CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Approximate Bayesian Computation with Sequential Surrogate Likelihoods

Master's thesis in Engineering Mathematics & Computational Science

FILIP WIKMAN

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2019

Master's thesis 2019

Approximate Bayesian Computation with Sequential Surrogate Likelihoods

FILIP WIKMAN

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2019

Approximate Bayesian Computation with Sequential Surrogate Likelihoods
FILIP WIKMAN

Supervisor: Umberto Picchini, Department of Mathematical Sciences
Examiner: Umberto Picchini, Department of Mathematical Sciences

Master's Thesis 2019
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2019

Approximate Bayesian Computation with Sequential Surrogate Likelihoods
FILIP WIKMAN
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

The purpose of this thesis was to implement, analyze, and possibly expand a Bayesian inference method related to approximate Bayesian computation (ABC). This method was initially suggested by the supervisor and was given the working name *approximate Bayesian computation with sequential surrogate likelihoods* (ABC-SSL). The underlying idea for the method was to replace ABC distances with predicted distances obtained using some regression technique, thus circumventing generation of synthetic datasets from the Bayesian model. These predictions would then be improved in a sequential manner, leading to a significant decrease of computational cost for parameter inference.
The literature on ABC was studied in search of similar techniques with the intent of finding suitable methods to be compared to ABC-SSL in a simulation study. Gaussian process regression was chosen to model the distances due to the need for flexibility. An interpretation and generalization of the preliminary ABC-SSL method was given, relating it to some of the methods found in the literature. The simulation study was constructed with three examples, including one of the standard models in the ABC literature, the g-and-k distribution. These examples were chosen to give an understanding of potential use of the method. Due to lack of promising results of these numerical results, the complexity of the tested models were kept low.
No conclusive evidence was found for the inference method to be suitable for practical use in its current state due to questionable asymptotic properties and difficulties in finding appropriate surrogate models. One possible application is to use the proposed technique to find regions of suspected high posterior probability of the parameter space to be used in combination with traditional ABC methods. Another possibility is to consider Bayesian optimization problems, although such problems were not explicitly investigated in this thesis.

Keywords: Approximate Bayesian computation, ABC, Bayesian statistics, statistics, Gaussian process regression.

Acknowledgements

I would like to thank my supervisor Umberto Picchini, both for the initial idea for the project and for helpful insights along the way. I would also like to thank Samuel Wiqvist for some tips on Gaussian processes.

Filip Wikman, Gothenburg, January 2019

Contents

List of Figures

1

Introduction

1.1 History of ABC

The study of Approximate Bayesian Computation (ABC) has gained attention in the statistical community since the turn of the century. Its popularity stems from the wide range of applications, with notable examples such as population genetics [20]. The purpose of ABC is to infer parameters for a Bayesian model with intractable likelihood where one can sample from the likelihood using a simulator. There are multiple review articles on the topic which give a more complete view of ABC methods than what is presented here [13, 12].

The simplest form of ABC is known as Rejection ABC (R-ABC) and consists of generating parameter proposals from a prior distribution, simulating data using these parameters, and accepting the parameters as a sample from the posterior if the data is within a given "distance" from the observed data [26]. Even though convergence is guaranteed for virtually any model with finite dimensional data [1], this simple approach is often not feasible in practice due to the large number of samples required to achieve precise inference. The ABC framework consists of multiple algorithms all derived from these ideas with the goal of aleviating some of the problems with R-ABC. Notable examples include Markov chain Monte Carlo ABC (ABC-MCMC, [14]), Kernel ABC (K-ABC, [27]), Sequential Monte Carlo ABC (ABC-SMC, [25, 2]) and many more. Further difficulties arise when simulating from the model or computing distances is computationally expensive. As a work-around for this problem, Umberto Picchini proposed an algorithm which consists of "learning" the likelihood function with some type of regression, e.g. Gaussian process regression, and using this to sequentially improve the parameter proposals, thus reducing the number of model simulations needed.

The suggested working name for this method is ABC with Sequential Surrogate Likelihoods, or ABC-SSL. The method was inspired by Sequential Neural Likelihood (SNL) [17].

1.2 Aim of the Project

The main objective of this project is to implement, analyze and expand the suggested method of inference and compare it to similar methods in the literature, highlighting limitations and benefits. This is achieved through a simulation study including variants of ABC-SSL and a selection of other methods for reference.

The focus is on putting the proposed method into the context of ABC. Relevant methods for comparison include forms of ABC such as the widely popular ABC-SMC [25, 2] ABC-MCMC [14, 24] as well as SNL [17]. The ABC-SSL method was implemented and tested in a simulation study with three examples, including one of the standard models in the ABC literature, the g-and-k distribution.

1.3 Outline

In Chapter 2, we go through some methods from the ABC literature. Chapter 3 introduces the proposed method of inference. Chapter 4 contains three numerical experiments with the purpose of investigating the quality of inference in comparison to some related methods. Discussion and conclusions are found in Chapter 5.

2

Background/Theory

Bayesian statistics is a popular statistical paradigm stemming from the assignment of probabilities to parameters of statistical models, which could be said to model ones "belief" in said parameters. These parameter beliefs are then updated through the use Bayes theorem with account to observed data, or "evidence".

Throughout this thesis, we assume to be given a Bayesian model with observed data $x_0 \in \mathcal{X}$ generated from $x \sim f(\,\cdot\,|\,\theta)$ where $f(x_0\,|\,\theta)$ is the likelihood of parameters $\theta \in \Theta$. We also assume that a prior distribution π is given. In order to do parameter inference we are interested in the posterior distribution $\pi(\theta\,|\,x_0) \propto f(x_0\,|\,\theta)\pi(\theta)$. This can be used for estimation of quantities of interest $\mathbb{E}(\psi(\theta)\,|\,x_0) = \int_\Theta \psi(\theta)\ \mathrm{d}\pi(\theta\,|\,x_0)$ for some function $\psi : \Theta \to \mathbb{R}$.

2.1 Approximate Bayesian Computation

The traditional way of doing parameter inference for a Bayesian model is through the likelihood function. More complex models can sometimes have a likelihood function that is *intractable* by not being known in any useful analytic form or simply being prohibitively time-consuming to evaluate numerically. Methods of addressing such situations are sometimes referred to as likelihood-free inference. When the likelihood is intractable but one can simulate synthetic datasets x from the likelihood, it can be suitable to make use of approximate Bayesian computation (ABC).

To perform ABC for a Bayesian model with prior $\theta \sim \pi(\,\cdot\,)$ and data $x\,|\,\theta \sim f(\,\cdot\,|\,\theta)$, we select a summary statistic $S : \mathcal{X} \to \mathcal{S}$ with some metric ρ on $\mathcal{S}$. We denote the observed statistic $s_0 = S(x_0)$. The idea is to to approximate the posterior $\pi(\theta\,|\,s = s_0)$ with $\pi(\theta\,|\,s \approx s_0)$ through $\pi(\theta\,|\,s_0) \approx \pi_\delta(\theta\,|\,s_0) = \pi(\theta\,|\,\rho(s, s_0) \leq \delta)$ for some threshold $\delta > 0$. The distribution $\pi_\delta(\,\cdot\,|s_0)$ tends to the posterior when $\delta \to 0$, and to the prior when $\delta \to \infty$. Computational cost can increase drastically when $\delta \to 0$, so in practice we are forced to use a fixed δ that is taken to be as small as our computational resources and allow. When presenting the various ABC algorithms, it is assumed that the Bayesian model, observed data, summary statistics, and metric are fixed beforehand and are therefore not considered as input to the algorithms.

2.1.1 Rejection ABC

The simplest way of using ABC is through the original Rejection ABC algorithm from 1999 [26]. The procedure is shown in Algorithm 1. We assume that we have a joint density function $f_{s,\theta}(s,\theta)$ for $\theta \in \Theta \subset \mathbb{R}^p$, $s \in \mathbb{R}^q$ and wish to produce a sample from the ABC posterior $\theta_1^{(\delta)}, \ldots, \theta_n^{(\delta)} \sim \pi(\,\cdot \,|\, \rho(s, s_0) \leq \delta)$ where

$$\pi(\theta \,|\, \rho(s, s_0) \leq \delta) \propto \int_{\mathrm{B}_\delta(s_0)} f_{s,\theta}(s,\theta)\, \mathrm{d}s \tag{2.1}$$

where $\mathrm{B}_\delta(s_0)$ denotes the ball of radius δ around s_0.

Algorithm 1 R-ABC

Input: Sample size N, threshold δ.
Output: Sample $\theta_1, \ldots, \theta_N$ from the ABC posterior.
Set $n \leftarrow 0$.
while $n < N$ **do**
    Sample proposal $\theta' \sim \pi(\,\cdot\,)$.
    Simulate $x' \sim f(\,\cdot \,|\, \theta')$ from the model.
    $s' \leftarrow S(x')$
    if $\rho(s', s_0) \leq \delta$ **then**
        Accept proposal $\theta_n \leftarrow \theta'$
        $n \leftarrow n + 1$
    end if
end while

Assume that we wish to estimate $\psi_0 = \mathbb{E}(\psi(\theta) \,|\, x = x_0)$ for $\psi : \mathbb{R}^p \to \mathbb{R}$. Under sufficiency of S, we have $\psi_0 = \mathbb{E}(\psi(\theta) \,|\, s = s_0)$. Set $Y_n^{(\delta)} = \frac{1}{n}\sum_{j=1}^n \psi(\theta_j^{(\delta)})$. Theorem 1 shows that we have a type of convergence to ψ_0 as the sample size $n \to \infty$ and the threshold $\delta \to 0$ almost everywhere with respect to f_s which denotes the marginal distribution in s of $f_{s,\theta}$. One should note that the assumptions are quite weak and the assumption that we have a density $f_{s,\theta}$ is not necessary for this particular result [1].

