the simulated observations, to draw samples of size $m = 10,000$ for each $\varphi_i(\theta)$. Unlike the binomial, CIR and INAR examples where drawing posterior samples for the $\varphi_i(\theta)$, $i = 1, \ldots, n$ assuming $\epsilon = 0$ took a total of approximately, 1, 2 and 20 minutes respectively on a standard desktop machine, for this example doing so was computationally more demanding. However, since sampling in PW-ABC is embarrassingly parallel (see Sect. 3.1) we were able to draw the required samples in 32 hours on a 48 core machine.

To obtain pMCMC results we found it necessary to assume an error model for the observations, hence we assumed errors to be IID Gaussian with mean zero and standard deviation equal to 2. Results are displayed in Fig. 6, which shows plots for univariate and pairwise bivariate marginal posterior densities for the pMCMC results, and for the PW-ABC approximations, $\hat{\pi}^{g}(\theta \mid x)$ and $\hat{\pi}^{k}(\theta \mid x)$. Both of the PW-ABC approximations agree well with each other and with the pMCMC results for this example.

5 Conclusion and discussion

PW-ABC works by factorising the posterior density, for which targeting by ABC would entail a careful choice of $s(\cdot)$

Fig. 6 Results for the
Lotka–Volterra example of
Sect. 4.4, showing univariate
and bivariate marginal posterior
densities of θ based on a
posterior sample from a
pMCMC algorithm, and from
the Gaussian- and kernel-based
PW-ABC approximations,
$\hat{\pi}^{g}(\theta \mid x)$ and $\hat{\pi}^{k}(\theta \mid x)$. For the
kernel approximation we used
$q = 5$ as the smoothing
parameter in (13). The contours
shown in the bivariate plots are
those that contain 5 %, 10 %,
50 %, 90 % and 95 % of
probability mass
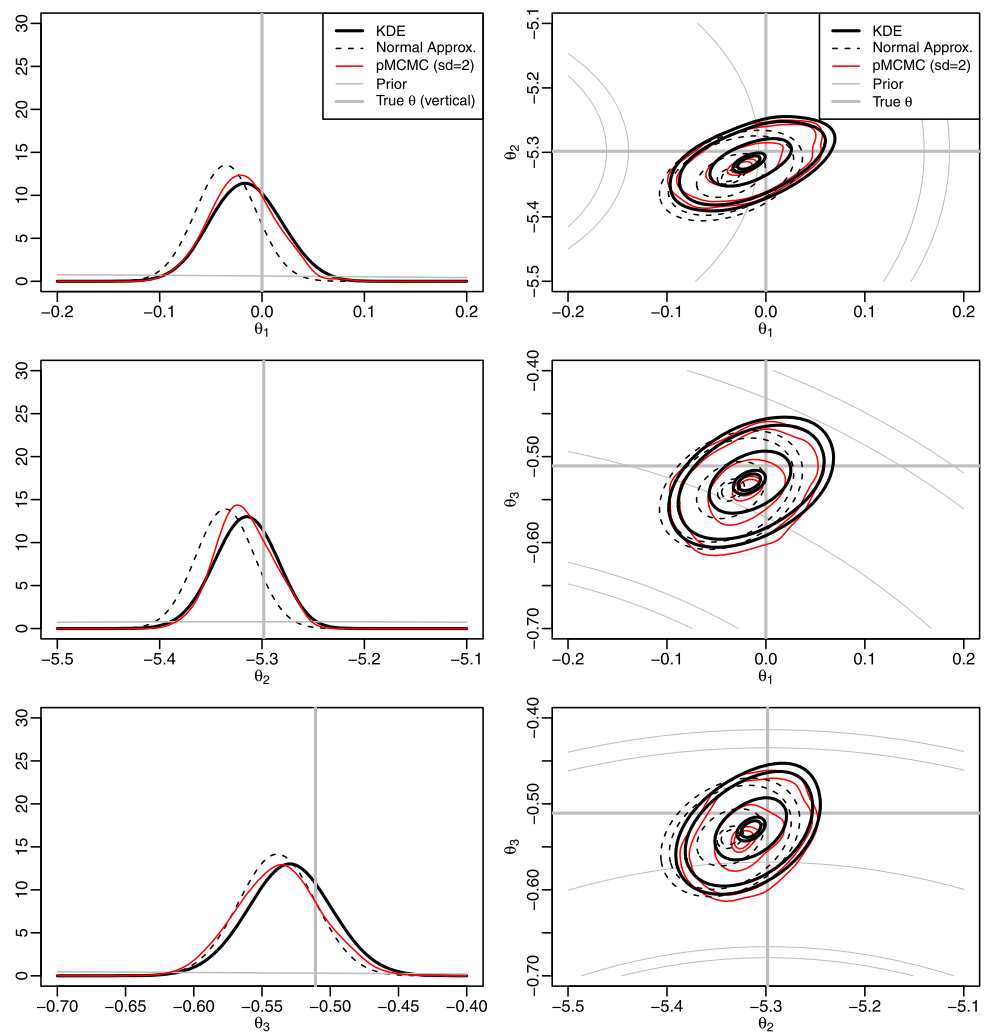
and/or a large tolerance ε, into a product involving densities $\varphi_i(\theta)$, each amenable to using ABC with $s = \text{Identity}(\cdot)$ and small or zero ε. Having sampled from each $\varphi_i(\theta)$ the question then becomes how to estimate $\pi(\theta \mid x)$ using these samples. In PW-ABC, we construct density estimates $\hat{\varphi}_i(\theta)$ of each $\varphi_i(\theta)$ then approximate $\pi(\theta \mid x)$ as the product of the $\hat{\varphi}_i(\theta)$. The approach of taking $\hat{\varphi}_i(\theta)$ to be Gaussian, with moments matched to the sample moments, is computationally cheap, and if the prior is also taken to be Gaussian then there is a closed form expression for the Gaussian posterior density and marginal likelihood, making calculations extremely fast. Taking $\hat{\varphi}_i(\theta)$ to be Gaussian is perhaps adequate in many applications: performance was strong in two of the four examples we considered. The poor performance in the CIR and INAR examples was due to skewness of at least some of the $\varphi_i(\theta)$. In the INAR example it is striking to see an effect so strong when the true posterior, and many of the $\varphi_i(\theta)$, are so close to Gaussian. Unfortunately, increasing the number, m, of ABC samples is no remedy to this problem: as $m \to \infty$, the normalised product of Gaussian

