APPROXIMATE BAYESIAN COMPUTATION FOR IMPLICIT STATISTICAL MODELS

AIDAN GLEICH

Abstract. This thesis covers the theory and application of a likelihood-free Bayesian inference method known as Approximate Bayesian Computation (ABC). Approximate Bayesian Computation has become an important method for estimating complex statistical models. The likelihood functions of such models are computationally expensive or impossible to compute, necessitating the use of likelihood-free inference methods. We begin with some fundamental concepts from the field of nonparametric statistics. We then introduce Bayesian inference before walking through an ABC algorithm. We conclude by using the algorithm to estimate a model from survival analysis.

Contents

Date: May 12, 2021.

This document is a senior thesis submitted to the Mathematics and Statistics Department of Haverford College in partial fulfillment of the requirements for a major in Mathematics.

1. Introduction

The likelihood function plays a central role in both frequentist and Bayesian statistics. Maximum likelihood estimation, a fundamental frequentist model fitting technique, requires maximizing the likelihood function over the parameter space. In Bayesian statistics, calculating the posterior density function requires the likelihood function, as do posterior sampling techniques such as Markov Chain Monte Carlo. While every model has an underlying likelihood function, it is not necessarily analytical or feasible to compute.

Restricting the choice of model by requiring an analytic likelihood function may reduce the ability to effectively model certain phenomena. However, without a likelihood function, techniques such as maximum likelihood estimation cannot be used. Thus, methods for carrying out statistical inference without a likelihood function are desirable. Such methods are referred to as likelihood-free.

Simulation is one approach to likelihood-free inference. If one can simulate data from a model for a given parameter value, then that simulated data can be compared to the observed data in an attempt to approximate the likelihood function. A parameter value that creates simulated data similar to the observed data is assumed to have a higher likelihood than a parameter value which creates simulated data quite different from the observed values. Approximate Bayesian Computation (ABC) methods use this concept to conduct Bayesian inference without requiring computation of a likelihood function.

The theoretical foundations of the algorithms draw heavily from the field of nonparametric statistics, which is where this paper will begin. It will then cover the basics of Bayesian estimation before introducing a simple ABC estimation algorithm. The paper concludes by applying the algorithm to an example from the field of survival analysis.

2. Kernel Density Estimation

2.1. **Random Variables.**

The theory of random variables forms the basis of the concepts discussed in this paper. The following subsection follows the structure of Chapter 1 of Severini [10] and Chapter 1 of Wasserman [11]. We begin with the concept of an **experiment**. An experiment is a random process with a well-defined set of potential outcomes, only one of which may occur each time the process is run [10]. We refer to the set of potential

outcomes of an experiment as the **sample space** Ω. Subsets of Ω are called **events** and will typically be denoted by A. Wasserman [11, §1.2] defines a **probability measure** P as follows:

Definition 2.1. A probability measure P is a function defined on a set $\mathcal{A} \subset 2^{\Omega}$ with the following properties:

- $P(\Omega) = 1$
- $P(A) \geq 0$ for all $A \in \mathcal{A}$
- If $A_1, A_2, \ldots$ are pairwise disjoint sets, then

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

A **probability space** is a triple $(\Omega, \mathcal{A}, P)$. Allow $\omega \in \Omega$ to be the outcome of an experiment. The following definition is based on the Severini's definition in [10, §1.2].

Definition 2.2. Given a probability space $(\Omega, \mathcal{A}, P)$, a **random variable** X is a function $X : \Omega \to \mathcal{X}$ where $\mathcal{X} \subseteq \mathbb{R}^n$ for some $n \in \mathbb{N}$.

Given a probability space $(\Omega, \mathcal{A}, P)$ and a random variable $X : \Omega \to \mathcal{X}$, the probability function P can be used to find the probabilities corresponding to X [10, §1.2]. Taking $A \subset \mathcal{X}$, we define P_X to be the probability function

$$P_X(A) = P\left(\{\omega \in \Omega : X(\omega) \in A\}\right)$$

When the distinction is necessary, the term **real-valued random variable** may be used to refer to a random variable whose codomain is a subset of $\mathbb{R}$. For random variables with codomain of dimension greater than 1, the term **random vector** may be used. In this paper, all random vectors will be column vectors. Any n-dimensional random vector X will be written as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

where $X_1, X_2, \ldots, X_n$ are real-valued random variables.

While the probability function P_X encodes all properties of the random variable X, it can be useful to limit the potential inputted sets A. One such limitation leads to the distribution function of X. The following definition paraphrases Severini [10, §1.4]

Definition 2.3. The **distribution function** of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ given by

$$F(x) = P(\omega \in \Omega : X(\omega) \leq x) = P(X \leq x), \qquad -\infty < x < \infty.$$

When X is a random vector of length n, we write

$$F(x_1, x_2, \ldots, x_n) = P\left(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\right).$$

If F is the distribution function of X, we write $X \sim F$.

Unless otherwise noted, we will assume that any function F is twice continuously differentiable.

2.2. **Families of Distributions.** We may define a **family of distributions** as a set of distribution functions where each function is indexed by a parameter $\theta \in \mathbb{R}^d$. For example, we say that a distribution function of the form

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

belongs to the family of exponential distributions with parameter $\lambda > 0$.

2.3. **Error of Estimators.** We will frequently be interested in estimating some quantity or quality of a random variable X using collected sample values of X. If we denote the true value of the quantity we are interested in as θ, we will denote an **estimator** of that quantity as $\hat{\theta}$. Because $\hat{\theta}$ is a function of the random sample $X_1, \ldots, X_n$, it is itself a random variable. We will define some quantities that can be used when discussing the effectiveness of an estimator $\hat{\theta}$

Definition 2.4. Suppose we are given an estimator $\hat{\theta}$ of a quantity θ where $\hat{\theta}$ is a function of $X_1, \ldots, X_n$ which are iid $F : \mathbb{R} \to [0, 1]$. We define the **bias** of $\hat{\theta}$ to be $E(\hat{\theta}) - \theta$. We will denote the bias of $\hat{\theta}$ as $\text{bias}(\hat{\theta})$. An **unbiased** estimator $\hat{\theta}$ is one whose bias is equal to zero.

Given an estimate $\hat{\theta}$ of a quantity θ, it would be useful to have a measure of how accurately $\hat{\theta}$ approximates θ on average. The **mean squared error** is one such measure.

Definition 2.5. Given an estimator $\hat{\theta}$ of some quantity θ, the **mean squared error** of θ is defined to be $E\left[(\hat{\theta} - \theta)^2\right]$.

The mean squared error is an important measure and the following decomposition will prove useful.

Theorem 2.6. *Given an estimator $\hat{\theta}$ of some quantity θ, the mean squared error of $\hat{\theta}$ can be written as follows:*

$$MSE = (bias(\hat{\theta}))^2 + var(\hat{\theta}).$$

Proof. First, note that $\text{bias}(\hat{\theta})^2 = (E(\hat{\theta}) - \theta)^2 = E(\hat{\theta})^2 - 2\theta E(\hat{\theta}) + \theta^2$. Expanding the square within the MSE produces the following:

$$\begin{aligned} E((\hat{\theta} - \theta)^2) &= E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) \\ &= E(\hat{\theta}^2) - 2E(\hat{\theta})\theta + \theta^2 \\ &= E(\hat{\theta}^2) - 2E(\hat{\theta})\theta + \theta^2 + (E(\hat{\theta})^2 - E(\hat{\theta})^2) \\ &= (E(\hat{\theta})^2 - 2E(\hat{\theta})\theta + \theta^2) + E(\hat{\theta}^2) - E(\hat{\theta})^2 \\ &= \text{bias}(\hat{\theta})^2 + \text{var}(\hat{\theta}) \end{aligned}$$

□

2.4. **Empirical Distribution.** Suppose we are interested in estimating the distribution function of a random variable X. Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from the distribution F. We can use the **empirical distribution** to approximate F. The following definition comes from Wasserman [11, §2.1].

Definition 2.7. Given a random sample $X_1, X_2, \ldots, X_n$ from the distribution F, the **empirical distribution** $\hat{F}_n$ is the distribution function that puts mass $\frac{1}{n}$ at each data point X_i. Formally,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\left(X_i \leq x\right)$$

where

$$I(X_i \leq x) : \Omega \to \{0, 1\} = \begin{cases} 1 & X_i(\omega) \leq x \\ 0 & X_i(\omega) > x \end{cases}$$

The empirical distribution is powerful because it allows the data to speak for itself. No assumptions about the form of the true distribution function are required; a random sample is the only component needed to construct the estimator. We can prove some important qualities of the empirical distribution. The following theorem comes from Wasserman [11, §2.1] while the proof is original.

