Modelling a Zombie Apocalypse with ABC

Dan Milner, Harry Tata, Hannah Sansford

May 30, 2022

1 Introduction

(General background info)Chaotic dynamic systems commensurate with ecology and epidemiology challenge conventional methods of statistical inference as oftentimes they have an intractable likelihood. It is unlikely that any natural/environmental process under investigation can be captured without error and all natural/environmental systems invariably suffer process stochasticity. Consequently, model complexity increases in a non-trivial way as each realisation of the model is essentially unique. For chaotic models this is especially true as process stochasticity induces divergence of paths generated using identical parameters and starting from the sme initial conditions (Fasiolo *et al*, 2016).

(Specific background Info) Zombies, particularly when instigating an apocalypse, exhibit just such chaotic behavior. Their aim is to kill, eat and infect people via a bit/bites, which leave an open wound with the zombie's saliva in and around it. This bodily fluid mixes with the blood, infecting the (previously susceptible) bitten individual and turns them into a zombie. Zombie apocalypses have been modeled on a number of previous occasions (Munz *et al*,) but given the seriousness of such a situation it is important that no stone is left un-turned.

(A description of the gap in our knowledge that this study is designed to fill) Statistical methods capable of dealing effectively with highly non-linear systems are not a trivial matter. The simulation-based methods, Sequential Monte Carlo (SMC), Approximate Bayesian Computation (ABC) and Synthetic Likelihood (SL), offer a solution through estimating a posterior via simulated data sets based on sample parameters taken from the prior distribution. Utilising the computational efficiency of SMC, this report aims to compute a 4-Class zombie apocalypse model by approximating the posterior using a progressively decreasing sequence of tolerances.

2 Model

We consider four basic classes:

- Susceptible (S)
- Infected (I)
- Zombie (Z)

- Removed (R)

Susceptibles class can become deceased through 'natural' causes (parameter δ). Removed class are those that have died either through a zombie attack or from natural causes. If a human has died in a zombie attack they can resurrect and become a zombie (parameter ζ). A Susceptible can become a zombie through transmission, i.e. an encounter with a zombie (parameter β). Zombies can, therefore, only come from two sources; either they are resurrected from the newly deceased or they are a susceptible that has become infected. Prior to becoming a zombie, the period of time for which a susceptibles is infected lasts approximately 24 hours. Infected individuals can still die of natural causes before becoming a zombie, otherwise they become a zombie.

The birth rate of the human population is assumed to constant, Π. Zombies can be defeated by removing the head or destroying the brain (parameter α).

The 4-Class model is thus:

$$
\begin{aligned}
S' &= \Pi - \beta SZ - \delta S \\
I' &= \beta SZ - \rho I - \delta I \\
Z' &= \beta SZ + \zeta R - \alpha SZ \\
R' &= \delta S + \alpha SZ - \zeta R
\end{aligned}
$$

3 Approximate Bayesian Computation

Likelihood-free inference methods are convenient for complex models, such as Epidemic models, where we do not have an explicit expression for the likelihood. In approximate Bayesian computation (ABC), simulation under the implicit model replaces computation of the likelihood. One can use simulations from the model for different parameter values to compare the simulated datasets with the observed data. These simulations are increasingly being used as training datasets for machine learning methods including deep neural networks (Jiang et al., 2017) and random forests (Raynal et al., 2019).

The use of ABC first became popular in the field of population genetics, where simulation from a range of models is possible, but the likelihood is intractable for realistic sized datasets. Pritchard et al. (1999) were the first to use ABC, conducting inference on human demographic history. The method has now been applied in various subject areas including systems biology (Liepe et al., 2014), climate modelling (Holden et al., 2018), astronomy (Hahn et al., 2017) and epidemiology (Minter and Retkute, 2019).

The goal of ABC is to find a posterior distribution for the parameters of the implicit model, explaining the complex and potentially high-dimensional dataset. The method is based on Bayesian statistics: updating our prior beliefs about the model parameters, where $\pi(\theta)$ is the prior, using the information from our simulations. Suppose the dataset consists of n observations $y_{obs} = (y_{obs,1}, ..., y_{obs,n})$. A typical ABC procedure would involve mapping the observations to a lower dimensional set of summary statistics, $s(y)$. The posterior is then proportional to the following elements

$$p_\epsilon(\theta, s|s_{obs}) \propto \pi(\theta) f_n(s|\theta) K_\epsilon(\|s - s_{obs}\|), \tag{1}$$

where $f_n(s|\theta)$ is the implicit density of the model, $K_\epsilon(x)$ is a kernel function with tolerance ϵ, and $\|\cdot\|$ is a distance metric (Beaumont, 2019). The kernel function enables us to include in our posterior density the parameter values that best approximate the observations. This idea of estimating the likelihood of parameters using simulations that are 'close' to the observed data dates back at least as far as Diggle and Gratton (1984).

