NTNU
Norwegian University of
Science and Technology

Master's thesis

2019

Benjamin Osei Arthur

Master's thesis

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences

Benjamin Osei Arthur

An Introduction to Approximate Bayesian Computation

April 2019

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

An Introduction to Approximate Bayesian Computation

Benjamin Osei Arthur

Statistics
Submission date: April 2019
Supervisor: Henning Omre Karl

Norwegian University of Science and Technology
Department of Mathematical Sciences

Summary

Approximate Bayesian Computation (ABC) methods is a technique used to make parameter inference and model selection of issues of intractable likelihood and complex models.

In this thesis, we briefly discuss the philosophy of Bayesian inference and elaborated more on the definition, implementation and demonstration of the three ABC algorithms. We wanted to know the efficiency of the ABC methods in computing the samples of posterior parameters compare to the analytically computation of the posterior parameters. The ABC algorithm is applied on two simple toy examples. In these toy examples, the posterior pdf is known before implementing the algorithm. We further compare the samples of posterior parameter values obtained using ABC to the true posterior and hence verify the accuracy of the algorithm.

Preface

This thesis is submitted in fulfillment of the requirements for the two year master of science (MSc) degree in Mathematics with specialization in statistics the Norwegian University of Science and Technology (NTNU).

I would like to express sincerest gratitude to my supervisor, Henning Omre, for his constant constructive criticism in this thesis. This has been invaluable and helped me to focus and improve the quality of my work. Also i would like to show appreciation to Maxime Conjard, for his assistance and advice in the quest of this work.

Lastly, I would also like to thank my family, friends and colleagues for their support throughout the course of my masters studies.

Contents

1 Introduction

Approximate Bayesian Computation (ABC) is a statistical technique and it was developed in past decades to make inference about parameters and for models selection in complex situations which often are encountered in population genetics. The Bayesian revolution, together with modern computers and powerful algorithms has allowed statistician to exploit Bayesian methodology in ecology, genetics and epidemiology.

Construction of models that describe our observations and that can directly simulate artificial data sets for given parameters can often be made. However, it is generally difficult to assess model parameters given a data set, that is computing the likelihood of the model. A naturally flexible structure within which to address these problems, is provided by the Bayesian paradigm. Certainly, the notion of simulating parameter values only really makes sense in a Bayesian approach. Since this approach allows a stochastic interpretation of the model parameters it is often straightforward to write a computer code to simulate data but difficult to work out the analytical likelihood function. Bayesian inference requires us to compute the likelihood function. In population ecology, genetics and epidemiology, a class of techniques, known as ABC has been developed to avoid the computation of likelihood in posterior distributions [1]. Bayesian methods are important not only because they circumvent the null hypothesis testing, but also because they allow for statistical inference. These ABC techniques complement the development of statistical inference in complex mathematical models.

1.1 Problem statement

Many areas in biological and environmental science encounters model selection challenges due to the complex and complicated nature of the models. Researchers are constantly dealing with this issue of selection and comparison, in particular when different complex stochastic models and reasonable selection criteria explains the data reasonably. In most situation, its quite a challenge to select models that are suitable among class of competing models and this often requires deeper understanding of the concept. Many techniques has been proposed in view of the above and arguably the most popular currently is the Bayesian approach. In the past decades, the Bayesian approach has found its use in many areas, among them are the model selection and statistical inference .

In the Bayesian paradigm, the best model strike the right balance between experience and goodness of fit. Several algorithms have been proposed for model selection based on Bayes concept, and Reversible-Jump Markov chain Monte Carlo (RJ-MCMC), Metropolis-Hasting Markov chain Monte Carlo (MH-MCMC) simulation and nonlinear filtering are examples of the popular algorithms used in these studies [2]. Essentially, the evaluation is based on maximum likelihood estimates and a penalty term to avoid complex models. The number of parameters in the model is often penalised. In many cases, the marginal likelihood estimation is made for each model separately, and the results are used to determine the plausibility of each model. This approach may cause problems when we are dealing with large data set which requires many parameters resulting in complex models. Computing the parameter likelihood in nonlinear cases is very difficult due to the non-Gaussian nature like multi-modality, of the phenomena.

The use of ABC algorithms was one of the recommended alternatives for models selection and parameter estimation in the Bayesian framework. If we compare ABC with the methods mentioned above, ABC is straightforward and general because it does not require extra evaluation criteria to differentiate between complex models. The model parameter inference can be made through evaluation of the similarity between the observed and simulated data.
Consider the variable x_d being the observational data set. In the Bayesian paradigm , the posterior probability density function (pdf) , contains all the information about the parameters of interest $\boldsymbol{\theta}$ and is defined as

$$p(\boldsymbol{\theta}|\mathrm{x}_d) = \frac{p(\mathrm{x}_d|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathrm{x}_d)} = \text{const} \times p(\mathrm{x}_d|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1}$$

where $p(\mathrm{x}_d|\boldsymbol{\theta})$ is the likelihood function , $p(\boldsymbol{\theta})$ is the prior pdf and $p(\mathrm{x}_d)$ is the marginal likelihood. Since the likelihood function $p(\mathrm{x}_d|\boldsymbol{\theta})$ may not be on a closed analytical form, as in population genetics, the ABC methods [3] uses a rejection technique to circumvent the computation of the likelihood function. If we observe $\mathrm{x}_d \sim p(\mathrm{x}_d|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ is the prior pdf of the parameter $\boldsymbol{\theta}$, then the original ABC algorithm jointly simulates

$$\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}), \quad \mathbf{x}_d^* \sim p(\mathbf{x}|\boldsymbol{\theta}^*),$$

and accept the simulated $\boldsymbol{\theta}_d^*$, if the simulated variable $\mathbf{x}_d^*$ is equal or close to the observed value, $\mathbf{x}_d$.

There are many algorithms that can be used to implement the ABC technique. We firstly define the ABC Rej algorithm which is the basis ABC algorithm and it is based on ideas from rejection and importance sampling. Secondly, the Markov Chain Monte Carlo (McMC) algorithm is defined, which keeps proposals within non-negligible posterior areas (regions) and lastly the ABC Population Monte Carlo (Population Monte Carlo (PopMC)) is presented.

The specific objectives of the study is:

1. Present the underlying ideas of the ABC parameter inference approach.

2. Define and discuss alternative ABC algorithms.

3. Implement and demonstrate alternative ABC algorithm.

The remaining of the study will be organized into four sections. Section two and three would entail the background theories and methodological issues respectively, whiles the two last sections would focus on simulation examples and conclusion of the work.

2 Basic Theory

The frequentist and Bayesian statistics differ in the interpretation of probability. For a frequentist, probability of an event is defined as the limit of the relative frequency of the occurrence of an event in a large number of trials. On the other hand, probability of an event in Bayesian context is defined as the plausibility of the event to occur, given the available information. Bayesian statistics do not consider probability as a frequency of occurrences but as a quantitative encoding of our knowledge about variables. Bayesian methods in particular, makes is possible to integrate scientific experience in the analysis by means of a prior model. Bayesian techniques may be applied to complicated and complex problems that conventional frequentist methods would find it difficult to handle [4].

2.1 Model Parameter Inference

In this section we formalize the difference between frequentist and Bayesian approaches to statistical inference. Consider the variable $\mathbf{X} : [X^1, ..., X^q] \in \Omega_{X^1} \times ... \times \Omega_{X^q} = \Omega_X$ being a q-variable and let $\boldsymbol{\theta} : [\theta^1, , ..., \theta^p] \in \Omega_{\theta^1} ... \Omega_{\theta^p} = \Omega_\theta$ be a p-variable vector called a model parameter. The model parameters $\boldsymbol{\theta}$ may represent the expectation and variance of a population from which $\mathbf{X}$ is a random variable.