Theorem 1 ([1, Proposition 3.1])**.** *Assume* $\mathbb{E}(|\psi(\theta)|) < \infty$. *Then, for* f_s*-a.e.* $s_0 \in \mathbb{R}^q$,

1. $\forall \delta > 0 \quad Y_n^{(\delta)} \longrightarrow \mathbb{E}(Y_n^{(\delta)})$ *a.s. as* $n \to \infty$;
2. $\forall n \in \mathbb{N} \quad \mathbb{E}(Y_n^{(\delta)}) \longrightarrow \mathbb{E}(\psi(\theta) \,|\, s = s_0)$ *as* $\delta \to 0$.

Although Theorem 1 seems to suggest that ABC works as soon as one can find a finite-dimensional sufficient statistic S, it is not clear how one should decrease δ. Some suggestions have been made in the literature for handling this problem, but it remains a challenge [1, 23]. Furthermore, it can be difficult in practice to find suitable S. There are techniques to aid in finding summary statistics for ABC but one often has to settle for approximate sufficiency and take into account the error that comes with it [16, 19].

2.1.2 Kernel ABC

Kernel ABC can be achieved by choosing some smoothing kernel K_δ with bandwidth δ and replacing the condition $\rho(s, s_0) \leq \delta$ in Algorithm 1 and instead accepting θ' as a sample from the posterior with probability proportional to $K_\delta(\rho(S(x'), s_0))$ after simulating synthetic data $x' \sim f(\cdot \,|\, \theta')$ from the model [3]. We use the abbreviation $\rho = \rho(S(x), s_0)$ which indicates that the ABC distances are the objects of most direct interest. We can consider the distances ρ as random variables with conditional distribution

$$\rho \,|\, \theta = \rho(S(x), s_0) \,|\, \theta \sim h(\cdot \,|\, \theta).$$

The goal is to obtain a sample from the ABC posterior

$$\pi_\delta(\theta \,|\, s_0) \propto \pi(\theta) \int_{\mathcal{X}} K_\delta(\rho(S(x), s_0)) f(x \,|\, \theta) \,\mathrm{d}x = \mathbb{E}_h(K_\delta(\rho) \,|\, \theta)\pi(\theta)$$

where $\mathbb{E}_h(K_\delta(\rho) \,|\, \theta)$ is the likelihood of θ.

K-ABC is a generalization of R-ABC since Algorithm 1 is retrieved by setting $K_\delta(\rho) = \mathbf{1}_{[0,\delta]}(\rho)$, called a uniform kernel. For the remainder of this thesis, we settle on using the uniform acceptance kernel, but we keep the representation of the ABC posterior in terms of K_δ for extensibility. Some discussion about other kernels can be found in Appendix A.

2.1.3 ABC-MCMC

Algorithm 1 is not always practically useful, since the acceptance probability can be extremely low for small δ. This is in part due to drawing the majority of parameter proposals from regions of low posterior density from the prior distribution. We would prefer for the algorithm to sample proposal parameters of high posterior density with higher frequency. This can be achieved through an adaptation of Markov chain Monte Carlo method to the ABC framework. Algorithm 2 presents ABC-MCMC with kernel K_δ (K-ABC-MCMC, see e.g. [24]). Theorem 2 states that the stationary distribution of the chain in θ is the ABC posterior. A proof can be found in Appendix A.

Theorem 2. *The stationary distribution of the Markov chain in θ from Algorithm 2 is $\pi_\delta(\theta \,|\, s_0) \propto \pi(\theta)\mathbb{E}_h(K_\delta(\rho) \,|\, \theta)$.*

2.1.4 ABC-SMC

One of the most popular ABC methods is called Sequential Monte Carlo ABC (ABC-SMC, [25]) which is presented in Algorithm 3. The idea of ABC-SMC is to use a sequence of proposal distributions π_r, $r = 1, \ldots, R$, which approximate the posterior distribution in order to iteratively improve the acceptance ratio $\mathbb{P}_r(\rho(s, s_0) \leq \delta_r)$

Algorithm 2 ABC-MCMC

1: **Input:** Chain length T, threshold δ, transition kernel $g(\theta' \mid \theta)$.
2: **Output:** Chain realization $\theta_1, \ldots, \theta_T$ with the ABC posterior as its stationary distribution.
3: Initialize (θ_1, x_1).
4: **for** $t = 1, \ldots, T$ **do**
5: Sample proposal $\theta' \sim g(\cdot \mid \theta_t)$.
6: Simulate $x' \sim f(\cdot \mid \theta')$ and set $\rho' \leftarrow \rho(S(x'), s_0)$.
7: Compute acceptance probability

$$\alpha \leftarrow \min\left(1, \frac{K_\delta(\rho')\pi(\theta')g(\theta_t \mid \theta')}{K_\delta(\rho_t)\pi(\theta_t)g(\theta' \mid \theta_t)}\right).$$

8: With probability α, set $(\theta_{t+1}, \rho_{t+1}) \leftarrow (\theta_t, x_t)$.
9: Otherwise set $(\theta_{t+1}, \rho_{t+1}) \leftarrow (\theta', \rho')$.
10: **end for**

while simultaneously decreasing the acceptance threshold δ_r. The idea of the algorithm is to maintain a "population" of parameter proposals that is updated sequentially. At each round r, new proposals are generated by drawing candidates from the previous population and perturbing them.

We set the initial proposal distribution π_0 to be the prior distribution. For rounds $r = 1, \ldots, R$, one instead chooses π_r as a mixture distribution by first sampling a proposal $\theta_{m'}^{(r-1)}$ from the population at round $r-1$ with probability proportional to $w_{m'}^{(r-1)}$ and perturbing it with some kernel $\theta' \sim g_r(\cdot \mid \theta_{m'}^{(r-1)})$, typically additive multivariate Gaussian noise. The weights are chosen to minimize the Kullback–Leibler divergence of the proposal distribution with respect to the target distribution as

$$w_m^{(r)} = \frac{\pi(\theta_m^{(r)})}{\sum_{j=1}^{M} w_j^{(r-1)} g_r(\theta_m^{(r)} \mid \theta_j^{(r-1)})}, \tag{2.2}$$

and with proposal distribution $g_r(\theta \mid \theta_m^{(r-1)}) = \mathrm{N}(\theta_m^{(r-1)}, 2\Sigma^{(r-1)})$ being the multivariate Gaussian centered in $\theta_m^{(r-1)}$ with variance being twice the sample variance of $(\theta_m^{(r-1)})_{m=1,\ldots,M}$, denoted $\Sigma^{(r-1)}$ [2, 8].

2.2 Gaussian Process Regression

Gaussian processes (GP) have proven to be a valuable modelling tool for regression, and their versatility has earned them considerable attention from the machine learning community. Below is a short presentation of the *Gaussian process regression* (GPR) based on the standard reference book by Rasmussen (2006, [22]).

Algorithm 3 ABC–SMC

Input: Number of rounds R, population size M.
Output: Sample $\theta_1, \ldots, \theta_M$ from the ABC posterior.
for r = 0, . . . , R **do**
  if $r > 0$ **then**
    Compute weights (2.2) for proposal distribution π_r.
    Set threshold δ_r and compute weights $w_1^{(r)}, \ldots, w_M^{(r)}$.
    Set parameters μ_r, Σ_r for perturbation kernel.
  end if
  $m \leftarrow 1$.
  while $m \leq M$ **do**
    Sample proposal $\theta \sim \pi_r(\,\cdot\,)$.
    Simulate synthetic data $x \sim f(\,\cdot \,|\, \theta)$.
    Compute $s \leftarrow S(x)$.
    if $\rho(s, s_0) \leq \delta_r$ **then**
      $\theta_m^{(r)} \leftarrow \theta$, $m \leftarrow m + 1$
    end if
  end while
end for

We say that a collection of random variables $\{f(x)\}_{x\in\mathcal{X}}$ is a Gaussian process on an index set $\mathcal{X}$ if $(f(x_1), \ldots, f(x_n))$ has multivariate normal distribution for any finite combination $x_1, \ldots, x_n \in \mathcal{X}$. Given a mean function $m : \mathcal{X} \to \mathbb{R}$ and non-negative and symmetric covariance kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ we denote $f \sim \mathrm{GP}(m, k)$ since these functions completely specify the process by

$$\boldsymbol{f} \sim \mathrm{N}(\boldsymbol{m}, \Sigma_k), \quad \boldsymbol{f} = (f(x_1), \ldots, f(x_n))^\intercal, \quad \boldsymbol{m} = (m(x_1), \ldots, m(x_n))^\intercal,$$

and $(\Sigma_k)_{i,j} = k(x_i, x_j)$.

Suppose we are given noisy observations of a Gaussian process $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$ where

$$\boldsymbol{y} \,|\, \boldsymbol{f} \sim \mathrm{N}(\boldsymbol{f}, \sigma^2 I_n), \quad \boldsymbol{f} \sim \mathrm{N}(\boldsymbol{m}, \Sigma_k),$$

with $\boldsymbol{m}$, $\boldsymbol{f}$ and Σ_k as above and where σ is the standard deviation of the normally distributed Gaussian noise, and $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix. We can predict the value at a new point $x_* \in \mathcal{X}$ by

$$f_* \sim \mathrm{N}(\boldsymbol{k}_*^\intercal(\Sigma_k + \sigma^2 I_n)^{-1}\boldsymbol{y},\, k(x_*, x_*) - \boldsymbol{k}_*^\intercal(\Sigma_k + \sigma^2 I_n)^{-1}\boldsymbol{k}_*).$$

where $\boldsymbol{k}_* = (k(x_*, x_1), \ldots, k(x_*, x_n))^\intercal$.

A typical example of covariance kernel in for a GP in $\mathbb{R}^d$ would be a squared exponential covariance kernel

$$k(x, x') = \eta_1 \mathrm{e}^{-(x-x')^\intercal(x-x')/2\eta_2}$$

for $x, x' \in \mathbb{R}^d$ some hyperparameters $\eta_1, \eta_2 > 0$. Hyperparameters can be chosen by maximization of the log marginal likelihood which is the conditional density

$$\log p(\boldsymbol{f}|x_1, \ldots, x_n) = -\tfrac{1}{2}\boldsymbol{f}^\intercal \Sigma_k^{-1}\boldsymbol{f} - \tfrac{1}{2}\log|\Sigma_k| - \tfrac{n}{2}\log 2\pi$$

with respect to η_1, η_2.

