Aalto University
School of Science
Bachelor's Programme in Science and Technology

A study on approximate Bayesian computation

Bachelor's Thesis

April 17, 2020

Thong Tran

Aalto University
School of Science
Bachelor's Programme in Science and Technology

ABSTRACT OF
BACHELOR'S THESIS

Author:	Thong Tran
Title of thesis:	A study on approximate Bayesian computation
Date:	April 17, 2020
Pages:	22
Major:	Data Science
Code:	SCI3027
Supervisor:	Professor Eero Hyvönen
Instructor:	Jukka Sirén, Postdoctoral Researcher (Department of Computer Science)

There are an increasing number of moderns applications relying on complex models whose likelihood functions cannot be evaluated. In this context, Approximate Bayesian computation (ABC) has emerged as a solution for making inference by using simulations to bypass the need of explicitly calculating likelihoods. In this study, we aim to investigate the effect of summary statistic, distance measure choice as well as dimension reduction usage. The experimentation utilizes Normal distributed data with conjugated prior and measures the performance of rejection ABC with respect to the mentioned factor using Wasserstein distance as the metric. Results from the experiment reflect the central importance of summary statistics in ABC as the quality of summary statistics dictates the performance of ABC in the study. While distance measure plays a more subtle role, it is still a factor that can significantly affect the outcome of ABC. The results also highlight the effectiveness of dimension reduction when implementing ABC.

Keywords:	Approximate Bayesian computation, Summary statistics, Distance measure, Dimension reduction, Linear regression, Likelihood-free inference, Wasserstein distance
Language:	English

Contents

1 Introduction

Statistical inference is the process of drawing conclusions about an unobserved population based on observed data, y_{obs}. The conclusions drawn often are the parameters θ of the underlying process leading to the data. Popularized in the later half of the 20th century, the Bayesian approach provides a new and intuitive framework for statistical inferences (Gelman et al., 2013).

In Bayesian inference, the complete knowledge about parameters of a model, $\theta \in \Theta$, is obtained by combining a prior distribution (believes about the parameters before seeing the data), $\pi(\theta)$, with observed data $y_{obs} \in Y$ through the likelihood function, $\pi(y_{obs}|\theta)$ (Young et al., 2005; Gelman et al., 2013). The result obtained from Bayes' Theorem is the posterior distribution:

$$\pi(\theta|y_{obs}) = \frac{\pi(\theta)p(y_{obs}|\theta)}{\int_\Theta \pi(\theta)p(y_{obs}|\theta)d\theta}$$

However, in many cases where the model is complex, the likelihood function can be intractable, and therefore, the posterior $\pi(\theta|y_{obs})$ cannot be evaluated. This challenge can be addressed by using a simpler model. While a model that has fewer parameters or is more tractable is more convenient to work with, it can potentially cause a loss of the desired details that a more complex model would have (Sisson et al., 2018).

Under this context, approximate Bayesian computation emerged as a new framework that approximates the inference process of the original model using simulation. Diggle and Gratton (1984) proposed that inference about θ can be made by comparing simulated data y with y_{obs}. When the simulated data is the same as the observed data, this method would produce an unbiased estimate of the likelihood function. However, the probability of an exact match is negligible or 0 in most applications. Therefore, a more practical alternative would be to consider the cases where the simulated data approximately matches the observed data. This is the basis for standard ABC methods (Blum and François, 2010; Drovandi et al., 2011; Fearnhead and Prangle, 2012; Sisson et al., 2018).

When dealing with high dimensional data, it is rare that ABC methods directly use the full data set since it it highly unlikely that $y \approx y_{obs}$ will occur due to the curse of dimensionality. Having an summary statistics that can retain sufficient information while significantly reduces dimension helps to attain better quality approximation for the same computational cost. Therefore, summary statistics as well as methods for dimension reduction have been a vital part in performing ABC (Sisson et al., 2018).

As comparing the simulated data with observe data is central to ABC, distance measure is also a factor to be considered. Though the importance of distance measure is often overlooked, it can significantly affect the quality and efficiency of ABC.

In this study we aim to assess the performance of the ABC with different combination of summary statistics and distance measure with or without the use dimension reduction. As a part of this study, we will examine the effects of the choice of summary statistics and distance measure has on how closely the result matches with that of the true posterior as well as the potential benefit of dimension reduction methods. This would be done by examining observed data generated from normal distribution.

The structure of this study is as follows. Section 2 gives some background information about Bayesian inference and approximate Bayesian computation. In Section 3, we conduct a literature review on existing summary statistics, dimension reduction methods and distance measures. We introduce the procedures for the experimentation in Section 4, and present the results in Section 5. Section 6 is for discussion of the results and conclusion.

2 Background

2.1 Bayesian Inference

In classical statistics, variable Y is considered to be random while parameter vector θ is considered to be fixed. On the other hand, in Bayesian inference, both Y and θ are treated as random variables with joint probability $p(\theta, y) = \pi(\theta)p(y|\theta)$ (Young et al., 2005). From this, probability statements that are conditional on y_{obs} such as posterior $\pi(\theta|y_{obs})$ or posterior predictive $p(y|y_{obs})$ can be produced.

There are generally three steps in Bayesian inference: constructing a joint probability model for both y and θ, calculating the posterior distribution $\pi(\theta|y_{obs})$, and evaluating the fit of the model. If necessary, one can modify the model and repeat the process (Gelman et al., 2013).