Theorem 2.8. *Let $X_1, X_2, \ldots, X_n$ be iid F and let $\hat{F}_n(x)$ be the empirical distribution. Then:*

$$E(\hat{F}_n(x)) = F(x) \qquad \textit{and} \qquad \textit{var}(\hat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n}$$

Proof. Suppose we have a random variable $X \sim F$ and n independent draws $X_1, X_2, \ldots, X_n$ from X. Because the indicator function $I(X \leq x)$ is a Bernoulli random variable with $p = F(x)$, we know it has mean $F(x)$ and

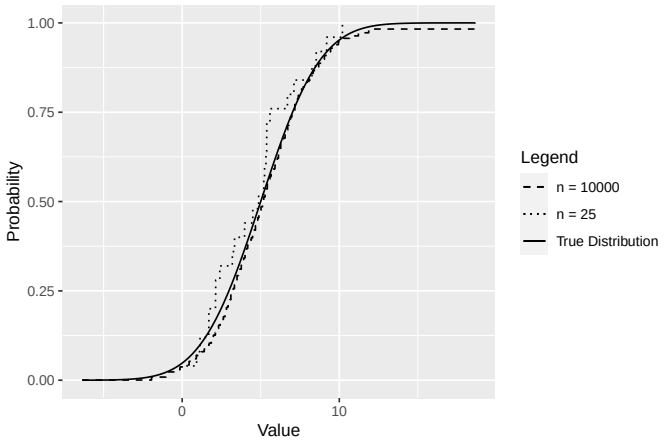

FIGURE 1. Empirical distribution for data simulated from an N(5,3) distribution. The true CDF is plotted in solid black.

variance $F(x)(1-F(x))$. Now we can find the expected value of the empirical distribution $\hat{F}_n$:

$$\begin{aligned} E(\hat{F}_n(x)) &= E(\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(I(X_i \leq x)) \\ &= \frac{1}{n}\sum_{i=1}^{n} F(x) \\ &= F(x) \end{aligned}$$

We can now find the variance:

$$\begin{aligned} \text{var}(\hat{F}_n(x)) &= \text{var}(\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} V(I(X_i \leq x)) \\ &= \frac{1}{n^2} nF(x)(1-F(x)) \\ &= \frac{F(x)(1-F(x))}{n} \end{aligned}$$

□

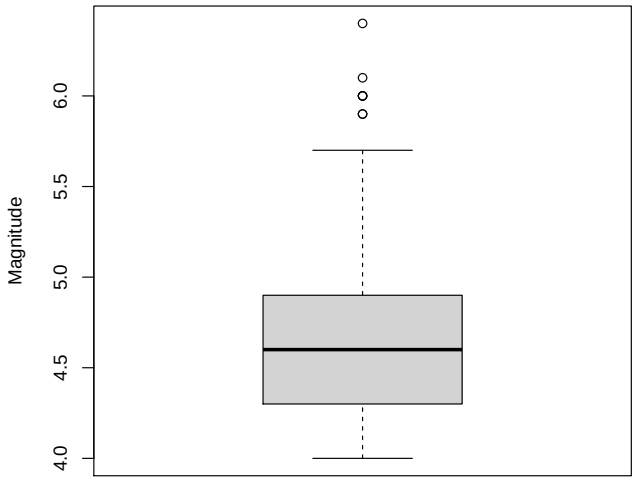

FIGURE 2. Earthquake magnitude data.

We can see empirical distributions from data simulated from an $N(5, 3)$ distribution (where 3 refers to the variance of the distribution) for different values of n plotted against the true CDF in Figure 1. As we should expect from the results of Theorem 2.8, the empirical distribution with higher n closely approximates the true distribution function while the empirical distribution with a small n is noisy.

The plot reveals the discrete nature of the empirical distribution. It is a step function with closed left endpoints and open right endpoints. While in some cases the discreteness may be undesirable if the underlying random variable is continuous, the empirical distribution remains a powerful tool for investigating the shape of both continuous and discrete distribution functions.

Let us look at an example of the empirical distribution using real data. We will look at the magnitude of 1000 earthquakes in Fiji. The data comes from Wasserman [11].

The data are plotted in Figure 2. The data does not appear to belong to an easily recognizable family of distributions on first glance. Thus, using the empirical distribution to approximate the distribution function may provide a more accurate estimate than assuming the data to follow a known distribution.

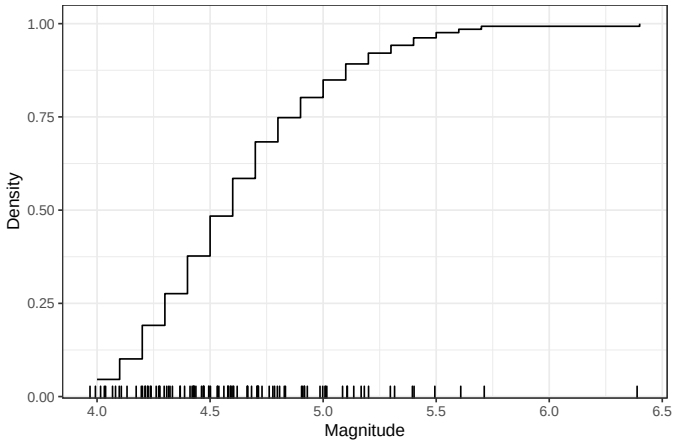

FIGURE 3. Empirical Distribution of Fiji earthquake magnitude data. Vertical lines represent data points and are offset slightly to avoid overlap. Note that the function is left continuous, meaning that each horizontal segment is closed at the left endpoint and open on the right.

The empirical distribution of the earthquake data is shown in Figure 3. We see that the empirical distribution is a step function. This may be undesirable if we believe that the true distribution function F is smooth, as would be the case for the distribution function of a continuous random variable. The next section describes a method to achieve a smooth estimate of the density function of the data.

2.5. **Kernel Density Estimation.** The following subsection loosely follows section 6.2 of Wasserman [11]. Suppose $X_1, \ldots, X_n$ are iid F where F is twice continuously differentiable and we are interested in estimating the probability density function $f = F'$. We can create an estimator $\hat{f}_n$ without assuming that f belongs to a specific family of distributions through the use of kernel functions. The following definition comes directly from Wasserman [11, §4.2].

Definition 2.9. A **kernel function** is any smooth function $K : \mathbb{R} \to \mathbb{R}$ such that $K(x) \geq 0$ and

$$\int_{\mathbb{R}} K(x)dx = 1$$
$$\int_{\mathbb{R}} xK(x)dx = 0$$
$$\int_{\mathbb{R}} x^2K(x)dx > 0$$

where $\int$ refers to the Lebesgue integral.

Throughout this paper, we will use the Gaussian kernel: $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. When necessary, we will show that our results are robust to the choice of kernel function. From Wasserman [11, §6.3], we can now define the estimator we will use:

Definition 2.10. Given a random sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, kernel function K, and a positive real number h, referred to as the **bandwidth**, the **kernel density estimator** $\hat{f}_n : \Omega \to [0, 1]$ is defined as

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right).$$

To better understand how kernel density estimation works, let us look at the term $K\left(\frac{X_i - x}{h}\right)$ in the definition of the KDE. Supposing that $X_i = x_i$, if the x is close to x_i, meaning that the term $x_i - x$ is close to zero, then $K(x_i - x)$ will be relatively large because $K(\cdot)$ is centered at zero. The bandwidth h scales the term $x_i - x$, in effect changing how close x_i must be to x in order for $K\left(\frac{X_i - x}{h}\right)$ to be relatively large. If h is large, $\frac{x_i - x}{h}$ will be brought close to zero, making the size of the term $x_i - x$ less important. When h is small (less than 1), $\frac{x_i - x}{h}$ will be made larger, making the size of $x_i - x$ more important in determining the magnitude of $K\left(\frac{X_i - x}{h}\right)$.

While the choice of kernel is usually unimportant, the choice of bandwidth is very important. Because of this, we need a metric to judge which choice of bandwidth is best for a particular set of data. We will use the expected value of the mean squared error, known as the **risk**, to evaluate the effectiveness of the estimator. We should investigate the properties of the kernel density estimator. Let us being with its expectation.

Lemma 2.11. *Given the Gaussian kernel function with variance σ_K^2 and bandwidth $h > 0$, the expected value of the kernel density estimator $\hat{f}(x)$ is given by*

$$E(\hat{f}(x)) = f(x) + \frac{1}{2} f''(x) h^2 \sigma_K^2 + O(h^4).$$

Proof. Recall that

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right).$$

Let us begin by finding the expected value of $K\left(\frac{X_i - x}{h}\right)$ for a given i. We will use the change of variables $u = \frac{x_i - x}{h}$, which means we can write x_i as

$x + uh$ and $hdu = dx_i$.

$$\begin{aligned} E\left(K\left(\frac{X_i - x}{h}\right)\right) &= \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x_i - x}{h}\right) f(x_i) dx_i \\ &= \int_{-\infty}^{\infty} \frac{1}{h} K(u) f(x + uh) h du \\ &= \int_{-\infty}^{\infty} K(u) f(x + uh) du \end{aligned}$$

We will now need to use a Taylor expansion of $f(x + uh)$ at $a = x$:

$$f(x + uh) = f(x) + f'(x)hu + \frac{1}{2} f''(x)h^2u^2 + \dots.$$

Recall that $\int K(x)\ dx = 1$ and $\int xK(x)\ dx = \int x^3 K(x)\ dx = 0$ due to the symmetry of the kernel function. We can now finish finding the expectation:

$$\begin{aligned} &E\left(K\left(\frac{X_i - x}{h}\right)\right) = \int_{-\infty}^{\infty} K(u) f(x + uh) du \\ &= \int_{-\infty}^{\infty} K(u) \left[f(x) + f'(x)hu + \frac{1}{2} f''(x)h^2u^2 + \cdots\right] du \\ &= \int_{-\infty}^{\infty} K(u) f(x)\ du + \int_{-\infty}^{\infty} K(u) u f'(x) h\ du + \int_{-\infty}^{\infty} K(u) u^2 f''(x) h^2\ du + O(h^4) \\ &= f(x) + \frac{1}{2} f''(x) h^2 \sigma_K^2 + O(h^4) \end{aligned}$$

□

We can now find the MSE of the kernel density estimator. The theorem comes Wasserman and the proof elaborates on a short, incomplete proof he provides [11, §6.3].