Define the statistical model,

$$\mathbf{X} \rightsquigarrow p(\mathbf{x}; \boldsymbol{\theta}) \quad \text{probability density/mass function,}$$

and assume that the set observations are available with outcomes

$$\begin{aligned} &\mathbf{X}_d : \mathbf{X}_1, ..., \mathbf{X}_n \quad \text{iid} \quad p(\mathbf{x}; \boldsymbol{\theta}) \\ &\mathbf{x}_d : \mathbf{x}_1, ..., \mathbf{x}_n \end{aligned}$$

For example $\mathbf{X}_d : X_1, ..., X_n$ could be the height of n final year statistic students at NTNU selected at random with outcome $\mathbf{x}_d : x_1, x_2, ..., x_n$. From the selection process and experience, we specify a model that is independent and normally distributed with mean μ and variance σ^2, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

The model parameter can be written as $\boldsymbol{\theta} = (\mu, \sigma^2)$ with a pdf,

$$X \rightsquigarrow p(x;\theta) = p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big) \tag{2}$$

and likelihood function,

$$\begin{aligned} L(\boldsymbol{\theta};\mathrm{x}_d) = p(\mathrm{x}_d;\boldsymbol{\theta}) &= \prod_{i=1}^{n} p(x_i;\mu,\sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big) \\ &= (2\pi\sigma^2)^{\frac{-n}{2}}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2) \end{aligned} \tag{3}$$

2.2 Classical Inference

The classical frequentist approach defines probability as a relative frequency of occurrence of a large number of trials in an experiment [5].
Considering the model parameter $\boldsymbol{\theta}$ as a fixed , unknown constant, the maximum likelihood (ML) method is often used to assess the parameter values that maximize the likelihood function [6].
The ML estimator for the model parameter $\boldsymbol{\theta}$ is defined as ;

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= argmax_{\boldsymbol{\theta}}\{p(\mathrm{x}_d;\boldsymbol{\theta})\} \\ &= argmax_{\boldsymbol{\theta}}\{\prod_{i=1}^{n} p(\mathrm{x}_i;\boldsymbol{\theta})\} \\ \hat{\boldsymbol{\theta}} &= \boldsymbol{\theta}_{ML}(\mathrm{x}_d) \end{aligned} \tag{4}$$

For a given $p(\mathrm{x};\boldsymbol{\theta})$ and observation set x_d one may define sufficient statistics $\mathrm{s}_\theta(\mathrm{x}_d) = [s_{\theta 1}(\mathrm{x}_d), ..., s_{\theta p}(\mathrm{x}_d)]$, [6] such that ,

$$p(\mathrm{x}_d;\boldsymbol{\theta}) = h_1(\mathrm{s}_\theta(\mathrm{x}_d);\boldsymbol{\theta})h_2(\mathrm{x}_d)$$

with $h_1(.)$ and $h_2(.)$ suitable functions, hence

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{ML}(\mathrm{x}_d) = \boldsymbol{\theta}_{ML}(\mathrm{s}_\theta(\mathrm{x}_d))$$

2.3 Bayesian Inference

In contrast, the Bayesian approach allows probability to represent subjective uncertainty or subjective belief [7]. To make inference about a model parameter $\boldsymbol{\theta}$, there is a need to have information or knowledge about the unknown $\boldsymbol{\theta}$ prior to obtaining the data.
From 2.1, we have a model $p(\mathbf{x};\boldsymbol{\theta})$ and there is a need to specify a prior pdf for $\boldsymbol{\theta}$. The distribution is called the prior pdf because this quantifies the uncertainty about $\boldsymbol{\theta}$ before the data is known [8].
Consider $\boldsymbol{\theta}$: $[\theta^1, , ..., \theta^p]$ as a random variable with user specified prior pdf, $\boldsymbol{\theta} \rightsquigarrow p(\boldsymbol{\theta})$ and define the posterior pdf as,

$$\begin{aligned}[\boldsymbol{\theta}|\mathbf{X}_d = \mathbf{x}_d] \rightsquigarrow p(\boldsymbol{\theta}|\mathbf{x}_d) &= const \times p(\mathbf{x}_d|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= const \times \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})\end{aligned} \tag{5}$$

For a given $p(\mathbf{x}|\boldsymbol{\theta})$ with sufficient statistics $\mathrm{s}_\theta(\mathbf{x}_d)$ one has

$$\begin{aligned}p(\boldsymbol{\theta}|\mathbf{x}_d) &= const \times h_1(s_\theta(\mathbf{x}_d)|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= p(\boldsymbol{\theta}|\mathrm{s}_\theta(\mathbf{x}_d))\end{aligned} \tag{6}$$

Usual Bayesian point estimators are defined by central tendency of the posterior pdf. Specific point estimates derived from the posterior distribution are:

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_{MAP} &= MAP[\boldsymbol{\theta}|\mathbf{x}_d] = MAP[\boldsymbol{\theta}|\mathrm{s}_\theta(\mathbf{x}_d)] \\ \tilde{\boldsymbol{\theta}}_E &= E[\boldsymbol{\theta}|\mathbf{x}_d] = E[\boldsymbol{\theta}|\mathrm{s}_\theta(\mathbf{x}_d)]\end{aligned} \tag{7}$$

2.4 Approximate Bayesian Computation ABC

The main focus of Bayesian statistics is the posterior distribution:

$$p(\boldsymbol{\theta}|\mathbf{x}_d) = \mathrm{const} \times p(\mathbf{x}_d|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{8}$$

Bayesian algorithm like McMC typically requires calculations of the likelihood $p(\mathbf{x}_d|\boldsymbol{\theta})$ for evaluation. This raises the question as to whether the algorithm can assess the posterior distribution without being able to calculate the likelihood.

The McMC algorithm can be developed to sample jointly over the parameters of interest and the variables [9]. With the use of ABC methods, the posterior pdf can be assessed even when the likelihood is not available for McMC simulation. There are two reasons, one mathematical and one computational, causing the likelihood function not to be available for McMC simulation. Mathematical reasons involve the functions being unavailable in closed form whereas the computational reasons are related to the expensive nature of simulating and calculating the likelihood function [9].

The ABC method, was initially mentioned in 1984 through a pedagogical and philosophical argument in [10] and later a generalized version of the method was developed in [3]. ABC method is used to substitute the calculation of likelihood function by an algorithm that simulates and produces an artificial data set $\mathbf{x}_d^*$, and calculates the distance between the simulated data $\mathbf{x}_d^*$ and observed data $\mathbf{x}_d$. The algorithm thereby generate a posterior parameter sets and some examples are found in [11] and [12]. It is common for statisticians to use the sum of squared residuals (SSR) as measure of discrepancy between the artificial data $\mathbf{x}_d^*$ and the observed data $\mathbf{x}_d$. Assessment of the model parameters is made from the posterior parameter set.

Classical and Bayesian inference require the likelihood model $p(\mathbf{x};\boldsymbol{\theta})$ to be available on explicit form, i.e with given $\mathbf{x}$ for a specific $\boldsymbol{\theta}$, the numerical value of $p(\mathbf{x};\boldsymbol{\theta})$ can be calculated.
It need not always be so, the model may be available only by a sample-generating-function,

$$\mathbf{X} = g(\boldsymbol{\theta}, \boldsymbol{\epsilon}) \tag{9}$$

with $g(.,.)$ being a suitable function or numerical model, and the sample-impulse $\boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_q] \rightsquigarrow Uni_q[0,1]$. Behind this sample-generating-function there will always exist a pdf model, although not necessary on analytical form.

The description above defines the ABC framework. Let the corresponding observation-set-generating-function be denoted :

$$\begin{aligned}\mathbf{X}_d &= [\mathbf{X}_1, ..., \mathbf{X}_n] = [g(\boldsymbol{\theta}, \boldsymbol{\epsilon}_1), ..., g(\boldsymbol{\theta}, \boldsymbol{\epsilon}_n)] \\ &= g_d(\boldsymbol{\theta}, \boldsymbol{\epsilon}_d)\end{aligned} \tag{10}$$

with $\boldsymbol{\epsilon}_d = (\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_n) \rightsquigarrow Uni_{nq}[0, 1]$.

The ABC approach aims as assessing the posterior pdf, defined by :

$$\begin{aligned}p(\boldsymbol{\theta}|\mathbf{x}_d) &= const \times p(\mathbf{x}_d|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= const \times p(g_d(\boldsymbol{\theta}, \boldsymbol{\epsilon}_d) = \mathbf{x}_d|\boldsymbol{\theta})p(\boldsymbol{\theta})\end{aligned} \tag{11}$$

with $\boldsymbol{\epsilon}_d \rightsquigarrow Uni_{np}[0, 1]$. Note that calculating $p(g_d(\boldsymbol{\theta}, \boldsymbol{\epsilon}_d) = \mathbf{x}_d|\boldsymbol{\theta})$ for a specific $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ may not be simple. For the discrete case $\mathbf{X} \in \Omega_X$ being a categorical sample space, assessing a consistent and unbiased estimator is simple however see Algorithm 1.