densities, itself Gaussian, in general does not converge to the Gaussian density closest in the Kullback–Leibler sense to the target density.

Two referees suggested the possibility of testing, across all of the $\varphi_i(\theta)$, whether a Gaussian approximation is appropriate. A wide literature exists on testing multivariate normality (see Székely and Rizzo 2005 for a recent contribution, plus many references therein to earlier work) and this seems a promising direction, but further work is needed to devise, and understand the properties of, a procedure based on applying these tests in the multi-testing setting of PW-ABC.

In terms of asymptotic performance, using the kernel approximation, $\hat{\varphi}_i^{k}(\theta)$, for $\hat{\varphi}_i(\theta)$ is preferable since, in this case, the estimated posterior density converges to the target as $m \to \infty$. The kernel approach is computationally more demanding, however, and its practical use is probably limited to problems in which θ has small dimension. It also requires a heuristic choice of a scalar smoothing parameter. The larger the value chosen for the smoothing parameter, the

more the posterior variance will be inflated; this said, however, in the examples we have considered we have found posterior inference to be fairly robust to the choice. A referee asked for guidance on how to choose m. It is difficult to offer general practical advice, because the m needed will depend on the dimension of θ, and on the number and nature of the $\varphi_i(\theta)$. The larger the better, of course; one possibility for checking whether m is large enough might be to use a resampling approach to confirm that the variance, under resampling, of the $\hat{\pi}^k(\theta \mid x)$ is acceptably small.

