

Dennis Prangle*, Paul Fearnhead, Murray P. Cox, Patrick J. Biggs and Nigel P. French

Semi-automatic selection of summary statistics for ABC model choice

Abstract: A central statistical goal is to choose between alternative explanatory models of data. In many modern applications, such as population genetics, it is not possible to apply standard methods based on evaluating the likelihood functions of the models, as these are numerically intractable. Approximate Bayesian computation (ABC) is a commonly used alternative for such situations. ABC simulates data x for many parameter values under each model, which is compared to the observed data x_{obs} . More weight is placed on models under which $S(x)$ is close to $S(x_{\text{obs}})$, where S maps data to a vector of summary statistics. Previous work has shown the choice of S is crucial to the efficiency and accuracy of ABC. This paper provides a method to select good summary statistics for model choice. It uses a preliminary step, simulating many x values from all models and fitting regressions to this with the model as response. The resulting model weight estimators are used as S in an ABC analysis. Theoretical results are given to justify this as approximating low dimensional sufficient statistics. A substantive application is presented: choosing between competing coalescent models of demographic growth for *Campylobacter jejuni* in New Zealand using multi-locus sequence typing data.

Keywords: ABC; model selection; sufficiency; *Campylobacter*; MLST; coalescent.

Text

*Corresponding author: Dennis Prangle, Department of Mathematics and Statistics, Lancaster University, UK, e-mail: d.b.prangle@reading.ac.uk

Paul Fearnhead: Department of Mathematics and Statistics, Lancaster University, UK

Murray P. Cox: Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; and Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

Patrick J. Biggs: Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Nigel P. French: Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; and Infectious Disease Research Centre, Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand

1 Introduction

The increasing availability of modern genetic data offers the possibility of learning more than ever before about the processes which generated it, for example the details of demographic change. However, for stochastic models that incorporate a high level of detail, it is impractical to evaluate numerically the probability of a dataset, preventing inference by standard likelihood-based methods. This has motivated the development of likelihood-free approaches, such as approximate Bayesian computation (ABC), which utilise the fact that simulating data from these models is relatively computationally cheap.

There is particular interest in using these methods to choose between explanatory models for observed data. However Robert et al. (2011) illustrated that, in the worst case, applying ABC to model choice problems can produce highly inaccurate results. This paper provides methods to address these concerns and improve the informativeness and efficiency of ABC model choice, while also scaling well with high dimensional data. We focus on a particular application, inferring the demographic history of *Campylobacter jejuni* in New Zealand from population genetic data. This will be described in detail later.

A simple ABC algorithm operates by simulating data sets x under various model and parameter pairs (\mathcal{M}, θ) . Pairs are accepted when x is sufficiently close to the observed data x_{obs} . This produces a sample of independent draws from an approximation to the Bayesian posterior distribution i.e. that of $\mathcal{M}, \theta | x$. Closeness is judged by the distance between vectors of *summary statistics*, $S(x_{\text{obs}})$ and $S(x)$. Previous work (e.g.

Blum, 2010; Fearnhead and Prangle, 2012) has shown that large approximation errors can be caused by high dimensional S . However, it is crucial that S is also informative about \mathcal{M} , as otherwise the problem of inaccurate results described by Robert et al. (2011) can occur.

This paper sets out a method to choose $S(x)$ for use in model selection. We give a theoretical result showing the existence of a low dimensional vector of statistics sufficient for model choice (under an appropriate definition given later). Our method aims to estimate such a vector. The idea is to use an extra simulation step to produce many (\mathcal{M}, θ, x) triples and then fit simple regression models of \mathcal{M} on x . The linear predictors fitted by the regressions are used, following an appropriate transformation, as S in a main ABC analysis. This S estimates the low dimensional sufficient statistics given by our theory. We refer to the approach as the *semi-automatic method* as it adapts the method of the same name in Fearnhead and Prangle (2012), which chooses S by regressing θ on x when the aim is inference of continuous parameters.

We expect that the targeted sufficient statistics are often complicated functions of the data which are hard to estimate globally. To make the task easier, we advise that the regressions are based on data simulated, within each model, from a limited subset of parameter values which is judged by preliminary analysis to hold most of that model's posterior mass (similar ideas are used in Blum and François, 2010; Fearnhead and Prangle, 2012; Fan et al., 2013). In other words, the simulation step mentioned above performs simulations from the models of interest following a truncation of their parameter supports. The resulting S can only be expected to perform well for choice between these truncated models. A separate theoretical contribution of the paper is to relate results from such a choice to the original model choice problem.

