

Model Selection for Simulator-based Statistical Models: A Kernel Approach

Takafumi Kajihara*^{1,2,5}, Motonobu Kanagawa*^{3,2}, Yuuki Nakaguchi¹, Kanishka Khandelwal¹, and Kenji Fukumizu^{4,2}

¹*NEC Corporation*

²*RIKEN AIP*

³*University of Tübingen and Max Planck Institute for Intelligent Systems*

⁴*The Institute of Statistical Mathematics*

⁵*National Institute of Advanced Industrial Science and Technology*

Abstract

We propose a novel approach to model selection for simulator-based statistical models. The proposed approach defines a mixture of candidate models, and then iteratively updates the weight coefficients for those models as well as the parameters in each model simultaneously; this is done by recursively applying Bayes' rule, using the recently proposed kernel recursive ABC algorithm. The practical advantage of the method is that it can be used even when a modeler lacks appropriate prior knowledge about the parameters in each model. We demonstrate the effectiveness of the proposed approach with a number of experiments, including model selection for dynamical systems in ecology and epidemiology.

1 Introduction

Computer simulation is nowadays a ubiquitous tool in modern scientific and engineering problems to study time-evolving complex phenomena, such as climate modeling, social science, ecology and epidemiology (Winsberg, 2010; Weisberg, 2012). In the language of statistics, computer simulation can be defined as sampling from a probabilistic model that describes the phenomena of interest. One practically important issue regarding simulation is how to select an appropriate model, if there exist multiple candidate models. This problem of *model selection* is challenging in various ways, in the context of computer simulation. From the viewpoint of statistics, a major obstacle comes from the incapability of using a likelihood function; this is because the conditional probability $P_m(y|\theta^m)$ of data y given model parameters θ^m cannot be calculated in general, as the data-generating process is often very complicated (where we use $m = 1, \dots, K$ for an indicator of one of K candidate models.) For

* Equal contribution. Correspondence to: Takafumi Kajihara <t-kajihara@ct.jp.nec.com>, Motonobu Kanagawa <motonobu.kanagawa@tuebingen.mpg.de>

instance, this is the case where simulation involves numerically solving differential equations describing the dynamical system of interest; see Sec. 4 for examples in ecology and epidemiology.

Approximate Bayesian Computation (ABC) is a simulator-based approach to Bayesian inference, which is applicable as long as sampling from a probabilistic model is possible Tavaré et al. (1997); Pritchard et al. (1999); Beaumont et al. (2002); Sisson et al. (2018). ABC has been used in model selection for simulator-based statistical models (Grelaud et al. 2009; Toni and Stumpf 2009; Toni et al. 2009; Pudlo et al. 2015; Sisson et al. 2018, Chapter 6); see Supplements for a brief review. Being a Bayesian approach, these ABC-based methods select a model by essentially comparing the *marginal likelihood* (or the *evidence*) $\int P_m(y|\theta^m)\pi_m(\theta^m)d\theta^m$ for each of the candidate models, provided a prior distribution $\pi_m(\theta^m)$ of the model parameters θ^m . Note that in the Bayesian approach, a “model” m is given by the tuple (P_m, π_m) including the prior $\pi_m(\theta)$, i.e., the hierarchical model of first generating parameters $\theta^m \sim \pi_m(\theta)$ and then data $y \sim P_m(y|\theta^m)$; the marginal likelihood $\int P_m(y|\theta^m)d\pi_m(\theta^m)$ represents how likely the tuple (P_m, π_m) is true.

The Bayesian approach, however, may not always be the best way in practice, if a modeler lacks appropriate prior knowledge about parameters in a model. For instance, consider a situation where a modeler only has vague prior knowledge about a certain parameter θ^m that it should be positive and “not very large,” say it cannot be larger than 1000, and suppose that she sets her prior $\pi_m(\theta^m)$ as the uniform distribution over $[0, 1000]$. But why should it be a uniform distribution? The choice of the uniform prior just reflects her lack of knowledge, and she might have chosen it because there was no better alternative. Therefore, the form of the prior $\pi_m(\theta^m)$ in this case is not fully informative, and thus it may not be the best idea to define the “model” as the tuple (P_m, π_m) and perform model selection by faithfully relying on the marginal likelihood $\int P_m(y|\theta^m)d\pi_m(\theta^m)$.

In this work, we rather focus on situations where a modeler is willing to define $P_m(y|\theta^m)$ as a “model” and is interested in selecting one model from candidates and estimating its parameters. For instance, this is the case where one is interested in predictions or forecasting of phenomena of interest; in such a case, a modeler is asked to select one model and its parameters that result in the best predictive performances. This is true in particular when simulations are computationally expensive, situations common when modeling complex phenomena (e.g., spatio-temporal modeling in climate science and multi-agent systems in social science). In such a case one can only afford to have one good model with tuned parameters. On the other hand, a (fully) Bayesian approach requires taking averages over candidate models and parameters therein, which is practically not possible.

In this paper, we propose a novel approach to model selection for simulation-based statistical models that simultaneously estimates model parameters. We propose to define a mixture of candidate models and associate each model with a weight coefficient, which intuitively represents how that model is likely to be true; those weights and the parameters in each model are simultaneously and iteratively updated, by recursively applying Bayes’ rule. This idea is implemented with the recently proposed kernel recursive ABC algorithm (Kajihara et al., 2018), a kernel-based method for point estimation with simulator-based statistical models.

This paper is organized as follows. We first explain the conceptual idea of our approach in Sec. 2, and then present a concrete algorithm based in Sec. 3. We report a number of experimental results to empirically assess the proposed approach in Sec. 4, including model selection for dynamical systems in ecology and epidemiology. A review of related approaches, details of experiments and additional results are provided in the Supplementary Materials.

2 Proposed approach: the conceptual idea

In this section we present the conceptual idea of our approach to model selection, based on which we propose a concrete algorithm in Sec. 3.

Assume that there are $K \geq 2$ candidate models. Each of the models $m \in \{1, \dots, K\}$ has a conditional distribution of observed data $y \in \mathcal{Y}$ given model parameters $\theta^m \in \Theta^m$, denoted by $P_m(y|\theta^m)$, where \mathcal{Y} and Θ^m are measurable spaces. We also assume that for each model m there is a prior distribution $\pi_m(\theta^m)$ for the model parameters. Suppose we are given observed data $y^* \in \mathcal{Y}$.

2.1 Mixture of candidate models

Our idea is to combine multiple models to form a mixture model. The proposed method then estimates the mixing coefficients, which represent the ‘‘probabilities’’ of the models being true, as well as the parameters in the models simultaneously.

Let $\phi^1, \dots, \phi^K \geq 0$ be mixing coefficients such that $\sum_{m=1}^K \phi^m = 1$ and that each coefficient ϕ_m is associated with the model $P_m(y|\theta^m)$ of the same model index $m = 1, \dots, K$. We introduce the notation Ψ to write all the mixing coefficients and the associated model parameters:

$$\Psi := ((\phi^m)_{m=1}^K, (\theta^m)_{m=1}^K) \in S_{K-1} \times \Theta_1 \times \dots \times \Theta_K,$$

where $S_{K-1} := \{(\phi^1, \dots, \phi^K) \in \mathbb{R}^K : \phi_1, \dots, \phi_K \geq 0, \sum_{m=1}^K \phi^m = 1.\}$ is the $K - 1$ simplex. That is, Ψ is defined on the product space of S_{K-1} and all the parameter spaces Φ_1, \dots, Φ_K . We then define a mixture of the candidate models as

$$P(y|\Psi) := \sum_{m=1}^K \phi^m P_m(y|\theta^m). \quad (1)$$

In other words, to generate y from this model, we first sample a model index m from the multinomial distribution with coefficients ϕ^1, \dots, ϕ^K , and then sample y from $P_m(y|\theta^m)$. For simplicity, we assume each model $P_m(y|\theta^m)$ has a density function $p_m(y|\Psi)$, and we denote by $p(y|\Psi)$ the resulting density function of (1). Note that in the setting of this paper $p(y^*|\Psi)$ cannot be evaluated, but in the sequel assume that we can obtain the resulting posterior densities in order to explain the idea.

2.2 Recursive Bayes update for the mixture

Our idea is to recursively apply Bayes’ rule to the same observed data y^* to obtain a sequence of posterior distributions of Ψ on the product space $S_{K-1} \times \Theta_1 \times \dots \times \Theta_K$. That is, we set the posterior distribution derived at one iteration as the prior for the next iteration.

For the first recursion $N = 1$, we define a prior distribution $\pi^1(\Psi)$ on $S_{K-1} \times \Theta_1 \times \dots \times \Theta_K$ as

$$\pi^1(\Psi) := \text{Dir}((\phi^m)_{m=1}^K | \alpha) \prod_{m=1}^K \pi_m(\theta^m), \quad (2)$$

where $\text{Dir}((\phi^m)_{m=1}^K | \alpha)$ denotes the Dirichlet distribution with concentration parameter $\alpha > 0$, and $\pi_m(\theta^m)$ is the prior of the parameter in each model. Then the posterior $p^1(\Psi|y^*)$ at the first recursion is given by Bayes’ formula

$$p^1(\Psi|y^*) \propto p(y^*|\Psi)\pi^1(\Psi).$$

For $N \geq 2$, assume now that the posterior $p^{N-1}(\Psi|y^*)$ at the $(N-1)$ -th recursion has already been obtained. Define a prior $\pi^N(\Psi)$ at the N -th recursion by this previous posterior:

$$\pi^N(\Psi) := p^{N-1}(\Psi|y^*). \quad (3)$$

Then the posterior distribution at the N -th recursion is again given by Bayes' rule

$$p^N(\Psi|y^*) \propto p(y^*|\Psi)\pi^N(\Psi). \quad (4)$$

Our motivation of considering the recursive Bayes updates may be understood by rewriting (4) as

$$p^N(\Psi|y^*) \propto (p(y^*|\Psi))^N \pi^1(\Psi).$$

This shows that, as $N \rightarrow \infty$, the powered likelihood term $(p(y^*|\Psi))^N$ gradually dominates the posterior $p^N(\Psi|y^*)$, and the effects of the prior $\pi^1(\Psi)$ gradually diminish. In fact, Lele et al. (2010, Corollary to Lemma A.2 in p.1624) implies that the power posterior $p^N(\Psi|y^*)$ degenerates to the Dirac distribution at the maximum likelihood point $\psi_{\max} := \operatorname{argmax}_{\Psi} p(y^*|\Psi)$ as $N \rightarrow \infty$, if it exists. Therefore, one may expect that the recursive Bayes updates provide a way of obtaining $\Psi = ((\phi^m)_{m=1}^K, (\theta^m)_{m=1}^K)$ such that the largest coefficient in $(\phi^m)_{m=1}^K$, say ϕ^{m^*} , indicates the ground-truth model and the associated parameters θ^{m^*} are those of maximum likelihood. This is an intuitive idea why we introduce the recursive Bayes updates; mathematically there is a subtle issue in this argument, which we will discuss in Sec. 3. We nevertheless provide one theoretical result below that indicates how the recursive Bayes updates take model complexities into account.

Proposition 1. *Let $p^N((\phi^m)_{m=1}^K|y^*)$ and $\pi^N((\phi^m)_{m=1}^K)$ be the marginal densities of $(\phi^m)_{m=1}^K$ induced from the posterior (4) and the prior (3) (or (2) if $N = 1$), respectively. Then for $N \geq 2$, we have*

$$p^N((\phi^m)_{m=1}^K|y^*) = C_N^{-1} \pi^N((\phi^m)_{m=1}^K) \times \sum_{m=1}^K \phi^m \int p_m(y^*|\theta^m) \pi^N(\theta^m | (\phi^i)_{i=1}^K) d\theta^m, \quad (5)$$

where $C_N := \int p(y^*|\Psi)\pi^N(\Psi)d\Psi$ is a normalization constant, and $\pi^N(\theta^m | (\phi^i)_{i=1}^K)$ is the conditional density of θ^m given $(\phi^i)_{i=1}^K$ induced from the prior (3) if $N \geq 2$, and $\pi^N(\theta^m | (\phi^i)_{i=1}^K) = \pi_m(\theta^m)$ if $N = 1$.

In Proposition 1, the identity (5) describes how the posterior density of $(\phi^m)_{m=1}^K$ changes from $\pi^N((\phi^m)_{m=1}^K) = p^{N-1}((\phi^m)_{m=1}^K|y^*)$ to $p^N((\phi^m)_{m=1}^K|y^*)$ as a result of one application of Bayes' rule. The key insight is that the integral

$$\int p_m(y^*|\theta^m) \pi^N(\theta^m | (\phi^i)_{i=1}^K) d\theta^m \quad (6)$$

can be regarded as the *marginal likelihood or model evidence* for the model m , for which the prior of the model parameter θ^m is updated to $\pi^N(\theta^m | (\phi^i)_{i=1}^K) = p^{N-1}(\theta^m | (\phi^i)_{i=1}^K, y^*)$. In fact, for $N = 1$ this interpretation is more evident, since in this case the integral is $\int p_m(y^*|\theta^m) \pi_m(\theta^m) d\theta^m$, the marginal likelihood of the model m with the original prior $\pi_m(\theta^m)$.