Algorithm 1 ABC discrete Case

```
Initiate :
$nx = 0$
for i=1,...,S do
    $\boldsymbol{\epsilon}_d^* \rightsquigarrow Uni_{nq}[0, 1]$
    $\mathbf{x}_d^* = g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}_d^*)$
    if $[\mathbf{x}_d^* = \mathbf{x}_d]$   $nx = nx + 1$
End do
$\hat{p}_s(g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}) = \mathbf{x}_d|\boldsymbol{\theta} = \boldsymbol{\theta}^*) = \frac{nx}{S}$
```

Then

$$\hat{p}_s(g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}_d) = \mathbf{x}_d|\boldsymbol{\theta}^*) \xrightarrow[s \to \infty]{\text{unbiased}} p(g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}_d) = \mathbf{x}_d|\boldsymbol{\theta}^*)$$

For the continuous case with $\mathbf{X} \subset \Omega_X$, being a continuous sample space assessment of a non-parametric estimate is also available. See Algorithm 2.

Algorithm 2 ABC continuous Case

```
Initiate
set $\tau \geq 0$ - tolerance
$nx\tau = 0$
for i=1,...,S do
    $\boldsymbol{\epsilon}_d^* \rightsquigarrow Uni_{nq}[0,1]$
    $\mathbf{x}_d^* = g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}_d^*)$
    if $[\|\mathbf{x}_d^* - \mathbf{x}_d\| < \tau]$   $nx\tau = nx\tau + 1$
End do
$\hat{p}_{s\tau}(g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}_d) = \mathbf{x}_d | \boldsymbol{\theta} = \boldsymbol{\theta}^*) = \frac{1}{2\tau} \cdot \frac{nx\tau}{S}$
```

Then

$$\hat{p}_{s\tau}(g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}_d) = \mathbf{x}_d | \boldsymbol{\theta}^*) \xrightarrow[s\to\infty, \tau\to 0]{\text{consistent}} p(g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}_d) | \boldsymbol{\theta}^*)$$

Hence, one may move $\boldsymbol{\theta}^*$ in the sample space Ω_θ and explore the posterior pdf $p(\boldsymbol{\theta}|\mathbf{x}_d)$ of interest. The usual ABC approach is not to numerically assess the posterior pdf , but rather sampling based inference.
Three sampling based ABC-estimators of the posterior pdf will be discussed and demonstrated.

- ABC Rejection (Rej) estimator
- ABC Markov chain Monte Carlo (McMC) estimator
- ABC Population Monte Carlo (PopMC) estimator

3 Algorithm Descriptions

We present the three alternative algorithm for assessing the ABC model parameter posterior pdf.

3.1 ABC Rejection (Rej) algorithm

The ABC rejection algorithm is the initial and basic ABC algorithm and it relies on ideas from importance sampling [10]. Assume that we have a perfectly observed system in which there is no latent variable layer, and represent the model parameter $\boldsymbol{\theta}$ by a prior $p(\boldsymbol{\theta})$. A simulation model for a new data set x_d^* is define by $p(\mathrm{x}|\boldsymbol{\theta})$. The following algorithm may be used to assess the posterior pdf. Firstly, simulate from the joint distribution $p(\boldsymbol{\theta}, \mathrm{x})$ by simulating $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$ and then $\mathrm{x}_d^* \sim p(\mathrm{x}_d|\boldsymbol{\theta}^*)$. Secondly, reject the proposed pair unless x_d^* matches the observed data x_d. The remaining $\boldsymbol{\theta}^*$ is a sample from the required posterior pdf.

Approximate rejection sampling

The ABC Rej algorithm accepts model parameter values $\boldsymbol{\theta}^*$ provided the associated simulated data x_d^* is ”sufficiently close” to the observed data x_d. The algorithm is specified in pseudo code in Algorithm 3,

Algorithm 3 ABC Rejection Algorithm

```
Given an observe data $\mathrm{x}_d$, we assume a model $\mathbf{X}_d = g_d(\boldsymbol{\theta}, \boldsymbol{\epsilon}_d)$.
Initiate:
    • $S$- number of generations.
    • $\rho(.,.)$- distance.
    • $\tau \geq 0$- tolerance.
for i=1,...,S do
    Generate $\boldsymbol{\theta}^* \rightsquigarrow p(\theta)$
    Generate $\boldsymbol{\epsilon}_d^* \rightsquigarrow Uni_{nq}[0, 1]$
    Calculate $\mathrm{x}_d^* = g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}_d^*)$
    if $\rho(\mathrm{x}_d^*, \mathrm{x}_d) < \tau$, keep $\boldsymbol{\theta}^*$ as sample from $p(\boldsymbol{\theta}|\mathrm{x}_d)$.
End do
```

To select the discrepancy function $\rho(.,.)$, the Euclidean distance is usually used as the norm for the rejection method though other norms can be used. The algorithm is ”exact” in the sense that it produces a representative realization from $p(\boldsymbol{\theta}|\mathrm{x}_d)$. Smaller choice of the tolerance level τ is preferred but if it is too small, the rejection rate will be high. This will be a challenge for high-dimension x_d, since a close match between the observed data x_d and the simulated data x_d^*, is highly unlikely.

In this situations other features that summarizes the information in the data x_d may be used in the discrepancy function. The features can be statistical dispersion (for example mean and variance) and some auto-correlation depending on the kind of problem at hand.
In simple cases, if a sufficient summary statistics $\mathrm{s}_\theta(\mathrm{x}_d)$ can be identified for $p(\boldsymbol{\theta}|\mathrm{x}_d)$ one may use $\rho(\mathrm{s}_\theta(\mathrm{x}_d), \mathrm{s}_\theta(\mathrm{x}_d^*)) < \tau$ for some given value of τ.

Example of ABC-Rej algorithm - Weibull distribution

Let $\mathbf{X}_d : X_1, X_2, ..., X_n$ being the sample of dielectric failures of a ceramic capacitor. The observational set is (iid) from a Weibull distribution. In order to compute the posterior pdf, we would need a prior to draw the simulated data x_d^*. In this example, uniform distribution is used. It is very important, when dealing with simulated-based inference to keep the dimension of the data low, in order for the method to work better. One may find the sufficient summary statistics for the observations to reduce the dimension of the sets of data. Depending on the type of distribution at hand, many people calculate the distance between the observed data x_d and simulated data x_d^*, when dealing with Rejection algorithm. In this Example, ABC-Rej algorithm is used and Euclidean distance ($\sqrt{(\bar{\mathrm{x}}_d^* - \bar{\mathrm{x}}_d)^2}$) plays an important role in the simulation process.

In our toy example, we create 30 iid observations from a Weibull pdf with parameters $\boldsymbol{\theta} = (2, 4)$ to obtain x_d and compute mean and standard deviation as summary statistics. This summary statistics helps reduce the computational burden associated with large data. We randomly draw sample parameters from a uniform pdf to find the simulated data x_d^*. Since the Weibull pdf has two parameters, we draw parameters from $\boldsymbol{\theta}_1 = Uni[0.01, 6]$ and $\boldsymbol{\theta}_2 = Uni[0.01, 10]$ for the shape and scale respectively. We choose a monotonically decreasing tolerance $\tau = (1, 0.5, 0.2, 0.05)$

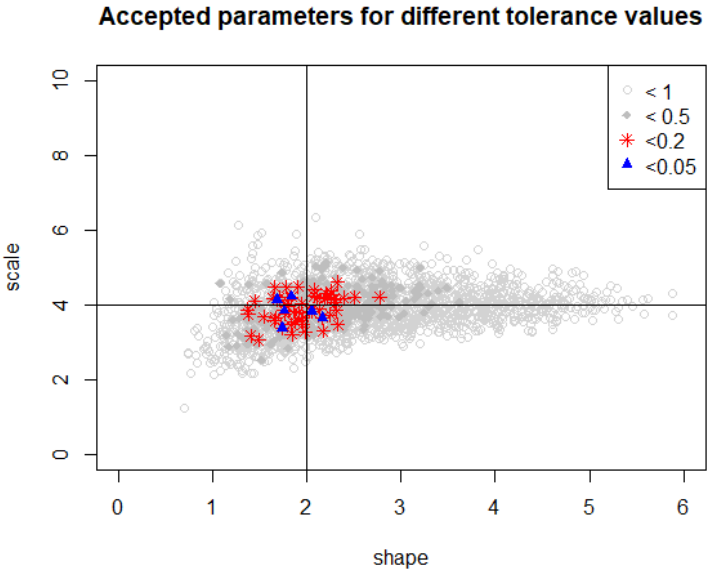

Figure 1: The approximate joint posterior distribution for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in the weibull pdf with samples size $N = 10^4 \times 2.0$ produced by ABC Rej iterations with $\tau_{=}(1, 0.5, 0.2, 0.05)$.