Our approach of performing regressions based on simulated data is similar to Estoup et al. (2012) who instead use linear discriminant analysis. We expect our other contributions would also be useful to this approach. An alternative approach is *subset selection* methods which attempt to select a subset of potential summary statistics $S'(x)$ which are approximately sufficient or optimise a related information theoretic quantity. Examples are Joyce and Marjoram (2008); Nunes and Balding (2010); Barnes et al. (2012b), the last of which has methods specifically for model choice. Subset selection methods seem best suited for problems with a small number of potential summary statistics as search costs are likely to scale badly. See Blum et al. (2013) for illustration in a parameter inference setting. Other work focuses on validating a particular choice of S . One approach is to run ABC analyses on a large number of simulated data sets to check whether S provides accurate results (Sjödin et al., 2012; Sousa et al., 2012). Marin et al. (2013) give a complementary approach, identifying necessary and sufficient properties of S for an ABC model choice analysis to be consistent in an asymptotic regime corresponding to highly informative data. Essentially, S must have different asymptotic means under the models. Given a choice of S , this property can be tested theoretically or through simulation. Validation techniques are useful, but not sufficient, to choose S for high dimensional genetic data where it is infeasible to compare all possible choices of S . Our contribution is a method which can be applied in this setting to propose good choices of S .

For computational efficiency, ideally the same ABC simulations would be used to provide inference on models and also their parameters. The method we present provides summary statistics suitable for model choice only. It would be desirable to augment them with informative summaries on model parameters, and we give an approach to do this that is specific to our main application. General methods are an interesting topic for future research.

The paper concentrates on providing and evaluating methodology for comparing up to three models, as this is currently the most common application of ABC model choice. There is scope to extend our methods to more models, as outlined in the discussion section.

The remainder of the paper is organised as follows. Section 2 describes ABC methods and our notation. Section 3 gives theoretical results on sufficiency. Section 4 explains our semi-automatic ABC method, and Section 5 illustrates it for simple examples. The main points of the application to *Campylobacter* data are given in Section 6, with further details supplied as supplementary material. The article concludes with a discussion in Section 7.

2 Background

Denote by \mathcal{M} a random variable which can take values $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M$, representing possible models. Let p_M be its prior mass function. In an abuse of notation \mathcal{M} will also denote a generic value of the variable, with usage clear from the context. Each model represents a joint distribution $\pi(x, \theta | \mathcal{M})$ on the data x and parameters $\theta \in \Theta$. This can be written as the product of prior and likelihood terms but we concentrate on the joint form for later convenience and to emphasise that the definition of a model includes a parameter prior. Note that it is possible for the parameters under each model to belong to different spaces, in which case Θ is their union, and that θ will also be used to denote both a random variable and generic value.

Bayesian inference concentrates on $\pi(\theta | x, \mathcal{M})$ – the posterior distribution of parameters under a specific model – and $\Pr(\mathcal{M} | x)$ – the posterior model probabilities. Inference on models can also be summarised using *Bayes factors* $B_{ij} = \pi(x | \mathcal{M}_i) / \pi(x | \mathcal{M}_j)$; the ratio of the *evidences* under models \mathcal{M}_i and \mathcal{M}_j . The Bayes factor does not involve p_M , but incorporating this information allows calculation of the ratio of posterior weights:

$$\Pr(\mathcal{M}_i | x) / \Pr(\mathcal{M}_j | x) = B_{ij} p_M(\mathcal{M}_i) / p_M(\mathcal{M}_j).$$

ABC is used in situations where it is possible to simulate $x | \mathcal{M}, \theta$ but evaluation of the density $\pi(x | \mathcal{M}, \theta)$ is impossible or impractically costly. A simple approach to ABC inference is Algorithm 1 (Grelaud et al., 2009).

Algorithm 1 Rejection sampling ABC incorporating model choice and parameter inference.

Input: Observed data x_{obs} , and a function $S(\cdot)$.
 A threshold $h \geq 0$ and a distance metric $d(\cdot, \cdot)$.
 An integer $N > 0$.

Iterate: For $i=1, \dots, N$

1. Simulate \mathcal{M}^* from $p_M(\mathcal{M})$.
2. Simulate θ^* from $\pi(\theta | \mathcal{M}^*)$.
3. Simulate x_{sim} from $\pi(x | \theta^*, \mathcal{M}^*)$.
4. Accept $(\mathcal{M}^*, \theta^*)$ if $d(S(x_{\text{obs}}), S(x_{\text{sim}})) \leq h$.