To provide an intuition about how the update of the posterior of $(\phi^m)_{m=1}^K$ occurs, let us assume that the updated prior in (6) is independent of the coefficients $(\phi^i)_{i=1}^K$, i.e., $\pi^N(\theta^m | (\phi^i)_{i=1}^K) = \pi_m^N(\theta^m)$, where $\pi_m^N(\theta^m)$ is the marginal density of θ^m induced from (3). Then the integral (6) is $\int p_m(y^*|\theta^m) \pi_m^N(\theta^m) d\theta^m$, and the update (5) becomes

$$p^N((\phi^m)_{m=1}^K|y^*) = C_N^{-1} \pi^N((\phi^m)_{m=1}^K) \times \sum_{m=1}^K \phi^m \int p_m(y^*|\theta^m) \pi_m^N(\theta^m) d\theta^m.$$

Assume $K = 2$, and $\int p_1(y^*|\theta^1)\pi_1^N(\theta^1)d\theta^1 > \int p_2(y^*|\theta^2)\pi^N(\theta^2)d\theta^2$, i.e., the marginal likelihood of the model 1 is larger than that of the model 2. Let $(\tilde{\phi}^1, \tilde{\phi}^2) \in S_1$. If $\tilde{\phi}^1 > \phi^1$ (and thus $\tilde{\phi}^2 < \phi^2$), then we have $\sum_{m=1}^2 \tilde{\phi}^m \int p_m(y^*|\theta^m)\pi_m^N(\theta^m)d\theta^m > \sum_{m=1}^2 \phi^m \int p_m(y^*|\theta^m)\pi_m^N(\theta^m)d\theta^m$. Recalling the notation $\pi^N((\phi^1, \phi^2)) = p^{N-1}((\phi^1, \phi^2)|y^*)$, this implies that $p^N((\tilde{\phi}^1, \tilde{\phi}^2)|y^*)/p^{N-1}((\tilde{\phi}^1, \tilde{\phi}^2)|y^*) > p^N((\phi^1, \phi^2)|y^*)/p^{N-1}((\phi^1, \phi^2)|y^*)$. In other words, if the marginal likelihood of the model 1 is larger than that of the model 2 at the current recursion, then the density $p^N((\tilde{\phi}^1, \tilde{\phi}^2)|y^*)$ of a point in the simplex $(\tilde{\phi}^1, \tilde{\phi}^2) \in S_1$ with larger coefficient $\tilde{\phi}_1$ for the model 1 becomes relatively higher from the previous one $p^{N-1}((\tilde{\phi}^1, \tilde{\phi}^2)|y^*)$. In this way, the mass of the posterior $p^N((\phi^m)_{m=1}^K|y^*)$ may move towards the model with larger marginal likelihoods, as N increases.

Note however that, in the above discussion, the assumption $\pi^N(\theta^m|(\phi^i)_{i=1}^K) = \pi_m^N(\theta^m)$ is valid for the initial update (i.e., $N = 1$), but in general this may not hold. Therefore the argument is correct for the initial update, but for the case $N > 1$ its validity is unclear. We leave theoretical analysis of the assumption $\pi^N(\theta^m|(\phi^i)_{i=1}^K) = \pi_m^N(\theta^m)$ (or the properties of $\pi^N(\theta^m|(\phi^i)_{i=1}^K)$ for $N \geq 2$) for future research.

3 Proposed approach: the concrete algorithm

To develop a concrete algorithm based on the above idea, we employ the kernel recursive ABC (KR-ABC) (Kajihara et al., 2018), a kernel-based method for point estimation with intractable likelihood. This method performs point estimation by recursively applying Bayes' rule in a reproducing kernel Hilbert space (RKHS). More specifically, it iterates the two steps: 1) posterior computation with kernel ABC, and 2) deterministic sampling from the posterior via kernel herding. There are two main reasons why we adopt this approach: i) To the best of our knowledge, this is the only existing method for recursive Bayes updates that can be used with intractable likelihood; ii) This method has a special property of being robust to misspecification of a prior distribution, which is the desideratum in our context.

Kernel mean embedding. Before describing our method, we need to explain *kernel mean embedding of distributions* (Smola et al., 2007), a framework for representing probability distributions in an RKHS. For details of the following concepts, see Muandet et al. (2017). Let \mathcal{X} be a measurable space, \mathcal{P} be the set of all probability distributions on \mathcal{X} , $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a measurable, bounded, positive definite kernel and \mathcal{H} be its RKHS. In this framework, each distribution $P \in \mathcal{P}$ is represented as a Bochner integral

$$\mu_P := \int k(\cdot, x)dP(x) \in \mathcal{H}, \quad (7)$$

which is referred to as the *kernel mean* (or the *mean embedding*) of P . The mapping $P \in \mathcal{P} \rightarrow \mu_P \in \mathcal{H}$ is one-to-one, if \mathcal{H} is large enough; in such a case k is called *characteristic*, an example being the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/\gamma^2)$ on $\mathcal{X} \subset \mathbb{R}^d$ (Fukumizu et al., 2008; Sriperumbudur et al., 2010). Thus if k is characteristic, kernel mean μ_P is uniquely associated with P and thus maintains all of its information. This fact justifies estimation of μ_P in place of P , and estimation of μ_P is often easier than that of P itself, thanks to the reproducing property of the RKHS \mathcal{H} . For instance, given an i.i.d. sample X_1, \dots, X_n from P , μ_P is estimated by $\hat{\mu}_P := \frac{1}{n} \sum_{i=1}^n k(\cdot, X_i)$ with convergence rate $\mathbb{E}[\|\hat{\mu}_P - \mu_P\|_{\mathcal{H}}] = O(n^{-1/2})$, a rate independent of the dimensionality of \mathcal{X} (Tolstikhin et al., 2017).

We are now ready to explain the proposed algorithm, which is summarized in Algorithm 1. For the ease of explanation, we may call $\mathcal{X} := S_{K-1} \times \Theta^1 \times \dots \times \Theta^K$ the *state space*, and $\Psi = ((\phi^m)_{m=1}^K, (\theta^m)_{m=1}^K) \in \mathcal{X}$ a *state*. In the algorithm, $n \in \mathbb{N}$ denotes the number of simulations in each recursion, and $N_{\text{iter}} \in \mathbb{N}$ the total number of recursions, which are specified by a user.

Algorithm 1 Model selection via KR-ABC

- 1: **Input:** Observed data $y^* \in \mathcal{Y}$, models $(P_m(y|\theta^m))_{m=1}^K$, priors $(\pi_m(\theta^m))_{m=1}^K$.
 - 2: **Setting:** Dirichlet concentration parameter $\alpha > 0$, number N_{iter} of iterations, number n of simulations per iteration, kernel k on $\mathcal{X} = S_{K-1} \times \Theta^1 \times \cdots \times \Theta^K$, kernel $k_{\mathcal{Y}}$ on \mathcal{Y} , regularization constant $\delta > 0$.
 - 3: **for** $N = 1, \dots, N_{\text{iter}}$ **do**
 - 4: **if** $N = 1$ **then**
 - 5: **for** $i = 1, \dots, n$ **do**
 - 6: *Sampling:* $(\phi_{N,i}^m)_{m=1}^K \sim \text{Dir}((\phi_m)_{m=1}^K | \alpha)$.
 - 7: *Sampling:* $\theta_{N,i}^m \sim \pi_m(\theta^m)$, $m = 1, \dots, K$.
 - 8: **end for**
 - 9: **end if**
 - 10: **for** $i = 1, \dots, n$ **do**
 - 11: *Sampling:* $y_{N,i} \sim \sum_{m=1}^K \phi_{N,i}^m P_m(y|\theta_{N,i}^m)$.
 - 12: **end for**
 - 13: *Kernel Bayes:* compute $\mathbf{w} = (w_1, \dots, w_n)^T \in \mathbb{R}^n$ by Eq.(10) using $G := (k_{\mathcal{Y}}(y_{N,i}, y_{N,j}))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ and $\mathbf{k}(y^*) := (k_{\mathcal{Y}}(y_{N,i}, y^*))_{i=1}^n \in \mathbb{R}^n$.
 - 14: *Deterministic sampling:* generate $(\Psi_{N+1,i})_{i=1}^n = ((\phi_{N+1,i}^m)_{m=1}^K, (\theta_{N+1,i}^m)_{m=1}^K)_{i=1}^n$ by kernel herding Eqs.(11) (12) using kernel k , $(w_i)_{i=1}^n$ and $(\Psi_{N,i})_{i=1}^n = ((\phi_{N,i}^m)_{m=1}^K, (\theta_{N,i}^m)_{m=1}^K)_{i=1}^n$.
 - 15: **end for**
 - 16: **Output:** Final states $(\Psi_{N_{\text{iter}}+1,i})_{i=1}^n = ((\phi_{N_{\text{iter}}+1,i}^m)_{m=1}^K, (\theta_{N_{\text{iter}}+1,i}^m)_{m=1}^K)_{i=1}^n \subset \mathcal{X}$.
-

Spaces and kernels. In our algorithm there are two kernels. One kernel, denote by $k_{\mathcal{Y}}$, is defined on the space \mathcal{Y} of summary statistics of observed data. The other kernel, denoted by k , is defined on the product space $\mathcal{X} = S_{K-1} \times \Theta^1 \times \cdots \times \Theta^K$. In particular we define k as a product kernel composed of a kernel $k_{S_{k-1}}$ on S_{k-1} and kernels k_m respectively defined on Θ^m . More precisely, for $\Psi = ((\phi^m)_{m=1}^K, (\theta^m)_{m=1}^K)$, $\tilde{\Psi} = ((\tilde{\phi}^m)_{m=1}^K, (\tilde{\theta}^m)_{m=1}^K) \in S_{K-1} \times \Theta^1 \times \cdots \times \Theta^K$, the kernel is defined as $k(\Psi, \tilde{\Psi}) = k_{S_{k-1}}((\phi^m)_{m=1}^K, (\tilde{\phi}^m)_{m=1}^K) \prod_{m=1}^K k_m(\theta^m, \tilde{\theta}^m)$. In the sequel let \mathcal{H} be the RKHS of this kernel k . In our experiments in Sec. 4, we focus on the case where each of \mathcal{Y} and $\Theta^1, \dots, \Theta^K$ is a subset of a Euclidean space, and we define $k_{\mathcal{Y}}$ and k_1, \dots, k_K as Gaussian kernels. While there are kernels designed for simplex such as information-diffusion kernels (Lafferty and Lebanon, 2005), we also use a Gaussian kernel as $k_{S_{k-1}}$; in our preliminary experiments we found the use of the information diffusion kernels somehow did not lead to better performance.

Initial prior sampling (Lines 5-8): For the first recursion $N = 1$, the algorithm generates initial states $(\Psi_{N,i})_{i=1}^n = ((\phi_{N,i}^m)_{m=1}^K, (\theta_{N,i}^m)_{m=1}^K)$ by randomly sampling from the prior $\pi^1(\Psi)$ in (2); $(\Psi_{N,i})_{i=1}^n$ provides an empirical approximation of $\pi^1(\Psi)$.

1. Bayes' rule via Kernel ABC (Lines 10-13): The aim of this step is to estimate the kernel mean

$$\mu_{p_N(\Psi|y^*)} := \int k(\cdot, \Psi) p_N(\Psi|y^*) d\Psi \quad (8)$$

of the posterior $p^N(\Psi|y^*)$ (4). To this end, we first obtain n pairs of state and simulated data $(\Psi_{N,i}, y_{N,i})_{i=1}^n$ by simulating data $y_{N,i}$ from state $\Psi_{N,i}$ using the mixture model (1) for $i = 1, \dots, n$. (If $N \geq 2$, assume that states $(\Psi_{N,i})_{i=1}^n$ have already been generated at the Step 2 of deterministic posterior sampling in the $(N-1)$ -th recursion, explained below.) Since $(\Psi_{N,i})_{i=1}^n$ approximately follow the prior $\pi^N(\Psi)$ (3), $(\Psi_{N,i}, y_{N,i})_{i=1}^n$ approximately follow the joint distribution $p(y|\Psi)\pi^N(\Psi)$. Note that the

posterior (4) is the conditional distribution of Ψ given y^* induced from this joint distribution. Therefore by estimating the conditional distribution of Ψ given y^* using $(\Psi_{N,i}, y_{N,i})_{i=1}^n$, one can estimate the posterior (4); this is the idea of ABC.