In Figure 1, the accepted $\boldsymbol{\theta}$-values for four values of τ are displayed. Note that, as the τ gets very small, the number of parameters that is accepted is decreased. In the limiting case where the $\tau \to 0$, and $n \to \infty$, the sample $\boldsymbol{\theta}^*$ would approach the correct posterior pdf $(\boldsymbol{\theta}|\mathbf{x}_d)$. In practice, we would get less and less points with high operational cost. Also, when the τ is large, we typically gets more $\boldsymbol{\theta}^*$ and a better approximation of the shape of the pdf, but the posterior estimate is severely biased. We can see that, when $\tau = 1$ most points are widely distributed around the true parameters. Further more, when $\tau = 0.05$, there is less number of parameter values $\boldsymbol{\theta}^*$ that is accepted and this reduce the precision of shape of the posterior and also a lot of information about the estimated parameter is lost. When $\tau = 0.2$, there are enough values $\boldsymbol{\theta}^*$ that are concentrated around the true parameter. This value of $\tau = 0.2$ gives a reasonable approximation of the shape of the posterior pdf.

3.2 ABC Markov Chain Monte Carlo (McMC) algorithm

The main purpose for McMC approach to ABC sampling is to keep proposals within non-negligible posterior regions. An important issue that needs to be considered when applying ABC methods is the ability to reject good proposals due to the strict matching condition of observed and simulated data in view of attaining accurate approximation. Focusing on the Metropolis-Hastings algorithm, ABC McMC approach is shown in Algorithm 4.

Algorithm 4 ABC McMC

```
Given an observe data set $\mathbf{x}_d$, we assume a model $\mathbf{X}_d = g(\boldsymbol{\theta}, \epsilon_d)$.
Initiate:
    • $S$- number of generations.
    • $\rho(.,.)$- distance.
    • $\tau \geq 0$- tolerance.
    • $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})$-proposal
Sample
    • $\boldsymbol{\theta}^o$ such that $p(\boldsymbol{\theta}^o) \geq 0$
for i=1,...,S do
    Set : $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}$
    Generate: $\boldsymbol{\theta}^* \rightsquigarrow q(\boldsymbol{\theta}|\boldsymbol{\theta}^i)$
    Generate: $\boldsymbol{\epsilon}_d^* \rightsquigarrow Uni_{nq}[0,1]$
    Calculate: $\mathbf{x}_d^* = g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}_d^*)$
    if  $\rho(\mathbf{x}_d^*, \mathbf{x}_d) \leq \tau$
        • Calculate :
        • $\alpha_* = min\big(1, \frac{p(\theta^*)}{p(\theta^i)} \times \frac{q(\theta^i|\theta^*)}{q(\theta^*|\theta^i)}\big)$
    With probability $\alpha_*$ set:
        • $\boldsymbol{\theta}^i = \boldsymbol{\theta}^*$
    End if
End do
```

An example of ABC McMC algorithm - Normal distribution

We apply the ABC McMC algorithm to assess the approximate likelihood values bases on simulated samples. We illustrate this in a toy example where, we have 20 iid observation from a Gaussian pdf with $\mu = 4.6$ and $\sigma = 1.9$ hence $\boldsymbol{\theta} = (4.6, 1.9)$. We use a prior pdf from a normal pdf with $(\mu, \sigma) = (4, 1.5)$, a distance measure $|\bar{\mathrm{x}}_d^* - \bar{\mathrm{x}}_d|$ and a tolerance $\tau = (0.1, 0.2)$ for mean and standard deviation respectively.

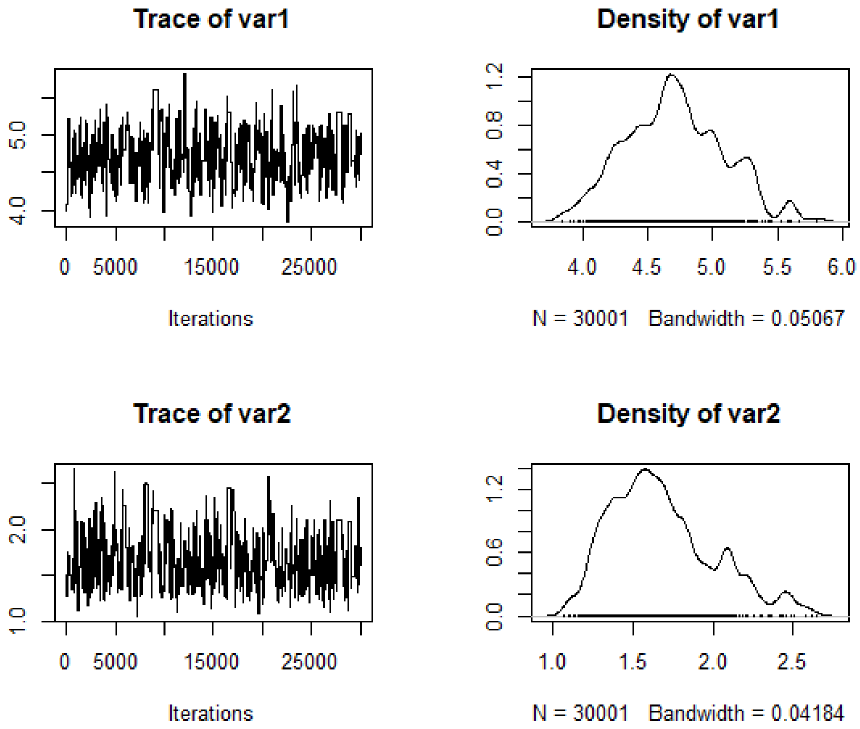

Figure 2: Estimates of the posterior distribution of two different parameters of a normal distribution by the ABC-MCMC algorithm. The trace plot and histogram represents the estimated parameter samples with size $N = 10^4 \times 3.0$ and tolerance values $\tau = (0.1, 0.2)$.

An example of the model parameter samples generated by the ABC McMC is shown in Figure 2. In the figure, there are two types of plots, which are the trace and the density plots for the posterior samples. From the trace plot, it is evident that the samples are all concentrating around the true parameters as expected and from the density plot we can see that the samples are approximately retrieving to the true parameter values in both values.

3.3 ABC Population Monte Carlo (PopMC) algorithm

The population Monte Carlo algorithm is an adaptive iteration important sampling technique, with important functions depending on previously generated samples [13]. There are two challenges that confront this method, which is the continuous nature of $\mathbf{x}_d$ and cases when the prior pdf is far from the posterior pdf, making it improbable to generate simulated data $\mathbf{x}_d^*$ that is close to the observed data $\mathbf{x}_d$. The ABC-PopMC can be used to solve the first problem, with the aid of the tolerance and discrepancy function, and the other problem is complex and needs more attention. Generally, when the target distribution is different from the prior distribution acquiring an accurate posterior is difficult and requires a lot of adjustment to the sampling and distance and tolerance. A special case of the ABC PopMC algorithm is when an algorithm with successive steps towards the posterior pdf is achieved by applying the weighted sampling from the set of parameter values whose distances between observed $\mathbf{x}_d$ and simulated $\mathbf{x}_d^*$ is smaller than a given threshold.

We initiate this algorithm by sampling N values from the prior $p(\boldsymbol{\theta})$ which is know as the particles. For each sample $\boldsymbol{\theta}^*$, we generate a simulated data $\mathbf{x}_d^*$ and calculate the distance between the observed and simulated data. We accept $\boldsymbol{\theta}^*$ if $\rho(\mathbf{x}_d^*, \mathbf{x}_d) \leq \tau_0$. In this initial step, we associate to each $\boldsymbol{\theta}^*$ the same weight , $w_{i,0} = 1/N$ for $i = 1, ..., N$.
In successive iterations , we perform sampling from a proposal distribution and re-weight the particle system so it targets the desired posterior pdf. For details see Algorithm 5.

3.4 Discrepancy function and Tolerance level of ABC

The main objective of the ABC algorithm is not finding point estimates of the parameters but instead to obtain samples from the posterior pdf. Recall that the posterior pdf of a parameter $\boldsymbol{\theta}$ is the distribution of that parameter conditioned on the observed data $\mathbf{x}_d$, $p(\boldsymbol{\theta}|\mathbf{x}_d)$.