The posterior is calculated by conditioning on the observed data y_{obs} with Bayes's Theorem:

$$\pi(\theta|y_{obs}) = \frac{\pi(\theta)p(y_{obs}|\theta)}{p(y_{obs})},$$

where:

$$p(y_{obs}) = \begin{cases} \int_\Theta \pi(\theta)p(y_{obs}|\theta)d\theta, & \theta \text{ continuous} \\ \sum_\Theta \pi(\theta)p(y_{obs}|\theta), & \theta \text{ discrete} \end{cases}$$

With fixed observed data, $p(y_{obs})$ can be considered as a constant, and we can rewrite the posterior density as:

$$\pi(\theta|y_{obs}) \propto \pi(\theta)p(y_{obs}|\theta)$$

This may be written in word as:

$$posterior \propto prior \times likelihood$$

This expression highlight the importance the likelihood function. It allows observed data to modify the prior knowledge of θ and influence the posterior (Box and Tiao, 2011; Gelman et al., 2013).

2.2 Approximate Bayesian computation

Typically, the goal of inference is to calculate integrals with the posterior density such as quantiles or predictive quantities (Gelman et al., 2013). However, the complexity of the posterior distribution makes such integral impossible to be calculated directly. Therefore, numerical methods such as the Markov Chain Monte Carlo (MCMC) or sequential Monte Carlo (SMC) are employed to produce the necessary integrals. These methods require that samples of θ can be drawn from the posterior distribution. For example, in the case of the MCMC algorithm, Metropolis-Hasting, the integral is evaluated by drawing θ from a Markov Chain with transition probability defined as $\alpha(\theta, \theta') = \min\{1, \frac{\pi(\theta'|y_{obs})q(\theta,\theta')}{\pi(\theta|y_{obs})q(\theta',\theta)}\}$, where $q(\theta, \theta')$ is some proposal density. Therefore, those methods still rely on the evaluation of the likelihood function (Sisson et al., 2018).

In practice, explicit evaluation of the likelihood function can be impossible or impractical when the model is complex. ABC provides a framework that utilize simulation to by pass evaluating the likelihood function. This can be done because, even when the likelihood function is intractable, simulating data from the model is often straightforward (Drovandi et al., 2011). First, distances between simulations y generated from $\theta \sim \pi(\theta)$ and y_{obs} is calculated. To reduce to computational cost, distances can be calculated between summary statistics instead of the data. Then, a set of θ that produces simulations close to y_{obs} is selected to form an approximated posterior (Blum et al., 2013).

Following Fearnhead and Prangle (2012), we can define the ABC posterior from the following components:

i) a function $S(\cdot)$ which transform n-dimensional data y to a lower d-dimensional summary statistic s,

ii) a distance measure $\|\cdot\|$,

iii) a standard smoothing kernel $K(u)$ where u is d-dimensional and $\int K(u)du = 1$, and

iv) a bandwidth, or scale parameter $h > 0$ of the kernel.

Algorithm 1 ABC Rejection Sampling Algorithm

Input: the prior $\pi(\theta)$ and a procedure for generating data from the model $p(y_{obs}|\theta)$,
$\pi(\theta|y_{obs}) > 0$,
a kernel $K_h(u)$ with bandwidth $h > 0$,
an integer $N > 0$.

Sampling procedure:

for $i = 1, \ldots, N$ **do**
repeat
Generate $\theta^{(i)} \sim \pi(\theta)$.
Generate $y \sim p(y|\theta^{(i)})$.
until $\theta^{(i)}$ is accepted with probability $K_h(\|y - y_{obs}\|)$
end for

Output: $\theta^{(1)}, \ldots, \theta^{(N)} \sim \pi_{ABC}(\theta|y_{obs})$

The full data set y_{obs} is replaced by the summary statistics $s_{obs} = S(y_{obs})$. From this, the posterior distribution $\pi(\theta|y_{obs})$ is approximated as $\pi(\theta|s_{obs}) \propto p(s_{obs}|\theta)\pi(\theta)$. An informative summary statistic s_{obs} yields a good approximation $\pi(\theta|s_{obs}) \approx \pi(\theta|y_{obs})$ as it retains a good amount of information of y_{obs}. An exact sufficient s_{obs} yields true posterior $\pi(\theta|s_{obs}) = \pi(\theta|y_{obs})$.

Furthermore, as $p(s_{obs}|\theta)$ is likely to be intractable when $p(y_{obs}|\theta)$ is intractable, we can construct the approximated posterior $\pi_{ABC}(\theta|s_{obs}) \approx \pi(\theta|s_{obs})$, where

$$\pi_{ABC}(\theta|s_{obs}) = \int \pi(\theta, s|s_{obs})ds \tag{1}$$

with

$$\pi(\theta, s|s_{obs}) \propto K_h(\|s - s_{obs}\|)p(s|\theta)\pi(\theta) \tag{2}$$

where $K_h(\cdot)$ is a smoothing kernel. The most basic kernel function is the indicator function $I(\|s - s_{obs}\| \leq h)$, which is equivalent to an "if $\|s - s_{obs}\| \leq h$" statement in algorithms. This, however, means that there is no discrimination between a θ that produces $\|s - s_{obs}\|$ close to 0 and a θ that produces $\|s - s_{obs}\|$ close to h. Therefore, it can be more efficient to use kernel functions that concentrate more at 0 and less further away. One simple example of such kernel is the triangular kernel $K_h(u) = (1 - |u/h|)I(|u/h| < 1)$ (Sisson et al., 2018).