Kernel ABC (Nakagome et al., 2013) instead estimates the kernel mean of this conditional distribution using $(\Psi_{N,i}, y_{N,i})_{i=1}^n$, and this results in the following estimator of the posterior kernel mean (8):

$$\hat{\mu}_{p_N(\Psi|y^*)} := \sum_{i=1}^n w_i k(\cdot, \Psi_{N,i}) \in \mathcal{H}, \quad (9)$$

$$\mathbf{w} := (w_1, \dots, w_n)^T := (G + n\delta I)^{-1} \mathbf{k}(y^*), \quad (10)$$

where $\mathbf{k}(y^*) := (k_{\mathcal{Y}}(y_{N,1}, y^*), \dots, k_{\mathcal{Y}}(y_{N,n}, y^*))^T \in \mathbb{R}^n$, $G := (k_{\mathcal{Y}}(y_{N,i}, y_{N,j}))_{i,j=1}^n \in \mathbb{R}^{n \times n}$, $\delta > 0$ is a regularization constant, and $I \in \mathbb{R}^{n \times n}$ is the identity. The consistency and convergence rates of (9) as $n \rightarrow \infty$, which require $\delta \rightarrow 0$ with appropriate speed, have been studied extensively in the literature; see Grünewälder et al. (2012); Fukumizu et al. (2013); Muandet et al. (2017).

2. Deterministic posterior sampling via kernel herding (Line 14): In this step, the algorithm generates new states $(\Psi_{N+1,i})_{i=1}^n$ from (9), so that $(\Psi_{N+1,i})_{i=1}^n$ approximately follow the posterior (4); these new states will be used in the next recursion to simulate observed data (Line 11). One approach for this purpose is *kernel herding* (Chen et al., 2010), a class of deterministic sampling method that is a greedy variant of Quasi Monte Carlo (Dick et al., 2013). Kajihara et al. (2018) showed that the use of kernel herding in the recursive Bayes updates leads to the robustness against a misspecified prior.

The concrete procedure, which is greedy, is as follows. The first state is given by

$$\Psi_{N+1,1} := \operatorname{argmax}_{\Psi \in \mathcal{X}} \sum_{i=1}^n w_i k(\Psi, \Psi_{N,i}), \quad (11)$$

where w_1, \dots, w_n are given by (10). Assume now that $\Psi_{N+1,1}, \Psi_{N+1,2}, \dots, \Psi_{N+1,t}$ for $1 \leq t < n$ have already been generated. Then the next state $\Psi_{N+1,t+1}$ is

$$\begin{aligned} \Psi_{N+1,t+1} := & \operatorname{argmax}_{\Psi \in \mathcal{X}} \sum_{i=1}^n w_i k(\Psi, \Psi_{N,i}) \\ & - \frac{1}{t+1} \sum_{i=1}^t k(\Psi, \Psi_{N+1,i}). \end{aligned} \quad (12)$$

In this way, one can obtain n states $(\Psi_{N+1,i})_{i=1}^n$. It can be shown that if k is shift-invariant (e.g., Gaussian), this procedure (11) (12) is equivalent to the greedy minimization of the RKHS distance $\left\| \hat{\mu}_{p_N(\Psi|y^*)} - \frac{1}{t} \sum_{i=1}^t k(\cdot, \Psi_{N+1,i}) \right\|_{\mathcal{H}}$ between $\hat{\mu}_{p_N(\Psi|y^*)}$ and the empirical average $\frac{1}{t} \sum_{i=1}^t k(\cdot, \Psi_{N+1,i})$ for $t = 1, \dots, n$ (Chen et al., 2010). In terms of this distance the convergence rate $O(n^{-1/2})$ is guaranteed for $\frac{1}{n} \sum_{i=1}^n k(\cdot, \Psi_{N+1,i})$ (Bach et al., 2012), which may hold even when the optimization problem (11) (12) are solved approximately (Lacoste-Julien et al., 2015; Kanagawa et al., 2016a). Since i) $\mu_{p_N(\Psi|y^*)}$ contains the all information of $p_N(\Psi|y^*)$, ii) $\hat{\mu}_{p_N(\Psi|y^*)}$ is a consistent estimator of $\mu_{p_N(\Psi|y^*)}$, and iii) $\frac{1}{n} \sum_{i=1}^n k(\cdot, \Psi_{N+1,i})$ is a convergent approximation of $\hat{\mu}_{p_N(\Psi|y^*)}$, we can say that the states $(\Psi_{N+1,i})_{i=1}^n$ approximately follow the posterior $p_N(\Psi|y^*)$; this is in the sense that the empirical average $\frac{1}{n} \sum_{i=1}^n f(\Psi_{N+1,i})$ converges to the expectation $\int f(\Psi) p_N(\Psi|y^*) d\Psi$ as $n \rightarrow \infty$ for all $f \in \mathcal{H}$ or for any f that can be approximated well by functions in \mathcal{H} (Kanagawa et al., 2016b).

Output (Line 16): After N_{iter} recursions, the algorithm yields the states $(\Psi_{N_{\text{iter}}+1,i})_{i=1}^n = ((\phi_{N_{\text{iter}}+1,i}^m)_{m=1}^K, (\theta_{N_{\text{iter}}+1,i}^m)_{m=1}^K))_{i=1}^n \subset \mathcal{X}$, which are an approximate sample of the posterior $p^{N_{\text{iter}}}(\Psi|y^*)$.

While there can be various ways to exploit these states, in this paper we propose to use the first state $\Psi_{N_{\text{iter}}+1,1} = ((\phi_{N_{\text{iter}}+1,1}^m)_{m=1}^K, (\theta_{N_{\text{iter}}+1,1}^m)_{m=1}^K)$ for the purpose of model selection and parameter estimation. This is motivated from Kajihara et al. (2018, Prop. 2), which shows that under certain assumptions the first state of kernel herding (11) converges to the mode of the posterior $p^{N_{\text{iter}}}(\Psi|y^*)$ as N_{iter} increases. Therefore $\Psi_{N_{\text{iter}}+1,1}$ can be regarded as an approximation of the mode of $p^{N_{\text{iter}}}(\Psi|y^*)$.

For instance, in our experiments we select the model m with the largest coefficient $\phi_{N_{\text{iter}}+1,1}^m$ as a “true” model, and the associated $\theta_{N_{\text{iter}}+1,1}^m$ as an estimate of the true parameter. We note that the magnitudes of the coefficients $(\phi_{N_{\text{iter}}+1,1}^m)_{m=1}^K$ are informative: For instance, if one of the coefficients is close to 1 and the rest is nearly to 0, then we may interpret the algorithm being confident that the ground-truth is the model associated with that largest coefficient; if all the coefficients are close to $1/K$, the algorithm’s bet is neutral.

Discussion. While the proposed method is an application of KR-ABC (Kajihara et al., 2018) to the mixture model (1), the principle we rely on is different from that in Kajihara et al. (2018). In Kajihara et al. (2018), the goal is not model selection, but parameter estimation for a *given* parametric model with intractable likelihood. The principle they rely on is Lele et al. (2010, Corollary to Lemma A.2 in p.1624), which claims that, if the likelihood function has a *unique* global maximum and satisfies other regularity conditions, then the recursive Bayes updates result in posteriors converging to the Dirac distribution at the maximum likelihood parameter. While this assumption may be reasonable in the context of Kajihara et al. (2018), it may not hold in our case: For the mixture model (1), there is in general no unique global maximum. Therefore, we cannot use the result of Lele et al. (2010) to analyze the convergence behavior of the posterior (4) as N goes to infinity.

While currently we do not have a complete characterization of the convergence property of the posterior (4), we attribute the model selection mechanism of our method to Prop. 1: As argued in Sec. 2.2, this preliminary result implies that the recursive Bayes updates for the mixture model (1) may automatically take the marginal likelihoods of the candidate models into account. We leave further theoretical analysis for future work, and empirically investigate the properties of the proposed method by experiments in Sec. 4.

4 Experiments

We report results of experiments conducted for empirically assessing the proposed approach. We first explain the setting common to all the experiments in Sec. 4.1. We then report results on model selection for polynomial regression models in Sec. 4.2, followed by results on dynamical systems in ecology (Sec. 4.3) and epidemiology (Sec. 4.4).

4.1 Experimental settings

Competing methods. We made comparisons with the following ABC-based methods: *ABC for Model Choice* (ABC-MC) (Grelaud et al., 2009), *ABC Sequential Monte Carlo SMC* (ABC-SMC) (Toni et al., 2009) and *ABC Random Forests* (ABC-RF) (Pudlo et al., 2015); see the Supplements for a review of these methods.

Hyper-parameters. For each method, we determined its hyper-parameters by adopting a cross-validation-like approach described in Park et al. (2016, Sec. 4), unless otherwise stated; see the Supplements for details. We set the Dirichlet concentration prior to be 0.01 in the proposed method for all the experiments in the main body.

Parameter estimation. For comparison, we also performed parameter estimation with the competing methods based on selected models; see the Supplements for details.

Evaluation measure. We performed experiments by varying the total number of simulations; the proposed method and ABC-SMC used 100 simulations for each iteration in all the experiments. To evaluate each method, we computed the average number of mistakes made in model selection, where the average is taken over 30 independent trials of an experiment; we refer to this as the *model error*. To further evaluate the quality in parameter estimation, we calculated the *data error* defined as follows: for each of 30 trials, using the estimated parameters in a selected model we simulated pseudo data and measure the Euclidean distance to the observed data; then the data error is the average over these Euclidean distances over 30 trials.

4.2 Polynomial regression models

We performed simple experiments using polynomial regression models. We first investigated the effects of the Dirichlet concentration parameter to the behavior of the mixing coefficients. The results suggest that the performance of the proposed method is stable w.r.t. the choice of the concentration parameter; see the Supplements for details.

Here we investigate the robustness of the proposed approach against misspecification of prior distributions of model parameters. This is motivated from Kajihara et al. (2018, Sec. 3), who argue the use of kernel herding leads to an auto-correction mechanism that makes the method robust against misspecification of priors.

Let $x_1, \dots, x_{25} \in [0, 5]$ be equally-spaced points such that $-1 = x_1 < x_2 < \dots < x_{25} = 5$. We consider two models: a 3rd polynomial model $y_i = \sum_{\ell=0}^3 a_\ell x_i + 3\varepsilon_i$ and a 4th order model $y_i = \sum_{\ell=0}^4 b_\ell x_i + 3\varepsilon_i$ ($i = 1, \dots, 25$), where $\varepsilon_1, \dots, \varepsilon_{25}$ are independent standard Gaussian noises and $(a_\ell)_{\ell=0}^4, (b_\ell)_{\ell=0}^4 \subset \mathbb{R}$ are the parameters in the models. We generated observed data $y^* := (y_1^*, \dots, y_{25}^*) \in \mathbb{R}^{25}$ from either of the two models, with parameters $a_0 = \dots = a_3 = 40$ or $b_0 = \dots = b_4 = 40$. We consider two cases where the prior distributions are (i) appropriate or (ii) misspecified. Specifically, for each of parameters $(a_\ell)_{\ell=0}^3$ and $(b_\ell)_{\ell=1}^4$, we set a prior distribution as the uniform distribution over $[30, 50]$ for (i) the appropriate case, and the uniform distribution over $[0, 30]$ for (ii) the misspecified case.

Figures 1 and 2 show the results for (i) the appropriate case, while Figures 3 and 4 show those for (ii) the misspecified case. For the appropriate case, all the methods correctly identified the true model. On the other hand, for misspecified case, only the proposed method was able to correctly identify the true model, for both cases where the ground truth was the 3rd order or 4th order. The propose method also outperformed the other methods in terms of data errors for all the experiments. This is the same for errors in extrapolation; see the Supplements.

4.3 Predator-prey models in ecology

We conducted experiments on dynamical systems for predator-prey models studied in ecology; this is a typical example of intractable likelihood models. Assume that there are two species, one being predators and the other being prey, and let $x(t) \geq 0$ and $y(t) \geq 0$ respectively represent the populations of prey and predators at time $t \in [0, T]$, where we set $T = 20$. In this section and Sec. 4.4, we use “dot” to denote derivatives w.r.t. t . (e.g., $\dot{x}(t)$ is the derivative of $x(t)$ w.r.t. t .) We consider the following two models defined by ordinal differential equations (ODE).