Another related practical question is how to choose ε if $\varepsilon = 0$ is not possible. In such a case, as with standard ABC approaches, there is a trade-off between making m large and making ε small. A reasonable heuristic to investigate the effects of non-zero ε would be to perform inference with a chosen ε and m, and then to keep m fixed and reduce ε (as discussed in Sect. 3.1, acceptable samples from the run with larger ε can be retained), and then check whether there is a marked difference in the posteriors for the different values of ε. Figure 2(b) shows for the CIR example, for instance, that there would be little difference between the posteriors for $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$. Such an approach could be applied iteratively, although for challenging problems—even using PW-ABC—the computational cost to maintain m samples as ε is decreased may prevent reaching a small enough ε that the posterior has "converged" to the true. Such a heuristic could be applied to standard ABC, of course, although PW-ABC has the advantage of enabling much smaller choices of ε.

The underlying idea in PW-ABC of replacing a high-dimensional ABC problem with multiple low-dimensional ones is also exploited in some sequential ABC algorithms; for example, Algorithm 4 in Fearnhead and Prangle (2012) (adapted from an algorithm by Wilkinson 2011b) uses ABC to incorporate observations from a Markov model sequentially, the ABC at each step involving a single data point conditional on the previous one, and where the posterior from one step is used as the prior for the next. In comparison with PW-ABC, such sequential algorithms have a potential advantage of progressively focusing computational effort on regions of parameter space with high posterior density, but on the other hand they are prone to problems with particle degeneracy, an issue that does not affect PW-ABC. Another major difference is that for sequential algorithms, samples at each step are dependent, so calculations are not "embarrasingly parallel", and nor is it so easy to reuse samples in the event of a retrospective decision to use a smaller ε or larger m; see Sect. 3.1.

A possibility that generalises the Gaussian and kernel approaches in PW-ABC, which we will explore in future work, is to let $\hat{\varphi}_i(\theta)$ be a mixture of, say, u Gaussians (see Fan et al. 2012 for an example of Gaussian mixtures being used in a related context). This encompasses (6) and (12) as special cases, with $u = 1$ and $u = m$ respectively. For a general

mixture model for $\hat{\varphi}_i(\theta)$, each of the component Gaussians is parameterised by a scalar weight, a mean vector and a covariance matrix which need to be determined. We would envisage regularising, e.g., by setting each covariance to be equal up to scalar multiplication, perhaps as for (12) taking the covariance proportional to the sample covariance, and then fitting each $\hat{\varphi}_i(\theta)$ based on the samples from $\varphi_i(\theta)$ using, say, an EM algorithm. This approach is a compromise between (6) and (12). It does not share the property of (12) that estimated densities converge to the true densities as $m \to \infty$, but on the other hand it is computationally much less involved and offers much extra freedom and flexibility over (6), particularly for dealing with multimodal densities. If u is taken sufficiently small then it may be feasible to work explicitly with the $(n-1)^u$-term resulting Gaussian mixture, $\prod \hat{\varphi}_i(\theta)$, enabling explicit calculations involving the posterior density, such as computing the marginal likelihood, analogous to (32), and direct sampling from the approximate posterior density (see Sect. 3.4).