Output: A set of accepted model and parameter pairs of the form $(\mathcal{M}^*, \theta^*)$.

Letting \mathbb{I} represent an indicator function, define

$$p_{\text{ABC}}(\mathcal{M} | S(x)) \propto p_M(\mathcal{M}) \int \pi(S(x) | \mathcal{M}) \mathbb{I}[d(S(x_{\text{obs}}), S(x)) \leq h] dx,$$

$$\pi_{\text{ABC}}(\theta | \mathcal{M}, S(x)) \propto \pi(\theta) \int \pi(S(x) | \theta, \mathcal{M}) \mathbb{I}[d(S(x_{\text{obs}}), S(x)) \leq h] dx.$$

Then the sample of (\mathcal{M}, θ) values output by Algorithm 1 is drawn from a distribution with conditionals $\pi_{\text{ABC}}(\theta | \mathcal{M}, S(x))$ and marginal $p_{\text{ABC}}(\mathcal{M} | S(x))$.

In the limit $h \rightarrow 0$, the ABC target distributions just defined converge on $\Pr(\mathcal{M} | S(x))$ and $\pi(\theta | \mathcal{M}, S(x))$. However, reducing h decreases the output sample size, increasing Monte Carlo approximation error. A *curse of dimensionality* result (Fearnhead and Prangle, 2012) shows that the rate of increase in error typically rises with the dimension of S (Blum, 2010 gives a related result on overall error.) This motivates a low dimensional S . It is also important that S is informative so that the limiting ABC targets approximate the posterior distributions $\Pr(\mathcal{M} | x)$ and $\pi(\theta | \mathcal{M}, x)$ well. Hence S is a crucial tuning choice.

In practice, the results of Algorithm 1 can be highly variable if some prior model masses are small. Algorithm 2 is a more stable alternative suggested by Grelaud et al. (2009). It samples \mathcal{M}^* values from a uniform distribution rather than p_M , and it is necessary to correct the results to take this into account. Let n_i be the number of occurrences of \mathcal{M}_i in the output sample. Then n_i/n_j is an estimator of the Bayes factor B_{ij} and $n_i p_M(\mathcal{M}_i) / \sum_{j=1}^M n_j p_M(\mathcal{M}_j)$ is an estimator of $\Pr(\mathcal{M}_i | S(x))$. The asymptotic results outlined above continue to hold (Grelaud et al., 2009), and the curse of dimensionality arguments of Fearnhead and Prangle (2012) apply with little modification.

Algorithm 2 A more stable modification of Algorithm 1.

As Algorithm 1 except:
 1. Set \mathcal{M}^* to $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M$ with equal probability.

More efficient ABC model choice algorithms have been proposed, mainly based on sequential Monte Carlo (SMC) (e.g., Toni and Stumpf, 2010; Del Moral et al., 2012). However, the tuning issues just described remain. The SMC algorithm of Toni and Stumpf (2010) is used later and described in the supplementary material (Section 1). Another approach to improve the quality of ABC results is to *post-process* them. This uses accepted parameters $\theta^{*,1}, \theta^{*,2}, \dots$, models $\mathcal{M}^{*,1}, \mathcal{M}^{*,2}, \dots$ and the corresponding simulations $x^{*,1}, x^{*,2}, \dots$. For parameter inference *regression adjustment* (Beaumont et al., 2002; Blum and François, 2010) fits a model $\theta = f(x, e)$, where f is a deterministic function and e a random residual, and outputs adjusted values $\theta^{*,i} = \hat{f}(x_{\text{obs}}, \hat{e}^i)$. Model choice results can be post-processed by fitting a multinomial regression model $\Pr(\mathcal{M} | x) = g(x)$ and returning $\hat{g}(x_{\text{obs}})$ (Beaumont, 2008).

3 Theory

A statistic $S(x)$ of data x is said to be *Bayes sufficient* for parameter θ if $\theta | S(x)$ and $\theta | x$ have the same distribution for any prior distribution and almost all x (Kolmogorov, 1942). This is a natural definition of sufficiency for ABC, as it shows that in an ideal ABC algorithm with $h \rightarrow 0$, the ABC target distribution equals the correct posterior when S is used. Throughout later sections of this paper we use “sufficient” to mean Bayes sufficient. Theorem 1 gives an alternative characterisation of Bayes sufficiency for \mathcal{M} in the setting described in Section 2.