Model 1: The most fundamental model is the Lotka-Volterra equations (Volterra, 1926; Lotka,

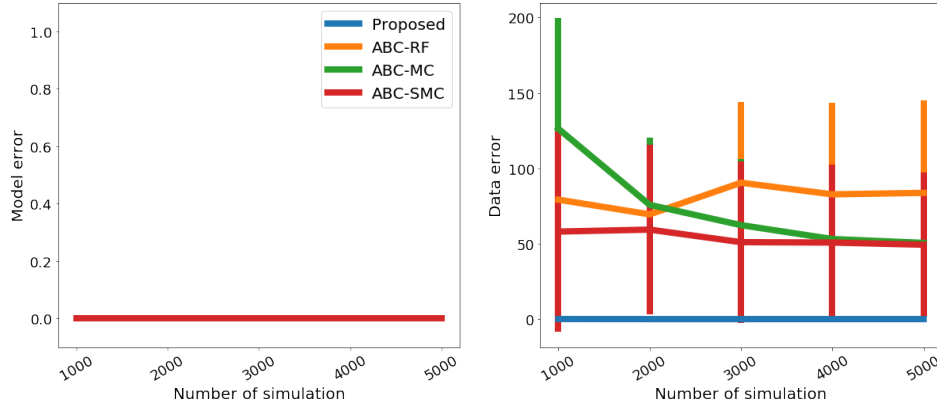


Figure 1: Model errors (left) and data errors (right) for varying total numbers of simulations in the appropriate case. The ground truth is the 3rd order model. (Sec. 4.2)

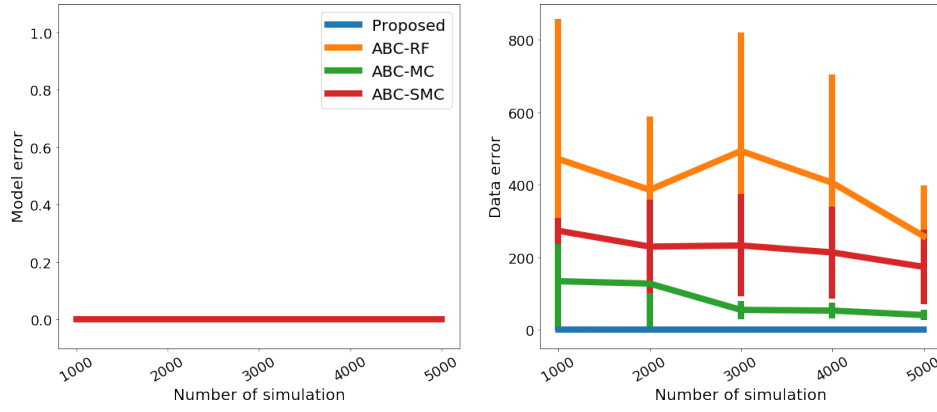


Figure 2: Model errors (left) and data errors (right) for varying total numbers of simulations in the appropriate case. The ground truth is the 4th order model. (Sec. 4.2)

1926):

$$\begin{aligned}\dot{x}(t) &= a_1x(t) - a_2x(t)y(t), \\ \dot{y}(t) &= -a_3y(t) + a_4x(t)y(t),\end{aligned}$$

where $a_1, a_2, a_3, a_4 \geq 0$ are the parameters of the model describing how the two species interact.

Model 2: The Lotka-Volterra model assumes that there are no intra-species interactions, but this may not be the case in reality: If the population of one species becomes too large, then there can be competition within that species for limited resources. Another model (Bazykin, 1998) taking into account such intra-species interactions is given by

$$\begin{aligned}\dot{x}(t) &= b_1x(t) - b_2x(t)y(t) - b_5x^2(t), \\ \dot{y}(t) &= -b_3y(t) + b_4x(t)y(t) - b_6y^2(t)\end{aligned}$$

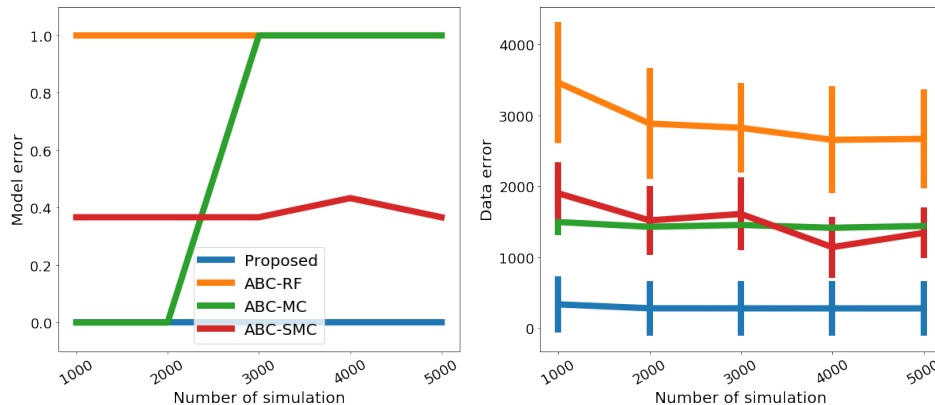


Figure 3: Model errors (left) and data errors (right) for varying total numbers of simulations in the misspecified case. The ground truth is the 3rd order model. (Sec. 4.2)

where $b_1, b_2, b_3, b_4 \geq 0$ are the parameters for inter-species interactions and $b_5, b_6 \geq 0$ are parameters representing intra-species competition rates.

For both models, we defined the initial values as $x(0) = 10$ and $y(0) = 5$. When generating observed data, we used as ground-truth either Model 1 with $a_1 = 1, a_2 = 0.1, a_3 = 1.5$ and $a_4 = 0.75$ or Model 2 with $b_1 = 1, b_2 = 0.1, b_3 = 1.5, b_4 = 0.75, b_5 = 0.01$ and $b_6 = 0.01$. We set the prior distribution of each parameter in $(a_i)_{i=1}^4$ and $(b_i)_{i=1}^6$ as the uniform distribution over $[0, 2]$.

To generate data from the ODEs, we used the SciPy ODE solver (`scipy.integrate.odeint`) with step size 1 and all other parameters being the default values; this results in $T \times 2 = 40$ dimensional summary statistics of observations. To make the problem more realistic, we added an independent standard Gaussian noise to each dimension of summary statistics, both for observed and simulated data. Trajectories from the two models are described in the Supplements.

The results are shown in Figures 5 and 6. Only the proposed method was able to reliably estimate the ground-truth model in both cases. Also in terms of data errors (and errors in extrapolation or prediction; see the Supplements), the proposed method significantly outperformed the competitors. We also conducted another experiment in a more difficult setting, and the proposed method performed the best; see the Supplements.

4.4 Dynamical systems for epidemics

In this experiment, we consider dynamical systems describing the spread of infectious diseases in a population of individuals. The most fundamental one is the SIR model (Anderson and May, 1992), which divides the population into susceptible (S), infected (I) and recovered (R) individuals. Let $S(t)$, $I(t)$, and $R(t)$ be the numbers of individuals in these respective groups at time $t \in [0, T]$, where we set $T = 70$. We consider the following three models defined by ODEs, following Toni and Stumpf (2009, Sec. 3.3). Below, the parameters $\alpha, \gamma, d, \nu \geq 0$ common for all the models respectively represent the rates for birth, infection, death, and recovery of individuals.

Model 1: The first model considers a situation in which a newly infected individual (I) can

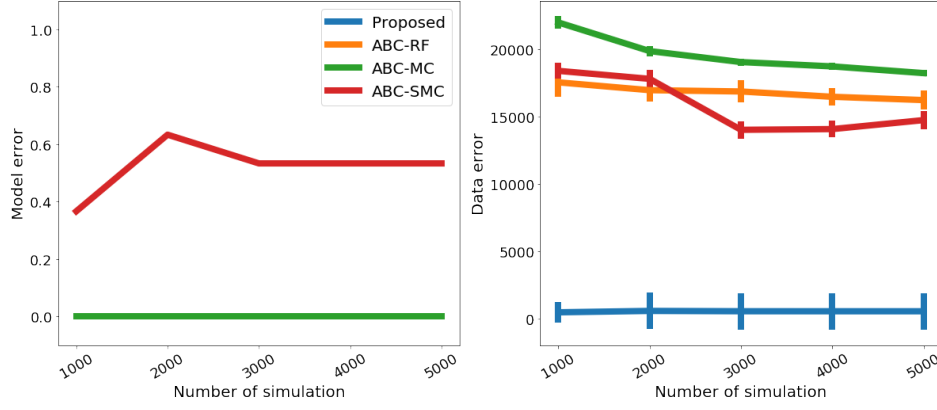


Figure 4: Model errors (left) and data errors (right) for varying total numbers of simulations in the misspecified case. The ground truth is the 4th order model. Note that in the left figure, the proposed method, ABC-RF and ABC-MC made no mistakes and thus the lines are overlapping. (Sec. 4.2)

immediately start infecting susceptible individuals (S):

$$\begin{aligned}\dot{S}(t) &= \alpha - \gamma S(t)I(t) - dS(t), \\ \dot{I}(t) &= \gamma S(t)I(t) - \nu I(t) - dI(t), \\ \dot{R}(t) &= \nu I(t) - dR(t).\end{aligned}$$

Model 2: This model takes into account the situation where a newly infected individual goes into an unobserved Latent (L) state that cannot infect other susceptible individuals. Thus the model includes a parameter $\delta \geq 0$ that represents the rate of transition from Latent to Infective state:

$$\begin{aligned}\dot{S}(t) &= \alpha - \gamma S(t)I(t) - dS(t), \\ \dot{L}(t) &= \gamma S(t)I(t) - \delta L(t) - dL(t), \\ \dot{I}(t) &= \delta L(t) - \nu I(t) - dI(t), \\ \dot{R}(t) &= \nu I(t) - dR(t).\end{aligned}$$

Model 3: The third model allows for a recovered patient to be susceptible once again at a rate $e \geq 0$:

$$\begin{aligned}\dot{S}(t) &= \alpha - \gamma S(t)I(t) - dS(t) + eR(t), \\ \dot{I}(t) &= \gamma S(t)I(t) - \nu I(t) - dI(t), \\ \dot{R}(t) &= \nu I(t) - (d + e)R(t).\end{aligned}$$

We set the initial values as $S(0) = 20$, $I(0) = 50$ and $R(0) = 0$ for all the models, and also $L(0) = 0$ for Model 2. When generating observed data, we set the parameters in the models as $\alpha = 0.5$, $\gamma = 0.001$, $d = 0.01$ and $\nu = 0.02$; also $\delta = 0.1$ (Model 2) and $e = 0.1$ (Model 3). For each of parameters we set a prior distribution as the uniform distribution over $[0, 1]$.

As in Sec. 4.3, we used the SciPy ODE solver with step size 1 with the default parameters to generate data from the ODEs, and this results in $T \times 3 = 210$ dimensional summary statistics of observations. We added an independent standard Gaussian noise to each dimension of summary statistics, both for observed and simulated data. Observed data from the three models are described in the Supplements.

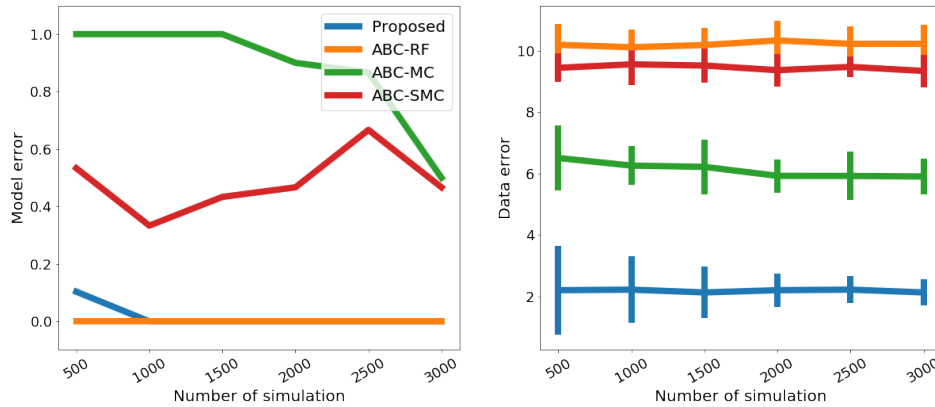


Figure 5: Model errors (left) and data errors (right) for varying total numbers of simulations. The ground-truth is Model 1. (Sec. 4.3)

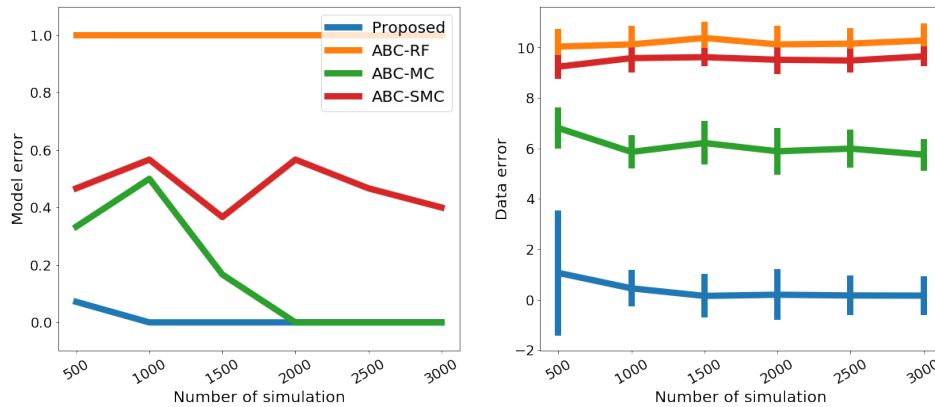


Figure 6: Model errors (left) and data errors (right) for varying total numbers of simulations. The ground-truth is Model 2. (Sec. 4.3)

The results are shown in Figures 7, 8 and 9. In terms of model errors, the competitive methods performed well for some cases but completely failed for other cases. For instance, ABC-SMC selected the correct model when the ground-truth was Model 1, but completely failed when the ground-truth was Model 3. On the other hand, the proposed performed reasonably well for all the cases. In terms of data errors (and extrapolation or prediction errors; see the Supplements), the proposed method significantly outperformed the other methods. Additional experimental results in a more difficult setting are reported in the Supplements.