Two approximations to the posterior distribution made by ABC methods to allow avoiding explicit evaluation of intractable likelihood functions were detailed by Blum et al. (2013). Firstly, the full data set is replaced with summary statistics. Secondly, $\pi(\theta|s_{obs})$ is replaced with $\pi_{ABC}(\theta|s_{obs})$ calculated in equation 1. With those approximations, ABC aims to approximate the true posterior of θ so that it can be used for inference about θ.

The approximation form will vary depending on $S(\cdot)$, $K(\cdot)$ and h. If $S(\cdot)$ is chosen as the identity function, we can obtain Algorithm 1, introduced by Pritchard et al. (1999).

Additionally, when $h \to 0$, equation (2) becomes

$$\begin{aligned}\lim_{h\to 0} \pi_{ABC}(\theta, y|y_{obs}) &\propto \lim_{h\to 0} K_h(\|y - y_{obs}\|)p(y|\theta)\pi(\theta) \\ &= \delta_{y_{obs}}(y)p(y|\theta)\pi(\theta).\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{h\to 0} \pi_{ABC}(\theta|y_{obs}) &\propto \int \delta_{y_{obs}}(y)p(y|\theta)\pi(\theta)dy \\ &= p(_{obs}|\theta)\pi(\theta).\end{aligned}$$

When $h \to 0$, the algorithm produces θ sample from the true posterior. Nonetheless, $h = 0$ is not practical as it corresponds with a zero acceptance rate for continuous distributions. For a more general summary statistics $S(\cdot)$, an importance sampling version for ABC can be implemented is as demonstrated in Algorithm 2. Unlike rejection sampling, importance sampling requires a proposal density $g(\theta)$ to draw θ from, and θ values that produce s close enough to s_{obs} are given weight $\frac{\pi(\theta)}{g(\theta)}$.

Algorithm 2 ABC Importance Sampling Algorithm

Input: the prior $\pi(\theta)$ and a procedure for generating data from the model $p(y_{obs}|\theta)$,
a proposal density $g(\theta)$ such that $g(\theta) > 0$ when $\pi(\theta|y_{obs}) > 0$,
a kernel $K_h(u)$ with bandwidth $h > 0$,
an integer $N > 0$.

Sampling procedure:

for $i = 1, \ldots, N$ **do**
Generate $\theta^{(i)} \sim g(\theta)$.
Generate $y \sim p(y|\theta^{(i)})$.
Weight $w^{(i)} = \frac{\pi(\theta^{(i)})}{g(\theta^{(i)})}$ is set for $\theta^{(i)}$ with probability $K_h(\|s - s_{obs}\|)$, else set $w^{(i)} = 0$
end for

Output: a set of $\{\theta^{(i)}\}_{i=1}^N$ and the corresponding weights $\{w^{(i)}\}_{i=1}^N$

As demonstrated in the ABC rejection sampling algorithm and the importance sampling algorithm, the two approximations detailed by Blum et al. (2013) circumvent the need to evaluate the intractable posterior and likelihood function. However, it should be noted that there is typically a trade-off between them. While large dimension of $S(\cdot)$ means that the quality of the first approximation $\pi(\theta|y_{obs}) \approx \pi(\theta|s_{obs})$ is good, it also reduce the efficiency of the kernel smoothing, making the second approximation (2) poorer. On the other hand, if $S(\cdot)$ has low dimension, which improves the bandwidth and the quality of the second approximation (2), loss of information from the mapping would make the first approximation poor.

There have been much work to address this trade-off by developing more efficient sampling algorithms such as Markov chain Monte Carlo and sequential Monte Carlo. However, having a low-dimensional and near-sufficient summary statistic $S(\cdot)$ remains to be crucial for achieving a good trade-off. Therefore, the choice of summary statistics is of central importance in ABC (Blum et al., 2013).

3 Summary statistics and distance measures

3.1 Summary statistics

3.1.1 Overview

When dealing with high dimensional data, it is rare that ABC methods directly use the full data set since it it highly unlikely that $y \approx y_{obs}$ will occur due to the curse of dimensionality. Having an summary statistics that can retain sufficient information while significantly reduces dimension helps to attain better quality approximation for the same computational cost. Therefore, reducing data to lower dimensional summary statistics has been a vital part in performing ABC (Prangle, 2015).

In practice, there usually is a trade-off between quality and computing cost in ABC. Applying summary statistics with low dimension or setting a large bandwidth reduce computational cost, but also lower the quality of the approximation (Blum et al., 2013). While having more complex or higher dimensional summary statistics and lower bandwidth preserve better information, it also leads to an increase in computational cost.

One important characteristic aside from dimension of summary statistics is sufficiency. In classical statistic, a summary statistic is sufficient if the conditional density $\pi(y|s, \theta)$ is invariant to θ. More relevant to ABC, a summary statistics is Bayes sufficient if $\theta|s$ has the same distribution as $\theta|y$ under any prior distribution and for almost any y. Concepts and methods stemming from sufficiency such as approximate and asymptotic sufficiency or comparing sufficiency results can also be useful in ABC (Prangle, 2015).

Naturally, a minimal sufficient statistic is the most efficient choice for ABC. However, in practice, it is often challenging or unachievable to determine sufficient summary statistics aside from the trivial full data set. Hence, there have been on going efforts to develop methods for constructing summary statistics for ABC from low-dimensional but insufficient summary statistics. Prangle (2015) divides summary statistic selection methods into the following strategies: subset selection (with regularization also included), projection, and auxiliary likelihood. All of those strategies require subjective inputs from the user.