Theorem 1 Let $T(x) = \{T_1(x), T_2(x), \dots, T_{M-1}(x)\}$ where

$$T_i(x) = \pi(x | \mathcal{M}_i) / \left[\sum_{j=1}^M \pi(x | \mathcal{M}_j) \right].$$

Then S is Bayes sufficient for \mathcal{M} if and only if there exists a function g such that $g[S(x)] = T(x)$ for almost all x .

Theorem 1 shows that for any situation with M models there are sufficient statistics for model choice of dimension $M-1$, namely the vector $T(x)$. Furthermore, vectors $S(x)$ which can be transformed to $T(x)$ are also sufficient. One particular sufficient choice of $S(x)$ used later is a vector of all Bayes factors under a one-to-one transformation. Additionally, we note that a sufficient $S(x)$ may contain summaries which do not contribute to $T(x)$ but are useful for parameter inference.

A similar result is Theorem 3a of Fearnhead and Prangle (2012), which shows that for continuous parameters θ , $S(x) = E(\theta | x)$ is an optimal choice to estimate parameter means in terms of minimising quadratic error loss. However this $S(x)$ is typically not sufficient for θ . Theorem 1 is a stronger result for the case of model choice (or, equivalently, for estimating discrete parameters) showing the existence of low dimensional vectors of sufficient statistics.

A sketch of the proof is now given, followed by the full details. The theorem states that Bayes sufficiency of $S(x)$ for \mathcal{M} is equivalent to there being a deterministic transformation from $S(x)$ to $T(x)$. The latter vector is $M-1$ posterior probabilities given observations x and uniform p_M . Under uniform p_M , conditioning \mathcal{M} on $S(x)$ satisfying this condition clearly recovers the posterior weights. Reweighting can be used to show that the posterior is also recovered under any other p_M . The converse can be shown by construction.

Proof of Theorem 1 Bayes sufficiency of $S(x)$ for \mathcal{M} is equivalent to the following being true for all i and p_M , and almost any x ,

$$\Pr(\mathcal{M}_i | S(x)) = \Pr(\mathcal{M}_i | x). \quad (1)$$

For convenience we introduce $\mathbf{p}=(p_M(\mathcal{M}_i))_{1 \leq i \leq M}$ to represent the prior mass function. Also, let $\mathbf{1}$ be a vector of M components equal to 1.

First assume S is Bayes sufficient for \mathcal{M} . Define $h_i(S(x), \mathbf{p}) = \Pr_p(\mathcal{M}_i | S(x))$ (i.e. the conditional probability under prior \mathbf{p}) and note $h_i(S(x), \mathbf{p}) = \Pr_p(\mathcal{M}_i | x)$. The required function is $g(S(x)) = (h_i(S(x), M^{-1}\mathbf{1}))_{1 \leq i \leq M-1}$.

It remains to prove Bayes sufficiency for \mathcal{M} given a function g of the form described in the theorem. Henceforth we consider only the case $\mathbf{p} = M^{-1}\mathbf{1}$, since in this case (1) is equivalent to $\Pr(x | \mathcal{M}_i) = k \Pr(S(x) | \mathcal{M}_i)$ for some constant k , and applying Bayes' theorem to this proves (1) for general \mathbf{p} . It also suffices to show that (1) holds for all $i < M$; the case $i = M$ follows as probabilities sum to 1. Fix some $i < M$ and define an indicator variable $Y = \mathbb{I}[\mathcal{M} = \mathcal{M}_i]$. Then $T_i(x) = \Pr(\mathcal{M}_i | x) = E[Y | x]$ and $\Pr(\mathcal{M}_i | S(x)) = E[Y | S(x)]$. To prove (1), we will show that these conditional expectations are almost always equal. Standard properties of conditional expectation give $E[Y | S(x)] = E[E\{Y | x\} | S(x)] = E[T_i(x) | S(x)]$. Finally, $E[T_i(x) | S(x)] = E[g_i(S(x)) | S(x)] = g_i(S(x)) = T_i(x) = E[Y | x]$ for almost all x as required, where $g_i(\cdot)$ represents the i th component of the $g(\cdot)$ function. \square

4 Method

The low dimensional sufficient statistics described by Theorem 1 are generally not available. However their existence motivates an approach of approximating them from simulated data, and then using these approximations as $S(x)$ within ABC. Algorithm 3 is an outline of an approach to do this. To run the algorithm, the user must choose the implementation details, such as how to perform Step 2, discussed below in Section 4.1, and the type of ABC analysis to use in Step 3 (e.g. rejection sampling or SMC). Due to these choices the method is referred to as “semi-automatic ABC”. When this algorithm (or Algorithm 4, given shortly) is used in later sections, full details of the implementation choices will be given.