5 Conclusions and future directions

For the problem of model selection for simulator-based statistical models, we proposed an approach based on a mixture of candidate models and recursive Bayesian updates, implemented using the kernel recursive ABC algorithm. Empirical results suggest the validity of the proposed approach.

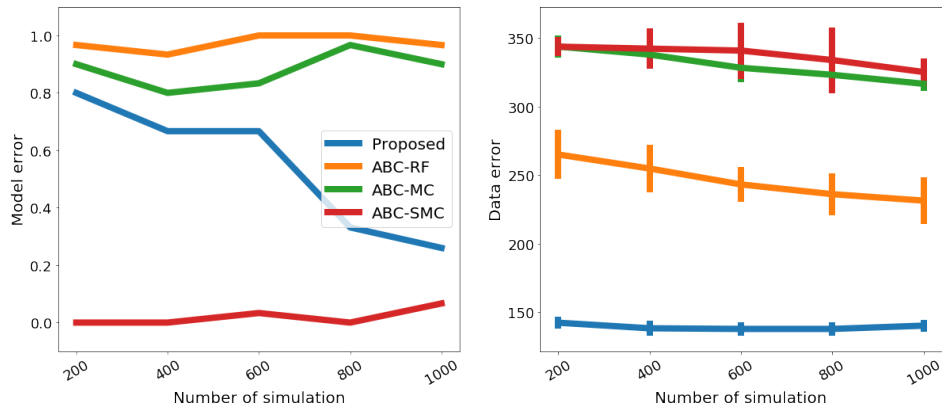


Figure 7: Model errors (left) and data errors (right) for varying total numbers of simulations. The ground-truth is Model 1. (Sec. 4.4)

An obvious topic for future research is theoretical analysis: statistical convergence properties of the proposed algorithm should be revealed mathematically. This also applies to the kernel recursive ABC algorithm itself, for which convergence guarantees have not been established. Another important direction is to investigate applications to large scale problems. This includes for instance multi-agent systems and spatio-temporal modeling involving partial differential equations; these have been used extensively in computer simulation in real-world problems, such as social science and climate modeling. We hope that our work serves as one important step towards solving technical challenges arising in these problems.

Acknowledgements

We thank Keisuke Yamazaki for fruitful discussions. MK acknowledges support by the European Research Council (StG Action 757275 PANAMA). KF has been supported in part by JSPS KAKENHI 26280009.

References

- Anderson, R. M. and May, R. M. (1992). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.
- Bach, F., Lacoste-Julien, S., and Obozinski, G. (2012). On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Conference on Machine Learning (ICML2012)*, pages 1359–1366.
- Bazykin, A. D. (1998). *Nonlinear Dynamics of Interacting Populations*. World Scientific.
- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.
- Chen, Y., Welling, M., and Smola, A. (2010). Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 109–116. AUAI Press.

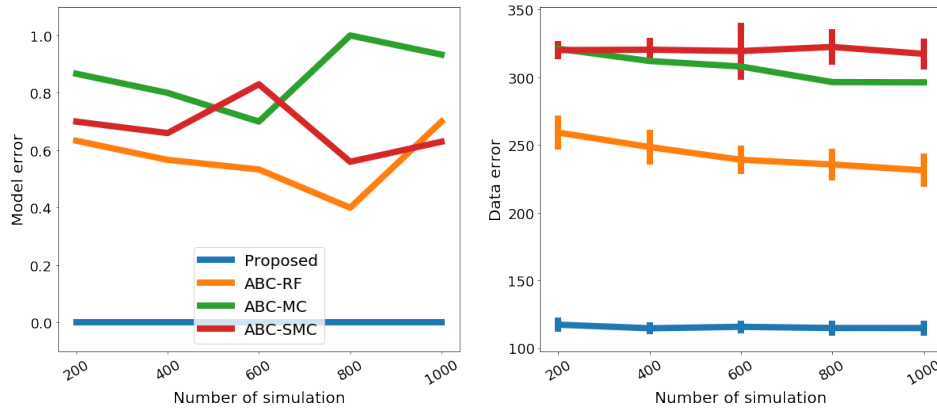


Figure 8: Model errors (left) and data errors (right) for varying total numbers of simulations. The ground-truth is Model 2. (Sec. 4.4)

Dick, J., Kuo, F. Y., and Sloan, I. H. (2013). High dimensional numerical integration - The Quasi-Monte Carlo way. *Acta Numerica*, 22(133-288).

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496.

Fukumizu, K., Song, L., and Gretton, A. (2013). Kernel Bayes’ rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783.

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40.

Grelaud, A., Robert, C. P., Marin, J.-M., Rodolphe, F., Taly, J.-F., et al. (2009). ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 4(2):317–335.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. (2012). Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning (ICML2012)*, pages 1823–1830.

Kajihara, T., Kanagawa, M., Yamazaki, K., and Fukumizu, K. (2018). Kernel recursive ABC: Point estimation with intractable likelihood. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2400–2409, Stockholmsmässan, Stockholm Sweden. PMLR.

Kanagawa, M., Nishiyama, Y., Gretton, A., and Fukumizu, K. (2016a). Filtering with state-observation examples via kernel monte carlo filter. *Neural Computation*, 28(2):382–444.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. (2016b). Convergence guarantees for kernel-based quadrature rules in misspecified settings. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 3288–3296. Curran Associates, Inc.

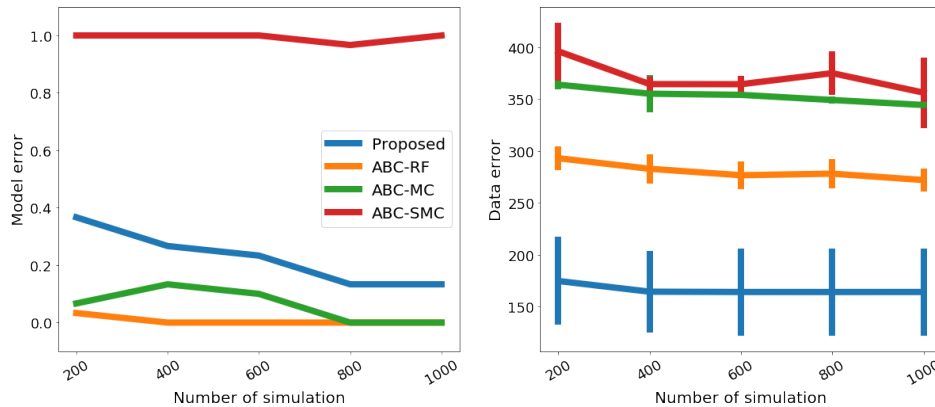


Figure 9: Model errors (left) and data errors (right) for varying total numbers of simulations. The ground-truth is Model 3. (Sec. 4.4)

- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.
- Lacoste-Julien, S., Lindsten, F., and Bach, F. (2015). Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In Lebanon, G. and Vishwanathan, S. V. N., editors, *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 544–552. PMLR.
- Lafferty, J. and Lebanon, G. (2005). Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163.
- Lele, S. R., Nadeem, K., , and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492):1617–1625.
- Lotka, A. J. (1926). Elements of physical biology. *Science Progress in the Twentieth Century (1919-1933)*, 21(82):341–343.
- Muandet, K., Fukumizu, K., Sriperumbudur, B. K., and Schölkopf, B. (2017). Kernel mean embedding of distributions : A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141.
- Nakagome, S., Fukumizu, K., and Mano, S. (2013). Kernel approximate Bayesian computation in population genetic inferences. *Statistical Applications in Genetics and Molecular Biology*, 12(6):667–678.
- Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian Computation with kernel embeddings. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 398–407. PMLR.
- Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

- Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. (2015). Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866.
- Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2016). ABC random forests for Bayesian parameter inference. *arXiv preprint arXiv:1605.05537*.
- Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Science & Business Media.
- Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC.
- Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer.
- Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561.
- Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.
- Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. (2017). Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47.
- Toni, T. and Stumpf, M. P. (2009). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110.
- Toni, T., Welch, D., Strelkova, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.
- Volterra, V. (1926). Variations and fluctuations of the number of individuals in animal species living together. *Animal Ecology*, pages 409–448.
- Weisberg, M. (2012). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.
- Winsberg, E. (2010). *Science in the Age of Computer Simulation*. University of Chicago Press.
- Yıldırım, S., Singh, S. S., Dean, T., and Jasra, A. (2015). Parameter estimation in hidden Markov models with intractable likelihoods using sequential Monte Carlo. *Journal of Computational and Graphical Statistics*, 24(3):846–865.

Supplementary Materials

A Proof of Proposition 1

Proof. Below $\pi^N((\theta^m)_{m=1}^K | (\phi^m)_{m=1}^K)$ denotes the conditional density of $(\theta^m)_{m=1}^K$ given $(\phi^m)_{m=1}^K$ that is induced from $\pi^N((\phi^m)_{m=1}^K, (\theta^m)_{m=1}^K)$.

$$\begin{aligned}
 & p^N((\phi^m)_{m=1}^K | y^*) \\
 = & \int p^N((\phi^m)_{m=1}^K, (\theta^m)_{m=1}^K | y^*) d\theta^1 \dots d\theta^K \\
 = & \int C_N^{-1} p(y^* | (\phi^m)_{m=1}^K, (\theta^m)_{m=1}^K) \pi^N((\phi^m)_{m=1}^K, (\theta^m)_{m=1}^K) d\theta^1 \dots d\theta^K \\
 = & \int C_N^{-1} \left(\sum_{\ell=1}^K \phi^\ell p_\ell(y^* | \theta^\ell) \right) \pi^N((\phi^m)_{m=1}^K) \pi^N((\theta^m)_{m=1}^K | (\phi^m)_{m=1}^K) d\theta^1 \dots d\theta^K \\
 = & C_N^{-1} \pi^N((\phi^m)_{m=1}^K) \sum_{\ell=1}^K \phi^\ell \int p_\ell(y^* | \theta^\ell) \pi^N((\theta^m)_{m=1}^K | (\phi^m)_{m=1}^K) d\theta^1 \dots d\theta^K \\
 = & C_N^{-1} \sum_{\ell=1}^K \phi^\ell \int p_\ell(y^* | \theta^\ell) \pi^N(\theta^\ell | (\phi^m)_{m=1}^K) d\theta^\ell.
 \end{aligned}$$

This proves the claim for the case $N \geq 2$. For $N = 1$ the claim follows from the definition of $\pi^1((\phi^m)_{m=1}^K, (\theta^m)_{m=1}^K)$ in (6) of the main text. \square

B A review of existing methods based on ABC

We provide a brief review of existing approaches to which we compare our method in the experiments. To this end, for simplicity we restrict ourselves to the case where there are two candidate models, each of which consists of a conditional distribution $P_m(y|\theta^m)$ on \mathcal{Y} given $\theta^m \in \Theta^m$ and a prior distribution $\pi_m(\theta^m)$ on Θ^m , where \mathcal{Y} and Θ^m are measurable spaces and $m = 1, 2$. Assume that $P_m(y|\theta^m)$ has a density function $p_m(y|\theta^m)$. Let $y^* \in \mathcal{Y}$ be observed data.