From numerous simulations, we obtain $\boldsymbol{\theta}^*$ as a sample from the posterior pdf using a pre-define distance $\rho(\mathbf{x}_d^*, \mathbf{x}_d)$ between the observed and simulated data. For a given tolerance level τ, the posterior pdf $p(\boldsymbol{\theta}|\rho(\mathbf{x}_d^*, \mathbf{x}_d) \leq \tau) = p(\boldsymbol{\theta}|\mathbf{x}_d)$ [3].

It is often convenient to define $\rho(\mathbf{x}_d^*, \mathbf{x}_d)$ as a distance between summary statistics, lets say $s(\mathbf{x}_d^*)$ and $s(\mathbf{x}_d)$. Taking an example, we let $s(.)$ be the sample mean, so we can say $\rho(\mathbf{x}_d^*, \mathbf{x}_d) = (\bar{\mathbf{x}_d^*} - \bar{\mathbf{x}_d})^2$ is the squared distance between the sample means. However, some statistics carry more information about the parameter that others.

If $\rho(s(\mathbf{x}_d^*), s(\mathbf{x}_d))$ are chosen as the difference between the sufficient statistic for $p(\boldsymbol{\theta}|\mathbf{x}_d)$, then the approximation given by an ABC algorithm will be exact when $\tau \rightarrow 0$ [14]. Many computational challenges occur when the tolerance threshold τ is very small, which may be resolved by using a sequence of tolerance criteria. The number of iterations in the ABC algorithm, will depend of the sequence of tolerance criteria. If the tolerance is too small the algorithm will results in high rejection rate and poor assessment of the posterior pdf. The main challenge, is to set values of the tolerance level τ that balance the approximation to the posterior pdf against the rejection rate.

Algorithm 5 ABC Popmc Algorithm

```
Given an observe data set $\mathbf{x}_d$, we assume a model $\mathbf{X}_d = g(\boldsymbol{\theta}, \boldsymbol{\epsilon}_d)$.
Initiate:
    • $S$- number of generations.
    • $\rho(.,.)$- distance.
    • $\tau \geq 0$- tolerance.
    • $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*; \boldsymbol{\Sigma}_\theta^+)$ - transition probability
for i=1,...,N do
    (A) Label
    Generate $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$
    Generate $\boldsymbol{\epsilon}^*_{\mathbf{d}} \rightsquigarrow Uni_{nq}[0, 1]$
    Calculate $\mathbf{x}^*_d = g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}^*_d)$
    if $[\rho(\mathbf{x}^*_d, \mathbf{x}_d) > \tau]$ go to (A) else
    Set $\boldsymbol{\theta}_{o,i} \leftarrow \boldsymbol{\theta}^*$
    Set $w_{o,i} \leftarrow \frac{1}{N}$
End do
Particle set : $\boldsymbol{\theta}_o : \{\boldsymbol{\theta}_{o,i}; \quad i = 1, ..., N\}$
Weight set: $\mathbf{w}_o : \{w_{oi}; \quad i = 1, ..., N\}$
Calculate : $\boldsymbol{\Sigma}_o$ -Var , set $\boldsymbol{\theta}_o$ with weight $\mathbf{w}_o$.
for t=1,...,S do
    for i=1,...,N do
        (B) Label
        Generate :$\boldsymbol{\theta}^+ \rightsquigarrow$ set $\boldsymbol{\theta}_{j-1}$ with probability $\mathbf{w}_{j-1}$
        Generate : $\boldsymbol{\theta}^* \rightsquigarrow q(\boldsymbol{\theta}|\boldsymbol{\theta}^+, 2\boldsymbol{\Sigma}_{j-1})$
        Generate : $\boldsymbol{\epsilon}^*_d \rightsquigarrow Uni_{nq}[0, 1]$
        Calculate : $\mathbf{x}^*_d = g_d(\boldsymbol{\theta}^*, \boldsymbol{\epsilon}^*_d)$
        if $[\rho(\mathbf{x}^*_d, \mathbf{x}_d) > \tau]$ go to (B) else
        Set $\boldsymbol{\theta}_{j,i} = \boldsymbol{\theta}^*$
        Set $w^+_{j,i} = \frac{p(\boldsymbol{\theta}_{j,i})}{\sum_{j=1}^N w_{(j-1),k} q\left(\boldsymbol{\theta}_{j,i}|\boldsymbol{\theta}_{(j-1),k}; 2\boldsymbol{\Sigma}_{j-1}\right)}$
    End for
    Calculate :$w_{j,i} = w^+_{j,i} / \sum_{k=1}^N w^+_{j,k}$
    Particle set : $\boldsymbol{\theta}_j : \{\boldsymbol{\theta}_{j,i}; \quad i = 1, ..., N\}$
    Weight set: $\mathbf{w}_j : \{w_{j,i}; \quad i = 1, ..., N\}$
    Calculate : $\boldsymbol{\Sigma}_j$ -Var of set $\boldsymbol{\theta}_o$ with weight $\mathbf{w}_o$
End for
Assess $p(\boldsymbol{\theta}|\mathbf{x}_d)$ from set $\boldsymbol{\theta}_s$ with weights $\mathbf{w}_s$.
```

4 Experiment and Analysis

In this section, we apply the various ABC algorithms to two different cases, discrete and continuous respectively. In the previous chapter, we defined and discussed the algorithms.

4.1 Multinomial distribution

The multinomial distribution is the generalization of the binomial distribution to situations where each trial has $k > 2$ possible outcomes and it is denoted $Mnom(\boldsymbol{\pi}, n)$ with n being number of trials and $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)$ being the model parameters.
An experiment is multinomial if,

- consist of n independent trials
- each trial results in one of the mutually exclusive events $E_1, ..., E_k$.
- The event E_j occurs with a probability π_j,$j = 1, .., k$, with $\sum_i \pi_i = 1$

Let $\mathbf{X} = (X_1, X_2, ..., X_k)$ be the number of outcomes of each event. Even though the individual X_j are random, they sum to the number of trials,

$$\sum_{j=1}^{k} X_j = n$$

hence the X_j are negatively dependent. The pdf of $\mathbf{X}$ is given by the multinominal pdf,

$$\mathbf{X} \rightsquigarrow p(\mathbf{x}; \boldsymbol{\pi}) = \frac{n!}{x_1!...x_k!} \prod_{j=1}^{k} \pi_j^{x_j} \tag{12}$$

where $x_j \in \mathbf{N}_{\oplus}$ and $\sum_j x_j = n$

The marginal pdf of the multinomial pdf are binomial,$Bin(\pi_j, n)$,

$$X_j \rightsquigarrow p(x_j; \pi_j) = \binom{n}{x_j} \pi_j^{x_j} (1 - \pi_j)^{n - x_j}$$

Dirichlet Distribution

The Dirichlet distribution is a multivariate generalization of the beta distribution for $k > 1$. A Dirichlet distributed variable $\mathbf{X} = (X^1, ..., X^k)$ is distributed as $Dir(\boldsymbol{\alpha})$ which is parameterized by a vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$,$\alpha_i > 0$. The probability density is given as

$$\mathbf{X} \rightsquigarrow p(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{k} \alpha_j)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} x^{j(\alpha_j - 1)} \tag{13}$$

where $\mathbf{x} = (x^1, ..., x^k) \in \Omega_x$ such that $x^j \geq 0$ and $\sum_j x^j = 1$, while $\Gamma(.)$ is the gamma function.

The marginal distributions are beta distributions [15]:

$$\begin{aligned} X^i \rightsquigarrow p(x; \alpha) &= Beta(\alpha_i, \alpha_0 - \alpha_i) \\ &= \frac{\Gamma(\alpha_i + \alpha_0 - \alpha_i)}{\Gamma(\alpha_i)\Gamma(\alpha_0 - \alpha_i)} x^{\alpha_i - 1}(1 - x)^{(\alpha_0 - \alpha_i) - 1} \end{aligned}$$

where $0 \leq x \leq 1$ and $\alpha_0 = \sum_j \alpha_j$.