Theorem 2.12. *The mean squared error of the kernel density estimator at the point x is given by*

$$MSE_x = \frac{1}{4}\sigma_K^4 h^4 f''(x)^2 + \frac{f(x)\int K(t)^2 dt}{nh} + O\left(\frac{1}{n}\right) + O\left(h^6\right)$$

Proof. We begin by finding the variance of the kernel density estimator.

$$\begin{aligned}
\text{var}\left(\hat{f}(x)\right) &= \text{var}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right) \\
&= \frac{1}{h^2n^2}\text{var}\left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right) \\
&= \frac{1}{h^2n^2}\sum_{i=1}^{n}\text{var}\left(K\left(\frac{X_i - x}{h}\right)\right) \\
&= \frac{1}{h^2n}\text{var}\left(K\left(\frac{X_i - x}{h}\right)\right) \\
&= \frac{1}{h^2n}E\left(K\left(\frac{X_i - x}{h}\right)^2\right) - \frac{1}{h^2n}E\left(K\left(\frac{X_i - x}{h}\right)\right)^2 \\
&= \frac{1}{h^2n}E\left(K\left(\frac{X_i - x}{h}\right)^2\right) - \frac{1}{n}f(x)^2 + O\left(\frac{1}{n}\right)
\end{aligned}$$

For the last equality, we use Lemma 2.11 to write $E\left(K\left(\frac{X_i-x}{h}\right)\right) = f(x) + O\left(h^2\right)$ and the fact that $\frac{O(h^2)}{nh^2} = O\left(\frac{1}{n}\right)$. To finish finding an expression for the variance, we must find $E\left(K\left(\frac{X_i-x}{h}\right)^2\right)$. Also note that the $O\left(\frac{1}{n}\right)$ term can absorb the $\frac{1}{n}f(x)^2$ term, which we will use when we write the final expression at the end of the proof. We use the same Taylor polynomial technique from Lemma 2.11:

$$\begin{aligned}
\frac{1}{h^2n}E\left(K\left(\frac{X_i - x}{h}\right)^2\right) = \\
&= \frac{1}{h^2n}E\left(\int_{-\infty}^{\infty} K(u)^2\left[f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2 + \cdots\right]h\,du\right) \\
&= \frac{1}{hn}f(x)\int_{-\infty}^{\infty} K(u)^2du + O(h^2)
\end{aligned}$$

Thus, combining our expressions, we have that

$$\text{var}\left(\hat{f}(x)\right) = \frac{f(x)\sigma_k^2}{nh} + O\left(\frac{1}{n}\right).$$

Using Lemma 2.11, our expression for the variance, and Theorem 2.6, we arrive at the expression

$$\text{MSE}_x = \frac{1}{4}\sigma_K^4h^4f''(x)^2 + \frac{f(x)\int K(t)^2dt}{nh} + O\left(\frac{1}{n}\right) + O\left(h^6\right)$$

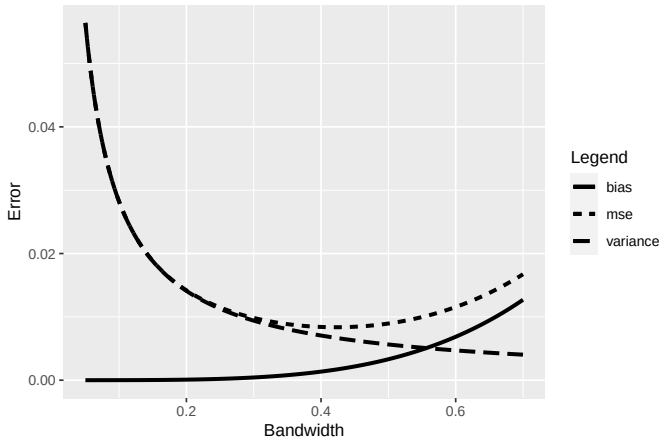

FIGURE 4. Bias-variance decomposition of mean squared error for the kernel density estimator of a N(5,3) with n=100.

□

Let us look at an example of decomposing the MSE of a kernel density estimator. Because our expression for the MSE relies on the true distribution functions f, we will need to explicitly define the distribution function of our random variable. Suppose that $X \sim N(5,3)$ so that $f(x) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-5}{3})^2}$. By simulating 100 independent draws from a $N(5,3)$ distribution, we can plot the integrated bias and variance as functions of the bandwidth h. In Figure 4, we can see the bias and variance of the kernel density estimator as a function of h in terms of the mean squared error. As we should expect, we see that the bias increases with h while the variance decreases with h.

Because the MSE depends of the true value of f, we will need to estimate it as well using cross-validation. The definition of the estimator comes directly from Wasserman [11, §6, 2].

Definition 2.13. Given a random sample $x_1, x_2, \ldots, x_n$ from the distribution F, the **cross-validation estimator of risk** is defined as

$$\hat{J}(h) = \int \hat{f}_n(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{(-i)}(x_i)$$

where $\hat{f}_{(-i)}$ is the kernel density estimator obtained after removing the observed value x_i.

Let us look at an example of the kernel density estimator. We will look at the calcium oxide levels within glass samples measured in weight percentages

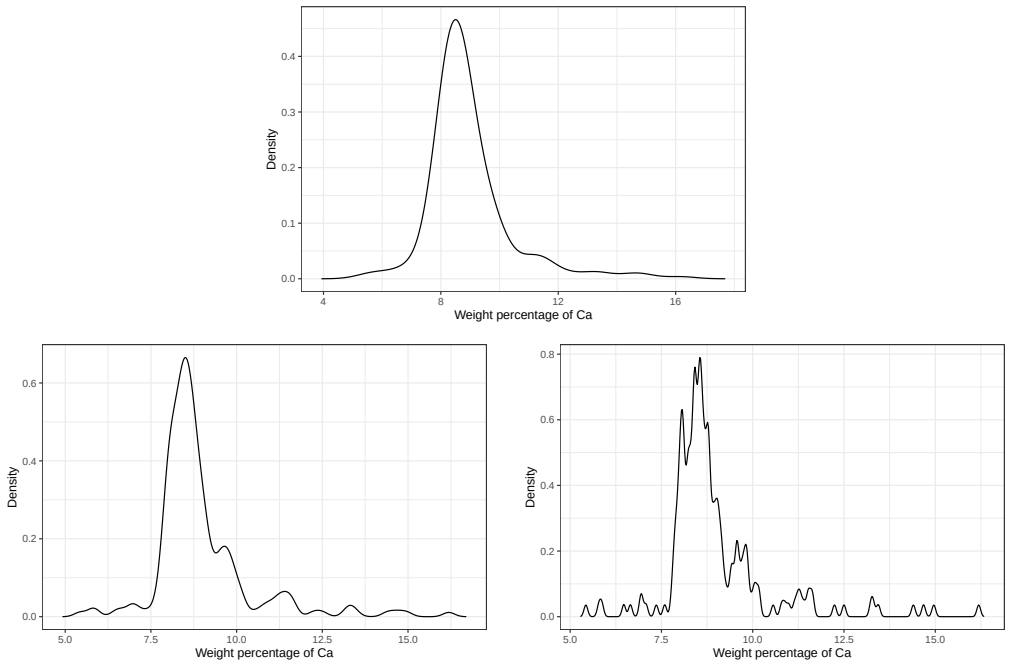

FIGURE 5. The top plot shows an over-smoothed kernel density estimator with bandwidth $h = 0.5$. The bottom left shows the kernel density estimator with the optimal bandwidth of $h = 0.17$ according to the cross-validation estimator. The bottom right shows an under-smoothed kernel density estimator with bandwidth $h = 0.05$.

using data obtained from the UCI Machine Learning Repository (Dua and Graff (2018) [5]).

In Figure 5 we plot three kernel density estimators: an oversmoothed, an undersmoothed, and the optimally smoothed according to the estimated MSE. Unlike our previous example, we do not have an expression for the true density function of the data. Thus, we need to estimate the MSE using the leave-one-out cross validation. Doing so produces the optimal bandwidth of approximately 0.17 seen in Figure 5.

3. Approximate Bayesian Computation

3.1. **Bayesian Inference.** The following discussion follows Section 1.3 of Gelman et al. [7] closely. Suppose we have a random variable $Y \sim F$ where F belongs to a family of distributions indexed by a d-dimensional parameter vector θ. Up until this point, we have assumed that θ is some fixed but unknown vector. Now, we will assume that θ is a random variable. Given

data, our goal is to make probability statements about θ that are conditional on the observed data. To do so, we begin by specifying a **prior distribution** of the random variable θ.

Definition 3.1. The **prior distribution** of a random vector θ is a family of distributions indexed by a vector of hyperparameters that summarizes the prior beliefs about the random vector θ.