ABC for Model Choice (ABC-MC) We first review the approach by Grelaud et al. (2009) called *ABC for Model Choice (ABC-MC)*. A standard Bayesian approach to model selection is to compute the Bayes factor (Kass and Raftery, 1995; Robert, 2007) defined as the ratio between the marginal likelihoods of the two models:

$$B_{1,2} := \frac{\int p_1(y^* | \theta^1) d\pi_1(\theta^1)}{\int p_2(y^* | \theta^2) d\pi_2(\theta^2)} = \frac{\Pr(1|y^*)\Pr(2)}{\Pr(2|y^*)\Pr(1)},$$

where $\Pr(m|y^*)$ denotes the posterior probability of model m being true, given observed data y^* , and $\Pr(m)$ is a prior probability of model m being true; the second identity follows from Bayes' rule. In ABC-MC, the Bayes factor is estimated using this second expression, estimating the two posterior probabilities $\Pr(m|y^*)$ by ABC.

Based on ABC, $\Pr(m|y^*)$ can be estimated in the following way. For $t = 1, \dots, n \in \mathbb{N}$, sample model index $m_t \sim \Pr(m)$ from the prior, sample model parameters $\theta_t^m \sim \pi_m(\theta^m)$, and then sample

pseudo-data $y_t \sim P_m(y|\theta^m)$; if $\|y_t - y^*\| < \varepsilon$ for a prespecified threshold $\varepsilon > 0$ and a distance $\|\cdot\|$ defined on \mathcal{Y} , accept y_t ; otherwise discard y_t . Denoting by $(\tilde{m}_t)_{t=1}^T$ the set of the model indices associate with accepted data, where $T \leq n$ is the number of acceptances, the posterior probability $\Pr(m|y^*)$ is then estimated by

$$\widehat{\Pr}_\varepsilon(m|y^*) := \frac{1}{T} \sum_{t=1}^T \mathbb{I}(m_t = m), \quad m = 1, 2,$$

where $\mathbb{I}(m_t = m) = 1$ if $m_t = m$ and $\mathbb{I}(m_t = m) = 0$ otherwise.

For the experiments in our paper, we employed the corresponding k -NN (nearest neighbors) version, as recommended by Grelaud et al. (2009).

ABC Sequential Monte Carlo (ABC-SMC) The method proposed by Toni et al. (2009), called *ABC Sequential Monte Carlo (ABC-SMC)*, is also a rejection-based ABC approach to Bayesian model selection, as for ABC-MC. Based on sequential Monte Carlo, the method iteratively updates the threshold ε and the associated proposal distributions in an adaptive manner.

ABC Random Forests (ABC-RF) The approach by Pudlo et al. (2015), termed *ABC Random Forests (ABC-RF)*, is based on interpretation of Bayesian model selection as a classification problem. As for ABC-MC, consider the following sampling procedure: For $t = 1, \dots, n \in \mathbb{N}$, sample model index $m_t \sim \Pr(m)$ from the prior, sample model parameters $\theta_t^m \sim \pi_m(\theta^m)$, and then sample pseudo-data $y_t \sim P_m(y|\theta^m)$. The posterior probability $\Pr(m|y^*)$ is then the conditional probability of model index m conditioning on observed data y^* that is induced from the joint distribution of $(m_t, y_t)_{t=1}^n$. Thus, the problem of model selection can be cast as estimation of this conditional probability, or more specifically as classification of observed data y^* to one of the model indices $m = 1, 2$ (or class labels) using $(m_t, y_t)_{t=1}^n$ as training data. ABC-RF solves this classification problem using random forests, thereby providing a way of Bayesian model selection. The essentially same idea can be used for parameter estimation in the setting of ABC, as proposed by Raynal et al. (2016).

C Detailed explanation of experimental settings

C.1 Hyper-parameter selection

For each method, we determined its hyper-parameters by adopting a cross-validation-like approach described in Park et al. (2016, Sec. 4), unless otherwise stated. That is, to evaluate one configuration of hyper-parameters, we first used 80% of the observed data for point estimation and then computed the discrepancy between the rest of the observed data and the ones simulated from point estimates; after applying this procedure to all candidate configurations, the one with the lowest discrepancy was finally selected.

For ABC-MC, the number k of k -nearest neighbours was selected from $\{100, 200, 300, 500, 800, 1000\}$ using the cross-validation like approach. For ABC-SMC, we used the median of pairwise distances between the observed and simulated data as an initial tolerance level. As a perturbation kernel, we used the zero-mean Gaussian distribution with variance 0.1, unless otherwise stated.

For ABC-RF, we set the number of trees to be 500. We set the number of bootstrapping samples to be the number of the sample points. We set the size of the subset of parameters to be considered at each node in a tree to be \sqrt{d} for classification and $d/3$ for regression, where d is the number of parameters. These configurations are those recommended by Pudlo et al. (2015).

For our method we defined a Gaussian kernel $\exp(-\|x-x'\|^2/\gamma^2)$ for each of observations, parameters and mixing coefficients. We selected the bandwidths of these kernels, denoted for now by $\gamma_{\mathcal{Y}}$, γ_{Θ} and γ_{ϕ} , as well as the regularization constant δ by the above cross-validation like approach. More specifically, for a constant $s > 0$ let $\gamma_{\mathcal{Y}} = s \times \text{med}(y_{N,1}, \dots, y_{N,n})$, $\gamma_{\Theta} = s \times \text{med}((\theta_{N,1}^m)_{m=1}^M, \dots, (\theta_{N,n}^m)_{m=1}^M)$, and $\gamma_{\phi} = s \times \text{med}((\phi_{N,1}^m)_{m=1}^M, \dots, (\phi_{N,n}^m)_{m=1}^M)$, where $\text{med}(x_1, \dots, x_n)$ denotes the median of pairwise distances between points x_1, \dots, x_n .

We selected this scaling constant s and the regularization constant δ by the cross validation-like approach, where candidates for s were logarithmically equally-spaced values between 2^{-2} and 2^2 and those for σ were logarithmically equally-spaced values between 2^{-4} and 2^2 .

C.2 Parameter estimation

For the purpose of comparison, we also performed parameter estimation with the competing methods as follows. We first chose a model $m \in \{1, \dots, M\}$ based on a model selection method, and then only retained the generated pairs (y_i^m, θ_i^m) of simulated data and parameters associated with the model m . For ABC-RF, we then performed regression by Random Forests from data to parameters based on the retained pairs (y_i^m, θ_i^m) . For ABC-MC and ABC-SMC, we estimated the posterior of the parameters based on the retained pairs (y_i^m, θ_i^m) , and then applied the mean shift algorithm (Fukunaga and Hostetler, 1975) to obtain MAP estimates for the parameters; for mean-shift, we used the implementation in the scikit-learn, with the bandwidth of KDE being specified as the median of all the pairwise distances of all samples.

D Additional experimental results

D.1 Mixing coefficients and the effects of the Dirichlet concentration parameter

We performed simple experiments to see how the mixing coefficients for candidate models in the proposed method behaves. To this end, we deal with model selection for a regression problem, where the 3rd order and 10th order polynomials are to be compared.

Let $x_1, \dots, x_{20} \subset [-1, 2]$ be equally-spaced points such that $-1 = x_1 < x_2 < \dots < x_{20} = 2$. The 3rd and 10th order polynomial models are then defined as $y_i = \sum_{\ell=0}^3 a_{\ell} x_i^{\ell} + \varepsilon_i$ and $y_i = \sum_{\ell=0}^{10} b_{\ell} x_i^{\ell} + \varepsilon_i$, respectively ($i = 1, \dots, 20$), where $(a_{\ell})_{\ell=0}^3 \subset \mathbb{R}$ and $(b_{\ell})_{\ell=0}^{10} \subset \mathbb{R}$ are the parameters in the models. For each parameter in $(a_{\ell})_{\ell=0}^3$ and $(b_{\ell})_{\ell=0}^{10}$ we defined a prior as the uniform distribution over $[-100, 100]$. When generating observed data $y^* := (y_1^*, \dots, y_{20}^*) \in \mathbb{R}^{20}$, we used either $a_0 = \dots = a_3 = 4$ (when the 3rd order is true) or $b_0 = \dots = b_{10} = 4$ (when the 10th order is true). We set the concentration parameter of the Dirichlet prior for ϕ_1 and ϕ_2 to be 100.

Figure 10 describes how the mixing coefficient for the ground-truth model changed as iteration proceeded, for both cases where the ground truth was the 3rd order (left) or the 10th order (right). The results show that for both cases the coefficients gradually converged towards the true model, suggesting the validity of our mixture approach.

We next investigated how the behavior of the proposed method changes when changing the concentration parameter of the Dirichlet prior α . Figure 11 shows the mixing coefficients after 30 iterations, when running our method for different values of α . The results show that for smaller α , the coefficients tend to be more confident about the true model (i.e., the coefficients tend to be close to 0 or 1). For instance, in Figure 11 (a) on the case where the truth is the 3rd order, the coefficient for

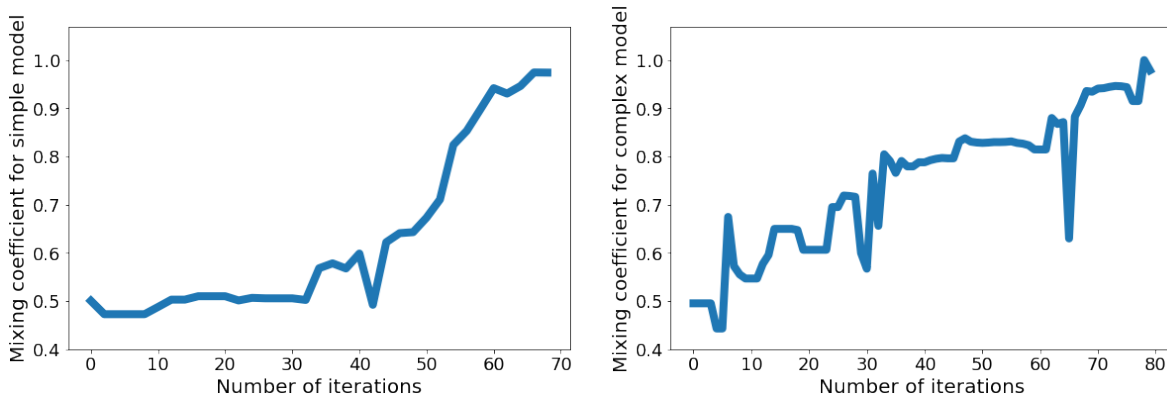


Figure 10: Demonstrations of how the mixing coefficient for the true model changes, as the iteration proceeds. Left: the ground-truth was the 3rd order model (simple); Right: the ground-truth was the 10th order model (complex). (Sec. D.1)

the 3rd order polynomial tends to 1 as α decreases (i.e., more confident the truth is the 10th-order), and approaches 0.5 as α increases (more neutral). This is consistent with the role of the concentration parameter for the Dirichlet prior: If α is small, then the mixing coefficients sampled from the Dirichlet prior tend to take values close to either of 0 or 1, while if α is large, then the sampled mixing coefficients tend to be close to 0.5.

We nevertheless found that, regardless of the choice of the concentration parameter, the mixing coefficients converge to either 0 or 1, as the iteration proceeds. This is described in Figure 12, which plots the number of iterations required for the mixing coefficients to reach the region $[0, 0.05]$ or $[0.95, 1]$. The result shows that a larger concentration parameter leads to rather conservative decisions regarding model selection, requiring a larger number of iterations for convergence.

D.2 Extrapolation errors in polynomial experiments (Sec. 4.2)

To evaluate the generalization performances of the parameters (and the models) estimated by each method, we also report extrapolation errors given as follows. Let $x_{26}, \dots, x_{30} \subset [5, 6]$ be equally-spaced points such that $5 = x_{26} < \dots < x_{30} = 6$. We then generated observed data $y^* := (y_{26}^*, \dots, y_{30}^*) \in \mathbb{R}^5$ from the ground-truth model evaluated at x_{26}, \dots, x_{30} . We calculated the corresponding data errors at x_{26}, \dots, x_{30} for each method, in the same way as the one explained in the main body.

We show the results for the appropriate prior distributions in Figure 13 and for the misspecified prior distributions in Figure 14. The proposed method outperformed the other methods in all the experiments. Notably, in the misspecified case, the errors of the other methods are huge, while those of the proposed method are negligible, verifying that the proposed method successfully selected the true model and estimated its parameters.

D.3 Predator-prey models (Sec. 4.3)

Dynamics of the two models. We show observed trajectories of the two models (with noises added) in Figure 15.