Algorithm 3 Outline of simple semi-automatic ABC for model choice. Full details of how the steps can be carried out are given in Sections 4.1 and 4.2. Uses in later sections describe the selected implementation choices in full.

-
1. Simulate a large number of (\mathcal{M}, θ, x) triples.
 2. Calculate $S(x)$ by estimating sufficient statistics from simulations.
 3. Perform the ABC analysis using $S(x)$.
-

Sufficient statistics are likely to be highly complicated functions of the data due to the complexity of the models, and thus hard to approximate. To make the task more tractable, we recommend some optional extra steps to give Algorithm 4. This simplifies the models by concentrating on the most likely parameter values. We view this as replacing the models $\pi(\theta, x | \mathcal{M}_i)$ with *truncated models*

$$\pi(\theta, x | \mathcal{M}_i') \propto \pi(\theta, x | \mathcal{M}_i) \mathbb{I}(\theta \in R_i), \quad (2)$$

where R_i is a *training region* for model \mathcal{M}_i . Calculation of S is performed using data simulated from the truncated models. The resulting S estimates sufficient statistics for the choice between the truncated rather than original models. Therefore the main ABC analysis must be performed between the truncated models, and, as will be shown in Section 4.2, the results can be used to estimate the model choice posterior for the original problem.

Algorithm 4 Outline of semi-automatic ABC for model choice with truncation steps. Full details of how the steps can be carried out are given in Sections 4.1 and 4.2. Uses in later sections describe the selected implementation choices in full.

-
1. Perform an ABC *pilot analysis* with ad-hoc summary statistics. Use the output for each model to choose training regions R_i of parameters which contain most of the posterior probability for each model \mathcal{M}_i .
 2. Simulate a large number of (\mathcal{M}, θ, x) triples using truncated models.
 3. Calculate $S(x)$ by estimating sufficient statistics from simulations.
 4. Perform the ABC *main analysis* using $S(x)$ and truncated models.
 5. Use truncation correction to estimate posterior probabilities.
-

The remainder of this section discusses the implementation of the steps in these algorithms in more detail. Performance is assessed through simulation examples in Sections 5 and 6.

4.1 Calculating summary statistics

This section describes a logistic regression based approach to estimating sufficient statistics from simulated *training data*. Consider first the case of two models \mathcal{M}_1 and \mathcal{M}_2 , which for this discussion may represent original or truncated models, with training data drawn from the joint distribution on (\mathcal{M}, x) , where $x=(x_1, x_2, \dots, x_p)$. Define $q(x)=\Pr(\mathcal{M}_1|x)$. This is clearly a sufficient statistic for \mathcal{M} . Logistic regression can be used to fit

$$\text{logit}q(x):=\log\{q(x)/[1-q(x)]\}=\beta_0+\sum_{i=1}^p\beta_i x_i. \quad (3)$$

The fitted $\hat{q}(x)$ is an estimate of a sufficient statistic. Note also that $q(x)/[1-q(x)]$ is the Bayes factor multiplied by a constant depending on the prior model weights.

To improve on the fit of (3) and cope with situations where x is very high dimensional or not a fixed-length vector, in practice we fit instead

$$\text{logit}q(x)=\beta^T f(x), \quad (4)$$

where $f(x)$ is a vector of transformations of x , including a constant term. This can perform initial dimension reduction and introduce non-linear functions of the data into the regression. Example choices of $f(\cdot)$ used later are order statistics of raw data or a large number of summaries of genetic sequence data used in previous literature, and also, in both cases, transformations of these (including a constant term). To assist in the choice of $f(\cdot)$, regression diagnostics can be used, for example to compare the quality of the logistic regression fits for some $f_1(\cdot)$ and $f_2(\cdot)$. The supplementary material (Section 6) gives examples in which cross-validation estimates of the deviance are compared.

Consider the case of choice between three models $\mathcal{M}_1, \mathcal{M}_2$ and \mathcal{M}_3 . Fix a pair of distinct models \mathcal{M}_i and \mathcal{M}_j , and consider the subset of training data made up of only the simulations from these models. Logistic regression can be used as above to estimate the probability of $\mathcal{M}_i|x$ under the (\mathcal{M}, x) distribution for this training data subset. This is repeated for all three pairs of distinct models, and results in a vector of one-to-one transformations of Bayes factors. This target was shown to be sufficient for \mathcal{M} in Section 3.

Alternatives to logistic regression and the case of $M>3$ models are discussed in Section 7.