For example, for a scalar random variable θ we may say that $\theta \sim \text{Beta}(\alpha, \beta)$. Here, $\text{Beta}(\alpha, \beta)$ is the prior distribution with α and β as hyperparameters. Such a prior distribution may be used if we believe that instances of θ are restricted to the interval $(0, 1)$. Specifying this prior produces a prior distribution function denoted by F_θ and prior density function denoted by f_θ. Note that these notations are not universal: they depend on the chosen prior distribution with the notation f being reserved for the probability density function of a continuous prior and p being reserved for the probability mass function of a discrete prior . The notation will be generalized in later sections.

To create our model, we must also specify a **sampling distribution**.

Definition 3.2. The **sampling distribution** of a Bayesian model is the probability distribution of the data conditional on θ.

Like with the prior distribution, the sampling distribution is specificied by the modeler. Continuing with our example from above, we specify the sampling distribution of the model to be $Y|\theta \sim \text{Bin}(n, \theta)$. Notice that the instance of θ specifies the success probability parameter of the Binomial distribution. This specification produces a cumulative distribution function $F_{Y|\theta}$ and probability mass function $p_{Y|\theta}$. The probability mass function (or in the case of a continuous random variable, the probability density function) is referred to as the **likelihood function** of the model.

Finally, given an observed value y of our random variable Y, we are interested in finding the **posterior distribution** of θ.

Definition 3.3. The **posterior distribution** of a Bayesian model is the distribution of θ conditional on observed data y.

The posterior density function is calculated using Bayes' Theorem. For our example, we write Bayes' Theorem as

$$f_{\theta|y} = \frac{p_{Y|\theta}(y)\ f_\theta(\theta)}{\int_0^1 p_{Y|\theta}(y) f_\theta(\theta) d\theta}.$$

Assuming $y = k$ to be the observed number of successes for an instance of our Binomial random variable Y, it can be shown that $\theta|y \sim \text{Beta}(\alpha+k, \beta+n-k)$ (see e.g. Section 8.3 of Blitzstein and Hwang [2]).

As we saw in the example above, some choices of prior and sampling distributions cause the posterior density function to belong to a recognizable family of distributions. Let us work through another such example.

3.2. **Normal Distribution with Known Variance.** Suppose we are interested in the density function of a random variable X. We believe that the distribution function of X belongs to the family of normal distributions with unknown parameter $\mu \in \mathbb{R}$ and known parameter $\sigma = 1$. To generalize our notation from above, we will use $\pi(\theta)$ to refer to the prior density function, $\pi(y|\theta)$ to refer to the likelihood function, and $\pi(\theta|y)$ to refer to the posterior density function. Using this notation, while less specific, avoids the need to use different notation when dealing with continuous versus discrete random variables.

Given θ, we have $\pi(y|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2}$. We set $\theta \sim N(0, 1.5)$. Our specification of $\theta \sim N(0, 1.5)$ gives us

$$\pi(\theta) = \frac{1}{\sqrt{3\pi}} e^{-\frac{1}{3}\theta^2}.$$

We see that the likelihood function $\pi(y|\theta)$ is a result of our modeling choices surrounding our beliefs about the random variable X. Our assumption that, given θ, the distribution function of X belongs to the family of normal distributions with parameters $\mu \in \mathbb{R}$ and $\sigma = 1$ leads directly to the form of $\pi(y|\theta)$. For a model where the distribution of the random variable Y comes from the family of Normal distributions with known variance σ indexed by the parameter θ and $\theta \sim \mathrm{N}(\mu_0, \tau_0^2)$, Section 2.5 of Gelman et al. [7] gives the posterior density as

$$\pi(\theta|y_1, \ldots, y_n) = \frac{1}{\tau_n \sqrt{2\pi}} \exp\left[-\frac{1}{2\tau_n^2}(\theta - \mu_n)^2\right]$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

Defining the **precision** as the inverse of the variance of a distribution aids in the interpretation of the formulas above. Because our sample is iid from a distribution with variance σ^2, we know that the variance of the sample mean is $\frac{\sigma^2}{n}$, implying that the precision of the sample mean of our data is $\frac{n}{\sigma^2}$. Thus, we see that the posterior precision is the sum of the prior precision and the precision of the sample mean of our data. As the number of observations increases, the posterior precision will increase. Additionally, we see that the denominator of the formula for μ_n is the posterior precision. The numerator is the sum of the prior mean and data mean weighted by their respective precisions.

In fact, μ_n is a weighted sum of the prior mean and the data average. Let us look at the weight on $\bar{y}$ as $n \to \infty$:

$$\lim_{n\to\infty} \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2}} = 1.$$

As the number of observations increases, more and more weight will be placed on the sample mean. As $n \to \infty$, the effect that $\bar{y}$ has on μ_n will drown out the effect of μ_0, leaving us with $\mu_n \to \bar{y}$.

3.3. **Summary and Sufficient Statistics.** In Section 3.2, we saw that the parameters of the posterior density relied only on the sample mean of the observed data, not on the full sample. The sample mean of the data is referred to as a **statistic** which we define using Section 5.2 of Casella and Berger [3].

Definition 3.4. Let $X_1, \ldots, X_n$ be a random sample and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable $Y = T(X_1, \ldots, X_n)$ is called a **statistic**.

A statistic can be thought of as a tool for reducing the dimension of the data. Some common examples, both from $\mathbb{R}^n \to \mathbb{R}$, are the sample mean $S(y) = \frac{1}{n}\sum_{i=1}^n y_i$ and the sample variance $S(y) = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$ where $\bar{y}$ is the sample mean.

Returning to our example in Section 3.2, the calculation of the posterior mean required only the one-dimensional sample mean instead of the full 10 dimensional sample. Reducing the data throws away information that the full data contained. This may be useful if the information being thrown away is useless for the particular question at hand. However, we could be throwing away valuable information by reducing our data. One way to differentiate between the type of information reported by different statistics is the concept of a **sufficient statistic**. We define a sufficient statistic in a Bayesian context using Section 2.5 of Gelman et al [7].

Definition 3.5. Given a model $p(y|\theta) \propto \pi(\theta)\pi(y|\theta)$, data y, and a statistic $T(Y)$, the statistic $T(Y)$ is a **sufficient statistic** if the posterior density $\pi(\theta|y)$ depends on y only through $T(y)$, that is if $\pi(\theta|y) = \pi(\theta|T(y))$.

Thus, $S(y) = \overline{y}$ where $S : \mathbb{R}^n \to \mathbb{R}$ is a sufficient statistic for the model in Section 3.2, since the posterior density depended on y only through $\bar{y}$.

3.4. **Approximate Bayesian Computation.** Suppose that we do not have an analytical form for the likelihood function $\pi(y|\theta)$. We would like to find a way to approximate the likelihood function in order to estimate the posterior distribution using Bayes' Theorem. If for a given θ we are able to simulate data from $\pi(y|\theta)$, then we can use the following definition from Fearnhead and Prangle [6] to approximate the posterior.

Definition 3.6. Given a data point y from the distribution F, a prior density $\pi(\theta)$ for a d-dimensional parameter vector θ, a likelihood function $\pi(y|\theta)$, a kernel function $K(x)$, and a bandwidth $h > 0$, we define an approximation to the likelihood as

$$p(\theta|y) = \int_{-\infty}^{\infty} \pi(y|\theta) K\left[(z-y)/h\right] dz.$$

We can now define the **ABC posterior** as

$$\pi_{ABC}(\theta|y) \propto \pi(\theta) p(\theta|y).$$

The expression $p(\theta|y)$ can be thought of as a weighted average. The term $K\left[(z-y)/h\right]$ uses kernel density estimation to measure how likely we are to observe z given y. Thus, we can think of the integral defining $p(\theta|y)$ as a weighted average of likelihood functions $\pi(z|\theta)$ where the weights correspond to how likely we are to observe z given y. If we set h to be large, then many values of z will produce relatively large values of the kernel density estimator $K\left[(z-y)/h\right]$. Thus, the integral defining $p(\theta|y)$ will place positive weight on many likelihood functions $\pi(z|\theta)$, creating an average that may be far away from the true likelihood function $\pi(y|\theta)$. However, if we choose h to be very small, then only values of z very close to y will produce relatively large values of $K\left[((z-y)/h\right]$. This will cause the weighted average to be closer to $\pi(y|\theta)$.

This reasoning is stated formally by Fearnhead and Prangle [6] when they claim that as $h \to 0$, $p(\theta|y) \to \pi(\theta|y)$. Because the ABC posterior is defined as being proportional to $\pi(\theta)p(\theta|y)$, if $p(\theta|y)$ converges to $\pi(y|\theta)$ as $h \to \theta$, then the ABC posterior will converge to a function proportional to $\pi(\theta)\pi(y|\theta)$, implying that it converges to the true posterior density function up to a proportionality constant as $h \to 0$.

The integral defining $p(\theta|y)$ contains the likelihood function, making it appear useless as an estimator of the likelihood. Luckily, by using Monte Carlo simulation to simulate y_{sim} from $\pi(y|\theta)$, we can approximate $p(\theta|y)$ (and therefore the ABC posterior) without needing to calculate $\pi(y|\theta)$ directly. In order to better understand the usefulness of the ABC posterior, let us explore an algorithm for approximating the ABC posterior adapted from Prangle and Fearnhead [6].

Algorithm 3.7.