3.1.2 Subset selection

Subset selection methods attempt to select an informative subset from a set of candidate summary statistics $z = (z_1, z_2, ..., z_n)$. Some subset selection methods are: Approximate sufficiency, entropy/loss minimization, mutual information, and regularization.

Approximate sufficiency method, proposed by Joyce and Marjoram (2008), attempts to determine the optimal subset by adding or removing one summary statistic at a time and evaluate the effect on the ABC posterior. The incentive of the approach is if $S(\cdot)$ is minimally sufficient, adding more summaries will not change $\pi(\theta|s_{obs})$ while removing any will.

Nunes and Balding (2010) suggested a two-stage approach, entropy and loss minimization. The motivation is that the more informative the posterior is, the lower is its entropy. Summaries subset are first selected by minimizing entropy, then further optimize the chosen subset by minimizing root mean square loss.

Mutual information method was introduced by Barnes et al. (2012). The mutual information between $S(y)$ and θ is maximized when the $S(\cdot)$ is sufficient. With this, $S(\cdot)$ is sufficient when the Kullback-Leiber (KL) divergence of $\pi(\theta|s_{obs})$ and $\pi(\theta|y_{obs})$ is zero:

$$\int \pi(\theta|y_{obs}) \log(\frac{\pi(\theta|y_{obs})}{\pi(\theta|s_{obs})}) d\theta = 0.$$

From this, the subset is selected in a stepwise manner, adding a summary statistic to the subset until the improvement to the KL divergence is below a threshold.

Sedki and Pudlo (2012) and Blum et al. (2013) suggested regularization approaches that find an informative subset by fitting a linear regression with response θ and variables z. Variables are then selected stepwise by minimizing AIC or BIC.

The subset selection methods allow for intuitive interpretation of the results. As such, they are useful in making model improvements. However, all of the methods above can be extremely computationally expensive. The first three approaches requires ABC to be run many times (one for each subset of z). Therefore, they are only practical with rejection or importance sampling ABC, which allows reusing of simulated data sets. Furthermore, the number of test statistics in z in each method is also limited since the computational cost increase drastically with the size of z (Prangle, 2015).

3.1.3 Projection

Projection methods aims to find an informative projection of z using training data (θ, y) created by simulation. Similarly to the subset selection methods, projection methods also require a vector of candidate summary statistics $z(y) = (z_1, z_2, ..., z_n)$.

Wegmann et al. (2009) proposed to use partial least-squares (PLS) to choose informative statistics from z. The motivation for this method is that PLS can objectively reduce the dimension of the summary statistics while retaining the desirable amount of information. PLS aims to construct uncorrelated linear combinations u of covariates of z that have high covariance with parameters θ. The method chooses the first c components such that the root mean square error of the linear regression of θ from u is minimized.

Another method is introduced by Fearnhead and Prangle (2012), which uses a linear model to fit the training data: $\theta \sim N(Az + b, \Sigma)$. The motivation behind this method is that estimated parameters $\hat{\theta}$ from the linear regression is the optimal choice in term of quadratic loss. The method first generates a set of simulated parameters and data. From this, $\hat{\theta}$ are generated using linear regression and used as summary statistics. Alternatively, $\hat{\theta}$ can also be produced by methods such as lasso or canonical analysis. The quality of the linear regression estimate can be further improved by using BIC values to select good $z(y)$ features.

Aeschbacher et al. (2012) suggested using boosting to produce predictors $\hat{\theta}(y)$ of $\mathrm{E}(\theta|y)$ to be used as ABC summary statistics. The procedure is repeated for each component of θ. "Weak" estimators are added one at a time to concentrate on data that was fitted poorly in the previous step until a "strong" estimator is formed.

All three methods require training data (θ, y). Wegmann et al. (2009) do this is drawing θ from the prior and y from the model. Aeschbacher et al. (2012) perform a ABC using all data features and use the result as training data for boosting. Fearnhead and Prangle (2012) approach this differently by performing the pilot ABC using ad-hoc summary statistics to find the training region from which θ will be drawn from.

Projection methods allows for a wider space of potential summary statistics as there can be a larger number of features $z(y)$ and they are not limited to the subset of z. They also avoid the high computational cost that comes with repeating calculations as in subset selection. However, this also make the results more difficult to interpret and assess (Prangle, 2015).

3.1.4 Auxiliary likelihood

The auxiliary likelihood methods is an indirect approach that derive summary statistics from a suitably chosen simpler auxiliary model with parameters ϕ and tractable likelihood $p_A(y|\phi)$. The motivation for this class of methods is that the sufficiency of summary statistics derived from auxiliary models can be evaluated.

The indirect parameter estimates method (ABC-IP) defines the summary statistics as

the maximum likelihood estimator under the auxiliary model:

$$s = \hat{\phi}(y) = \text{argmax}_{\phi} p_A(y|\phi)$$

Gleim and Pigorsch (2013) show that if the generative model (the original model of interest) is nested in the auxiliary model, $\hat{\phi}(y)$ would be asymptotically sufficient for the generative model. Even though this is rare in practice, Bayesian consistency can be attained allowing for discrimination between data coming from different parameters θ.