4.2 Other steps

4.2.1 Pilot analysis

The pilot ABC analysis uses an ad-hoc choice of summary statistics S_{pilot} . The purpose of the pilot analysis is to roughly identify regions containing most of the posterior mass, so the procedure should be reasonably robust to the choice of S_{pilot} . Fearnhead and Prangle (2012) illustrate this argument by example. Validation tests could also be performed to test the quality of ABC output from analysing simulated data using S_{pilot} .

In our implementation the pilot uses an ABC model choice algorithm such as Algorithm 2. An alternative approach would be to perform a separate pilot run for each model, focusing only on finding training regions, rather than initial model choice analysis. We did not investigate this as a pilot analysis incorporating model choice has useful properties. The estimated posterior can serve as a verification that the final results appear sensible. Also, if the pilot results are sufficiently convincing in showing that certain models are incompatible with the data, they could be ruled out at this stage saving computational resources. (This assumes the goal is purely inference of model weights. For predictive analysis, models with low probabilities may make an important contribution.)

4.2.2 Training region choice

The training region R_i for model \mathcal{M}_i' should cover most of the posterior mass. Our implementation is to choose a hypercube, with the range of each parameter being the interval of sampled values.

4.2.3 Simulating data

We generate training data from the distribution on (\mathcal{M}, θ, x) defined by the priors and models (or truncated models). An alternative model distribution can be used without affecting the arguments in Section 4.1 showing that the fitted summary statistics are estimates of sufficient statistics. This would be useful if some prior model weights are too small to fit all regressions well.

4.2.4 Truncation correction

Results of the main ABC analysis choosing between truncated models can be used to estimate those for the original model choice problem by the following consequence of (2):

$$\pi(x | \mathcal{M}_i) = r_i \pi(x | \mathcal{M}_i'), \quad \text{where } r_i = \Pr(\theta \in R_i | \mathcal{M}_i) / \Pr(\theta \in R_i | x, \mathcal{M}_i).$$

That is, the evidence of \mathcal{M}_i equals that of \mathcal{M}_i' multiplied by r_i , the ratio of the prior and posterior probabilities of R_i . This allows estimation of Bayes factors or posterior probabilities for the original models given r_i values. As R_i is chosen with the aim of containing most of the posterior mass, we estimate its posterior probability by 1, giving an estimate $\hat{r}_i = \Pr(\theta \in R_i | \mathcal{M}_i)$. This prior probability can usually be calculated directly when R_i is a hypercube.

5 Examples

To illustrate our semi-automatic ABC method, we apply it to three simple binary model selection examples from the literature (Didelot et al., 2011; Marin et al., 2013), and extend one of these to a three model example. The examples are summarised in Table 1. The binary examples are the first two models in each letter group, and the 3 model example is the full A group. In each case the data are 100 independent draws from one of the models and the models have equal prior probabilities.

For each example, 100 test datasets were generated based on parameters drawn from the priors, with the numbers of datasets generated from each model as close to equal as possible. This section considers how well estimated posterior probabilities from various approaches match the true generating models. All these approaches use ABC rejection sampling (Algorithm 2) once S is selected.

Table 1 Models used in the examples of Section 5. For details of the g -and- k distribution see Rayner and MacGillivray (2002).

| Name | Model | Prior |
|------|----------------------------------|--|
| A1 | Poisson(θ) | $\theta \sim \text{Exponential}(1)$ |
| A2 | Geometric(θ) | $\theta \sim \text{Uniform}(0, 1)$ |
| A3 | Binomial(10, θ) | $\theta \sim \text{Beta}(1, 9)$ |
| B1 | Laplace ($\theta, 1/\sqrt{2}$) | $\theta \sim \text{Normal}(0, 2^2)$ |
| B2 | Normal($\theta, 1$) | $\theta \sim \text{Normal}(0, 2^2)$ |
| C1 | $gk(0, 1, 0, k)$ | $k \sim \text{Unif}(-0.5, 5)$ |
| C2 | $gk(0, 1, g, k)$ | $(g, k) \sim \text{Unif}([0, 4] \times [-0.5, 5])$ |