The Model

Let $\mathbf{X} = (X^1, X^2, X^3)$ be the number of balls that falls into three boxes with different probabilities and we fix the number of balls to n hence $\sum_i x_i = n$. We know that the probability for a ball to fall in each box is π_i with constraint $\sum_i \pi_i = 1$. In this case X_i are not independent and the joint probability of vector $\mathbf{x} = (x^1, x^2, x^3)$ is multinomial distributed. The likelihood of $\boldsymbol{\pi}$ based on $\mathbf{X}_d = \mathbf{X}$ is defined as the joint probability function of $X^1 = x^1, X^2 = x^2, X^3 = x^3$ which is ,

$$\begin{aligned} L(\boldsymbol{\pi}; \mathbf{x}_d) &= p(\mathbf{x}; \boldsymbol{\pi}) \\ &= [\frac{n!}{x^1! x^2! x^3!}] \prod_{i=1}^{3} \pi_i^{x^i} \\ &= n! \prod_{i=1}^{3} \frac{\pi_i^{x^i}}{x^i!} \end{aligned} \tag{14}$$

We further compute the log-likelihood function for Equation14

$$log\big\{L(\boldsymbol{\pi};\mathbf{x}_d)\big\} = \log(n!) - \sum_{i=1}^{3}\log(x^i!) + \sum_{i=1}^{3} x^i \log(\pi_i) \tag{15}$$

The MLE for a model parameter $\boldsymbol{\pi}$ is ;

$$\begin{aligned} \hat{\boldsymbol{\pi}} &= argmax_{\boldsymbol{\pi}}\big\{logL(\boldsymbol{\pi};\mathbf{x}_d)\big\} \\ \sum_i \hat{\pi} &= 1 \end{aligned}$$

We introduction the the Lagrange multiplier into the object function,

$$Q(\boldsymbol{\pi};\mathbf{x}_d) = \log(n!) - \sum_{i=1}^{3}\log(x^i!) + \sum_{i=1}^{3} x^i \log(\pi_i) + \lambda(1 - \sum_{i=1}^{3}\pi_i) \tag{16}$$

and we take the derivative of the object function, and set to zero,

$$\begin{aligned} \frac{dQ}{d\pi} &= \frac{x^i}{\pi_i} - \lambda = 0 \\ \frac{dQ}{d\lambda} &= 1 - \sum_{i=1}^{k}\pi_i = 0 \end{aligned} \tag{17}$$

We solve the system of equation and obtain the ML estimator for $\boldsymbol{\pi}$ as

$$\hat{\pi}_i = \frac{x^i}{n}; \quad i = 1, ..., 3$$

In order to find the Bayesian posterior pdf of the model parameters, we define a prior model for $\boldsymbol{\pi}$ and it would be appropriate to use Dirichlet distribution which is a conjugate prior of the multinomial distribution. The posterior pdf for the model parameters is ,

$$\begin{aligned} p(\boldsymbol{\pi}|\mathbf{x}) &\propto p(\mathbf{x}|\boldsymbol{\pi})p(\boldsymbol{\pi}) \\ &= n!\prod_{i=1}^{3}\frac{\pi_i^{x^i}}{x^i!} \times \frac{\Gamma(\sum_{i=1}^{3}\alpha_i)}{\prod_{i=1}^{3}\Gamma(\alpha_i)}\prod_{i=1}^{3}\pi_i^{\alpha_i-1} \\ &\propto \prod_{i=1}^{3}\pi_i^{x^i}\prod_{i=1}^{3}\pi_i^{\alpha_i-1} \\ &= \prod_{i=1}^{3}\pi_i^{\alpha_i-1+x^i} \end{aligned} \tag{18}$$

hence

$$p(\boldsymbol{\pi}|\mathbf{x}) \sim Dir(\alpha_p)$$

with $\alpha_p = (\alpha_{p1}, \alpha_{p2}, \alpha_{p3})$ and $\alpha_{pi} = \alpha_i + x^i, \quad i = 1, ..., 3$. Hence the marginal distribution for the posterior is

$$p(\pi_i; \alpha_{pi}) = Beta(\alpha_{pi}, \alpha_{p0} - \alpha_{pi})$$

where $\alpha_{p0} = \sum_i \alpha_{pi} = \alpha_0 + n$.

Since we are using Dirichlet prior which is a conjugate prior for the multinomial pdf, we would derive a Dirichlet posterior pdf. This makes estimation of the posterior pdf very simple by referring to the corresponding Dirichlet distribution.

ABC-Rej estimation of the posterior distribution

The Dirichlet posterior pdf of the parameter $\boldsymbol{\pi}$, with a Beta marginal distribution, could be used in statistical inference like hypothesis testing or confidence interval construction . Focus of this study is to evaluate the accuracy of the estimated of posterior pdf by using ABC algorithms.

We consider the multinomial distribution as a complex function $g(\boldsymbol{\theta}, \boldsymbol{\epsilon})$ for which the likelihood is difficult to compute, and assign a Dirichlet prior pdf to the model parameters $\boldsymbol{\theta} = \boldsymbol{\pi}$. Then we simulate data $\mathbf{x}_d^*$ for the Dirichlet prior $p(\boldsymbol{\theta})$ given multinomial $g(\boldsymbol{\theta}, \boldsymbol{\epsilon})$ and use the ABC-Rej algorithm to assess the ABC posterior pdf.

Note that in this specific case $\mathbf{X}_d \in \Omega_{\mathbf{x}} = \mathbf{N}^3_{\oplus}$, which is a discrete variable. Hence it is possible to use a identity tolerance function, entailing acceptance if $\mathbf{x}^*_d = \mathbf{x}_d$.

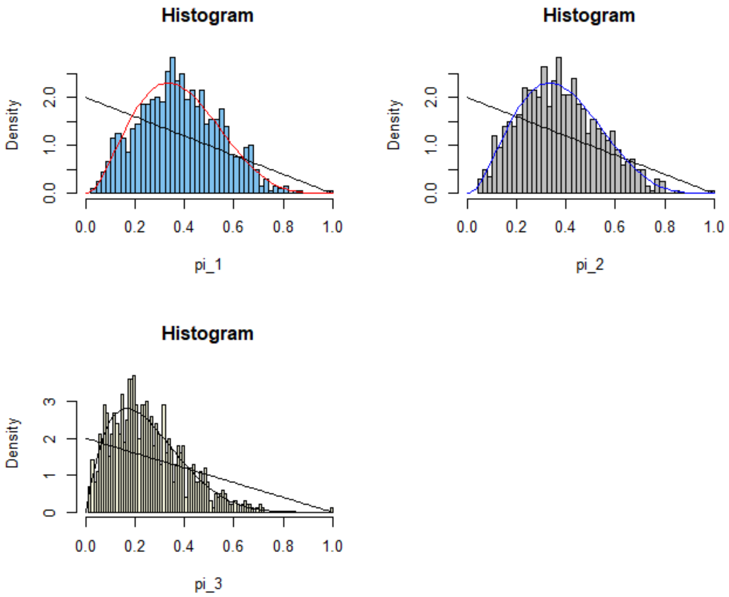

Figure 3: Density estimate of the posterior distribution for three different parameters of beta distribution , obtained through ABC-Rej using Multinomial distribution with 1 sample size and $N = 10^3 \times 1.0$ iterations and tolerance value $\tau = 0.1$.

The Findings

In the experiment the simulation of the model is done in one sample size, with two red balls, two blue balls and one green ball, $\mathbf{X} = (2, 2, 1)$, generated with three probability $\boldsymbol{\pi} = (0.3, 0.5, 0.2)$. The prior model of $\boldsymbol{\pi}$ is tri-variate Dirichlet with hyperparameters $\alpha = (1, 1, 1)$

To assess the posterior pdf, we sample $N = 1000$ values of $\boldsymbol{\pi}^*$ and generate the simulated data sets $\mathbf{x}^*_d$. The ABC Rej algorithm with a tolerance level $\tau = 0.1$ is used. Due to the design of the experiment we can calculate the posterior pdf analytically, and it is Dirichlet (α_p), with

$\alpha_p = (1+2, 1+2, 1+1)$. The posterior marginals are Beta distributed. Figure 3 display the three marginal pdfs of $\boldsymbol{\pi}$ which are beta distributed. The figures display the prior pdfs (solid line), posterior pdfs (coloured line) and the samples of the $[\boldsymbol{\pi}|\mathbf{x}_d]$ from the ABC Rej algorithm.

4.2 Pareto Distribution

The Pareto distribution which was defined by the Italian civil engineer, economist, and sociologist Vilfredo Pareto in the 19th century, is represented as $X \sim Par(\alpha, \beta)$ with shape and location parameters (α, β). Graphically, the Pareto distribution is skewed with heavy or slow decaying tails that is, most of the data is located in the tails, and is used for modelling of income and city population distributions. The pdf of Pareto distribution is ,

$$X \rightsquigarrow p(x;\alpha,\beta) = \frac{\beta\alpha^{\beta}}{x^{\beta+1}} I(x \geq \alpha), \tag{19}$$

with model parameters $\alpha \in \mathbf{R}_{\oplus}$ as known and $\beta > 2$. The parameter α is the minimum value of X and we fix it to α_0, while β is a positive parameter which determines the concentration of data towards the mode.