Introduced by Gleim and Pigorsch (2013), likelihood distance method (ABC-IL) is a variation of ABC-IP. ABC-IL uses the log likelihood ratio of the auxiliary model between the MLEs of observed data and simulated data as the distance measure:

$$\|\hat{\phi}(y), \hat{\phi}(y_{obs})\| = \log p_A(y_{obs}|\hat{\phi}(y_{obs})) - \log p_A(y_{obs}|\hat{\phi}(y))$$

$\hat{\phi}(y)$ should behaves such that $\|\hat{\phi}(y), \hat{\phi}(y_{obs})\| = 0$ if and only if $\hat{\phi}(y) = \hat{\phi}(y_{obs})$.

Gleim and Pigorsch (2013) also suggested taking the summary statistics as the score of the auxiliary likelihood under $\hat{\phi}(y_{obs})$:

$$s = \left(\frac{\delta}{\delta \phi_i} \log p_A(y|\phi) \Big|_{\phi=\phi_{y_{obs}}} \right)_{1 \leq i \leq p}$$

The motivation of the approach, ABC-IS, is that the score has the desired asymptotic properties while being computationally cheaper as the numerical optimization is done only once for $\hat{\phi}(y_{obs})$. The underlying assumption for this asymptotic sufficiency is that the data is more informative as $n \to \infty$.

The chosen auxiliary likelihood should have a reasonably small number of parameter. It should also allow fast and accurate calculation of the MLE or score. Lastly, it should have summary statistics that are informative of the generative model. This last criteria is harder to judge, and can be approached by goodness-of-fit tests or the BIC as suggested by Gleim and Pigorsch (2013).

Auxiliary likelihood methods also avoid the high computational cost of generating training data or repeatedly running ABC, and the task of choosing data features that subset selection and projection methods require. However, they are still met with a similar decision of auxiliary likelihood, which demands subject area knowledge.

3.2 Distance measures

While the summary statistic has been a subject of central importance in ABC, there has been less consideration on the role of the distance measure $\|\cdot\|$. Distance measures can have a considerable impact on both the efficiency and the quality of ABC.

Some of the most popular distance measures in ABC can be written in the form

$$\|s - s_{obs}\| = (s - s_{obs})^T \Sigma^{-1} (s - s_{obs}).$$

Different choices of Σ will result in different distance measures. Specifying Σ as the identity matrix gives us the Euclidean distance, Σ as the a diagonal matrix gives weighted Euclidean distance, and Σ as the full covariance matrix gives the Mahalanobis distance. The choice of Σ can have significant impact on the quality of the posterior approximation and should be made with regards to the summary statistics. For example, if the summary statistics has significantly different scalings, it should be more reasonable to choose Σ as the diagonal or the full covariance matrix instead of the identity matrix; if there are also known relationships between summary statistics, choosing Σ as the full covariance matrix can be more optimal (Sisson et al., 2018).

4 Methods

In this study, the performance of ABC was investigated under varying factors: summary statistics, distance measures, dimension reduction usage. The dimension reduction method of choice for the study is the projection method using linear regression introduced by Fearnhead and Prangle (2012). Section 4.1 outlines the steps of the experiment. Section 4.2 describes the data samples as well as the simulation process. Section 4.3 provides more information regarding the choice of summary statistics and distance measures. Section 4.4 introduces Wasserstein distance as a measure of the quality of the ABC posterior. Section 4.5 gives some further practical details of the experimentation.

4.1 Experimentation procedures

The experiment for this study followed the steps below for each ABC run:

1) Draw sample data and compute the true posterior from drawn sample data,

2) Generate simulations and compute summary statistics,

3) Dimension reduction by linear regression,

4) Calculating distances $\|s - s_{obs}\|$,

5) Compute ABC posteriors using rejection sampling,

6) Calculate Wasserstein distances between the ABC posterior and the true posterior for each combination of summary statistic set and distance measure.

For each combination of statistics set and distance measures, simple rejection sampling ABC without dimension reduction was also done to provide a base line for measuring performance of the projection method.

4.2 Data and simulation details

The data samples $y_{obs} = (y_1, ..., y_n)$ are generated with $y_i \sim \mathrm{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown model parameters. The choice of prior is based on the conjugate prior distribution for normal data presented by Gelman et al. (2013). This choice of conjugated prior allows the true posterior to be generated and used as the reference to determine the performance of ABC. Accordingly, prior for σ^2 is the scaled inverse-$\chi^2(\nu_0, \sigma_0^2)$ distribution. The conditional distribution for μ given σ^2 is $\mathrm{N}(\mu_0, \sigma^2/\kappa_0)$.

The simulation process has the following steps. First, σ^2 is drawn from $\mathrm{Inv} - \chi^2(\nu_0, \sigma_0)$. Then, μ is drawn from the conditional distribution $\mathrm{N}(\mu_0, \sigma^2/\kappa_0)$. Finally, the simulated data is drawn from $\mathrm{N}(\mu, \sigma^2)$.