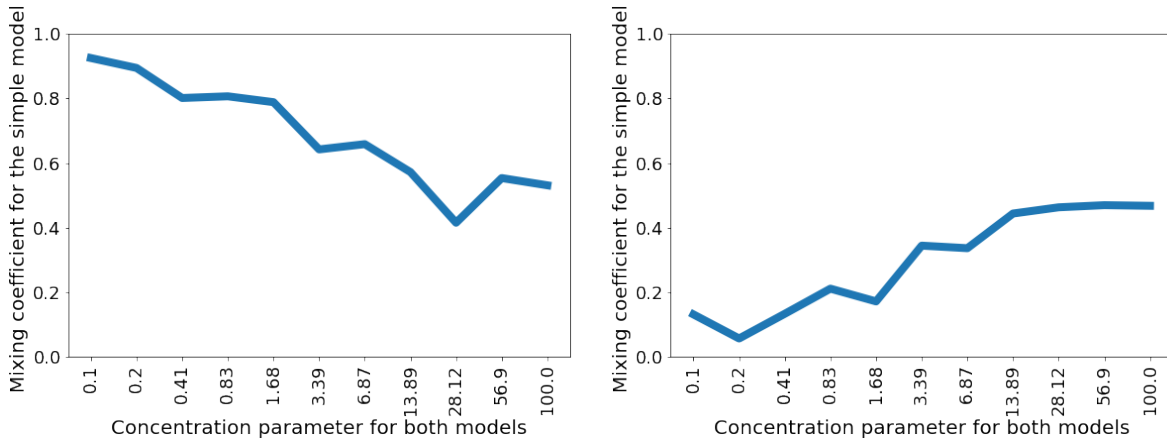


Figure 11: The coefficients for the simpler model (i.e., the 3rd order polynomial) after 30 iterations, for different values of the concentration parameter α . Left: the truth was the 3rd order model; Right: the truth was the 10th polynomial model. (Sec. D.1)

Experiments in a more difficult setting. To make the problem more challenging, we assumed the initial population sizes $x(0)$ and $y(0)$ to be unknown, and let them as parameters in the models to be estimated. We defined a wide prior distribution for each of $x(0)$ and $y(0)$ as the uniform distribution over $[0, 2000]$. We set $T = 30$, and removed the first 10 values $x(0), \dots, x(9)$ and $y(0), \dots, y(9)$ from observed data. In preliminary experiments we found that, because of the wide prior distribution for the initial values, the variance of the generated data could become so huge that any of the methods did not work well. We therefore applied the arctangent transformation to the observed and simulated data to stabilize the large variance, following Yildirim et al. (2015, p.854).

Figures 16 and 17 respectively show the results for the two cases where the ground truth is Model 1 or Model 2. ABC-SMC and ABC-MC estimated the true model correctly only in one of the two cases. In other words, those methods favor one model exclusively. Only the proposed method was able to estimate the true model reliably for both cases. The proposed method also outperformed other methods in terms of data errors for both cases.

Extrapolation errors To evaluate the predictive performances of the parameters (and the models) estimated by each method, we also report extrapolation errors, given as follows. Extrapolation errors were evaluated at equally-spaced 5 points $t = 21, 22, \dots, 25$ for the setting of the main body, and $t = 31, 32, \dots, 35$ for the difficult setting. Results for the setting of the main body are shown in Figure 18, and those for the difficult setting are shown in Figure 19. The proposed method outperformed the other methods in all the experiments.

D.4 Dynamical systems for epidemics (Sec. 4.4)

Dynamics of the three models. We show observed trajectories of the three models (with noises added) in Figure 20.

Experiments in a more difficult setting. To make the problem more challenging, we treated the initial values $S(0)$, $I(0)$ and $R(0)$ (and $L(0)$ for Model 2) as unknown parameters in the models to be

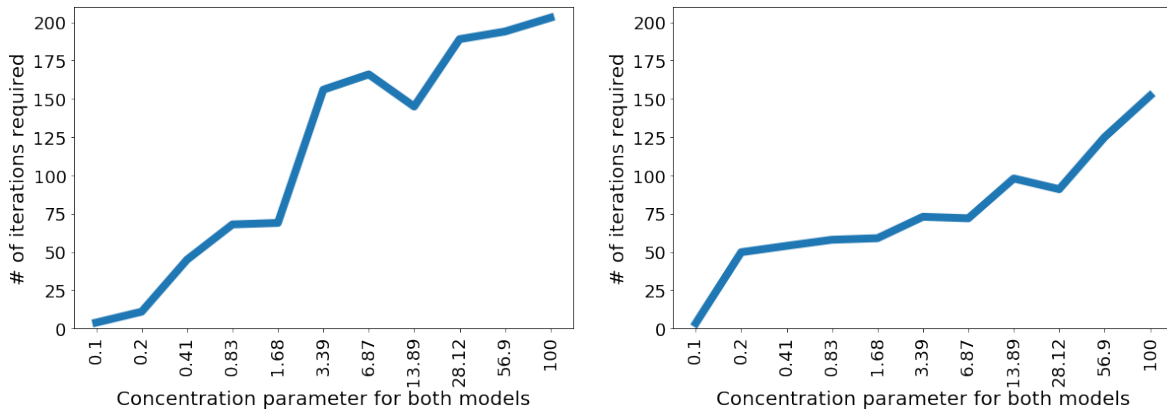


Figure 12: The number of recursive Bayes’ updates required to make the mixing coefficient larger 0.95 for a true model over a varying concentration parameter α . Left: the truth is the 3rd order model; Right: the truth is the 10th order model. (Sec. D.1)

estimated. We set $T = 80$, and removed the first 10 values $S(0), \dots, S(9), I(0), \dots, I(9), R(0), \dots, R(9)$ from observed data. We defined a wide prior for each of $S(0), I(0), R(0)$ and $L(0)$ as the uniform distribution over $[0, 500]$. As none of the methods was able to identify a true model if the three candidate models were jointly compared, we report only the case where the second and third models are compared. The results are shown in Figure 21 and 22. ABC-RF and ABC-MC identified correctly the ground-truth model only for one of the two cases, and failed for the other case. Only the proposed method was able to identify the true model reliably for the both cases.

Extrapolation errors. The extrapolation errors were evaluated at equally-spaced 15 points $t = 71, 72, \dots, 85$ for the setting of the main body, and at $t = 81, 82, \dots, 95$ for the difficult setting defined above. Results for the setting of the main body are shown in Figure 23, and those for the difficult setting are shown in Figure 24. The proposed method outperformed the other methods for most of the experiments.

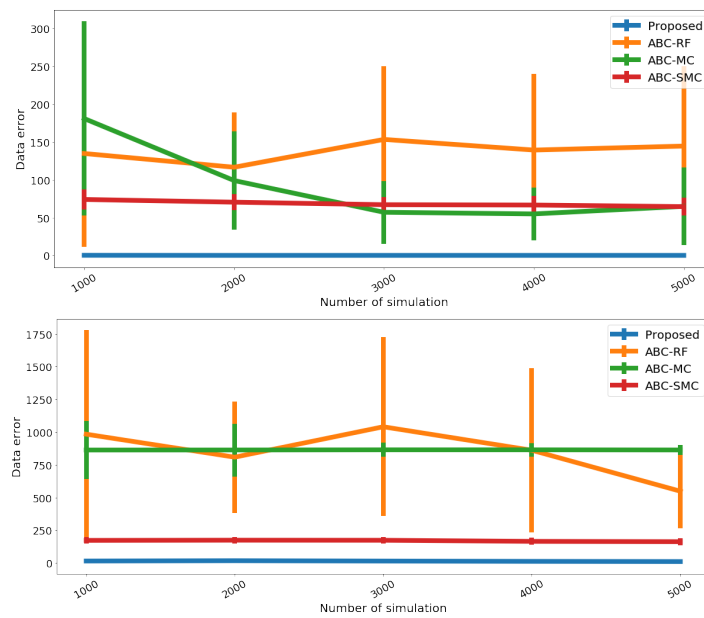


Figure 13: Extrapolation errors for polynomial experiments with appropriate prior distributions. Top: the ground truth is the 3th order polynomial model; Bottom: the ground truth is the 4th order polynomial model.

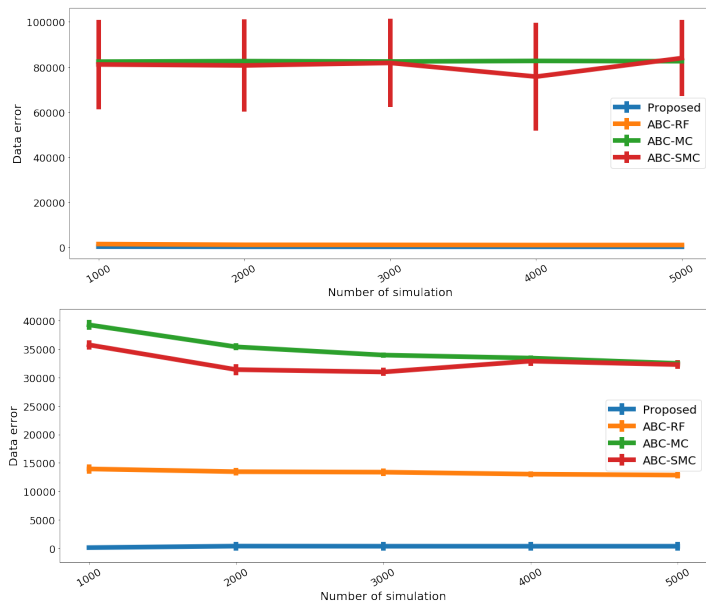


Figure 14: Extrapolation errors for polynomial experiments with misspecified prior distributions. Top: the ground truth is the 3th order polynomial model; Bottom: the ground truth is the 4th order polynomial model.

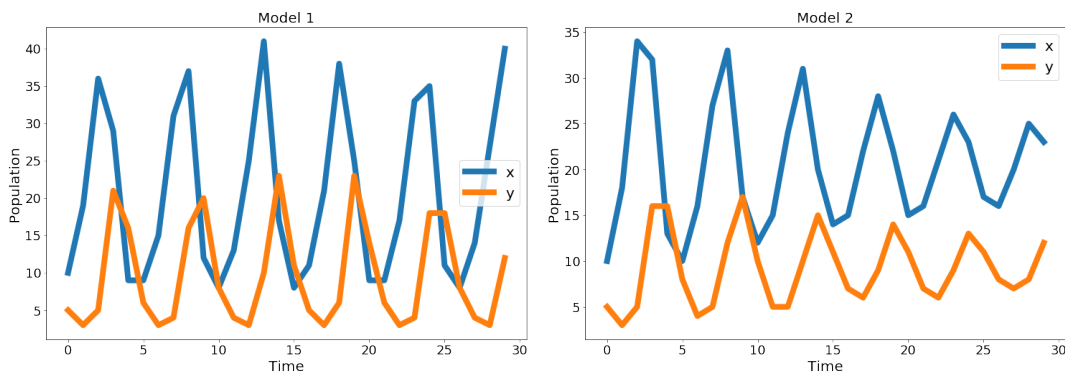


Figure 15: Observed trajectories (with noises) of the two populations (predators $y(t)$ and prey $x(t)$) generated with the true parameters. Left: Model 1; Right: Model 2.

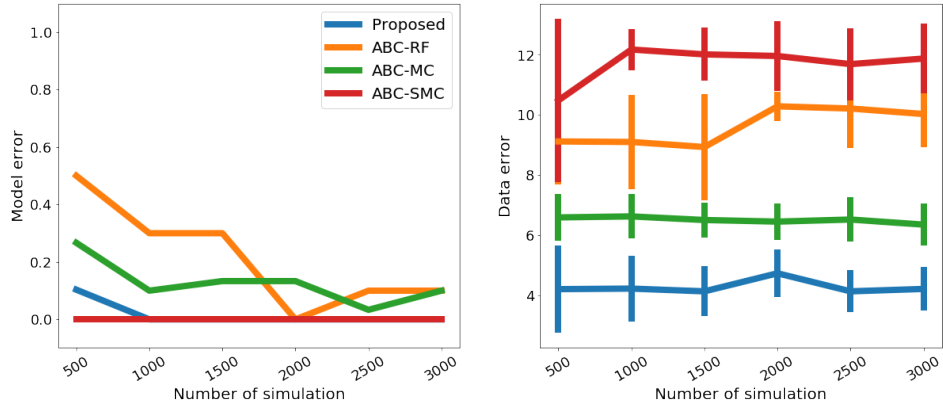


Figure 16: Model errors (left) and data errors (right) in the predator-prey experiments on the difficult setting. The ground-truth is Model 1.