3

Sequential Surrogate Likelihoods

In order to make parameter inference with ABC methods it is often necessary to simulate a large number of synthetic datasets from the model. If this task is computationally demanding, e.g. for some complex dynamical systems in biology, one is forced to limit the the number of simulations, which in the worst case scenario could render such methods useless. In particular, if there is very little prior knowledge about parameters of the model of interest, one is condemned to simulate synthetic datasets with little information about the posterior distribution. The same effects could be caused by computational cost due to the computation of summary statistics or ABC distances [4]. This motivates the search of methods of decreasing the number of simulations needed for successful parameter inference. The hope is to extract as much information about the posterior as possible from each computed ABC distance. *Sequential neural likelihoods* is an example of a technique that can be used to address this problem, and it is one of the main inspirations for the method we propose here [17].

3.1 Main Algorithm

In Section 2.1, we saw that the goal of the ABC algorithms was to obtain a sample from the ABC posterior distribution $\pi_\delta(\theta \,|\, s_0) \propto \pi(\theta) L_\delta(s_0 \,|\, \theta)$ where the likelihood $L_\delta(s_0 \,|\, \theta)$ is given by

$$L_\delta(s_0 \,|\, \theta) = \mathbb{E}_h(K_\delta(\rho) \,|\, \theta). \tag{3.1}$$

If we knew the likelihood in an explicit form, we could target the ABC posterior distribution directly with MCMC, as opposed to using ABC-MCMC which can have low acceptance probability for small thresholds δ and can suffer when simulation or computation of ABC distances is expensive. It is typically not the case that $\pi_\delta(\theta \,|\, s_0)$ is known since not much is known about $h(\,\cdot \,|\, \theta)$, but we can introduce a model for the distances $\hat{\rho} \,|\, \theta \sim \hat{h}(\,\cdot \,|\, \theta)$, such that we can compute the corresponding ABC posterior density $\hat{\pi}_\delta(\theta \,|\, s_0) \propto \hat{L}_\delta(s_0 \,|\, \theta)\pi(\theta)$ analytically. We refer to $\hat{h}(\,\cdot \,|\, s_0)$ as the *surrogate model* and $\hat{L}_\delta(s_0 \,|\, \cdot\,)$ as the *surrogate likelihood*.

We propose ABC with sequential surrogate likelihoods (ABC-SSL) in Algorithm 4. This algorithm differs slightly from the one that was initially proposed, which can be seen in Algorithm 5 in Appendix B. These differences consist of removing

some details and a generalization by explicitly targeting ABC posteriors given by the randomness in the surrogate model.

Algorithm 4 ABC-SSL

1: **Input:** Number of rounds R, sample size M, family of surrogate models.
2: **Output:** ABC posterior sample $\theta_{R,1}, \ldots, \theta_{R,M}$ under the surrogate model.
3: Initialize $\pi_0 \leftarrow \pi$, $\mathcal{D}_0 \leftarrow \{\}$.
4: **for** $r = 1, ..., R$ **do**
5: **for** $m = 1, ..., M$ **do**
6: Sample $\theta_{r,m} \sim \pi_{r-1}(\,\cdot\,)$ (e.g., by MCMC).
7: Simulate $x_{r,m} \sim f(\,\cdot \,|\, \theta_{r,m})$.
8: Compute $\rho_{r,m} \leftarrow \rho(S(x_{r,m}), s_0)$.
9: **end for**
10: Let $\mathcal{D}_r \leftarrow \mathcal{D}_{r-1} \cup \{(\rho_{r,m}, \theta_{r,m}) : m = 1, \ldots, M\}$.
11: Fit $\hat{\rho} \,|\, \theta, \mathcal{D}_r \sim \hat{h}_r(\,\cdot \,|\, \theta)$ on $\mathcal{D}_r$.
12: Set threshold δ_r.
13: Set $\pi_r(\theta) \propto \pi(\theta) \mathbb{E}_{\hat{h}_r}(K_{\delta_r}(\hat{\rho}) \,|\, \theta)$.
14: **end for**

Much like in ABC-SMC, we update the proposal distribution $\pi_r(\theta)$ sequentially in an attempt to approximate the posterior distribution. We maintain a set of training data $\mathcal{D}_r = \{(\rho_{t,m}, \theta_{t,m}) : m = 1, \ldots, M,\ t = 1, \ldots, r\}$ on which we train our model $\hat{\rho} \,|\, \theta, \mathcal{D}_r \sim \hat{h}_r(\,\cdot \,|\, \theta)$ each round by choosing $\hat{h}_r(\,\cdot \,|\, \theta)$ from a predetermined family of distributions. We set the proposal distribution to be the ABC posterior under the surrogate model

$$\pi_r(\theta) \propto \hat{L}_r(\theta \,|\, s_0)\pi(\theta) = \mathbb{E}_{\hat{h}_r}(K_{\delta_r}(\hat{\rho}) \,|\, \theta)\pi(\theta),$$

after first choosing a suitable threshold δ_r.

In Line 6 of Algorithm 4, we want to generate M samples from the proposal distribution $\pi_r(\,\cdot\,)$. There are a multitude of Monte Carlo methods for sampling from a given distribution, but it can be difficult to find a suitable method with little prior knowledge of the target distribution. We choose to use the *Adaptive Metropolis* algorithm [10] with a large chain length N and pick a sample from the posterior by thinning the tail to achieve a reasonable effective sample size. For simplicity, we decide on computing the acceptance ratio α for the last M elements of the chain and then draw M points from the last $\lfloor M/\alpha \rfloor$ elements of the floor at uniform.

3.2 Implementation

Since extremely large distances are not very informative about the posterior distribution, it seems appropriate to suppress these by modeling the distances in log-scale.

Let us choose the model

$$\log(\hat{\rho}) \,|\, \theta, \mathcal{D}_r \sim \mathcal{N}(\mu_r(\theta), \sigma_r^2(\theta)), \tag{3.2}$$

that is, we let $\log(\hat{\rho})$ be normally distributed, conditionally on θ, with conditional mean $\mu_r(\theta)$ and conditional variance $\sigma_r^2(\theta)$. We can fit models of this form using, e.g., linear regression or Gaussian process regression. It could be practical in some situations to consider a model with no noise. This can be achieved in either case for this model by setting $\sigma_r^2(\theta) = 0$. Since we only consider the uniform kernel $K_\delta(\rho) = \mathbf{1}\{\rho \leq \delta\}$ we get a resulting surrogate likelihood

$$\hat{L}_r(s_0 \,|\, \theta) \propto \mathbb{P}_{\hat{h}_r}(\hat{\rho} \leq \delta_r \,|\, \theta) = \Phi\Big(\frac{\log(\delta_r) - \mu_r(\theta)}{\sigma_r(\theta)}\Big)$$

where Φ denotes the cumulative distribution function of the standard normal distribution.

Having decided on a family of surrogate models of the form (3.2), we need to find a way of fitting such a model on data $\mathcal{D}_r$. A simple idea is to fit a linear model on $\mathcal{D}_r$. Here we choose to do polynomial regression, so that the mean $\mu_r(\theta)$ is a polynomial of degree d. In this case, we will have constant noise $\sigma_r^2(\theta) = \sigma_r^2$ which we can choose to ignore by setting $\sigma_r^2 = 0$.

If we choose surrogate model by training a Gaussian process $\log(\hat{\rho}) \sim \mathrm{GP}(m_r, k_r)$ with observational noise on $\mathcal{D}_r$, for some mean function m_r and covariance kernel k_r we get $\mu_r(\theta)$ and $\sigma_r^2(\theta)$ that vary in θ where. The posterior variance $\sigma_r^2(\theta)$ will be larger in regions of parameter space with fewer observations which can be interpreted as uncertainty caused by lack of evidence. We train the GP surrogate model by choosing a constant mean and setting the covariance kernel to be of the squared exponential type. It should be noted that the choice of covariance kernel here is quite arbitrary and may affect the inference. To mitigate this, we optimize the hyperparameters by maximizing the marginal likelihood every few rounds as the model is trained. There are software packages available that can automate the task of training, predicting and optimizing hyper-parameters with gradient based optimization method for some common choices of mean function and covariance kernel [21].

4

Numerical Experiments and Results

We start off with a simple example with one single parameter in Section 4.1, move on to a common test model from the ABC literature in Section 4.2 and finish with an example with real world data in Section 4.3.

4.1 Example 1: Explicit Distribution of Distances

Consider a Bayesian model with prior $\theta \sim \text{Uni}([-1, 1])$ and ABC distances with distribution $\rho \,|\, \theta \sim \text{N}(\theta^2, 0.1)$. The reason to consider a model like this is that the distances can easily be modeled as a Gaussian process and estimated with GPR. It is also possible to find the true ABC posteriors as well as the ABC posteriors under the surrogate models which allows for good visualizations.

We run the ABC-SSL algorithm for $R = 10$ rounds and use GPR for the surrogate model. Each round, $M_r = 30$ training points are sampled from $\pi_r(\,\cdot\,)$ and included into the training data. The training data was initialized with $M_0 = 100$ parameter proposals and generated ABC distances. After each round, the threshold was updated to be the smallest ABC distance in the training data

$$\delta_{r+1} = \min\{\, \rho_{t,m} : t = 1, \ldots, r,\, m = 1, \ldots, M_t \,\}.$$

We use uniform kernel $K_\delta(\rho) = \mathbf{1}(\rho \leq \delta)$ leading to the ABC posterior $\pi_\delta(\theta \,|\, s_0) \propto \mathbb{E}_h(K_\delta(\rho) \,|\, \theta) = \mathbb{P}_h(\rho \leq \delta \,|\, \theta)$, which implies

$$\pi_\delta(\theta \,|\, s_0) = \frac{\Phi\big(\dfrac{\log(\delta) - \theta^2}{0.1}\big)}{\displaystyle\int_{-1}^{1} \Phi\big(\frac{\log(\delta) - \theta^2}{0.1}\big)\, \mathrm{d}\theta}.$$

By Theorem 4 in Appendix A, the posterior distribution $\pi(\,\cdot \,|\, s_0)$ is a point mass in $\theta = 0$.