From the conjugate prior, the true posterior is computed as follow:

$$\begin{aligned}\sigma^2|y &\sim \mathrm{Inv} - \chi^2(\nu_n, \sigma_n^2)\\ \mu|\sigma^2 &\sim \mathrm{N}(\mu_n, \sigma^2/\kappa_n),\end{aligned}$$

where,

$$\begin{aligned}\kappa_n &= \kappa_0 + n\\ \nu_n &= \nu_0 + n\\ \mu_n &= \frac{\kappa_0}{\kappa_n}\mu_0 + \frac{n}{\kappa_n}\bar{y}_{obs}\\ s^2 &= \sum (y_i - \bar{y}_{obs})^2\\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y}_{obs} - \mu_0)^2.\end{aligned}$$

4.3 Choice of summary statistics and distance measures

For this experiment set up, four different sets of summary statistics were chosen. The first set is the set of sufficient summary statistics, sample mean and variance. The second set of summary statistic is a set of quantiles from 5% to 95% with 5% interval, which is informative, but high dimensional. The third set of summary statistics is a set of sample minimum and maximum values, which is highly non-robust. The last set of summary statistics is a combination of statistics, including mean and variance, minimum and maximum, and quantiles with 10% interval. This set mimics summary statistic sets in practice, where there can be both reliable and unreliable summary statistics.

Distance measures for this experiment include the three cases described in Section 3.2, Euclidean, weighted Euclidean, and Mahalanobis distance. More specifically, we set Σ for the weighted Euclidean distance as the diagonal of the covariance matrix, making it standardized Euclidean. The covariance matrix was calculated from the data generated by the simulations.

4.4 Wasserstein distance

In this study, Wasserstein distance is the choice of metric to measure the performance of different combinations of test statistics, distance measures and ABC methods. This is achieved by computing the distance between the ABC posterior and the true posterior for each test case.

Wasserstein, or earth mover distance is a metric for measuring distance between probability distributions on a given metric space M (Bernton et al., 2019). Informally, if the distributions are viewed as piles of dirt, Wasserstein distance is the minimum cost of turning one pile into the other by moving an amount of dirt by some distance.

The p-Wasserstein distance between distribution u and v belonging in the set of all couplings on $M \times M$, $\Gamma(u, v)$, is defined as

$$W_p(u, v)^p = \inf_{\gamma \in \Gamma(u,v)} \int_{M \times M} \rho(x, y)^p \mathrm{d}\gamma(x, y).$$

For the simple case of univariate distributions such as the posteriors in this experiment, the 1-Wasserstein distance can be written as $\int_{-\infty}^{\infty} |U - V|$, where U and V are the respective cumulative distribution functions.

The correlation between the similarity of two distributions and the Wasserstein distance between them is demonstrated visually in Figure 1 in Section 5.

4.5 Implementation practicality

For this study, the experiment outlined in Section 4.1 was run 100 times. The sample data y_{obs} has 1000 iid observations generated from $\mathrm{N}(0, 1)$. The parameters chosen for the scaled Inv-χ^2 prior is $\nu_0 = 16$ and $\sigma_0^2 = 0.5$. For each run, 1 million sets of μ, σ^2, and simulated data of 1000 data points was generated. The bandwidth h is not explicitly defined, but was set as the 0.3% quantile of the set of distances $\|s - s_{obs}\|$. With the triangular kernel, the acceptance rate for the ABC was between 0.03% and 0.1%.

5 Results

The statistics set of sample mean and variance performed the best and most consistently, with mean Wasserstein distance between 0.05 and 0.056. On the other hand, the sample minimum and maximum had poor performances across all test case, with mean Wasserstein distances from 0.247 to 0.253. ABC with quantiles and mixed summary statistics performances showed more dependence on other factors, as shown in Figure 2 and Table 1 .

For the data model in this study, Euclidean distance produced the best results in most cases, except in the case of mixed summary statistics without dimension reduction. For most test cases, results of Mahalanobis distance closely matched those of standardized Euclidean distance. However, Mahalanobis distance gave very poor results for quantiles and mixed statistics without dimension reduction.

Overall, ABC with dimension reduction by linear regression showed significant improvement for quantiles and mixed summary statistics and marginal effect for sufficient statistics. Notably, the dimension reduction method produced vast improvement for the cases with Mahalanobis distance mentioned above.

Table 1: Mean Wasserstein distance between ABC posterior and true posterior

Method	Summary statistics	Euclidean	Standardized Euclidean	Mahalanobis
Without dimension reduction	Mean and variance	0.0516	0.0558	0.0554
	Quantiles	0.0701	0.0814	0.3048
	Minimum and maximum	0.2488	0.2488	0.2476
	Mixed	0.0854	0.0646	0.2306
Dimension reduction with linear regression	Mean and variance	0.0508	0.0532	0.0557
	Quantiles	0.0611	0.0641	0.0645
	Minimum and maximum	0.2508	0.2528	0.2476
	Mixed	0.0545	0.0582	0.0586