Several further generalisations of PW-ABC are possible. In (1), each of the $n - 1$ factors $\pi(x_i \mid x_{i-1}, \theta)$, $i = 2, \ldots, n$ is the likelihood for a single data point conditional on the previous. An alternative possibility is to factorise the likelihood into fewer factors, with each corresponding to a "block" of multiple observations, e.g., $\pi(x_{i+v_i}, x_{i+v_i-1}, \ldots, x_i \mid x_{i-1}, \theta)$ for some choice of v_i, and the factorised likelihood becomes a product over the relevant subset of $i = 2, \ldots, n$. To an extent this potentially reintroduces difficulties that with PW-ABC we sought to avoid, namely lower acceptance rates leading to a possible need to use a summary statistic and non-zero tolerance (and the ensuing ABC error they bring). On the other hand, we might expect, owing to the central limit theorem, that a factor $\varphi_i(\theta)$ which depends on several data points will be closer to Gaussian than a factor dependent on only a single data point, and hence that (6) and (12) (especially the former) will perform better.

If using larger "blocks" of data in the factorisation makes it necessary to use a non-zero tolerance $\varepsilon > 0$ (or if $\varepsilon > 0$ is necessary even when using a single observation per factor) then there are theoretical advantages to using what Fearnhead and Prangle (2012) call "noisy ABC". In the context of this paper, noisy ABC would involve replacing the summary statistic $s(\cdot)$ with a random variable $s'(\cdot)$ which has density uniform on a ball of radius ε around $s(\cdot)$. Using noisy ABC ensures that, under mild regularity conditions, as $n \to \infty$, the posterior converges to a point mass at the true parameter value; see Sect. 2.2 of Fearnhead and Prangle (2012).

Recently, we have learnt of an interesting paper by Barthelmé and Chopin (2011) who have developed an approach termed *Expectation Propagation-ABC* (EP-ABC) that shares similarities with ours. EP-ABC is an ABC adaptation of the Expectation Propagation approach developed

by Minka (2001). EP-ABC uses a factorisation of the posterior (Eq. (1.2) in Barthelmé and Chopin 2011) analogous to our factorisation (2), and it involves a Gaussian approximation to the density of each factor analogous to (6). But then EP-ABC proceeds rather differently: instead of drawing ABC samples for, say, the ith factor by sampling from the prior, EP-ABC draws samples from an iteratively updated pseudo-prior. The pseudo-prior is a Gaussian approximation to the component of the posterior that involves all the data *except* those pertaining to the ith factor. The use of the pseudo-prior offers a high acceptance rate in the ABC sampling and so EP-ABC can potentially lead to an extremely fast approximation to the full posterior $\pi(\theta \mid x)$. A disadvantage is that conditions sufficient for the convergence of EP-ABC (or even the simpler deterministic EP) are not known. Also, as with using PW-ABC with (7), since EP-ABC uses a Gaussian approximation for each factor, it is potentially ill-suited to problems with complicated (e.g. multimodal or otherwise non-Gaussian) likelihoods; convergence of the product density is not assured to any "optimal" approximation to the target posterior. A promising direction for future work will be to investigate adapting the EP-ABC idea of sampling from a pseudo-prior to the ideas in this paper of using kernel (or Gaussian mixture) density estimates for each likelihood factor.

Appendix 1

Expression for $B_{j_2,\ldots,j_n}$, $a_{j_2,\ldots,j_n}$, and $w_{j_2,\ldots,j_n}$ in (18), analogous to (8)–(10), are as follows:

$$B_{j_2,\ldots,j_n} = \left(\sum_{i=2}^{n} H_i^{-1} \right)^{-1},$$

$$a_{j_2,\ldots,j_n} = B_{j_2,\ldots,j_n} \left(\sum_{i=2}^{n} H_i^{-1} \theta_{i(j_i)}^* \right),$$

$$w_{j_2,\ldots,j_n}$$
$$= m^{(1-n)} \det(2\pi B_{j_2,\ldots,j_n})^{1/2} \prod_{i=2}^{n} \det(2\pi H_i)^{-1/2}$$

$$\times \prod_{s=2}^{n} \prod_{t>s}^{n} \exp\left(-\frac{1}{2} \left(\theta_{s(j_s)}^* - \theta_{t(j_t)}^* \right)^T R_{st} \left(\theta_{s(j_s)}^* - \theta_{t(j_t)}^* \right) \right),$$

$$R_{st} = H_s^{-1} B_{j_2,\ldots,j_n} H_t^{-1}.$$

Expressions for $B'_{j_2,\ldots,j_n}$, $a'_{j_2,\ldots,j_n}$, and $w'_{j_2,\ldots,j_n}$ in (22) are given respectively by the right-hand sides of (20), (21), and (32) with B replaced by $B_{j_2,\ldots,j_n}$, a replaced by $a_{j_2,\ldots,j_n}$, and w replaced by $w_{j_2,\ldots,j_n}$.