Figure 4.1 shows the results of running ABC-SSL for 10 rounds. It should be noted that not much improvement in the surrogate model nor in the ABC posterior is seen after the first few rounds.

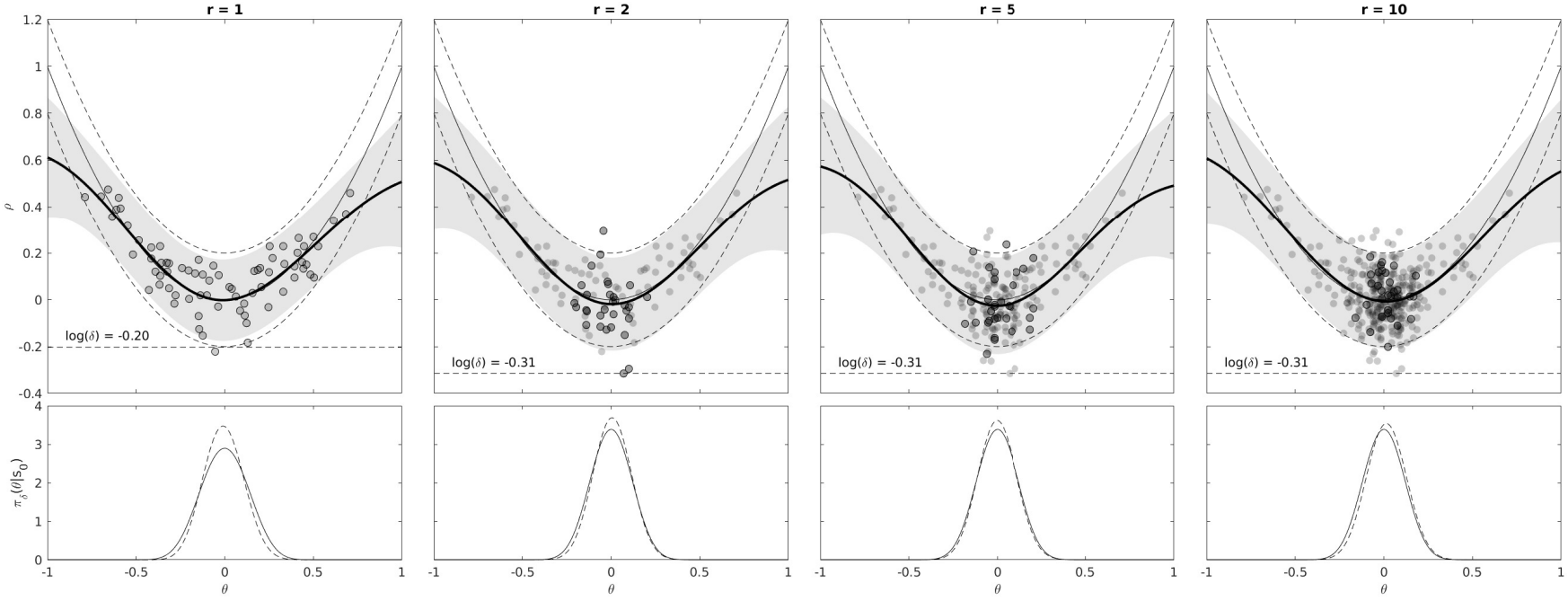


Figure 4.1: Result of ABC-SSL after 1, 2, 5, and 10 rounds of ABC-SSL. The thick line is the predicted mean $\mu_r(\theta)$ and the shaded area shows confidence bounds $\mu_r(\theta) \pm 2\sigma_r(\theta)$. The quadratic curves show the corresponding curves under the true model. Training data is shown as points and the points that are new for the round are indicated with black circles. The lower plots show the ABC posterior under the true model and under the surrogate model (dashed).

4.2 Example 2: The g-and-k Distribution

The g-and-k distribution is a simple distribution with intractable likelihood which is often used to test ABC methods [7]. The distribution is defined by its quantile function

$$q(r) = A + B\Big(1 + 0.8\frac{1 - \mathrm{e}^{-gz(r)}}{1 + \mathrm{e}^{-gz(r)}}\Big)(1 + z(r)^2)^k z(r), \quad r \in [0, 1],$$

where $z(r)$ is the r-quantile of the standard normal distribution. Samples can be generated by sampling from a normal distribution and plugging the value into this defining formula.

We define our model by letting the data x consist of $n = 2000$ independent draws from the g-and-k distribution with parameters $\theta = (A, B, g, k)$. As observed data, we generate from the model with true parameters $\theta_0 = (3, 1, 2, 0.5)^\intercal$. We choose prior distribution $\theta \sim \mathrm{Uni}([-10, 10] \times [0, 10] \times [-10, 10] \times [0, 10])$. As summary statistics, we let $S(x)$ be robust moment estimates $s_1, \ldots, s_4$ computed from the octiles $q_1, \ldots, q_7$ of x as

$$s_1 = q_1, \quad s_2 = q_6 - q_2, \quad s_3 = (q_6 + q_2 - 2q_4)/s_2, \quad s_4 = (q_7 - q_5 + q_3 - q_1)/s_2,$$

which is a common choice for this model [6]. To choose a suitable metric ρ, we first sample 1000 parameter points from the prior distribution, simulate data and compute summary statistics for each parameter. We set ρ to be the scaled Euclidean distance where each summary statistic s_i is scaled with the component-wise inverted median absolute deviations computed from the simulated summary statistics. This scaling can avoid the situation where one summary statistic dominates the others leading to a less informative metric [18].

We now present the results of running the proposed method for $R = 15$ rounds with $M = 1000$ samples from each round. The surrogate likelihood is chosen from the posterior mean after using Gaussian process regression for the surrogate model. For comparison, we also run the algorithm with a 4th degree polynomial as the mean of the surrogate model. At each round, we sample $M = 100$ parameters from the approximate posterior induced by our surrogate likelihood using MCMC. In Figure 4.3, we see the final approximate posterior distributions. Figure 4.2 shows the marginal distributions of the training data from rounds 1 through 15. The ABC posteriors shown are correspond to the same metric ρ (scaled Euclidean), summary statistics and ABC threshold δ, so the inaccuracies stem mainly from modelling error. That is, the surrogate model $\log(\rho) \,|\, \theta \sim \mathrm{N}(\mu(\theta), \sigma(\theta)^2)$ was not optimal. From Theorem 4, it is clear that the distances can not follow this type of model since it would imply that the posterior distribution is degenerate, which, judging by Figure 4.3, it is not. The surrogate model is particularly unreasonable for parameters close to the true posterior, since the tails of a normal distribution are too light. This is illustrated in Figure 4.4 where ABC distances were drawn from the true distance distribution $h(\,\cdot \,|\, \theta)$ for a few choices of θ. Finally, one should note that although introducing a surrogate model significantly decreased the precision of the inference,

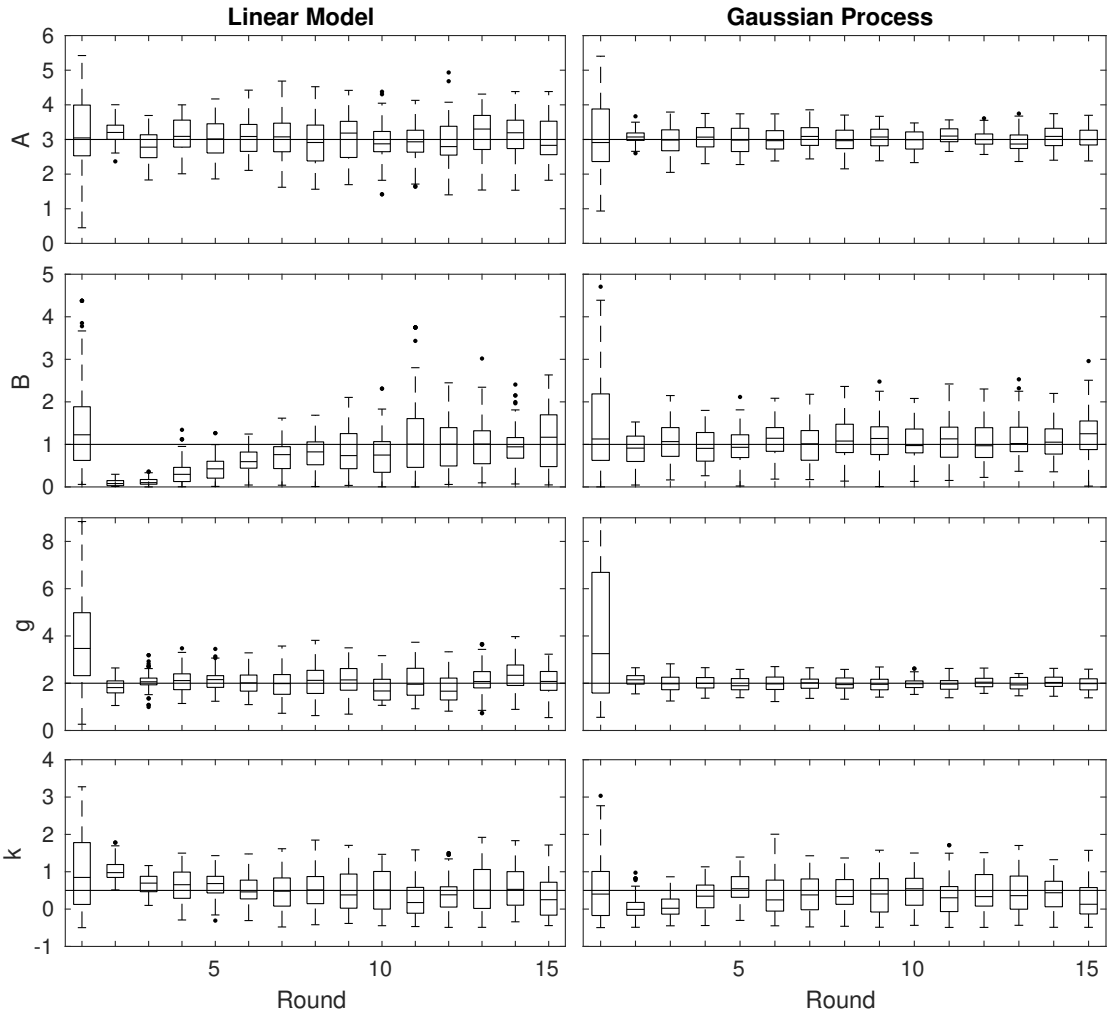


Figure 4.2: Marginal distributions of the training data for each round of ABC-SSL for the g-and-k model.