Input - *a random sample* $y_1, \ldots, y_n$*; a function* $S(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$*; a density kernel* $K(\cdot) : \mathbb{R}^m \to [0,1]$ *with bandwidth* $h > 0$*; an integer* $N > 0$

Initialize - *define* $s_{obs} = S(y_{obs})$

Iterate - *for* $i = 1, \ldots, N$*:*

step 1: simulate θ^i *from* $\pi(\theta)$
step 2: simulate y_{sim} *from* $\pi(y|\theta^i)$ *and set* $s = S(y_{sim})$
step 3: with probability $K((s-s_{obs})/h)$ *accept* θ^i*; otherwise, reject* θ^i

Output - *a set of accepted values of* θ^i.

Step 2 and step 3 of Algorithm 3.7 approximates $p(\theta|y)$ and therefore allow us to approximate the posterior density without an analytic form for the likelihood function. They do so by checking how "close" the simulated data is to the observed value using a specified distance metric. A likely value of θ^i, i.e. a value of θ^i such that $\pi(y|\theta^i)$ is relatively large, is assumed to produce simulated data close to the observed data. Such an assumption can be reasonable, as the likelihood function is in part measuring the likelihood of observing y given θ. If it is likely to observe y given θ, then it should be likely to simulate y from the model given θ.

Notice how this process relates to the integral defining $p(\theta|y)$. In Algorithm 3.7 we use $K((s - s_{\text{obs}})/h)$ as the probability of accepting θ^i whereas in the definition of $p(\theta|y)$ we use the kernel density estimator to decide the weights within the average of likelihood functions. Both processes rely on measuring "closeness" using kernel density estimation to create an estimate of the true likelihood function, but $p(\theta|y)$ constructs a weighted average by directly computing the likelihood function while Algorithm 3.7 relies on Monte Carlo simulation to set the probability of acceptance in an attempt to approximate the likelihood function.

However, the estimated posterior distribution produced by Algorithm 3.7 is not guaranteed to converge to the true posterior distribution. Here, convergence refers to the empirical distribution function produced by the accepted values of θ^i in Algorithm 3.7 converging pointwise to the true posterior distribution function. According to Fearnhead and Prangle [6], no such convergence is guaranteed. A lack of convergence theorems is unfortunate and requires that any user of Algorithm 3.7 understand it deeply before applying it to a problem. However, as we will show through multiple examples, with careful application the algorithm can and does succeed in estimating posterior distributions.

According to Fearnhead and Prangle [6], the most common implementation of Algorithm 3.7 uses a uniform distribution function for $K(\cdot)$. That is, we define

$$K(x) = \begin{cases} \frac{1}{2\epsilon} & \text{if } x \in [-\epsilon, \epsilon] \\ 0 & \text{otherwise} \end{cases}.$$

To use such a kernel, a distance metric in $\mathbb{R}^m$ must be chosen. The most common choices are the distance metrics corresponding to the ℓ_1 norm and ℓ_2 norm. Recall that for m dimensional vectors s and s_{obs}, the distance metric

implied by the ℓ_1 norm is given by

$$\|s - s_{\text{obs}}\|_1 = \sum_{j=1}^{m} |s_j - s_{\text{obs},j}|.$$

The distance metric implied by the ℓ_2 norm is given by

$$\|s - s_{\text{obs}}\|_2 = \sum_{j=1}^{m} (s_i - s_{\text{obs},i})^2.$$

Once a distance metrice has been chosen, using a uniform kernel results in a deterministic accept/reject scheme where all values of θ^i that produce simulated data y_{sim} such that $K(\|s - s_{\text{obs}}\|) > 0$ are accepted. Despite the fact that $\|s - s_{\text{obs}}\| \geq 0$, we define $K(\cdot)$ to be positive on the interval $[-\epsilon, \epsilon]$ to ensure it has mean zero in order to satisfy Definition 2.9. We will now provide an example of such an implementation.

3.5. **ABC Approximation for Normal Prior and Likelihood.** Given a random variable $X \sim \text{N}(\mu, 1)$, we would like to learn about the parameter μ. Using the standard notation $\theta = \mu$, we begin by assuming $\theta \sim \text{N}(0, 1.5)$, giving us the prior density function

$$\pi(\theta) = \frac{1}{1.5\sqrt{2\pi}} e^{-\frac{1}{4.5}\theta^2}.$$

The likelihood function is given by

$$\pi(y|\theta) = \Pi_{i=1}^{n} \pi(y_i|\theta).$$

Notice that this is the same model as the example in Section 3.1. Our observed data $y = (y_1, \ldots, y_{10})$ is simulated from a $\text{N}(1, 1)$ distribution. We will use Algorithm 3.7 to estimate the posterior density. We first must choose $K(\cdot)$. We define

$$K(x) = \begin{cases} \frac{1}{4} & \text{if } x \in [-2, 2] \\ 0 & \text{otherwise} \end{cases}.$$

We set $N = 1\,000\,000$. For now, we set $S(\cdot)$ to be the identity function. Let us walk through one iteration of the algorithm to better understand what is going on. First, we simulate θ^i from $\pi(\theta)$. Next, we simulate $y_{\text{sim}} = \{x_{\text{sim},1}, x_{\text{sim},2}, \ldots, x_{\text{sim},10}\}$, a vector of 10 random draws from the likelihood function $\pi(y|\theta^i)$. Then, we calculate

$$K\left(\|y_{\text{sim}} - y\|_2\right) = \begin{cases} \frac{1}{4} & \text{if } \|y_{\text{sim}} - y\|_2 \leq 2 \\ 0 & \text{else} \end{cases}.$$

If $K\left(\|y_{\text{sim}} - y\|_2\right) = \frac{1}{4}$, we accept and store θ^i, otherwise we reject θ^i. In effect, our choice of $K(\cdot)$ means that we will accept all θ^i such that the simulated data drawn from $\pi(y|\theta^i)$ is within a distance of 2 from our observed

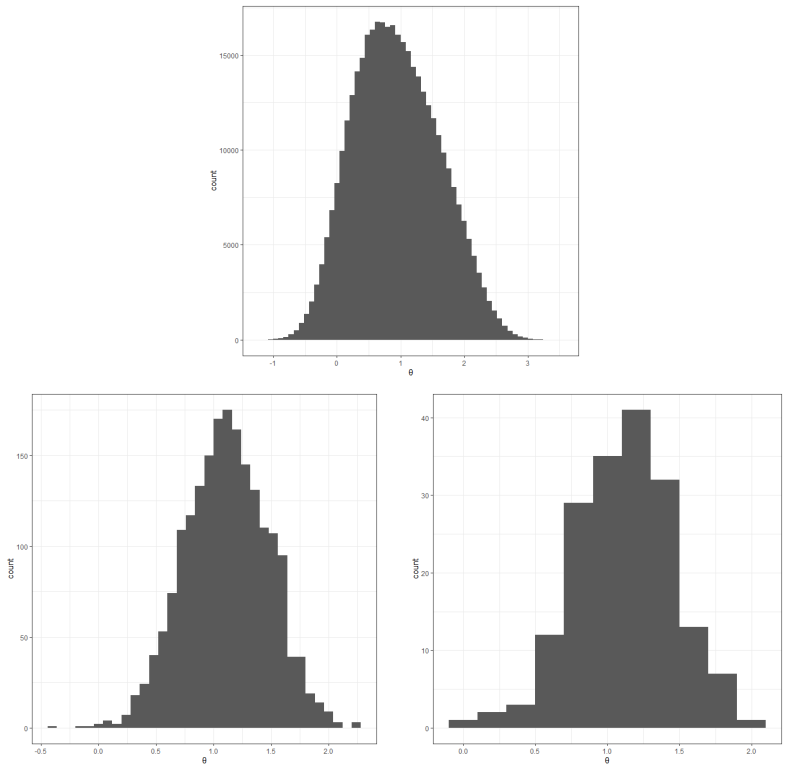

FIGURE 6. The top plot shows a wide approximation of posterior density with $\epsilon = 5$ not being restrictive enough. The bottom left shows a good approximation of the posterior density with $\epsilon = 2$. The bottom right shows a rough approximation with $\epsilon = 1.2$ being too small.

data. Under this interpretation, the maximum distance that y_{sim} can be from the observed data y for θ^i to be accepted is referred to as ϵ. In this case, $\epsilon = 2$.

We will focus on how the choice of ϵ affects the approximation of posterior density. In Figure 6, we show histograms of accepted values of θ for three choices of ϵ: 5, 2, and 1.5. When $\epsilon = 5$, the histogram shows the distribution of the $380,707$ θ^i's corresponding to simulated data within a distance of 5 from the observed data. The graph appears to be approximately normal. The sample standard deviation of the accepted values of θ is 0.671. When $\epsilon = 2$, the the graph also appears approximately normal but it is much narrower than when $\epsilon = 5$ with a standard deviation of 0.362. Only 1959 θ^i's are accepted here. When $\epsilon = 1.2$, the approximation becomes quite noisy with only 24 θ^i's being accepted. The graph still looks somewhat normal but is

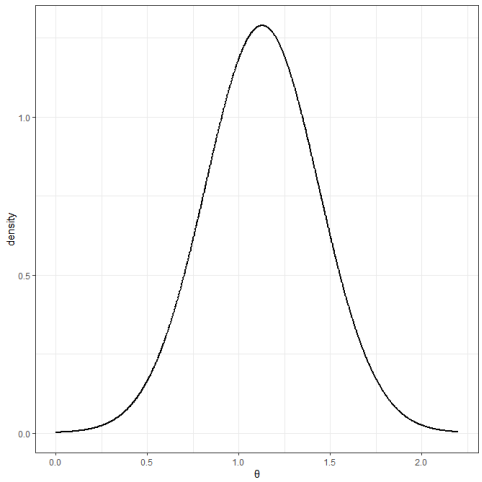

FIGURE 7. A plot of the true posterior density function of θ.