5.1.1 Binary model selection

The semi-automatic ABC method of Algorithm 4 was implemented starting with a pilot analysis using $S_{10}(x)=(x^{(5)}, x^{(15)}, \dots, x^{(95)})$ where $x^{(i)}$ is the i th order statistic. Model choice summary statistics were fitted as described in Section 4.1 using $f(x)=(1, x^{(1)}, x^{(2)}, \dots, x^{(100)})$. No other summaries were added for parameter inference. The analysis used 2×10^4 simulations, one quarter for the pilot and the rest used for both summary statistic fitting and the main analysis. The pilot and main analysis both accepted 100 simulations. Some alternative ABC analyses on the data were performed, each using the same total number of simulations and acceptances. Firstly, the analysis was repeated using Algorithm 3. Secondly, standard ABC analyses were performed with Algorithm 2 using (a) $S=S_{10}$ (b) particular choices of S used in Marin et al. (2013); 4th and 6th moments for B, 10% and 90% quantiles for C. All ABC analyses used the following distance metric

$$d(x, y) = \left[\sum_{i=1}^p (x_i - y_i)^2 / \hat{\sigma}_i^2 \right]^{1/2}, \quad (5)$$

i.e., Euclidean distance between p -dimensional summary statistics normalised by estimated standard deviations, $\hat{\sigma}_i$. The latter were estimated from the simulated data.

Figure 1 shows estimated posterior probabilities for S_{10} and Algorithm 4. Numerical summaries of estimation quality are given in Table 2. This reports the mean entropic loss (Robert, 1996),

$$-\frac{1}{n} \sum_{i=1}^n \log \hat{\Pr}(\mathcal{M}_{0,i} | x_{\text{obs},i}), \quad (6)$$

the mean negative log probability of the correct models $\mathcal{M}_{0,1}, \dots, \mathcal{M}_{0,n}$ estimated from the corresponding simulated datasets $x_{\text{obs},1}, \dots, x_{\text{obs},n}$ where in this case the number of test datasets, n , is 100. Also reported is the misallocation rate; the proportion of datasets where the highest weighted model was not the correct model. Our method provides an improvement in all scenarios, although this is modest for example C. The use of the truncation steps from Algorithm 4 is shown to sometimes be crucial; when Algorithm 3, which omits these, is used instead, the results for example C are the worst of all methods. However the effect is problem

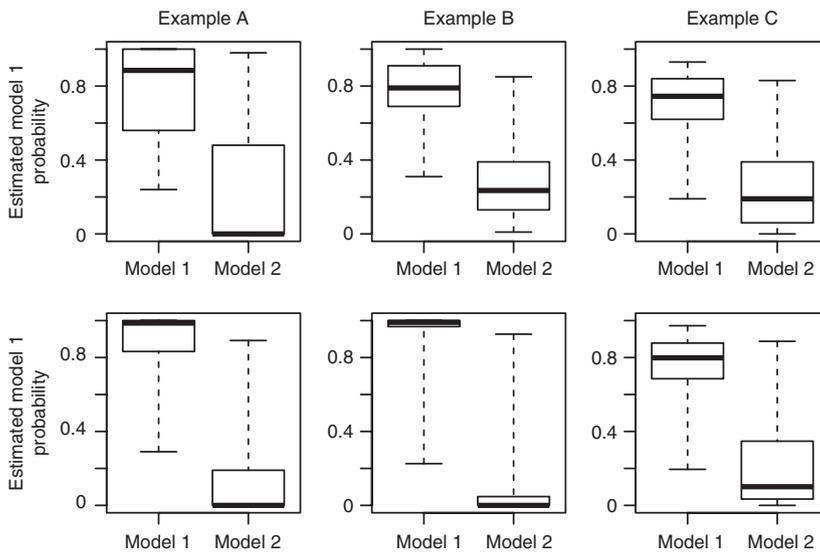


Figure 1 Boxplots of posterior probabilities of model 1 estimated by ABC (without post-processing) for 100 simulated datasets in each of three binary model comparison examples. The boxplots show quartiles of the values. Within each graph results are split by which model generated the data. The top row uses $S=S_{10}$, and the second row chooses S by semi-automatic ABC (Algorithm 4; see text for full details of its implementation). The columns represent three model choice examples detailed in Table 1.

Table 2 Mean entropic loss and misallocation rate (in brackets) from several ABC analyses of 100 simulated datasets in each of four model comparison examples, detailed in Table 1. The final row shows values under the exact posterior, where these are available, for comparison. Full details of how the algorithms were implemented for these examples are given in the text.

| Summary statistics | Example | | | |
|--------------------|-------------|-------------|----------------|-------------|
| | Binary A | Binary B | Binary C | 3 models |
| S_{10} | 0.330 (17%) | 0.335 (11%) | 0.430 (16%) | 0.707 (39%) |
| From literature | – | 0.553 (25%) | 0.409 (20%) | – |
| From Algorithm 3 | 0.302 (14%) | 0.135 (5%) | ∞ (21%) | 0.659 (42%) |
| From Algorithm 4 | 0.198 (15%) | 0.139 (7%) | 0.384 (14%) | 0.589 (33%) |
| Posterior | 0.198 (12%) | 0.156 (8%) | – | 0.581 (36%) |

dependent; in example B it made little difference. Exact posterior calculations are possible for examples A and B (for Laplace marginal likelihood calculations see Robert, 2014), and in both cases Algorithm 4 provides comparable results.