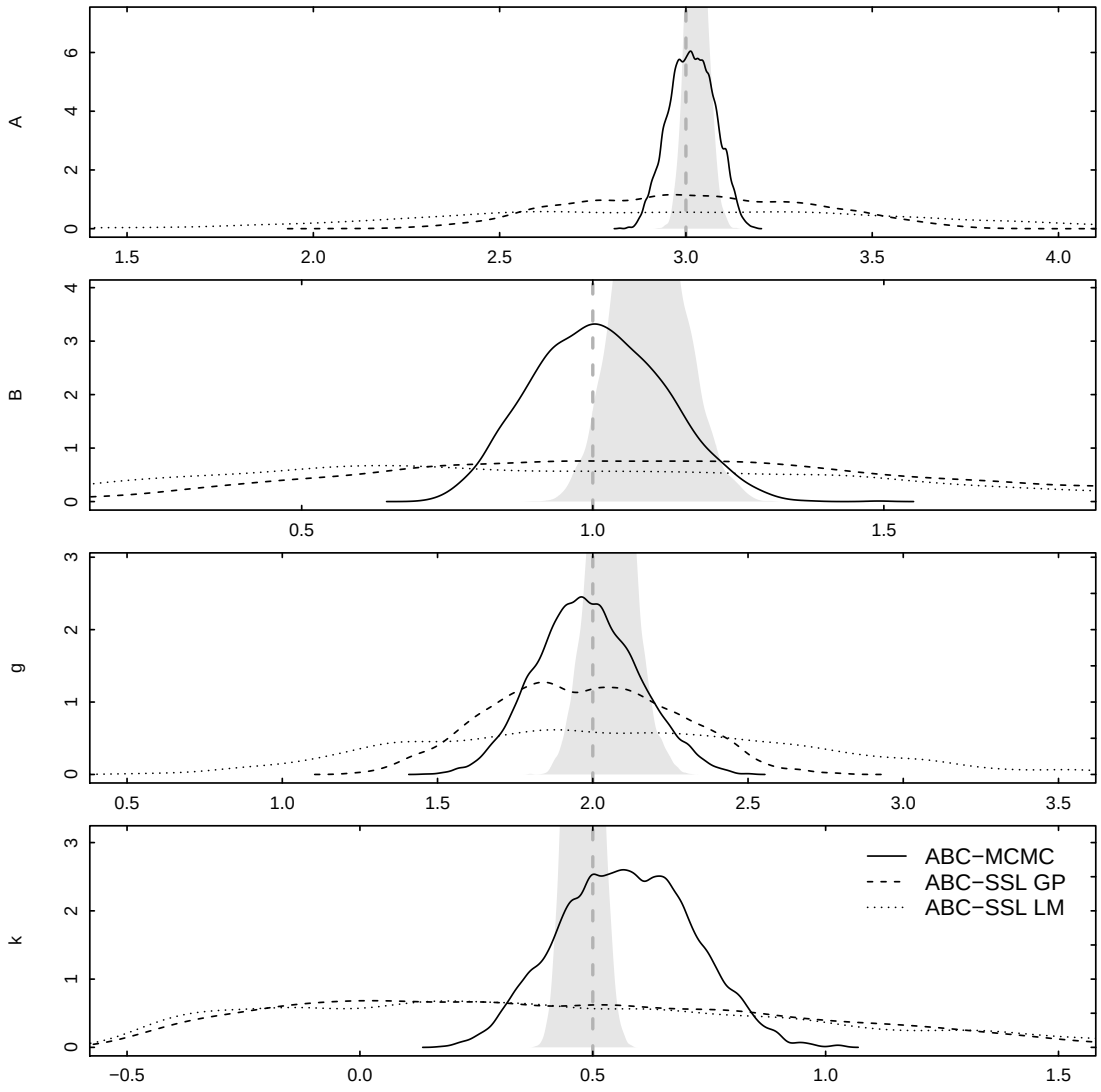


Figure 4.3: Approximate marginal posterior densities corresponding to ABC threshold $\delta = 0.125$ for the g-and-k model compared to the true posterior (shaded). The vertical lines indicate the parameter value used to generate the observed data.

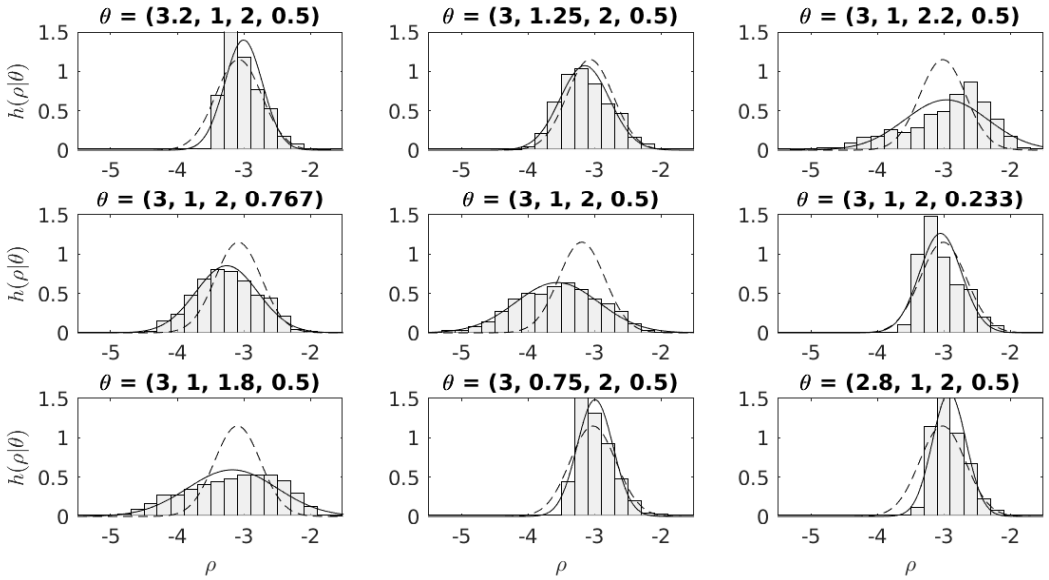


Figure 4.4: Surrogate model fit for the g-and-k model. The histograms show distances computed from repeated model simulations at the true parameters (middle) and two approximate posterior standard deviations away in each parameter component. On top of the histograms are the distance distributions under the final surrogate model $\hat{h}_R(\,\cdot\,|\,\theta)$ (dashed) as well MLE normal densities based on the histogram data.

the number of simulations of the model needed was only 2500 in the end. This could have been a good trade-off if the simulations were extremely computationally expensive.

Table 4.1: Concentrations of sulfur dioxide (µg/m^3) in the air measured at Marylbone Road, London, between 1998–01–01 and 2005–06–23.

Concentration	0	3	5	8	11	13	16	19	21
Count	2180	8196	9032	8300	7432	6530	5172	4049	2997
Concentration	24	27	29	32	35	37	40	43	>43
Count	2237	1576	1105	782	506	363	233	189	656

4.3 Example 3: Real-world Data Example

We wish to fit a g-and-k model to 61535 sulfur dioxide concentrations measured at Marylebone Road, London, between 1998–01–01 and 2005–06–23 made available through the Automatic Urban and Rural Network [5]. To model these data as i.i.d. observations from a g-and-k distribution serves little practical purpose, but it constitutes a suitable example due to the computational cost that comes with the large data size [11]. We use the same setup as in the previous example but we let the summaries $S(x)$ be the octiles $q_1, \ldots, q_7$ of x and we let ρ be the Euclidean metric without any scaling of the summaries. As prior distribution, we choose $\theta \sim \text{Uni}([0, 20], [0, 20], [-5, 5], [0, 5])$.

The data only takes certain integer values, some of which can be seen in Table 4.1. The peculiar values are likely to have been caused by nearest integer rounding and conversion $x_i = \text{round}(w\,\text{round}(z_i))$, $i = 1, \ldots, 61535$, where z_i are the true concentrations, x_i are the observations in the model, and w is some conversion factor (units omitted). From the full data, after assuming that $3 = \text{round}(w)$, the conversion factor w can be inferred to lie in $[133/50, 149/56]$, say $w = 2.6605$.