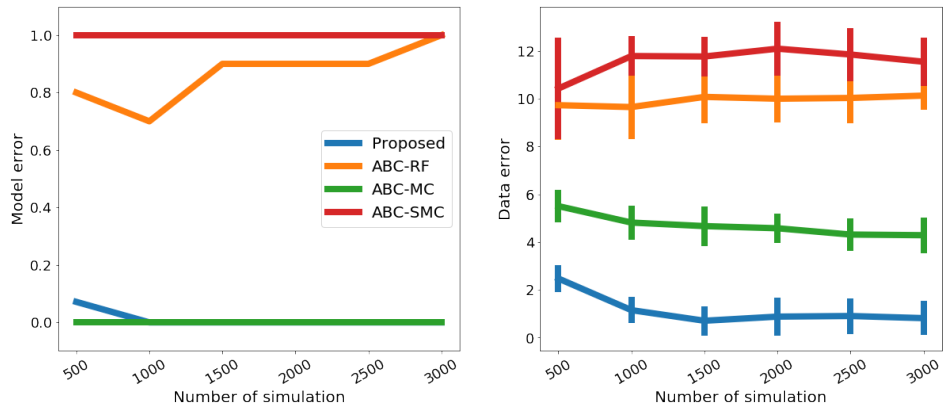


Figure 17: Model errors (left) and data errors (right) in the predator-prey experiments on the difficult setting. The ground-truth is Model 2.

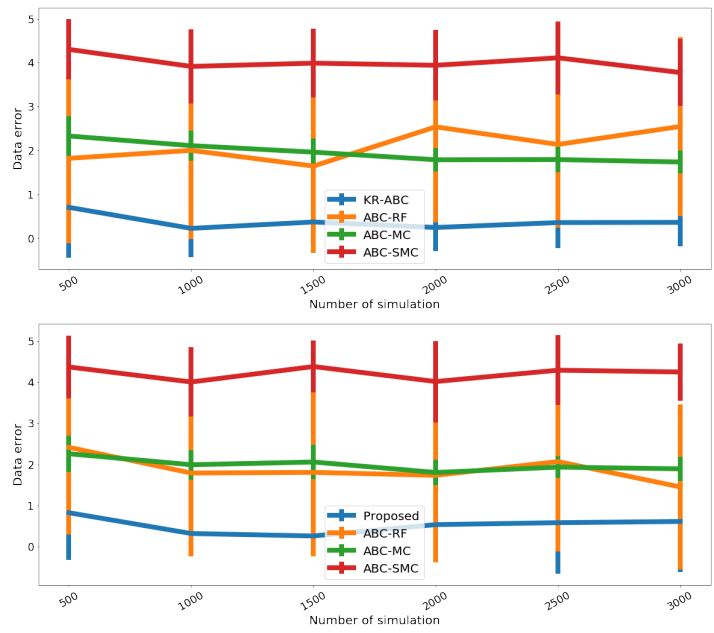


Figure 18: Extrapolation errors in the predator-prey experiments for the setting of the main body. Top: the ground-truth is Model 1; Bottom: the ground-truth is Model 2.

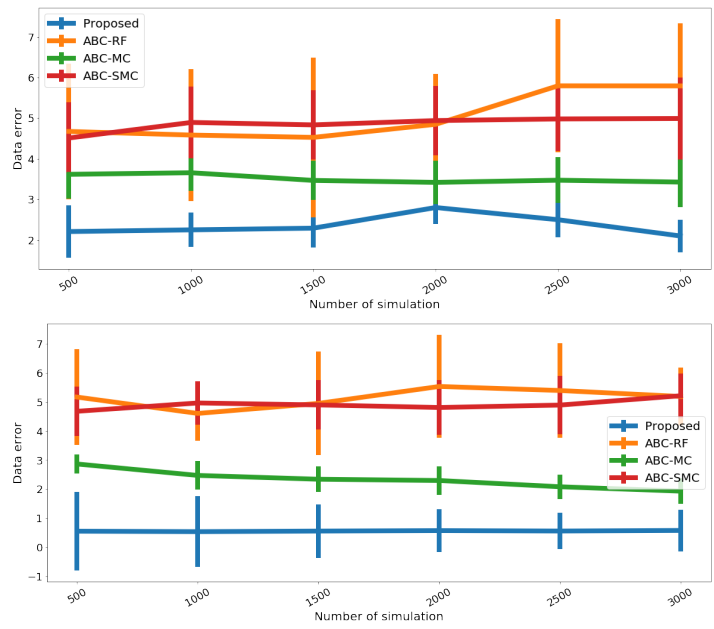


Figure 19: Extrapolation errors in the predator-prey experiments for the difficult setting. Top: the ground-truth is Model 1; Bottom: the ground-truth is Model 2.

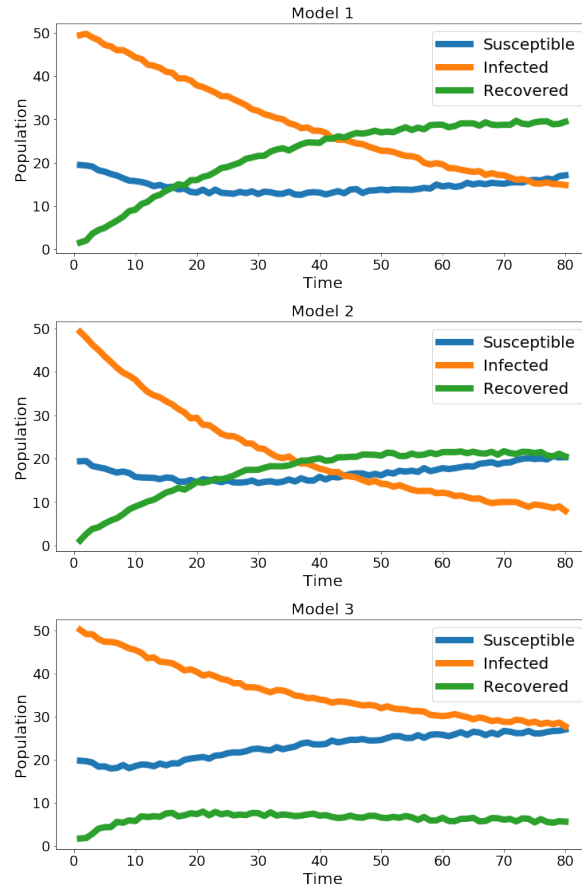


Figure 20: Observed trajectories from Model 1 (Top), Model 2 (Middle) and Model 3 (Bottom) with the true parameters in epidemics experiments. As explained in the main body, independent noises are added in each trajectory.

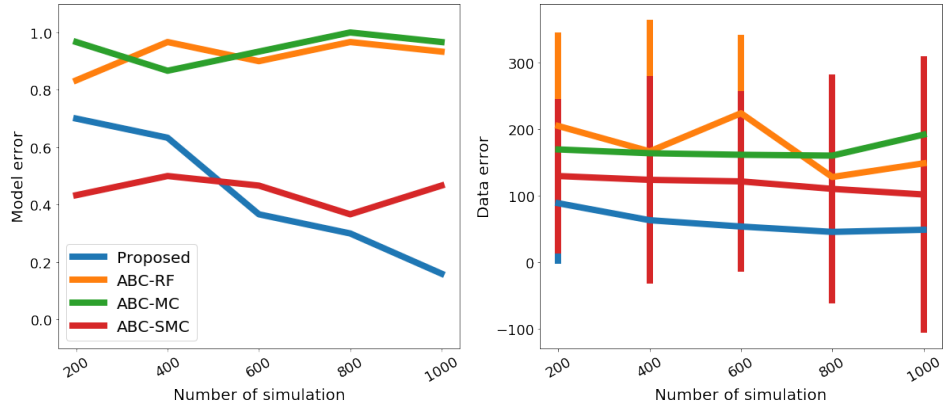


Figure 21: Model errors (left) and data errors (right) in the epidemics experiments for the difficult setting. The ground-truth model is Model 2.

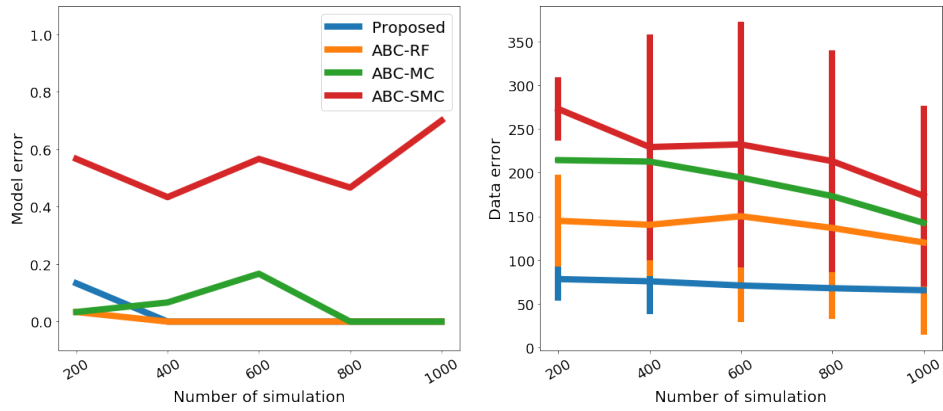


Figure 22: Model errors (left) and data errors (right) in the epidemics experiments for the difficult setting. The ground-truth model is Model 3.

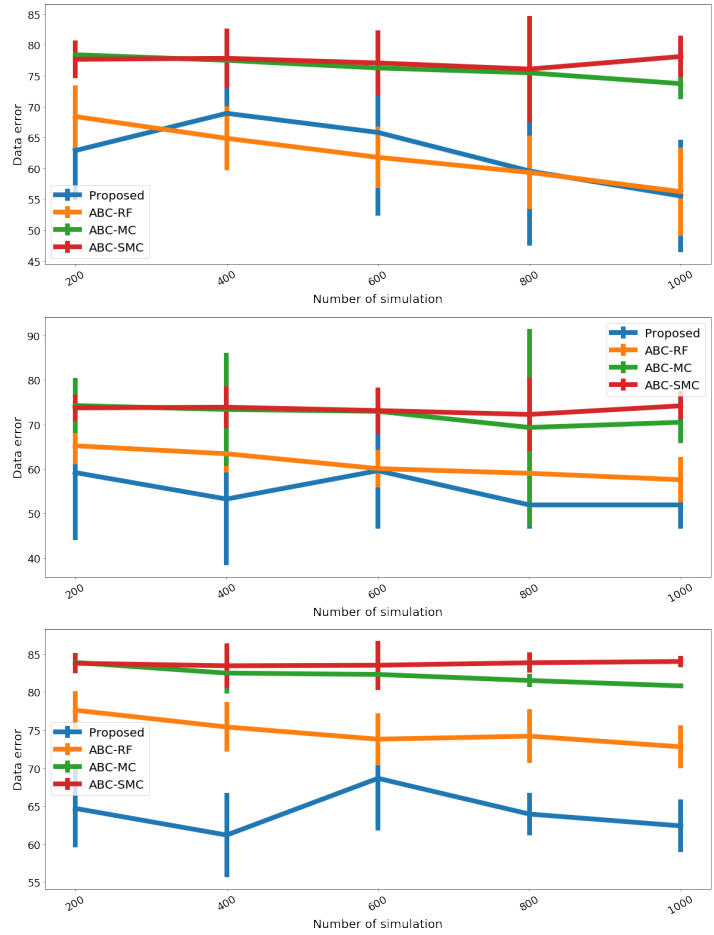


Figure 23: Extrapolation (or prediction) errors in the epidemics experiments, for the cases where the ground-truth is Model 1 (top), Model 2 (middle) and Model 3 (bottom). The settings (such as the prior distributions) are the same as in the main body.

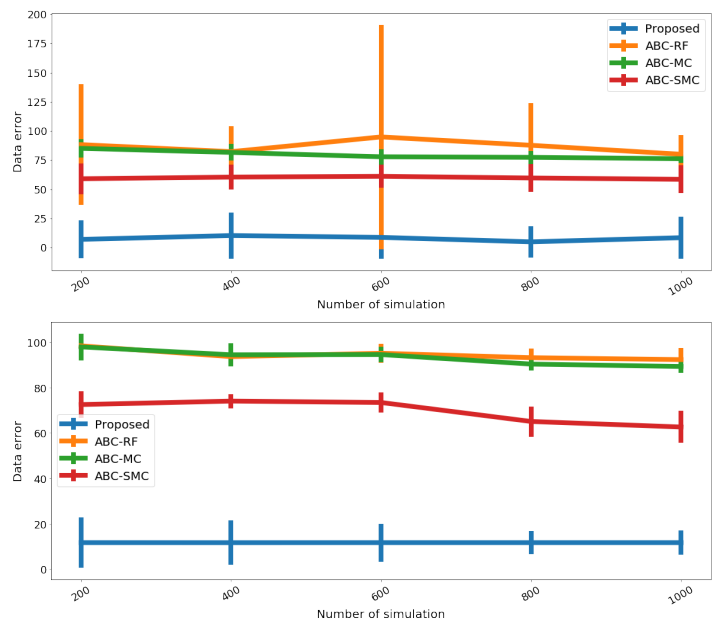


Figure 24: Extrapolation (or prediction) errors in the epidemics experiments, for the cases where the ground-truth is Model 2 (top) and Model 3 (bottom). The setting is the difficult setting defined in the Supplementary Materials.