Gamma Distribution

The Gamma distribution $Gam(\lambda, \kappa)$ is one of the widely use distribution for reliability and life testing analysis [16]. It is also related to the beta distribution and arises naturally in the processes for which the waiting times between Poisson distributed events are relevant. The pdf is given as ,

$$X \rightsquigarrow p(x;\lambda,\kappa) = \frac{\lambda^{\kappa}}{\Gamma(\kappa)} x^{k-1} \exp(-\lambda x) I(x \geq 0) \tag{20}$$

for parameters $\lambda \in R_{\oplus}$ and $\kappa \in R_{\oplus}$.

The Model

Consider $\mathbf{X} = (X_1, ..., X_n)$ being a sample of income from a population. The observation set is independent and identically distributed (iid) from a Pareto pdf with model parameter (α_0, β), and α_0 fixed. The likelihood function for the Pareto distribution parameter β, given the outcome $\mathbf{x}_d$: $x_1, ..., x_n$ is ;

$$\begin{aligned} L(\beta; \mathbf{x}_d) &= p(\mathbf{x}_d; \beta) \\ &= \prod_{i=1}^{n} \frac{\beta\alpha_0^\beta}{x_i^{\beta+1}} I(x_i \geq \alpha_0) \\ &= \beta^n \alpha_0^{n\beta} \big[\prod_{i=1}^{n} x_i\big]^{-(\beta+1)} I(x_{(1)} \geq \alpha_0) \end{aligned} \tag{21}$$

where $x_{(1)} = min\{x_1, ..., x_n\}$, define the log-likelihood function,

$$logL(\beta; \mathbf{x}_d) = n\beta log\alpha_0 + nlog\beta - (\beta+1)\sum_{i=1}^{n} logx_i + logI(x_{(1)} \geq \alpha_0) \tag{22}$$

The MLE of the model parameter β is

$$\hat{\beta} = argmax_\beta \big\{logL(\beta; \mathbf{x}_d)\big\}$$

We compute the estimate for β, by taking the derivative of log-likelihood and set it equal to zero,

$$\begin{aligned} \frac{d}{d\beta}(\beta|x_d) &= nlog\alpha_0 + n\frac{1}{\beta} - \sum_i logx_i = 0 \\ \hat{\beta} &= \big[\frac{1}{n}\sum_i logx_i - log\alpha_0\big]^{-1} \end{aligned} \tag{23}$$

The ML estimator for β is

$$\hat{\beta} = \big[\frac{1}{n}\sum_i logX_i - log\alpha_0\big]^{-1}$$

From this ML estimator we observe that the sufficient statistic for β based on $\mathbf{x}_d$ is

$$s_\beta(\mathbf{x}_d) = \sum_i logx_i$$

In order to identify the Bayesian posterior pdf of the model parameter β in the Pareto pdf, we use the conjugate prior Gamma model. This is mathematically expressed as,

$$\begin{aligned} p(\beta|\mathbf{x}_n) &\sim p(\mathbf{x}_d|\beta) \times p(\beta) \\ &= \beta^n \alpha_0^{n\beta} \big[\prod_{i=1}^{n} x_i\big]^{-(\beta+1)} I(x_{(1)} \geq \alpha_0) \times \frac{\lambda^\kappa}{\Gamma(\kappa)} \beta^{\kappa-1} \exp(-\lambda\beta) \\ &\sim \beta^n \alpha_0^{n\beta} \big[\prod_{i=1}^{n} x_i\big]^{-(\beta+1)} \beta^{\kappa-1} \exp(-\lambda\beta) \\ &= \beta^{\kappa+n-1} \exp\{n\beta \log \alpha_0\} \exp\{-(\beta+1) \sum_i \log x_i\} \exp\{-\lambda\beta\} \\ &= \beta^{\kappa+n-1} \exp\{-(\lambda + \sum_i \log x_i - n \log \alpha_0)\beta\} \end{aligned} \tag{24}$$

hence

$$p(\beta|\mathbf{x}_d) \sim Gam(\lambda_p, \kappa_p)$$

Hence the Gamma pdf is a conjugate prior model for iid samples from the Pareto pdf with given α_0 , with posterior model parameters $\lambda_p = \lambda + \sum_{i=1}^{n} \log x_i - n \log \alpha_0$ and $\kappa_p = \kappa + n$.
Hence the corresponding sufficient statistics $s_1(\mathbf{x}) = \sum_i \log x_i$, as we observed from the ML-estimator above.

ABC-MCMC assessment of posterior pdf

To illustrate the ABC-MCMC approach, we may consider inference in modelling of Auto insurance coverage. In this study we focus on the accuracy of the estimated posterior pdf. We assume that the underlying distribution for the auto insurance coverage follows a Pareto distribution with shape parameter $\beta = 2$ and scale parameter $\alpha_0 = 1$.
To generate a data set, we simulated 30 insurers ($n = 30$), from a Pareto distribution with the above parameter values. It is important to specify a good prior distribution $p(\boldsymbol{\theta})$ and Gamma pdf is used as the prior. Even good proposals are often rejected due to strict matching condition of observed data $\mathbf{x}_d$ and simulated data $\mathbf{x}_d^*$ in order to obtain a reliable approximation.

We consider the Pareto distribution, with $\alpha_0 = 1$ given as a complex function $g(\boldsymbol{\theta}, \boldsymbol{\epsilon})$for which the likelihood is difficult to compute, and assign a Gamma prior pdf to the model parameter $\boldsymbol{\theta} = \beta$. The Pareto distribution defines $\mathbf{X}_d = g_d(\boldsymbol{\theta}, \boldsymbol{\epsilon})$. The model parameters in the prior Gamma pdf is $(\kappa, \lambda) = (4, 5)$. In the ABC McMC algorithm, the squared distance between sufficient statistics $s(\mathbf{x}_d) = \sum_{i=1}^{30} \log \mathbf{x}_i$ was used as distance function, and different tolerances are used in different runs.

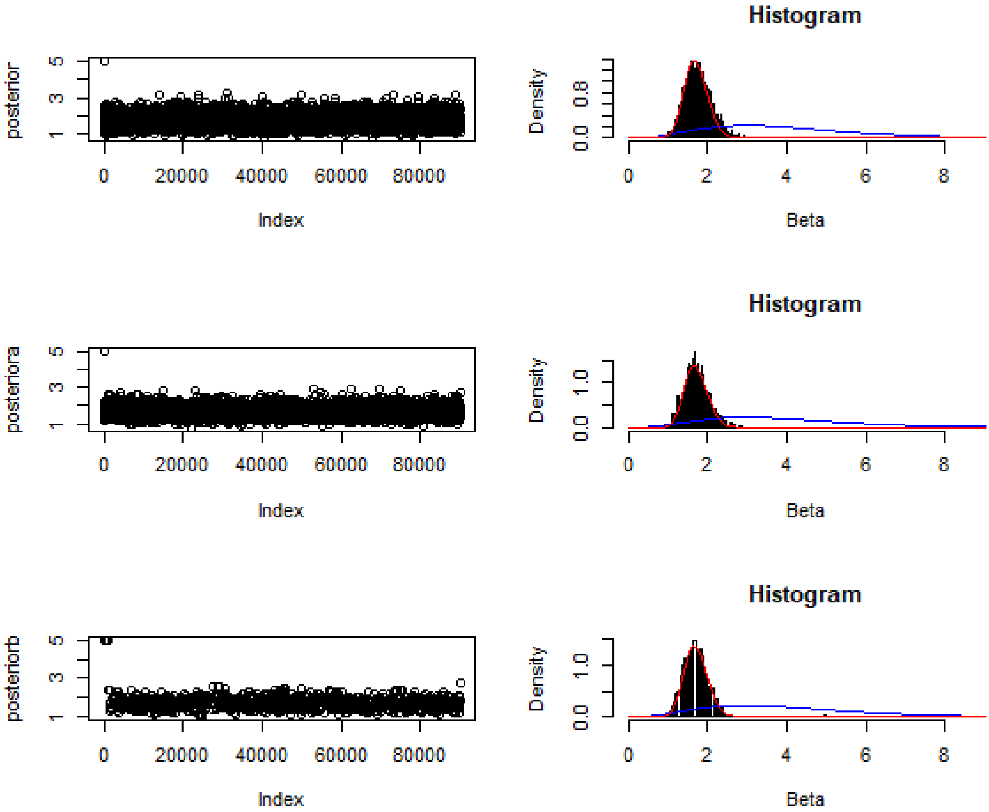

Figure 4: The posterior distribution of β at three tolerance level $\tau = (0.1, 0.05, 0.01)$ of a Pareto-Gamma distribution, obtained through ABC-McMC algorithm. On the right-hand side, the histogram represents the estimated posterior distribution, the red and blue lines indicates the true posterior distribution and the prior distribution respectively. The trace plot on the left-hand side represent values of a Gamma distribution

The Findings

Figure 4 displays the results of three runs for different tolerance values $\tau = (0.1, 0.05, 0.01)$. The prior pdf (blue line) and posterior pdf (red line) are both Gamma pdfs. The histogram displays samples of the posterior parameter values obtained using ABC-MCMC. The sample traces in the left displays demonstrate that convergence is obtained. The trace plot clearly shows that, when the tolerance values are decreasing, convergence to the true parameter is fast resulting to a better approximation. Furthermore, the figure shows that there is no significant different between the plots if τ is below 0.05.