This view of the data gives rise to an alternative model, namely a multinomial model parametrized by the g-and-k distribution parameters $\theta = (A, B, g, k)$,

$$(y_0, \ldots, y_J) \,|\, \theta \sim \text{Multinomial}\left(p_0(\theta), p_1(\theta), \ldots, p_J(\theta)\right)$$

where y_j is the bin count for bin j and

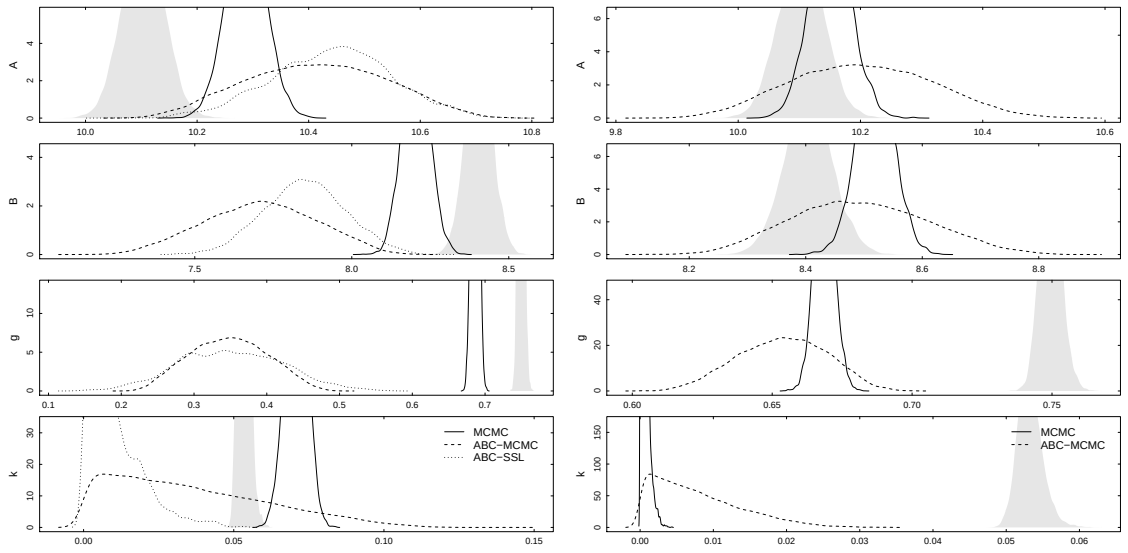
$$p_j(\theta) = \mathbb{P}_\theta(\hat{x}_{j-1} < x \leq \hat{x}_j), \quad j = 1, \ldots, J, \quad p_0(\theta) = 1 - \sum_{j=1}^{J} p_j(\theta),$$

with bin edges $\hat{x}_0 = 0$, $\hat{x}_j = (2j - 1)w/2$, $j = 1, \ldots, J$. The log-likelihood can be written $\ell(\theta) = \sum_{j=1}^{J} y_j \log(p_j(\theta))$ so there is no need for ABC although computation of the bin probabilities $p_j(\theta)$ requires inversion of the g-and-k quantile function. Nevertheless, for comparison, we employ ABC for this model by letting the summaries $S(y)$, $y = (y_1, \ldots, y_J)$, be be the bin counts after first combining bins to reduce computational burden, ending up with bin edges $kw/2$ for $k = 0, 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 23, \infty$. We do not scale the statistics in the ABC distances.

Algorithm 4 was run for $R = 8$ rounds with a GP surrogate model. The ABC threshold δ was decreased to be in the 0.1 quantile of the training data distances from the previous round. The parameter samples in the training data for each round can be seen in Figure 4.7. In Figure 4.5, we see the ABC posterior densities compared to the true posterior densities found with MCMC. For ABC-MCMC, the ABC threshold δ_r was chosen to achieve an acceptance probability of approximately 15%. The difference in the location of the posterior densities in parameters g and k for the multinomial model with and without grouping of bins can be attributed to the influence of very large concentrations x_i. In many cases the true density lies far out in the tail of the corresponding ABC densities which is partially due to the large sample size leading to narrow posterior distributions. Little weight should be given the ostensible similarity of the posterior densities for ABC-SSL and ABC-MCMC since they do not use the same ABC thresholds.

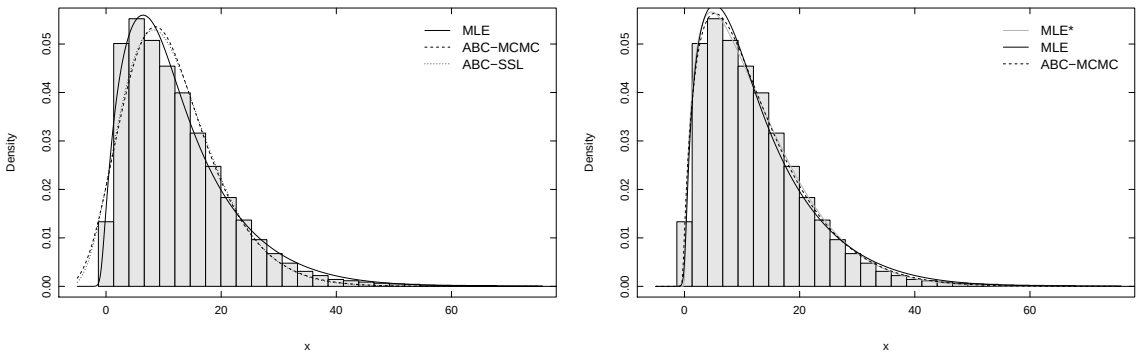
Figure 4.6 shows the g-and-k densities compared to the observed data. For the ABC methods, point-estimates were taken using minimum mean square error estimates (MMSE) (or posterior mean). The ABC methods for the raw g-and-k model seem have produced point-estimates that are slightly skewed in relation to the observed data. This could be a result from non-informative ABC summaries or the misspecification of the model leading to poor inference. It should be noted that the g-and-k distribution is distributed over $\mathbb{R}$ so it is not a suitable choice for non-negative data despite the figures suggesting a decent fit.

In this example, the computational benefit of using a surrogate model for the ABC distances is not justifiable, since the computational burden can be addressed by mindful modelling choices. An important take-away is that it is often preferable with exact inference for an approximate model instead of approximate inference for a more exact model.



(a) g-and-k model **(b)** Multinomial model

Figure 4.5: Approximate marginal posterior densities for the models for the UK air data. The shaded regions show the posterior density for the multinomial model without grouping of bins which are included for reference.



(a) g-and-k model **(b)** Multinomial model

Figure 4.6: Histograms of sulfur dioxide concentrations with g-and-k densities on top for parameters estimated using various techniques. Point estimates for ABC were chosen as MMSE. The light gray curve in (b) corresponds to MLE without grouping of the bins.

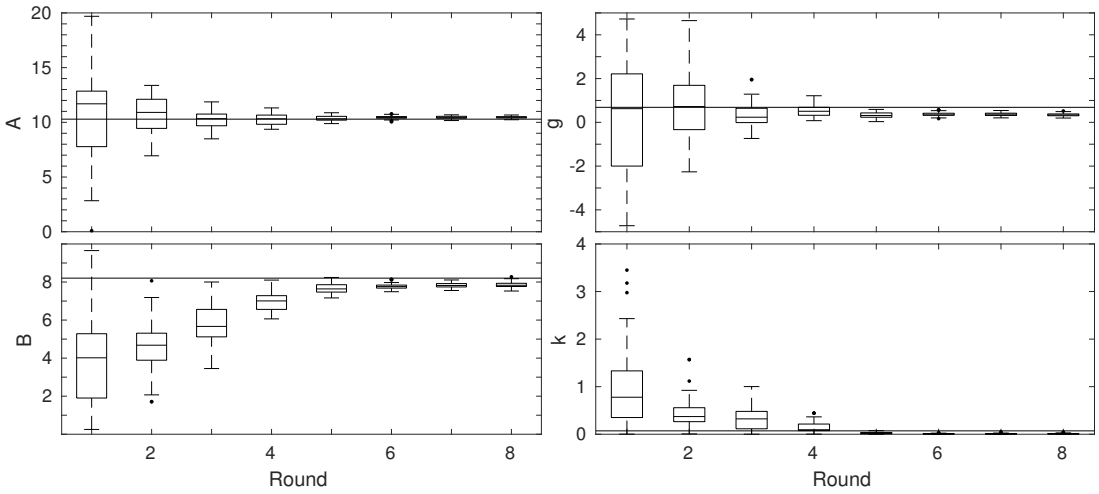


Figure 4.7: Marginal distributions of the training data samples for each round of ABC-SSL with GP surrogate model for the UK air data example. The horizontal lines show the parameter MLE for the raw g-and-k model.

5

Conclusions

5.1 Discussion

The main difficulty with ABC-SSL lies in finding a suitable surrogate model $\hat{\rho} \mid \theta, \mathcal{D}_r \sim \hat{h}_r(\cdot \mid \theta)$. The Gaussian process surrogate model works well enough for the examples shown here, but this is not the case in general. This can easily be seen since Theorem 4 in Appendix A shows that the ABC posterior under the surrogate model degenerates to a point weight. One should take this into account by not selecting too small thresholds δ. In the examples we did this somewhat carelessly by setting δ to be some quantile of the observed distances generated from the model. The GPR framework provides more flexibility than the simplistic models chosen here, so it could still be more useful. Perhaps a technique such as autoregressive flow which was proposed to use for SNL could be more suitable as a model for ABC distances [17].

One should note that the normality assumption that comes with using a linear model or Gaussian process regression is not properly justified theoretically. Ideally one should be able to refer to a central limit theorem to use such a model. This is possible in some cases when modelling the summary statistics instead since these can be constructed to be approximately normal (with the potential loss of descriptive power) [28, 15].

Under some regularity assumptions, the GP posterior mean will converge to the conditional expectation $\mathbb{E}_h(\log(\rho) \mid \theta)$ [22]. Unfortunately, this is not enough for this surrogate model to be suitable with much generality. One could instead model acceptances, which will have Bernoulli distribution conditional on θ. In this case the posterior mean is all that is needed to compute the approximate posterior, but the problem with doing so is that one also disregards much of the information in the distances.

If one instead considers deterministic simulator models, using the GP posterior mean would suffice, but in this case, something like a Bayesian optimization approach could be more suitable [9].

5.2 Conclusions

The method was found to approximately locate the posterior distribution for the considered models with significantly fewer synthetic dataset simulations than traditional ABC, but was not able to match the precision of the ABC posterior due to inaccurate surrogate models.

Further insight is needed to achieve reliable inference with this method of inference in practice, but it can be used to find an initial estimate of a posterior distribution which can speed up the inference with traditional ABC methods.