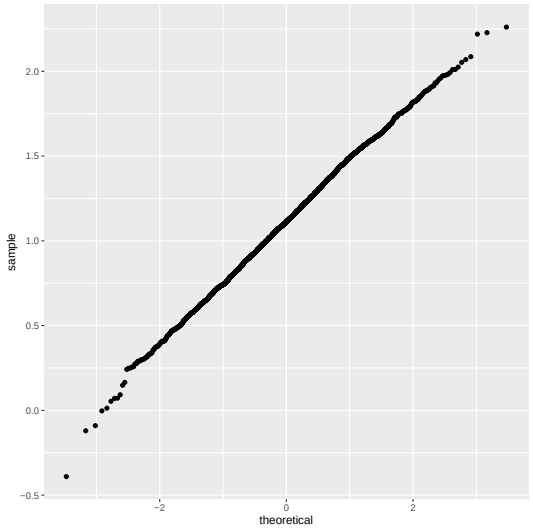

FIGURE 8. A normal quantile plot of the accepted values of θ when $\epsilon = 2$.

much rougher than the previous two and the sample standard deviation has increased to 0.432.

We can compare our approximations of the posterior density when $\epsilon = 2$ to the true posterior density given in section 2.5 of Gelman et al. [7]. Using the formula from Section 3.2, we find that the true posterior distribution is $\theta|y \sim \mathrm{N}(1.128, 0.095)$.

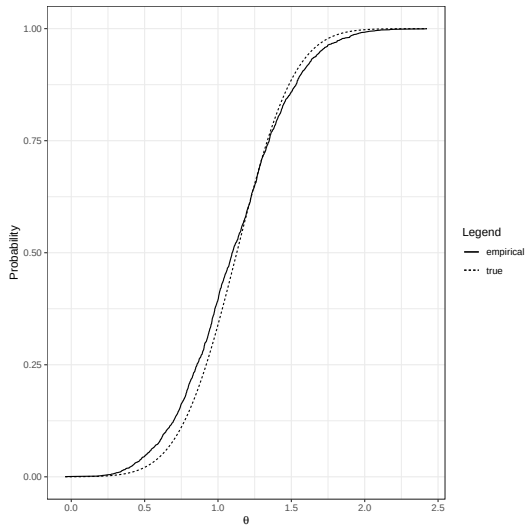

FIGURE 9. Empirical distribution of accepted parameter values of θ plotted against the true posterior distribution function of θ.

Comparing Figure 7 to Figure 6 for the case when $\epsilon = 2$, we see the shape of the histogram resembles the shape of the posterior density quite closely. The quantile plot given in Figure 8 that the distribution of accepted θ's when $\epsilon = 2$ is approximately normal. Moreover, the mean of the accepted values of θ is 1.113 and the standard deviation is 0.361. These are quite close to the true values of 1.128 for the mean and 0.309 for the standard deviation of the true posterior density.

Using the empirical distribution, we can compare our approximation of the posterior distribution to the true distribution. The empirical distribution is useful here because there is no need to choose a parameter such as bin width or bandwidth to create the estimate of the posterior distribution. Additionally, unlike when comparing the histogram of accepted values to the true density function, we can directly compare the output of the empirical distribution to the output of the true distribution function, allowing us to plot the two functions on top of each other. We do so in Figure 9, which reveals that our empirical distribution resembles the true distribution closely.

Remember from Section 3.2 that the posterior distribution of our model depends on y only through $\bar{y}$. We can apply this fact to our ABC approximation algorithm in Section 3.2 by setting $S(y) = \bar{y}$. Doing so reduces our distance calculation to 1 dimension instead of 10 dimensions, the benefits of which will be seen in our discussion of the Curse of Dimensionality.

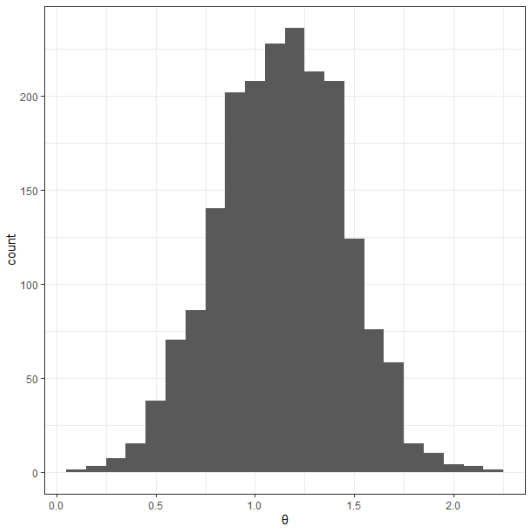

FIGURE 10. A histogram of the accepted values of θ when the algorithm from Section 3.2 is used with the sample mean instead of the full data. Here, $\epsilon = 0.005$.

In Figure 10, we see that the histogram of accepted values of θ when $\epsilon = 0.005$ closely resembles Figure 6 when $\epsilon = 2$. Thus, even though we have reduced the dimensionality of step 3 from ten to one, we receive similar results. Notice that our choice of ϵ has decreased drastically from 2 to 0.005.

3.6. **Curse of Dimensionality.** We now provide the motivation for using summary statistics within our ABC methods instead of the full data. The general idea behind the curse of dimensionality, a term first used by Richard Bellman [1], is that as the dimension of data increases, the volume of the space increases exponentially, causing massive sample sizes to be required in order to avoid sparseness. The following example is inspired by Section 2.5 of Hastie, Tibshirani, and Friedman [8, § 2.5].

Consider the p-dimensional hypercube with each axis ranging from $[0, 1]$. That is, each edge l of our cube has length 1. We will notate this cube by $[0, 1]^p$. Suppose we want to inscribe a smaller hypercube that contains a fraction r of the volume of our hypercube $[0, 1]^p$. The volume of a p dimensional hypercube with edge length l is l^p. Thus, in order to capture r of the volume, the edge length l of our inner cube must be $r^{1/p}$.

Notice that the edge length increases exponentially as the dimension increases. For $p = 3$, we have that $l = 0.46$, meaning that the cube must cover roughly half the range of each axis in order to capture only $\frac{1}{10}$ of the volume of the primary cube. Moving to $p = 10$, we have that $l = 0.80$, meaning that we must cover 80% of the range of each axis in order to capture only 10% of

the volume. The ranges required to capture relatively small percentages of the volume reveal how sparse the space becomes in high dimensions.

Remembering that the ℓ_1 norm can be used to define a cube of points within a certain distance of a central point, the previous example reveals how difficult it can become to define local areas within high dimensional spaces. Returning to our example in 10 dimensions, if one wanted to create an average of the data contained within 10% of the volume of the hypercube, one would need to define the average over 80% of each axis, which can hardly be defined as a "local" average.

While the ℓ_1 norm struggles in high dimensions, it could be that the commonly used ℓ_2 norm provides better results. However, the ℓ_2 norm actually performs much worse in higher dimensions than the ℓ_1 norm as can be seen in the following example. Suppose we have a cube with side length l We can inscribe a sphere of radius $\frac{1}{2}l$ within our cube. In a space of dimension n, the cube has volume l^n and the sphere has volume $\frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}(\frac{1}{2}l)^n$. Thus, the fraction of the volume of the cube contained within the sphere is

$$\frac{\pi^{n/2}\frac{1}{2^n}l^n}{\Gamma(\frac{n}{2}+1)l^n} = \frac{\pi^{n/2}}{2^n\Gamma(\frac{n}{2}+1)}.$$

As $n \to \infty$, the product $2^n\Gamma(\frac{n}{2}+1)$ outgrows $\pi^{n/2}$. Thus, as $n \to \infty$, the volume of the cube contained within the sphere goes to zero. Therefore, the sphere captures even less data than our cube as n increases. For this reason, we avoid using the ℓ_2 norm in high dimensional spaces.

Our final example of the Curse of Dimensionality uses the multivariate standard normal distribution. Suppose we have independent random variables $Y_1, \ldots, Y_p$ where each $Y_i \sim \mathrm{N}(0,1)$. Let us write $Y = (Y_1, \ldots, Y_p)$, meaning that $Y \sim \mathrm{N}(0, I)$ where 0 is the 0 vector of length p and I is the $p \times p$ identity matrix. The mode of the density of Y is located at the origin. We see that the sum of the squared random variables is distributed chi-squared with p degrees of freedom. Formally, we have $Y_1^2 + \ldots + Y_p^2 \sim \text{Chi-square}(p)$. Notice that the sum of the squared random variables is the squared ℓ_2 norm. Thus, the expected distance of Y from the mode of its density as measured by the ℓ_2 norm is the square root of the dimension of Y, i.e. $\sqrt{p}$.

We can see the expected Euclidean distance from the mode as a function of the dimensions in Figure 11.

The curse of dimensionality, and sparse data in general, is particularly salient for ABC methods because they rely directly on the distances between data points when estimating the posterior density function. A high dimensional parameter space or high dimensional data could require impossibly large simulations in order to produce points close enough together to build a reasonable approximation. Thus, we have the motivation for summary statistics as a dimension reduction tool.