As the tolerance value ϵ tends to zero, the simulations we accept are closer to the true data and the approximate posterior becomes closer to its true distribution. Unfortunately, in order to simulate the same number of points at a smaller tolerance level often requires many more simulations, and the simplest algorithms can quickly become computationally inefficient. Luckily, we can harness techniques such as Markov chain Monte Carlo (Marjoram et al., 2003; Wegmann et al., 2009) and sequential Monte Carlo (Beaumont et al., 2009; Drovandi and Pettitt, 2011; Del Moral et al., 2012; Lenormand et al., 2013) to improve the computational efficiency of ABC.

3.1 Rejection algorithm

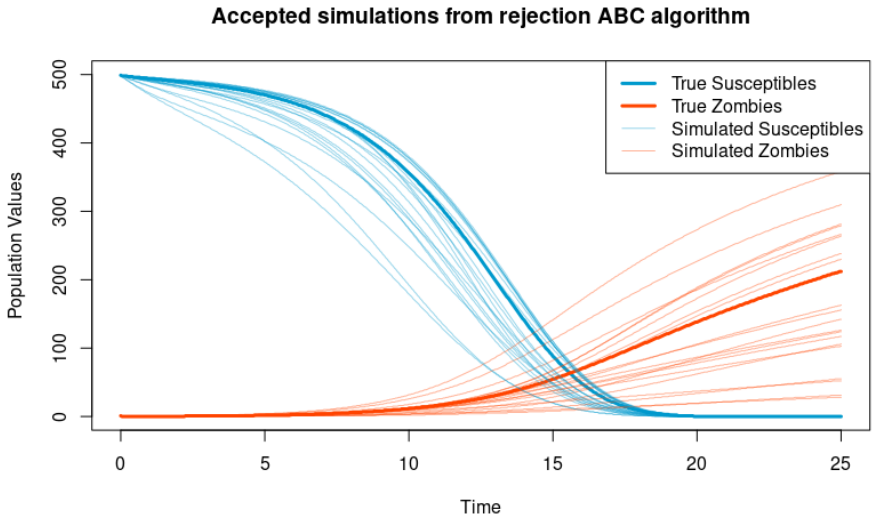
Instead of pre-defining the tolerance level ϵ, it is sometimes beneficial to simulate a set number of datasets n, of which we retain the proportion p closest to the observed data. This version of the rejection algorithm is preferential in the scenario where one is unsure of a suitable tolerance to use a priori. The algorithm used in this report is outlined below:

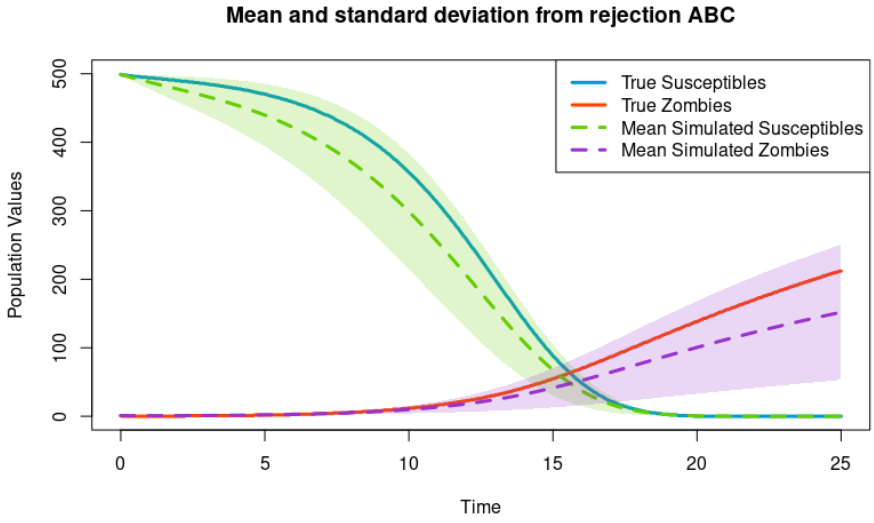
Algorithm 1: ABC Rejection

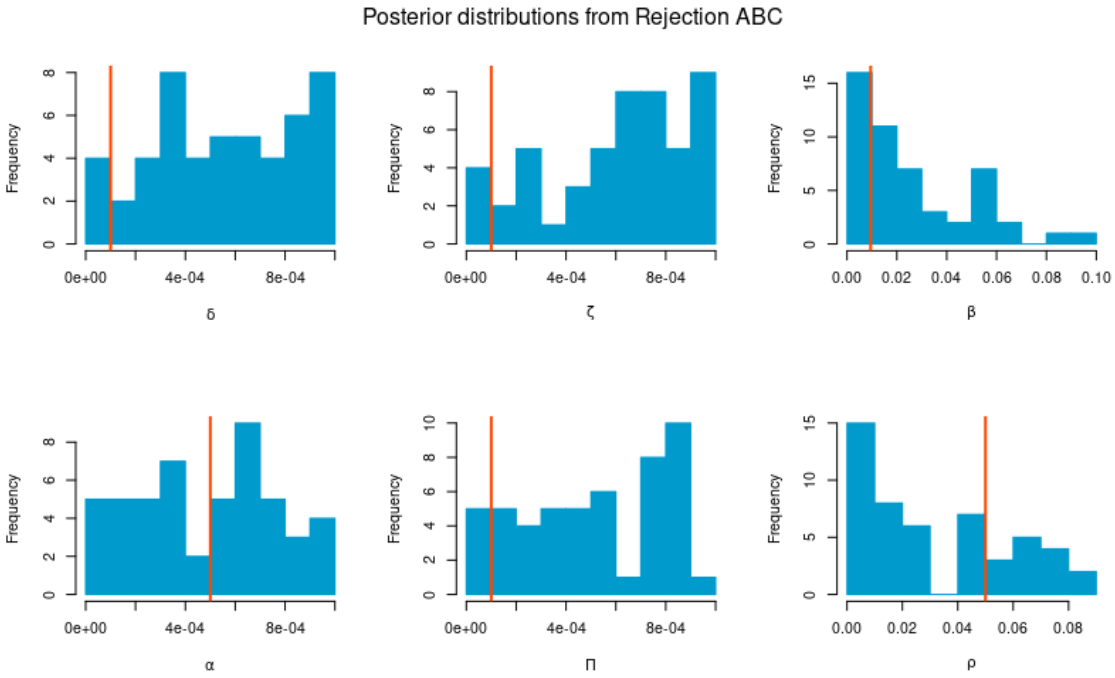
for $i \leftarrow 1$ **to** n **do**
  Sample $\theta_i \sim \pi(\theta)$;
  Simulate data $y_i \sim f(y|\theta_i)$;
  Transform data y_i into summary statistics s_i (optional);
  Calculate distance $\rho(s_i, s_{obs})$
end
Retain proportion p with smallest distance

Hence, in simple terms, the rejection algorithm consists of sampling parameters from a prior distribution and simulating datasets from a model using these parameters. One can then, optionally, transform these datasets into summary statistics, before retaining the proportion p closest to the observations. The accepted sample of parameters then approximates the posterior distribution.

3.2 Application to Zombie model







4 Sequential Monte Carlo

Sequential Monte Carlo (SMC) methods applied to ABC are concerned with reducing the number of model runs in order to achieve a certain quality of posterior approximation. We have now seen that the simple ABC rejection algorithm is very computationally demanding, which often limits applications to simple models; the number of simulations required to sample the entire parameter space grows exponentially with the number of parameters (Beaumont, 2010). SMC methods aim to progressively approximate the posterior using a decreasing sequence of tolerance levels $\{\epsilon_1, ..., \epsilon_T\}$. The idea is to sample from areas of the parameter space with a higher likelihood, rather than systematically from the whole space, in order to gain computational efficiency.

The basis of SMC algorithms is to construct a sequence of target distributions, that one propagates a set of particles through in order to move from the initial distribution to the final distribution.

4.1 Population Monte Carlo

The first stage of the PMC ABC method of Beaumont et al. (2009) is identical to that of rejection ABC in Algorithm 1, giving the first stage approximation to the posterior distribution. One then resamples parameter values from a density kernel

$$q(\theta) = \sum_{j=1}^{N} w_j^{(t-1)} K(\theta | \theta_j^{(t-1)}; \tau_t^2),$$

with N being the number of samples in the first, and subsequent, iterations. A popular choice is a Gaussian kernel and Beaumont et al. (2009) shows that a good choice of τ is twice the empirical