Bibliography

[1] Stuart Barber, Jochen Voss, Mark Webster, et al. The rate of convergence for approximate bayesian computation. *Electronic Journal of Statistics*, 9(1):80–105, 2015.

[2] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[3] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[4] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Inference in generative models using the wasserstein distance. *arXiv preprint arXiv:1701.05146*, 2017.

[5] David C Carslaw and Karl Ropkins. openair — an r package for air quality data analysis. *Environmental Modelling & Software*, 27–28(0):52–61, 2012.

[6] Christopher C Drovandi and Anthony N Pettitt. Likelihood-free bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, 55(9):2541–2556, 2011.

[7] Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

[8] Sarah Filippi, Chris P Barnes, Julien Cornebise, and Michael PH Stumpf. On optimality of kernels for approximate bayesian computation using sequential monte carlo. *Statistical applications in genetics and molecular biology*, 12(1):87–107, 2013.

[9] Michael U Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *The Journal of Machine Learning Research*, 17(1):4256–4302, 2016.

[10] Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[11] George Karabatsos and Fabrizio Leisen. Bayes calculations from quantile implied likelihood. *arXiv preprint arXiv:1802.00796*, 2018.

[12] George Karabatsos, Fabrizio Leisen, et al. An approximate likelihood perspective on abc methods. *Statistics Surveys*, 12:66–104, 2018.

[13] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

[14] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[15] Edward Meeds and Max Welling. Gps-abc: Gaussian process surrogate approximate bayesian computation. *Uncertainty in Artificial Intelligence*, 30:593–602, 2014.

[16] Matthew A Nunes and David J Balding. On optimal selection of summary statistics for approximate bayesian computation. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

[17] George Papamakarios, David C Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. *arXiv preprint arXiv:1805.07226*, 2018.

[18] Dennis Prangle et al. Adapting the abc distance function. *Bayesian Analysis*, 12(1):289–309, 2017.

[19] Dennis Prangle, Paul Fearnhead, Murray P Cox, Patrick J Biggs, and Nigel P French. Semi-automatic selection of summary statistics for abc model choice. *Statistical applications in genetics and molecular biology*, 13(1):67–82, 2014.

[20] Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[21] Carl Edward Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (gpml) toolbox. *Journal of machine learning research*, 11(Nov):3011–3015, 2010.

[22] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian process for machine learning*. MIT press, 2006.

[23] Daniel Silk, Sarah Filippi, and Michael PH Stumpf. Optimizing threshold-schedules for sequential approximate bayesian computation: applications to molecular systems. *Statistical applications in genetics and molecular biology*, 12(5):603–618, 2013.

[24] Scott A Sisson and Yanan Fan. Likelihood-free markov chain monte carlo. *Handbook of Markov Chain Monte Carlo. Chapman and Hall*, 2011.

[25] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

[26] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

[27] Richard David Wilkinson. Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12(2):129–141, 2013.

[28] Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102, 2010.

A

Theoretical Details

A.1 Limit Distributions for the Considered Surrogate Models

In order to analyze the ABC posteriors under the surrogate model (3.2), we need a simple lemma about the tail behavior of the cumulative distribution function $\Phi : \mathbb{R} \to (0, 1)$ of the standard normal distribution.

Lemma 3. *Let $a_1, a_2 > 0$, $b_1, b_2 \in \mathbb{R}$. Then*

$$\frac{\Phi(a_1 x + b_1)}{\Phi(a_2 x + b_2)} \longrightarrow 0$$

as $x \to -\infty$ if one of the following is true:

1. *$a_1 > a_2$;*
2. *$a_1 = a_2$ and $b_1 < b_2$.*

Proof. Integration by parts gives that

$$\begin{aligned}\int_{-\infty}^{x} \mathrm{e}^{-y^2/2}\,\mathrm{d}y &= \int_{-\infty}^{x} \Big(-\frac{1}{y}\Big)\Big(-y\mathrm{e}^{-y^2/2}\Big)\,\mathrm{d}y \\ &= -\frac{1}{x}\mathrm{e}^{x^2/2} - \int_{-\infty}^{x} \frac{1}{y^2}\mathrm{e}^{-y^2/2}\,\mathrm{d}y\end{aligned}$$

and, similarly

$$\Big|\int_{-\infty}^{x} \frac{1}{y^2}\mathrm{e}^{-y^2/2}\,\mathrm{d}y\Big| \leq \frac{1}{|x|^3}\int_{-\infty}^{x}\Big(-y\mathrm{e}^{-y^2/2}\Big)\,\mathrm{d}y = \frac{1}{|x|^3}\mathrm{e}^{-x^2/2}.$$

Hence $\Phi(x) = \frac{1}{\sqrt{2\pi}|x|}\mathrm{e}^{-x^2/2} + \mathrm{O}\big(|x|^{-3}\mathrm{e}^{-x^2/2}\big)$. Furthermore,

$$\begin{aligned}\frac{\Phi(a_1x+b_1)}{\Phi(a_2x+b_2)} &= \frac{|a_1x+b_1|^{-1}\mathrm{e}^{-(a_1x+b_1)^2/2} + \mathrm{O}(|a_1x+b_1|^{-3}\mathrm{e}^{-(a_1x+b_1)^2/2})}{|a_2x+b_2|^{-1}\mathrm{e}^{-(a_2x+b_2)^2/2} + \mathrm{O}(|a_2x+b_2|^{-3}\mathrm{e}^{-(a_2x+b_2)^2/2})} \\ &= \frac{|a_2x+b_2|}{|a_1x+b_1|}\mathrm{e}^{(a_2^2-a_1^2)x^2/2}\mathrm{e}^{(a_2b_2-a_1b_1)x}\mathrm{e}^{(b_2^2-b_1^2)/2}\frac{1+\mathrm{O}(|a_1x+b_1|^{-2})}{1+\mathrm{O}(|a_2x+b_2|^{-2})},\end{aligned}$$

so $\Phi(a_1x+b_1)/\Phi(a_2x+b_2) \to 0$ as $x \to -\infty$ if $(a_2^2-a_1^2)x^2/2 + (a_2b_2-a_1b_1)x \to -\infty$ as $x \to -\infty$. This occurs if and only if one of the conditions hold, so the proof is done. □

Consider Algorithm 4 with ABC distances $\log(\rho) \,|\, \theta \sim \mathrm{N}(\mu(\theta), \sigma(\theta))$ for continuous functions μ, σ. The following theorem states that under some simple assumptions, the posterior distribution corresponding to these

Theorem 4. *Consider the ABC posterior distribution corresponding to the surrogate model* (3.2)*,*

$$\pi_\delta(\theta \,|\, s_0) = \frac{\Phi\big(\frac{\log(\delta)-\mu(\theta)}{\sigma(\theta)}\big)\pi(\theta)}{\int_\Theta \Phi\big(\frac{\log(\delta)-\mu(\theta)}{\sigma(\theta)}\big)\pi(\theta)\,\mathrm{d}\theta}.$$

for continuous functions μ, σ.

(i) Take constant $\sigma(\theta) = \sigma_0$, and let $\mu(\theta_0) = \mu_0$ be a unique minimum with $\pi(\theta_0) > 0$ such that for all $\varepsilon > 0$ there exists $\mu_1 > \mu_0$ such that $\mu(\theta) \geq \mu_1$ whenever $\theta \in \Theta \setminus \mathrm{B}_\varepsilon(\theta_0)$. Then $\pi_\delta(\cdot|s_0)$ converges weakly to a point weight in θ_0.

(ii) Let $\mu(\theta) \geq \mu_0$ and let $\sigma(\theta_0) = \sigma_0$ be a unique maximum with $\pi(\theta_0) > 0$ such that for all $\varepsilon > 0$ there exists $\sigma_1 < \sigma_0$ such that $\sigma(\theta) \leq \sigma_1$ whenever $\theta \in \Theta \setminus \mathrm{B}_\varepsilon(\theta_0)$. Then $\pi_\delta(\cdot|s_0)$ converges weakly to a point weight in θ_0.

Proof. In both cases we show that for arbitrarily small $\varepsilon > 0$,

$$\frac{\mathbb{P}_\delta(\theta \notin \mathrm{B}_\varepsilon(\theta_0))}{\mathbb{P}_\delta(\theta \in \mathrm{B}_\varepsilon(\theta_0))} \to 0$$

as $\delta \to 0$, which ensures the sought convergence.

(i) Fix arbitrary $\varepsilon > 0$ and take μ_1 as in the statement. Choose $\mu_2 \in (\mu_0, \mu_1)$ and take $\varepsilon' \in (0, \varepsilon)$ such that $\mu(\theta) < \mu_2$ for all $\theta \in \mathrm{B}_{\varepsilon'}(\theta_0)$. Such μ_2 exists by

continuity of μ. We get

$$\begin{aligned}\frac{\mathbb{P}_\delta(\theta \notin \mathrm{B}_\varepsilon(\theta_0))}{\mathbb{P}_\delta(\theta \in \mathrm{B}_\varepsilon(\theta_0))} &= \frac{\int_{\Theta\setminus\mathrm{B}_\varepsilon(\theta_0)} \pi_\delta(\theta \,|\, s_0)\ \mathrm{d}\theta}{\int_{\mathrm{B}_\varepsilon(\theta_0)} \pi_\delta(\theta \,|\, s_0)\ \mathrm{d}\theta} \\ &= \frac{\int_{\Theta\setminus\mathrm{B}_\varepsilon(\theta_0)} \Phi\big(\frac{\log(\delta)-\mu(\theta)}{\sigma_0}\big)\pi(\theta)\ \mathrm{d}\theta}{\int_{\mathrm{B}_\varepsilon(\theta_0)} \Phi\big(\frac{\log(\delta)-\mu(\theta)}{\sigma_0}\big)\pi(\theta)\ \mathrm{d}\theta} \\ &\leq \frac{\Phi\big(\frac{\log(\delta)-\mu_1}{\sigma_0}\big)\int_{\Theta\setminus\mathrm{B}_\varepsilon(\theta_0)} \pi(\theta)\ \mathrm{d}\theta}{\Phi\big(\frac{\log(\delta)-\mu_2}{\sigma_0}\big)\int_{\mathrm{B}_{\varepsilon'}(\theta_0)} \pi(\theta)\ \mathrm{d}\theta} \longrightarrow 0\end{aligned}$$

as $\delta \to 0$ by Lemma 3.