Appendix 2

Proposition 1 *Let* $I = \mathbb{1}\{\theta^* \text{ is accepted}\}$ *be the indicator function of whether an ABC draw* θ^* *is accepted. The acceptance probability is*

$$\mathbb{P}(I = 1) = V \tilde{\pi}_{\text{ABC}}(x)$$

where $\tilde{\pi}_{\text{ABC}}(x)$ *is the marginal likelihood of the implied ABC posterior.*

Proof Recall from Sect. 2.4 that $K_{\varepsilon,p}(z) = V^{-1} \mathbb{1}\{\|z\|_p \leq \varepsilon\}$ and $\tilde{\pi}_{\text{ABC}}(x \mid \theta) = \int \pi(y \mid \theta) \times K_{\varepsilon,p}(y - x) \mathrm{d}y$ is the implied ABC likelihood approximation. Then

$$\mathbb{P}(I = 1) = \int_{\theta} \mathbb{P}(I = 1, \theta) \mathrm{d}\theta$$
$$= \int_{\theta} \pi(\theta) \mathbb{P}(I = 1 \mid \theta) \mathrm{d}\theta$$
$$= \int_{\theta} \pi(\theta) \left\{ \int_y \pi(y \mid \theta) \mathbb{1}\{\|y - x\|_p \leq \varepsilon\} \mathrm{d}y \right\} \mathrm{d}\theta$$
$$= \int_{\theta} \pi(\theta) \left\{ \int_y \pi(y \mid \theta) V K_{\varepsilon,p}(y - x) \mathrm{d}y \right\} \mathrm{d}\theta$$
$$= \int_{\theta} V \pi(\theta) \tilde{\pi}_{ABC}(x \mid \theta) \mathrm{d}\theta$$
$$= V \tilde{\pi}_{\text{ABC}}(x). \qquad \square$$

An estimator of $\tilde{\pi}_{\text{ABC}}(x)$ is hence $V^{-1} \hat{\mathbb{P}}(I = 1)$, where $\hat{\mathbb{P}}(I = 1)$ is the empirical proportion of ABC draws which are accepted.

References

Aït-Sahalia, Y.: Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. Econometrica **70**, 223–262 (2002)

Al-Osh, M.A., Alzaid, A.A.: First-order integer-valued autoregressive (INAR(1)) process. J. Time Ser. Anal. **8**(3), 261–275 (1987)

Andrieu, C., Doucet, A., Holenstein, R.: Particle Markov chain Monte Carlo methods. J. R. Stat. Soc., Ser. B, Stat. Methodol. **72**(3), 269–342 (2010)

Barthelmé, S., Chopin, N.: Expectation-Propagation for Summary-Less, Likelihood-Free Inference. J. Acoust. Soc. Am. ArXiv e-prints (2011, to appear)

Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate Bayesian Computation in population genetics. Genetics **162**(4), 2025–2035 (2002)

Blum, M.G.B., François, O.: Non-linear regression models for approximate Bayesian computation. Stat. Comput. **20**, 63–73 (2010)

Boys, R.J., Wilkinson, D.J., Kirkwood, T.B.: Bayesian inference for a discretely observed stochastic kinetic model. Stat. Comput. **18**(2), 125–135 (2008)

Cox, J.C., Ingersoll, J.E., Ross, S.A.: A theory of the term structure of interest rates. Econometrica **53**(2), 385–407 (1985)