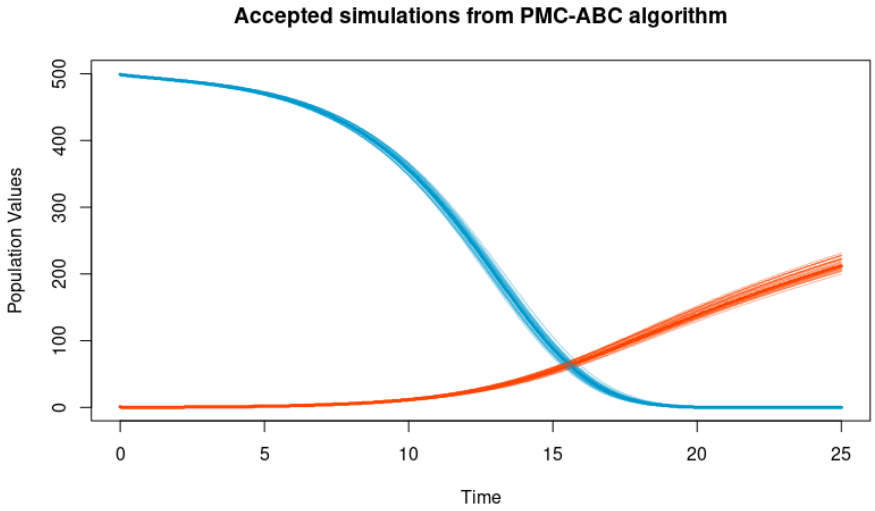
variance of the simulated parameters in the previous iteration. The bias that is induced by sampling from a proposal distribution, rather than the prior, is fixed by assigning an importance weight to each particle, as outlined in Algorithm 2 below.

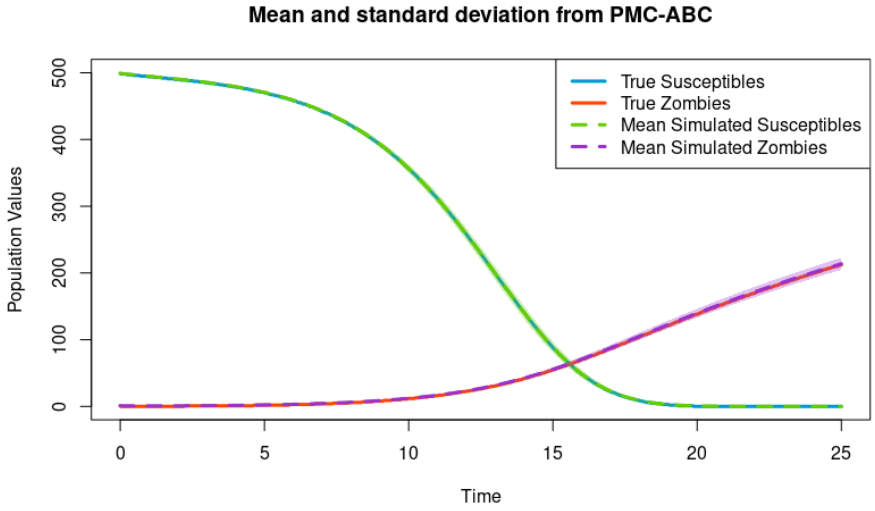
Algorithm 2: Population Monte Carlo ABC (PMC)