(ii) By the same argument as above, we take arbitrary $\varepsilon > 0$ and take $\sigma_1 < \sigma_0$ as in the statement. Likewise, we choose $\sigma_2 \in (\sigma_1, \sigma_0)$ and take $\varepsilon' > 0$ such that $\sigma(\theta) > \sigma_2$ for all $\theta \in \mathrm{B}_{\varepsilon'}(\theta_0)$. Let $\mu_1 = \sup_{\theta\in\mathrm{B}_{\varepsilon'}(\theta)} \mu(\theta)$. Without loss of generality, we assume that $\log(\delta) < \mu_0$. We get

$$\begin{aligned}\frac{\mathbb{P}_\delta(\theta \notin \mathrm{B}_\varepsilon(\theta_0))}{\mathbb{P}_\delta(\theta \in \mathrm{B}_\varepsilon(\theta_0))} &= \frac{\int_{\Theta\setminus\mathrm{B}_\varepsilon(\theta_0)} \Phi\big(\frac{\log(\delta)-\mu(\theta)}{\sigma(\theta)}\big)\pi(\theta)\ \mathrm{d}\theta}{\int_{\mathrm{B}_\varepsilon(\theta_0)} \Phi\big(\frac{\log(\delta)-\mu(\theta)}{\sigma(\theta)}\big)\pi(\theta)\ \mathrm{d}\theta} \\ &\leq \frac{\Phi\big(\frac{\log(\delta)-\mu_1}{\sigma_1}\big)\int_{\Theta\setminus\mathrm{B}_\varepsilon(\theta_0)} \pi(\theta)\ \mathrm{d}\theta}{\Phi\big(\frac{\log(\delta)-\mu_2}{\sigma_2}\big)\int_{\mathrm{B}_{\varepsilon'}(\theta_0)} \pi(\theta)\ \mathrm{d}\theta} \longrightarrow 0\end{aligned}$$

as $\delta \to 0$ by Lemma 3.

□

A.2 Other ABC Kernel Functions

Below we see the surrogate likelihood under the model (3.2) for some of the most common choices of ABC kernel K_δ. For the sake of readability, abbreviate $\mu = \mu_r(\theta)$, $\sigma^2 = \sigma_r^2(\theta)$ and $\rho = \hat{\rho}$. We will now see that many of the popular ABC posteriors have analytical form under this model.

A.2.1 Uniform Kernel

Letting $K_\delta(\rho) = \mathbf{1}\{\rho \leq \delta\}$ is the simplest choice of ABC kernel which makes for a good example. We get approximate likelihood

$$L_r(s_0 \,|\, \theta) = \mathbb{P}(\log(\rho) \leq \log(\delta)) = \Phi\Big(\frac{\log(\delta) - \mu}{\sigma}\Big).$$

If we set $\sigma = 0$, we instead get

$$L_r(s_0 \,|\, \theta) = \mathbf{1}\{\mu \leq \log(\delta)\}.$$

That is, assuming a uniform prior, the approximate posterior will be a uniform distribution on some subset of Θ. In the case when $\mu_r(\theta)$ is a polynomial of degree k and $\Theta \subset \mathbb{R}$, the subset will be a finite collection of intervals whose ends are the zeroes of a polynomial of degree k, so at most $k/2+1$ intervals for intervals of finite length. This severely limits the class of possible target distributions.

A.2.2 Epanechnikov Kernel

Let $K_\delta(\rho) = (1 - \rho^2/\delta^2)\mathbf{1}\{\rho \leq \delta\}$. The resulting surrogate likelihood can be found by using the substitution $z = (\log(\rho) - \mu - \sigma^2)/\sigma$,

$$L_r(s_0 \,|\, \theta) = \Phi\Big(\frac{\log(\delta) - \mu}{\sigma}\Big) - \frac{1}{\delta^2}\mathrm{e}^{2(\mu+\sigma^2)}\Phi\Big(\frac{\log(\delta) - \mu - 2\sigma^2}{\sigma}\Big).$$

When we set $\sigma = 0$, we instead have

$$L_r(s_0 \,|\, \theta) = \Big(1 - \frac{\mathrm{e}^{2\mu}}{\delta^2}\Big)\mathbf{1}\{\rho \leq \delta\}.$$

A.2.3 Gaussian Kernel

Let $K_\delta(\rho) = \mathrm{e}^{-\rho^2/(2\delta^2)}$. The surrogate likelihood is found by integrating

$$L_r(s_0 \,|\, \theta) = \int_0^\infty \frac{1}{\rho\sigma\sqrt{2\pi}} \exp\Big(-\frac{\rho^2}{2\delta^2} - \frac{(\log(\rho) - \mu))^2}{2\sigma^2}\Big)\mathrm{d}\rho$$

since ρ is log-normally distributed. This integral is not straight-forward to compute, so to target this distribution one could instead sample distances from the surrogate model, which leads to ABC-MCMC but without sampling from the model. If we instead consider $\sigma = 0$, we get

$$L_r(s_0 \,|\, \theta) = \exp\Big(-\frac{\mathrm{e}^{2\mu}}{2\delta^2}\Big).$$

A.3 Stationary distribution of ABC-MCMC

Here follows a proof of Theorem 2 which states that the stationary distribution of the chain in Algorithm 2 is the ABC posterior.

Proof. Consider the Markov Chain in (θ, ρ) from Algorithm 2. Let $t((\theta', \rho') \,|\, (\theta, \rho))$ be the transition density of the chain described above. Fix $(\theta', \rho'), (\theta, \rho)$. Without loss of generality, we assume that

$$K_\delta(\theta)\pi(\theta)g(\theta' \,|\, \theta) \leq K_\delta(\theta')\pi(\theta')g(\theta \,|\, \theta').$$

We shall see that this chain fulfills the detailed balance condition with stationary distribution $\pi_\delta(\theta, \rho \,|\, s_0) \propto K_\delta(\rho)q(\rho \,|\, \theta)\pi(\theta)$. By the assumption, we have $t((\theta', \rho') \,|\, (\theta, \rho)) = q(\rho' \,|\, \theta')g(\theta' \,|\, \theta)$, whereas

$$t((\theta, \rho) \,|\, (\theta', \rho')) = q(\rho \,|\, \theta)g(\theta \,|\, \theta')\frac{K_\delta(\rho)\pi(\theta)g(\theta' \,|\, \theta)}{K_\delta(\rho')\pi(\theta')g(\theta \,|\, \theta')}.$$

In combination, we get

$$t((\theta', \rho') \,|\, (\theta, \rho))\pi_\delta(\theta, \rho \,|\, s_0) = t((\theta, \rho) \,|\, (\theta', \rho'))\pi_\delta(\theta', \rho' \,|\, s_0).$$

Hence $\pi(\theta, \rho \,|\, s_0)$ is the stationary distribution fo the chain in (θ, ρ). Integrating out ρ gives the marginal distribution in θ, $\pi_\delta(\theta \,|\, s_0) \propto \pi(\theta)L(s_0 \,|\, \theta)$ as claimed. □

B

Original Algorithm Outline

Below is the algorithm outline suggested by Umberto Picchini.

Algorithm 5 ABC-SSL original outline.

Input: observed s_0. Number of rounds R, simulations per round N. A positive integer $M \ll N$. $\mathcal{D} = \{\}$.
Output: M draws from the approximate posterior $\hat{p}(\theta \mid s_0)$.
Initialization:
for $m = 1 : M$ **do**
    sample $\theta_m \sim \pi(\theta)$, simulate $x_m \sim f(x \mid \theta_m)$, obtain $s_m \leftarrow S(x_m)$ and $\rho_m \leftarrow \rho(s_m, s_0)$. Set $\mathcal{D} \leftarrow \mathcal{D} \cup (\theta_m, \rho_m)$
end for
Train $q_\phi(s_0 \mid \theta) \leftarrow \mathrm{GP}_{\phi,s_0}(\theta)$ on $\mathcal{D}$ to obtain $\hat{\phi}$ and set $\hat{\pi}_0(\theta \mid s_0) \propto q_{\hat{\phi}}(s_0 \mid \theta)\pi(\theta)$.
for $r = 1, \ldots, R$ **do**
    for $n = 1, \ldots, N$ **do**
        propose $\theta_n^* \sim g(\theta \mid \theta^\#)$, predict ρ_n^* from the fitted GP and form $k_n^* \equiv K_\delta(\rho_n^*)$, then accept $\theta_n^* \sim \hat{\pi}_{r-1}(\theta \mid s_0)$ with probability

$$\alpha = \min\left(1,\ \frac{k_n^*}{k^\#} \times \frac{\pi(\theta^*)}{\pi(\theta^\#)} \times \frac{g(\theta^\# \mid \theta^*)}{g(\theta^* \mid \theta^\#)}\right)$$

    and if accepted set $k^\# \leftarrow k_n^*$ and $\theta^\# \leftarrow \theta_n^*$.
    end for
    Disregard the initial $N - M$ draws and run the following loop on the remaining M draws.
    for $m = 1, \ldots, M$ **do**
        simulate $x_m \sim f(x \mid \theta_m)$, obtain $s_m \leftarrow S(x_m)$ and $\rho_m = \rho(s_m, s_0)$. Set $\mathcal{D} \leftarrow \mathcal{D} \cup (\theta_m, \rho_m)$
    end for
    re-train $q_\phi(s_0 \mid \theta) \leftarrow \mathrm{GP}_{\phi,s_0}(\theta)$ on $\mathcal{D}$ to obtain $\hat{\phi}$ and set $\hat{\pi}_r(\theta \mid s_0) \propto q_{\hat{\phi}}(s_0 \mid \theta)\pi(\theta)$.
end for