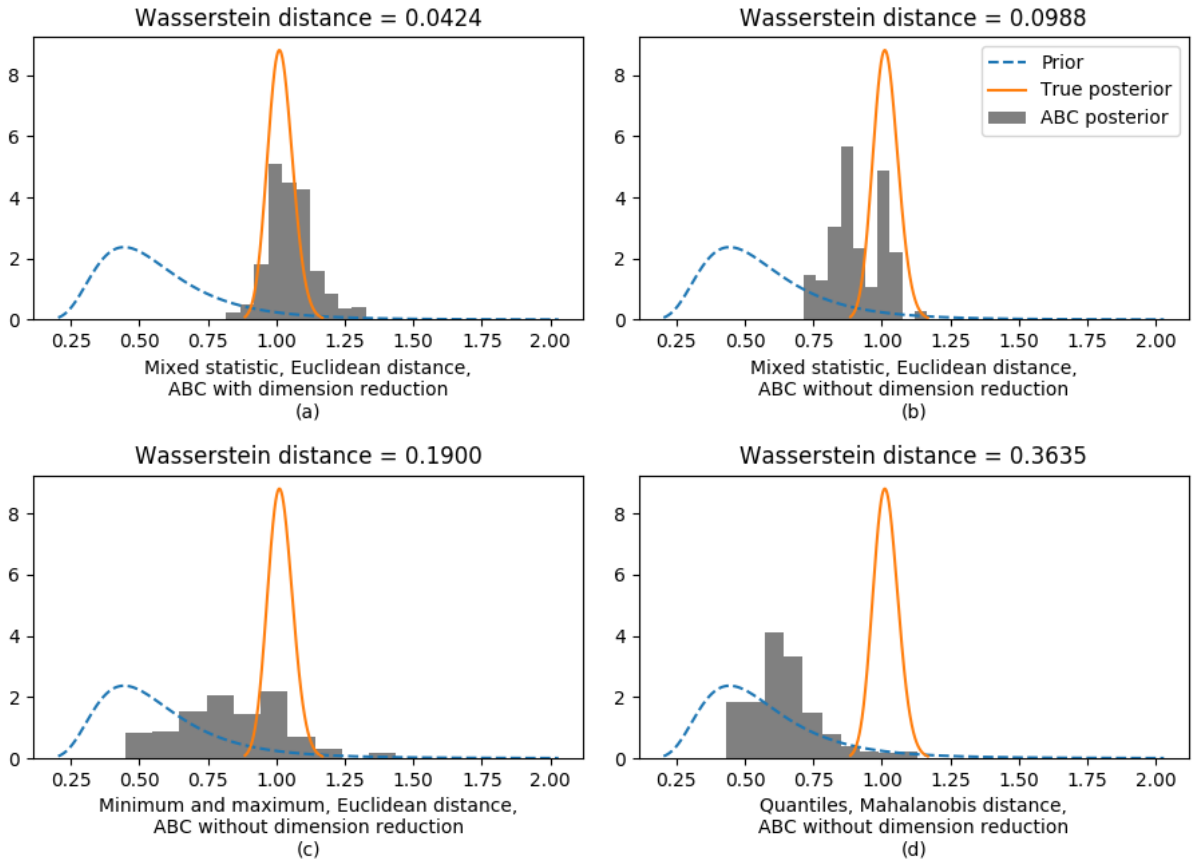

Figure 1: Various ABC posteriors (histograms) with different level of fit measured by Wasserstein distance, true posterior (solid line), and prior (dashed line). The posteriors were obtained in a single run of ABC