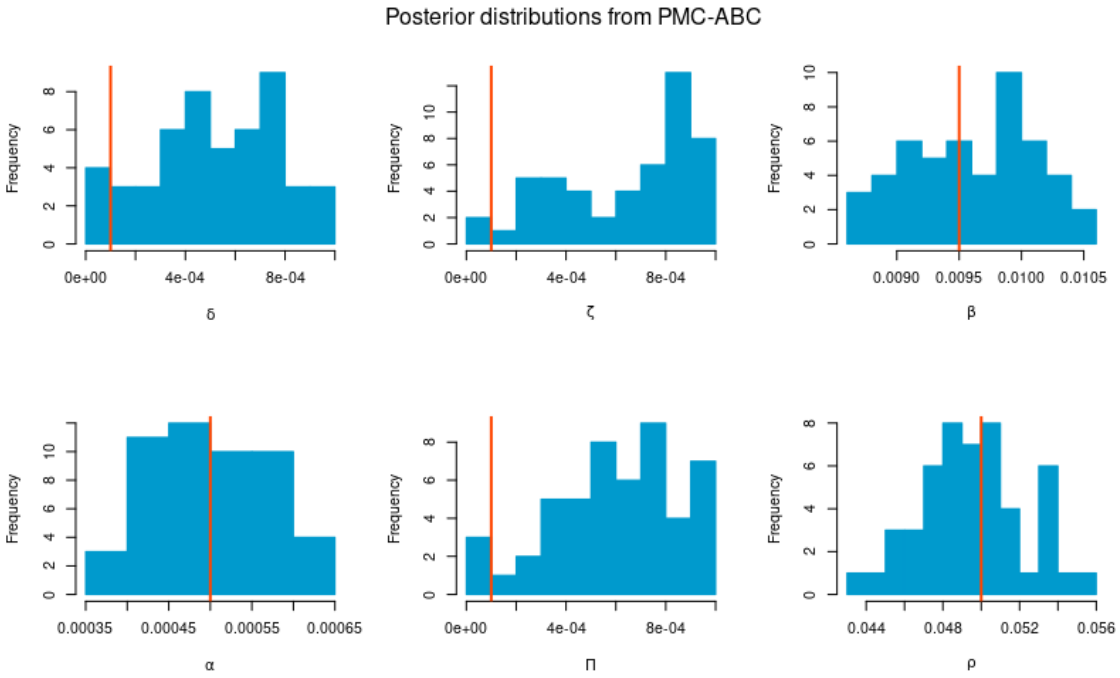
Given a decreasing sequence of tolerance levels $\{\epsilon_1, ..., \epsilon_T\}$;
for $t = 1$ **do**
  for $i = 1$ *to* N **do**
    until $\rho(y, y_{obs}) < \epsilon_1$;
    Simulate $\theta_i^{(1)} \sim \pi(\theta)$ and $y \sim f(y|\theta_i^{(1)})$
  end
  Set $w_i^{(1)} = 1/N$
end
for $t = 2$ *to* T **do**
  for $i = 1$ *to* N **do**
    until $\rho(y, y_{obs}) < \epsilon_t$;
    Sample θ_i^* from $\theta_j^{(t-1)}$ with probabilities $w_j^{(t-1)}$;
    Generate $\theta_i^{(t)} \sim K(\theta|\theta_i^*; \tau_t^2)$ and $y \sim f(y|\theta_i^{(t)})$
  end
  Set $w_i^{(t)} \propto \pi(\theta_i^{(t)}) / \sum_{j=1}^{N} w_j^{(t-1)} K(\theta_i^{(t)}|\theta_j^{(t-1)}; \tau_t^2)$;
  Set τ_{t+1}^2 as twice the weighted empririal variance of the $\theta_i^{(t)}$'s
end

A drawback of this method is that it requires the user to pre-define a decreasing sequence of tolerance levels $\{\epsilon_1, ..., \epsilon_T\}$. A poor choice of sequence could detriment the possible benefits of the importance sampling procedure. In the population genetic example conducted by Beaumont et al. (2009), ϵ_1 is based on a preliminary simulation as the 0.1 quantile and they perform four iterations with $\epsilon_2 = 0.75\epsilon_1$, $\epsilon_3 = 0.9\epsilon_2$ and $\epsilon_4 = 0.9\epsilon_3$.

4.2 Application to Zombie model







References

Beaumont, M. A. (2010). Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406.

Beaumont, M. A. (2019). Approximate bayesian computation. *Annual Review of Statistics and Its Application*, 6(1):379–403.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate bayesian computation. page arXiv:0805.2256.

Blum, M. G. and François, O. (2010). Non-linear regression models for approximate bayesian computation. *Statistics and computing*, 20(1):63–73.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential monte carlo method for approximate bayesian computation. *STATISTICS AND COMPUTING*, 22(5):1009–1020.

Diggle, P. J. and Gratton, R. J. (1984). Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):193–227.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate bayesian computation. *Biometrics*, 67(1):225–33.

Fearnhead, P. and Prangle, D. (2012). 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.

Hahn, C., Vakili, M., Walsh, K., Hearin, A. P., Hogg, D. W., and Campbell, D. (2017). Approximate bayesian computation in large-scale structure: constraining the galaxy–halo connection. *Monthly Notices of the Royal Astronomical Society*, 469(3):2791–2805.

Holden, P. B., Edwards, N. R., Hensman, J., and Wilkinson, R. D. (2018). Abc for climate: dealing with expensive simulators. *Handbook of approximate Bayesian computation*, pages 569–95.

Jiang, B., Wu, T.-y., Zheng, C., and Wong, W. H. (2017). Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, pages 1595–1618.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

Lenormand, M., Jabot, F., and Deffuant, G. (2013). Adaptive approximate bayesian computation for complex models. *Computational Statistics*, 28(6):2777–2796.

Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C. P., and Stumpf, M. P. (2014). A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Minter, A. and Retkute, R. (2019). Approximate bayesian computation for infectious disease modelling. *Epidemics*, 29:100368.

Nunes, M. A. and Balding, D. J. (2010). On optimal selection of summary statistics for approximate bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 9(1).

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Mol Biol Evol*, 16(12):1791–8.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2019). Abc random forests for bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *Genetics*, 182(4):1207–18.