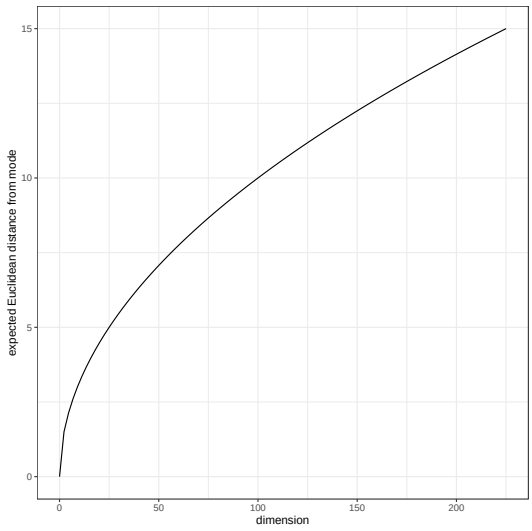

FIGURE 11. Expected Euclidean distance of a draw from a standard multivariate normal distribution as a function of dimension.

4. Application

4.1. **Survival Analysis.** Our example involves modeling the survival time of beetles after being given a dose of insecticide. Before introducing the model, we need to cover some basic concepts from survival analysis, the field of statistics that focuses on modeling the duration of time until a certain event occurs. In our case, the even of interest is the death of a beetle. We use the non-negative continuous random variable T to represent the time until death occurs. We denote the cumulative distribution function of T by $F(t)$ and assume that $f(t) = F'(t)$ exists. Here, $F(t)$ provides the probability that death has occurred by time t. Thus, the *survival function*

$$S(t) = 1 - F(t) = \int_t^\infty f(t)dt$$

gives the probability of survival up to time t.

Given the distribution function F, we can calculate the probability of death in the interval $(t, t+h)$ as $F(t+h)-F(t) = S(t)-S(t+h)$. This is the probability of survival up to time $t+h$ conditional on survival to time t. Taking the limit as $h \to 0$ produces the *hazard function*, given by $h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)}$. Therefore, we see that the hazard function provides the likelihood of the instantaneous occurrence of death at time t conditional on survival up until time

t. Unlike the density function $f(t)$ which provides the unconditional likelihood of death at time t, the hazard function conditions on survival up until time t, which is why the survival function $S(t)$ appears in the denominator.

4.2. **Model.** We will be modeling an experiment described in Hewlett [9] where beetles received a dose of insecticide and the time until death was recorded. Diggle [4] provides a frequentist model for the experiment. We will first provide Diggle's model before transitioning it to a Bayesian model. We will then estimate the Bayesian model using ABC.

A beetle receives an initial does of insecticide z_0. A series of independent releases V_i at random times T_i remove the insecticide from the beetle's system. We use $N(t)$ to denote the number of releases T_i that occur at or before time t. The amount of insecticide in a beetle's system at time t is given by

$$Z(t) = \max\left\{0, z_0 - \sum_{i=1}^{N(t)} V_i\right\}.$$

We specify $N(t)$ to be distributed Poisson with mean αt, the V_i to be iid exponential with parameter β, and the hazard function to be given by $h(z,t) = (\gamma + pt)z$. Notice that the hazard function depends on time only through z, the amount of insecticide inside of the beetle's system.

4.3. **Estimation.** We want to estimate the parameters α, β, γ, and p. However, the model does not produce an analytical likelihood function. Thus, we will estimate the parameters using Algorithm 3.7. We begin by placing priors on our parameters.

The parameter β implies that the expected release of insecticide at each T_i is $\frac{1}{\beta}$. Thus, we should ensure that $\frac{1}{\beta} \leq z_0$, otherwise our model would imply that the initial release of toxin has positive probability of being greater than the initial dose of toxin. We do not have much information about the minimum expected release for each T_i. Diggle's estimates for $\frac{1}{\beta}$ are all above 0.1. We do not want to rule out the possibility that $\frac{1}{\beta}$ can take on small values but we also believe that instances of $\frac{1}{\beta}$ are most likely to fall in the interval $(0.1, 0.2)$. We set the prior density of β to be

$$f(x) = \begin{cases} 4e^{-4(x-5)} & x \geq 5 \\ 0 & x < 5 \end{cases}.$$

Notice that this is an exponential distribution with parameter 4 that has been shifted to the right. Instead of being positive on the interval $(0, \infty)$, it is positive on the interval $(5, \infty)$. Thus, we allow for the possibility of instances of $\frac{1}{\beta}$ being small while placing the upper bound at 0.2.

The estimates from Diggle [4] for the parameter α fall between 0.2 and 0.8. We know that α can only take positive values since it determines the

Time interval (days)	*Observed frequencies*	
0–1	3	
1–2	11	
2–3	10	
3–4	7	
4–5	4	
5–6	3	5
6–7	2	
7–8	1	
8–9	0	
9–10	0	3
10–11	0	
11–12	1	
12–13	1	
No response	101	
Total	144	

FIGURE 12. Observed death frequencies of beetles with initial does of $0.2mg/cm^2$ of insecticide from Diggle [4].

parameter of a Poisson distribution, which requires positive parameters. We want to allow for the possibility that α is greater than one, so we choose the prior distribution Uniform$(0, 2)$.

Both γ and ρ have little intuitive meaning besides their occurrence in the hazard function. The hazard function must be positive, implying that both parameters should be set as positive random variables. All of Diggle's estimates are less than 1, so we will set both prior distributions as Uniform$(0, 1)$.

We can now run Algorithm 3.7. Our observed values from Diggle [4] are presented in Figure 12. When calculating distances between observed and simulated values, we group the data the same as Figure 12. We run 10^6 iterations of our algorithm.

To construct the estimate of the posterior density, we choose an ϵ such that we accept all parameter draws associated with simulated data that was less than ϵ distance away from the observed data. We choose $\epsilon = 12$ in order to keep approximately 0.1% of the parameter draws.

Table 1 provides summary statistics of the accepted values for $\epsilon = 12$. The histograms of accepted values are presented in Figure 13. Increasing the number of iterations does not significantly impact the shape of the histograms, revealing that 10^5 iterations is enough to stabilize the variation in the posterior estimates due to the simulation.

The histogram of α appears approximately normal with most of the mass laying in the interval $(0.6, 0.9)$. The histograms of γ and p look similar. They have means of around 2 and are slightly right skewed. In comparison to their shared prior distribution, both posterior distributions have shifted slightly to the right but still resemble a Gamma$(1.5, 1)$. Table 1 shows how the summary statistics have shifted from prior to posterior. The histogram

Parameter	Mean	Median	Mode	Standard Deviation
α	0.671 (0.5)	0.656 (0.5)	0.623 (0.5)	0.117 (0.288)
β	5.294 (5.25)	5.214 (5.173)	5.690 (5)	0.281 (0.25)
γ	1.885 (1.5)	1.690 (1.183)	2.216 (0.5)	1.149(1.225)
p	2.025 (1.5)	1.782 (1.183)	1.453 (0.5)	1.283 (1.225)

TABLE 1. Summary statistics for values of accepted parameter values with corresponding values for the prior distributions in parentheses.

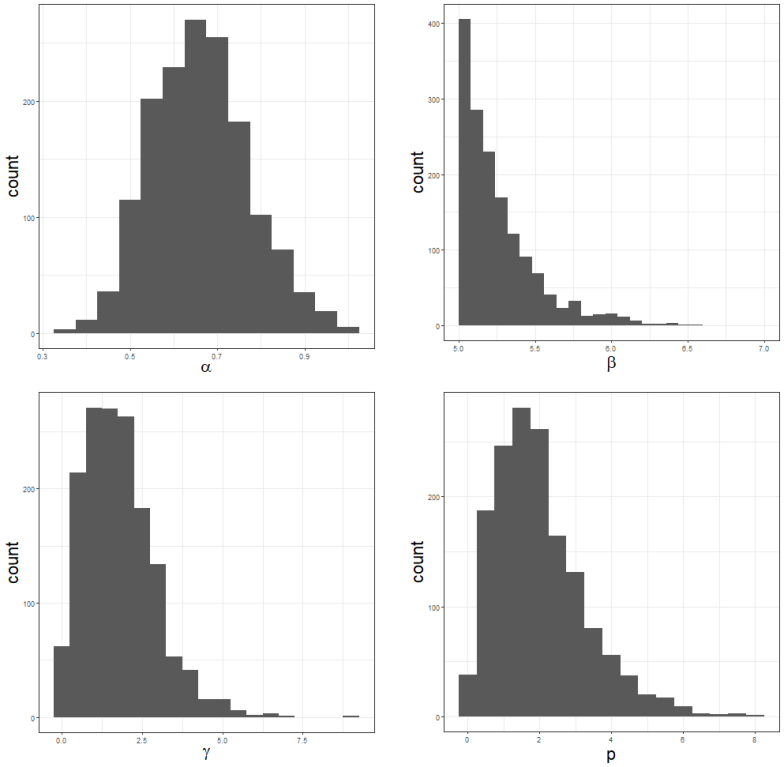

FIGURE 13. Histograms for the accepted parameter draws with $\epsilon = 12$.

for β has the same shape as the prior. The prior has a mean of 5.25 and standard deviation of 0.25 while our accepted values of β have a sample mean of 5.253 and sample standard deviation of 0.2501. Thus, while the shapes of the distributions of the accepted values of the other parameters differ greatly

from the prior distributions, the distribution of accepted values of β appear almost identical to the prior.