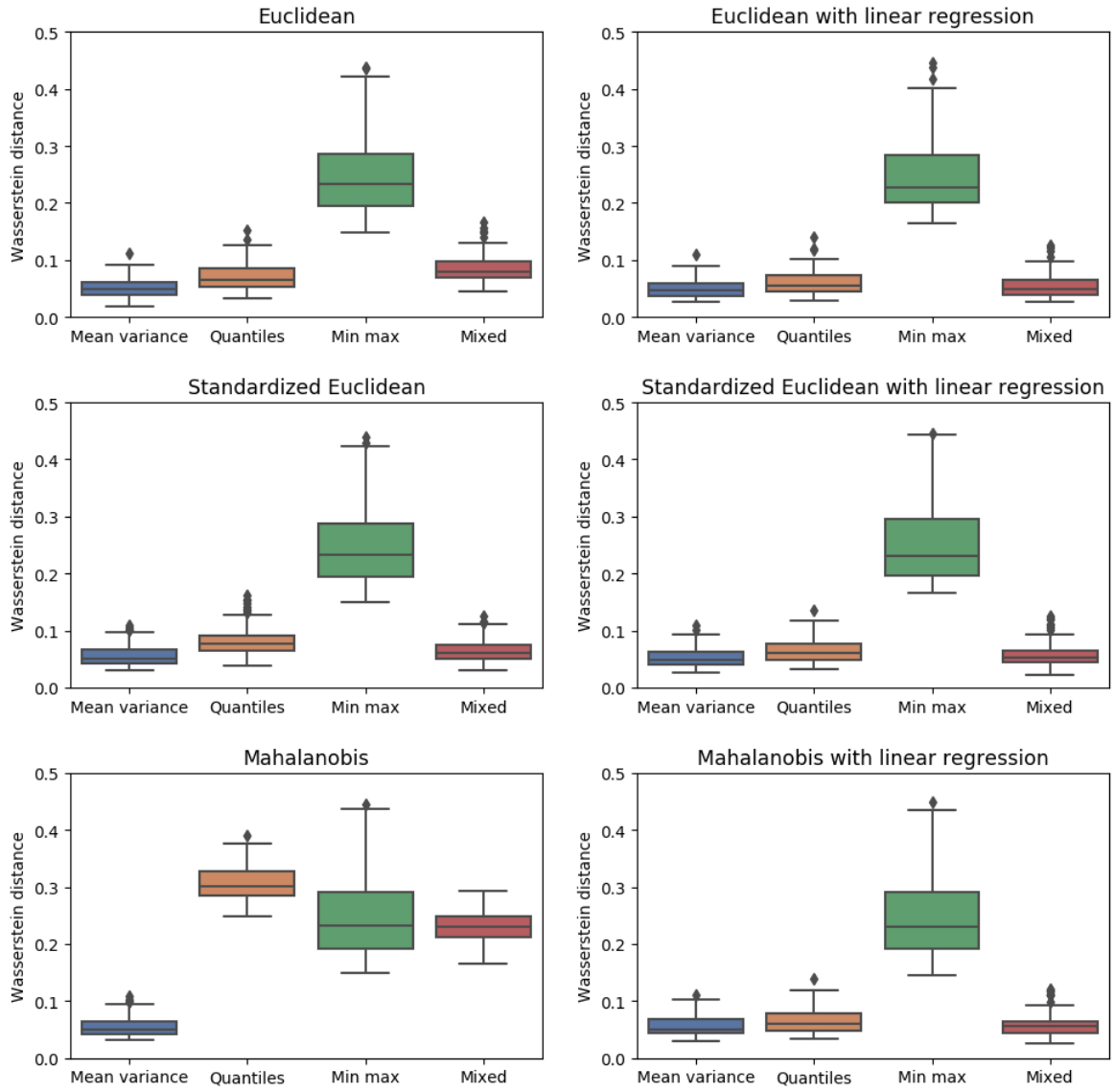

Figure 2: Wasserstein distance distributions of all the test cases. Left column shows the baseline ABC results and the right column shows the results from ABC with linear regression dimension reduction. Rows illustrate results obtained with Euclidean distance (top), standardized Euclidean distance (middle), and Mahalanobis distance (bottom).

6 Discussion and conclusion

The choice of statistics has the most substantial impact on the quality of ABC. As expected, the sufficient statistic set of sample mean and variance yielded the results closest to the true posterior regardless of other factors. Informative statistic sets such as quantiles or mixed statistics yielded results that show some bias to the left toward the prior such as in Fig. 1(b). This bias can be attributed to the high dimensionality of the two statistic sets. Non-robust statistic set of sample minimum and maximum yielded uninformative results that are clearly biased toward the prior as shown in Fig. 1(c).

Dimension reduction was shown to have significant impact on the quality of ABC posterior. More specifically, the method proposed by Fearnhead and Prangle (2012) improved the Wasserstein distance of mixed statistics to close to that of sufficient statistics. This is achieved with a relatively low additional computational cost of the linear regression step, taking 20 to 30 seconds in each ABC run that was approximately 600 to 700 seconds long.

For the notable cases of quantiles and mixed statistic sets combined with Mahalanobis distance, the most probable cause would be the high correlation of statistics. As the quantile statistics are highly correlated, the calculation of the inverse covariance matrix required for the Mahalanobis distance becomes unreliable, making the distance measure less effective.

In this experiment, the effects of distance measure were shown to be less pronounced compared to that of summary statistics, although this could partly be attributed to the simplicity of the chosen data model. More complex models with parameters of various scales and correlations can reveal more about the effectiveness of different measure distances.

In this study, the performance of ABC was investigated under various contexts. Through the experiment, the central importance of summary statistics is reaffirmed. Dimension reduction with linear regression was proven to be an effective method to improve the performance of ABC, especially when sufficient statistics cannot be determined, and sets of informative but high dimension statistics has to used. The study also showed potential drawbacks of Mahalanobis distance when applied to high dimensional and highly correlated summary statistics.

Finally, there are a few direction for this study to be extended. More complex models can be chosen for the experiment to gain more meaningful insight to the impact of distance measure. Moreover, factors such as sample size and number of simulations per run can be varied. From this, we can have a more realistic view of how ABC perform under additional constrains such as small sample size or limited computing power.

Acknowledgements

I would like to thank Jukka Sirén for providing invaluable guidance throughout the whole process of writing this thesis.

References

Simon Aeschbacher, Mark A Beaumont and Andreas Futschik. A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, 192(3): 1027–1047, 2012.

Chris P Barnes, Sarah Filippi, Michael PH Stumpf and Thomas Thorne. Considerate approaches to constructing summary statistics for ABC model selection. *Statistics and Computing*, 22(6):1181–1197, 2012.

Espen Bernton, Pierre E Jacob, Mathieu Gerber and Christian P Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269, 2019.

Michael GB Blum and Olivier François. Non-linear regression models for Approximate Bayesian Computation. *Statistics and computing*, 20(1):63–73, 2010.

Michael GB Blum, Maria Antonieta Nunes, Dennis Prangle, Scott A Sisson et al. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013.

George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

Peter J Diggle and Richard J Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):193–212, 1984.

Christopher C Drovandi, Anthony N Pettitt and Malcolm J Faddy. Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):317–337, 2011.

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.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Alexander Gleim and Christian Pigorsch. Approximate Bayesian computation with indirect summary statistics. *Draft paper: http://ect-pigorsch. mee. uni-bonn. de/data/research/papers*, 2013.

Paul Joyce and Paul Marjoram. Approximately sufficient statistics and Bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

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.

Dennis Prangle. *Handbook of approximate Bayesian computation*, chapter 5. Chapman and Hall/CRC, 2015.

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.

MA Sedki and P Pudlo. Contribution to the discussion of Fearnhead and Prangle (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B*, 74:466–467, 2012.

Scott A Sisson, Yanan Fan and Mark Beaumont. *Handbook of approximate Bayesian computation*, chapter 1. Chapman and Hall/CRC, 2018.

Daniel Wegmann, Christoph Leuenberger and Laurent Excoffier. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182(4):1207–1218, 2009.

G Alastair Young, George Albert Young, Thomas A Severini, RL Smith, Robert Leslie Smith et al. *Essentials of statistical inference*. Cambridge University Press, 2005.