Dean, T., Singh, S.: Asymptotic behaviour of approximate Bayesian estimators. Technical report, University of Cambridge (2011)

Dean, T.A., Singh, S.S., Jasra, A., Peters, G.W.: Parameter Estimation for Hidden Markov Models with Intractable Likelihoods. ArXiv e-prints (2011)

Durham, G.B., Gallant, A.R.: Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. J. Bus. Econ. Stat. **20**, 297–338 (2002)

Fan, Y., Nott, D.J., Sisson, S.A.: Approximate Bayesian computation via regression density estimation. Technical report (2012). arXiv:1212.1479

Fearnhead, P., Prangle, D.: Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. RSS Series B (2012). doi:10.1111/j.1467-9868.2011.01010.x

Fukunaga, K.: Introduction to Statistical Pattern Recognition. Electrical Science Series. Academic Press, San Diego (1972)

Gabriel, E., Wilson, D.J., Leatherbarrow, A.J., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Fearnhead, P., Hart, C.A., Diggle, P.J.: Spatio-temporal epidemiology of campylobacter jejuni enteritis, in an area of northwest England, 2000–2002. Epidemiol. Infect., **138**, 1384–1390 (2010)

Golightly, A., Wilkinson, D.J.: Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. Interface Focus **1**(6), 807–820 (2011)

Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Probability and Mathematical Statistics: a Series of Monographs and Textbooks. Academic Press, London (1979). ISBN 0-12-471250-9

Marin, J.M., Pudlo, P., Robert, C., Ryder, R.: Approximate Bayesian computational methods. Stat. Comput. **22**(5), 1009–1020 (2012)

Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S.: Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **100**(26), 15324 (2003)

McKenzie, E.: Discrete variate time series. In: Stochastic Processes: Modelling and Simulation. Handbook of Statist, vol. 21, pp. 573–606. North-Holland, Amsterdam (2003)

McKinley, T., Cook, A., Deardon, R.: Inference in epidemic models without likelihoods. Int. J. Biostat. **5**, 24 (2009)

Minka, T.P.: Expectation propagation for approximate Bayesian inference. In: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, UAI'01, pp. 362–369. Morgan Kaufmann, San Francisco (2001). ISBN 1-55860-800-1. http://dl.acm.org/citation.cfm?id=647235.720257

Moriña, D., Puig, P., Ríos, J., Vilella, A., Trilla, A.: A statistical model for hospital admissions caused by seasonal diseases. Stat. Med. **30**(26), 3125–3136 (2011)

Neal, P., Subba Rao, T.: MCMC for integer-valued ARMA processes. J. Time Ser. Anal. **28**(1), 92–110 (2007)

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W.: Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16**(12), 1791–1798 (1999)

Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall, London (1986)

Sudderth, E.B., Ihler, A.T., Freeman, W.T., Willsky, A.S.: Nonparametric belief propagation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. **1**, 605 (2003)

Székely, G.J., Rizzo, M.L.: A new test for multivariate normality. J. Multivar. Anal. **93**(1), 58–80 (2005)

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface **6**, 187–202 (2009)

Van Kampen, N.G.: Stochastic Processes in Physics and Chemistry. North-Holland Personal Library, Amsterdam (2007)

Wand, P., Jones, C.: Kernel Smoothing. Monographs on Statistics and Applied Probability. Chapman & Hall, London (1995)

Wilkinson, D.J.: Stochastic Modelling for Systems Biology, 2nd edn. Chapman & Hall, London (2011a)

Wilkinson, D.J.: Parameter inference for stochastic kinetic models of bacterial gene regulation: a Bayesian approach to systems biology. Bayesian Stat. **9**, 679–690 (2011b)

Wilkinson, R.D.: Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Stat. Appl. Genet. Mol. Biol. (2013). Available online as arXiv:0811.3355. doi:10.1515/sagmb-2013-0010