ABC-Popmc assessment of posterior pdf

To illustrate the ABC-popmc approach, we use the same set-up as in the previous. We draw enough samples from the prior and check the reduction rate using $\rho(\mathbf{x}_d^*, \mathbf{x}_d) \leq \tau$ at a reasonable value of τ. If we consider a tolerance of $\tau = 1$, this would help us to get a sufficient proposals because of less rejection , but the resulting posterior estimates will not be accurate. Our objective is to reduce the τ so we can get sufficient proposal that will move to become a good posterior. In this work , we are fortunate to have a conjugate prior which means we have a closed-form expression for the posterior pdf and can compare our results with the truth.

The Findings

Figure 5 represents the results from assessment of posterior pdf of the parameter β by ABC Popmc algorithm with tolerance $\tau = (0.1, 0.01)$. The prior Gamma pdf (blue line) and posterior Gamma pdf (red line) are analytically computed. The posterior samples of the paramter values are displayed in the histogram. The left display contains trace plots of the sample. From the two trace plots , its can be seen that large τ give wide spread of samples compared to smaller τ. This indicates that, when the tolerance is very small, there is a small variance of the samples. Also an increase in the particle size would results in a better approximation and estimation.

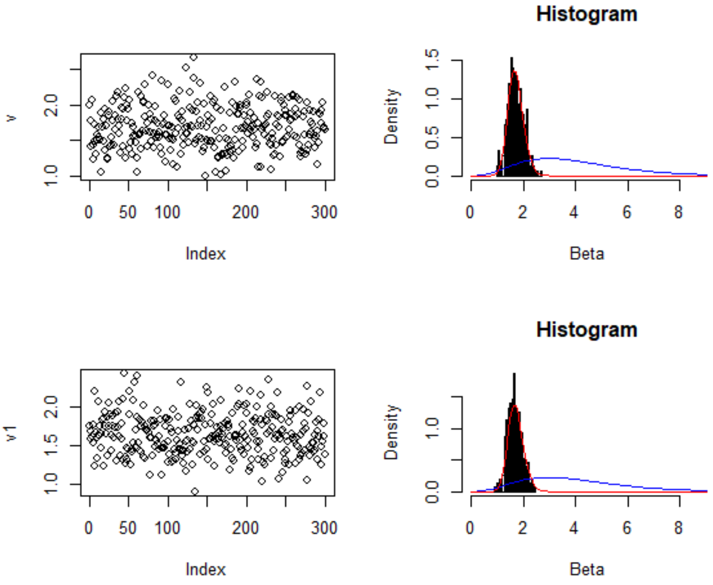

Figure 5: The estimated posterior for parameter β in a Pareto pdf at tolerance level $\tau = (0.1, 0.01)$, sample size $N = 30$ and particle size $S = 300$. The posterior distribution is obtained by the ABC-popmc algorithm

5 Conclusion

This section presents the conclusions drawn from the findings as well as recommendations for future works .

Final remarks

We presented an approach for Bayesian analysis named Approximate Bayesian Computation (ABC). Conventional methods of parameter estimation or inference are applicable if the likelihood function is available and easy to compute. In recent times, due to modernization and population growth, there have being increase in data size and complex models in areas of genetic, ecology and epidemiology, resulting to exploration of larger parameter sets and making it complicated to make inference about them.

In this work, we define and demonstrate three kinds of ABC methods and related algorithms namely, ABC-Rej algorithm, ABC McMC algorithm and ABC PopMC algorithm. We demonstrated the power of these algorithms in posterior pdf estimation, assuming there is an intractable likelihood. We implement this algorithms and demonstrate them on simple toy examples. ABC Rej algorithm was used to estimate parameter of a Multinomial pdf and a Dirichlet prior pdf. Also the two other algorithms were used to estimate the parameter of a Pareto pdf with a Gamma prior pdf.

In spite of the fact that, ABC approach provides ways to bypass intractable likelihood functions, this approach comes with a cost. The ABC is more computational expensive than the standard Bayesian methods. A typical example is seen in the multinomial pdf where an increase in observation from 1 to 100 increases the computation time from 7 seconds to 2 minutes respectively.

There are some limitations associated with this approach of parameter inference and some identified in this work is as follows:
The precise assessment of the posterior requires either setting the tolerance level to very low (approaching zero) or increase the number of simulation with low tolerance. Also summary statistics selected for the ABC acceptance is highly empirical. Therefore an automated approach with approximate sufficiency would be attractive for non-standardized and complex set-

tings.

Recommendations

Future analysis into ABC methods is needed for more conclusive results. The development of ABC methods may solve the issues of finding posterior with intractable likelihood. The validity of the methods is still in question. Attaining convergence requires increasing the sample size and letting the tolerance approach zero, which is unpractical. We recommend a fixed error bound for positive tolerance and the use of a finite sample size to improve the implementation and assessment of the method. Also, comparing of ABC methods with other approximation based inference like variational Bayesian methods and expectation propagation should be explored. We recommend a dimensional reduction technique and use of summary statistics n standardized settings..

References

[1] M. A. Beaumont, "Downloaded from www.annualreviews.org by Arizona State University on 08/14/11. For personal use only," *Annu. Rev. Ecol. Evol. Syst*, vol. 41, pp. 379–406, 2010.

[2] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," Tech. Rep. 4, 1995.

[3] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, "Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites," *Mol. Biol. Evol*, vol. 16, no. 12, pp. 1791–1798, 1999.

[4] J. M. Bernardo Bayesian Statistics and J. M. Bernardo, "BAYESIAN STATISTICS *," tech. rep.

[5] M. Botje, "Introduction to Bayesian Inference," tech. rep.

[6] G. Casella and R. Berger, *Statistical Inference*. Duxbury Press, Pacific Grove,363, 2001.

[7] A. Eshky, "Bayesian Methods of Parameter Estimation," tech. rep.

[8] T. S. Jayanta K. Ghosh, Mohan Delampady, *An Introduction to Bayesian Analysis (Theory and Methods)*. Sprinter Science + Business Media ,LLC, 2006.

[9] C. R. N. W. Christopher Drovandi, Gentry White, *Computational Bayesian Statistics*. Statistics and Operation Discipline, 2015.

[10] D. Rubin, *Bayesianly justifiable and relevant frequency calculations for the applied statistician*. Annals Statistics 12(4) 1151-1171, 1984.

[11] Z. R. . S. R. Malmberg, K. J., *Turning up the noise or turning down the volume? on the nature of the impairment of episodic recognition memory by Midazolam*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 540–549, 2004.

[12] L. D. R. D. C. . F. M. Nosofsky, R. M., *Short-term memory scanning viewed as exemplar-based categorization*. Psychological Review, 118, 280–315, 1984.

[13] J. M. M. C. P. R. O. Cappe, A. Guillin, "Population monte carlo," *Journal of Computational and Graphical Statistics*, vol. 13 (4), pp. 907–929, 2004.

[14] M. A. Beaumont, *Approximate Bayesian computation in evolution and ecology*. Annual Review of Ecology, Evolution, and Systematics, 41, 379–406, 2010.

[15] Farrow.Malcolm, *MAS3301 Bayesian Statistics*. Newcastle University, 2013.

[16] J. F. Lawless, "Statistical Models and Methods for Lifetime Data Second Edition," tech. rep.