To test how well our algorithm has estimated the model, let us use the mean of the estimated posterior distributions to simulate the experiment. We can then compare the results of the simulated experiment to the results of the observed experiment. We use parameter values $\alpha = 0.7$, $\beta = 5.25$, $\gamma = 1.9$, and $p = 2$. We simulate the experiment $10,000$ times and take the average. The computed frequencies corresponding to Table 12 are

$$[3.07, 10.56, 7.45, 4.24, 2.26, 1.20, 0.62, 0.31, 0.15, 0.08, 0.04, 0.39, 102.62].$$

Our results are nearly identical to Table 12, showing that our model reproduces the results of the experiment well. The results do not change significantly if we use the median as opposed to the mean of the estimated posterior distributions.

We can also simulate the experiment using the point estimates from Diggle (1984) [4] to compare the results of the frequentist and Bayesian approaches. As we did using the results from our Bayesian method, we simulate the experiment $10,000$ times and take the average to arrive at the following frequencies:

$$[0.17, 1.23, 2.82, 3.92, 4.24, 3.88, 3.11, 2.38, 1.72, 1.18, 0.79, 0.53, 2.50, 115.53].$$

We see that the frequencies do not correspond to the frequencies from the experiment given in Table 12. Thus, our ABC method produces estimates that more accurately reflect the results of the experiment than the frequentist method used in Diggle (1984) [4].

The results of our application reveal the power of ABC. Without using a likelihood function, we have fit a model that produces data nearly identical to the data from the observed experiment. While further analysis is necessary to show that the model is viable, for this example our only goal is to test whether ABC can produce posterior estimates whose corresponding data generating process reflects the observed data. To this end, the similarity between the frequencies produced using the means of our posterior estimates and the observed data reveals the success of our algorithm.

5. Conclusion

The best model may be one without a likelihood function. Combining nonparametric statistical methods and Monte Carlo simulation, Approximate Bayesian Computation allows for the posterior density to be estimated without use of a likelihood function. As with most statistical algorithms, ABC struggles in higher dimensions. While we have shown that the effective use of summary statistics can assist in overcoming the curse of dimensionality, no choice of summary statistics is guaranteed to produce convergence. While the lack of convergence theorems is undesirable, through examples we have

shown that thoughtful application of ABC can produce accurate estimates of posterior densities.

Acknowledgments

I would like to thank Lynne Butler for all of her help and support throughout this entire project. This thesis would not have been finished without her help. I also owe her for introducing me to probability and statistics and for her continual guidance as I pursue both topics further.

I would like to thank all of the wonderful professors who have helped me grow during my time at Haverford. I chose to be a math major largely due to Charlie Cunningham's excellent introduction to linear algebra during my freshman year. Josh Sabloff not only taught me analysis but also taught me how to see my own potential. Weiwen Miao taught me many of the concepts which appear in this thesis.

I owe practically everything to my parents. Their constant support and love were vital to the completion of this thesis. I cannot thank them enough.

If I tried to thank all of my peers and friends that I've met at Haverford individually, this acknowledgment section would end up longer than the rest of the document. Before limiting myself to those I've met within the math department, I'd like to say that almost all of my favorite memories from Haveford involve Katie Leiferman; as one of my first and closest friends at the school, she has shaped my experience here into four of the best years of my life. Some of my best friends at this school have also been my collaborators in my math classes here: Jake Van Wiggeren, Matt Katz, Eric Beery, and Nathan Akerhielm are a few of the many. I'd like to thank them for all of the good times both in and out (mostly out) of the classroom.

References

[1] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton Legacy Library. Princeton University Press, 1961.

[2] J.K. Blitzstein and J. Hwang. *Introduction to Probability*. CRC Press, 2014.

[3] George Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series. Brooks/Cole Publishing Company, 1990.

[4] Peter J. Diggle and Richard J. Gratton. "Monte Carlo Methods of Inference for Implicit Statistical Models". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 46.2 (1984), pp. 193–227.

[5] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017.

[6] Paul Fearnhead and Dennis Prangle. "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 ().

[7] A. Gelman et al. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, 2013.

[9] P.S. Hewlett. "Time from dosage to death in beetles, Tribolium castaneum, treated with pyrethrins or DDT, and its bearing on dose-mortality relations". In: *Journal of Stored Products Research* 10.1 (1974), pp. 27–41.

[10] Thomas A. Severini. *Elements of Distribution Theory*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2005.

[11] L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York, 2006.

Appendix A. Code

```
########################################
# function to simulate one iteration of the experiment. For a
# given number of beetles, function determines if and when each
# beetle dies.
########################################

# total_beetle: positive integer, refers to total number of beetles
# the in experiment
# alpha_i, beta_i, gamma_i, p_i: positive reals, parameter values
# z_0: positive real, initial dose of insecticide

# returns length 14 vector where first 13 values correspond
# to the number of beetles that died on that day and final
# value corresponds to the number of beetles that survived.
function beetle_sim(total_beetle, alpha_i, beta_i, gamma_i, p_i, z_0)

    results = zeros(14)

    S_prev = 1

    for i = 1:total_beetle
        # number of releases occuring up to or on 13th day
        V_count = rand(Poisson(alpha_i * 13))

        # if no releases, count beetle as survived
        if V_count == 0
            results[14] = results[14] + 1
            continue
        end

        # draw inter-arrival times for releases
        T_inter = rand(Exponential(1/alpha_i), V_count)

        # set amount of insecticide
        z_i = z_0

        # death status of beetle
        death = false
```

```
# time in loop begins at 0
current_t = 0

# calculate probability of death for each inter-release
# period. Then remove amount of toxin released at end of
# period and move on to next period
for j = 1:(V_count + 1)

    # last iteration, for time between last release and
    # end of 13th day
    if j == V_count + 1

        # prob of death in inter-release interval
        death_p = z_i * gamma_i + p_i * z_i *
                (2*current_t*(13 - T_inter[j-1]) +
                (13 - T_inter[j-1])^2) / 2

        if death_p > rand(Uniform(0,1))
            # if the beetle dies, set death true and
            # break out of for loop
            death = true
            # add to death counter of final interval
            results[13] = results[13] + 1
        end

        break
    end

    # update time for next loop
    current_t = sum(T_inter[1:j])

    # make sure we haven't gone above 13 days
    if current_t > 13
        break
    end

    S_cur = exp(-(gamma_i * current_t * z_i +
                        z_i * p_i * current_t^2 / 2))
```

```
            death_p = S_prev - S_cur

            # set S_prev for next iteration
            S_prev = S_cur

            # use Uniform to generate random deaths
            if death_p > rand(Uniform(0,1))
                # if the beetle dies, set death true and
                # break out of for loop
                death = true
                # add to death counter
                index = convert(Int,round(current_t, RoundUp))

                results[index] = results[index] + 1
                break
            end

            # remove insecticide released at end of this period
            z_i = z_i - rand(Exponential(1 / beta_i))

            # stop loop if negative insecticide
            if z_i <= 0
                break
            end

        end

        if death == false
            results[14] = results[14] + 1
        end


    end

    return results

end
```

```
# ABC simulation function for the beetle experiment. Takes as inputs
# original data from experiment, number of iterations,
# prior distributions, and initial
# dose of insecticide. For each iteration it records the parameter
# values drawn from the priors and the distance of the simulated
# data from the observed data. Returns a dataframe with those values
# for each iteration.


function abc_sim(data, z_0, n, beetle_n, alpha_prior,
                                 beta_prior, gamma_prior, p_prior)
    results = DataFrame( alpha = Float64[], beta = Float64[],
                      gamma = Float64[], p = Float64[],
                      dist = Float64[])

    # these lines put data in same format given in Diggle 1984
    sum_1 = sum(data[6:7])
    sum_2 = sum(data[8:13])
    data = vcat(data[1:5], sum_1, sum_2, data[14])
    for i = 1:n

        # draw instances of parameters from prior distributions
        alpha_i = rand(alpha_prior)
        beta_i = rand(beta_prior) + 5
        gamma_i = rand(gamma_prior)
        p_i = rand(p_prior)

        # run the experiment simulation
        sim_i = beetle_sim(beetle_n, alpha_i, beta_i, gamma_i, p_i,z_0)

        # convert data into form from Diggle 1984
        sum_i1 = sum(sim_i[6:7])
        sum_i2 = sum(sim_i[8:13])
        sim_i  = vcat(sim_i[1:5], sum_i1, sum_i2, sim_i[14])

        # calculate distance b/t simulated and observed data
        dist = norm(sim_i - data, 1)

        # record results
        push!(results, [alpha_i, beta_i, gamma_i, p_i, dist])
```

```
    end
    return results
end

# script to run the ABC estimation algorithm
using Distributions, Random
using LinearAlgebra
using DataFrames
using Plots

beta_p = Exponential(0.25)
alpha_p = Uniform(0,1)
gamma_p = Gamma(1.5,1)
p_p = Gamma(1.5,1)

data = [3, 11, 10, 7, 4, 3, 2, 1, 0,0, 0, 1, 1, 101]

abc_sim(data, 0.2, 10^5, 144, alpha_p, beta_p, gamma_p, p_p)
```