We attempted to apply post-processing by the method of Beaumont (2008). For example A this was usually not possible either because there was no variation in the accepted summaries, which were discrete in this case, or because all acceptances were for a single model. For the other examples, it had little effect on entropic loss or misallocation rate, so these are not reported.

5.1.2 Selection between three models

Algorithms 3 and 4 were implemented as for the two model examples, with the addition that three summary statistics were fitted, corresponding to three pairs of models. Figure 2 plots exact posterior probabilities against ABC estimates, and shows that Algorithm 4 performs better than the comparison analysis using $S=S_{10}$. This is confirmed by the quantitative summaries in Table 2, which also shows that Algorithm 4 outperforms

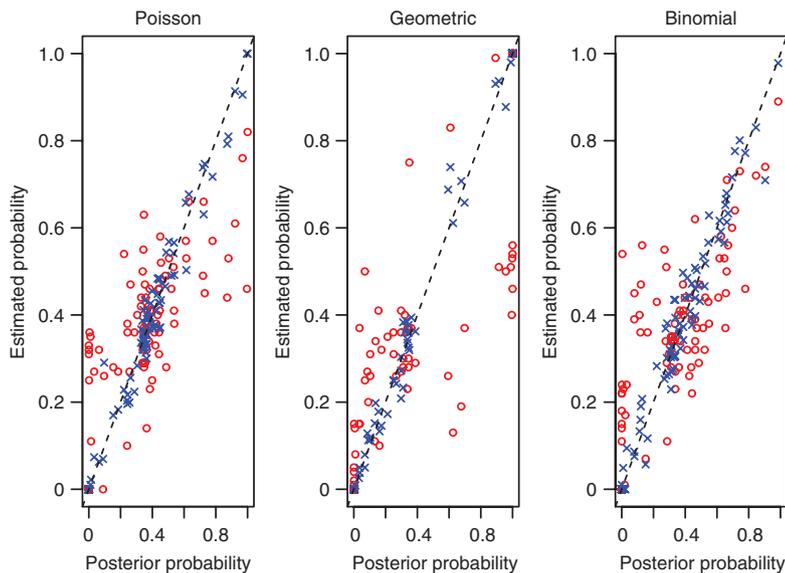


Figure 2 Plots of true posterior model weight against ABC estimates for 100 simulated datasets in a three model example. Circles are for $S=S_{10}$ and crosses for S chosen by semi-automatic ABC (Algorithm 4; see text for full details of its implementation).

Algorithm 3 and achieves comparable results to the true posterior values. Post-processing results are not shown because, as mentioned above, they could usually not be calculated for this example.

6 Application

Campylobacter jejuni and *Campylobacter coli* are bacterial pathogens that are a major cause of human gastroenteritis around the world (Humphrey et al., 2007). They are considered commensals of a wide variety of animals, including poultry, ruminants and wild birds, and human infection occurs as a result of ingesting contaminated food or drinking water and via direct contact with animal faeces (Savill et al., 2003). New Zealand has very high rates of campylobacteriosis and an investigation into the source of human infection (Mullner et al., 2009) has generated a large dataset of isolates from humans and animals that have been characterized by multilocus sequence typing, MLST (Dingle et al., 2001). The dataset of *C. jejuni* and *C. coli* isolates from New Zealand has been used to inform control policy (Sears et al., 2011) and to estimate evolutionary parameters, such as the rates of mutation and recombination (Yu et al., 2012). We focus on the question of demographic history, which is of particular interest in New Zealand due to the relatively recent colonization by man and the unique pattern of animal introductions (both wildlife and domestic animals) (Atkinson and Cameron, 1993). We ask: can we detect historic growth in the effective population size, and if so, does it correspond to a particular historical event? The relative isolation of this location means that neglecting ongoing exchange with outside populations is reasonably realistic. MLST data are available for over 3000 isolates from a variety of hosts.

We present our methods and results below, with a discussion given in Section 6.4. Further details are provided as supplementary material.

6.1 Models and priors

We modelled the *C. jejuni* data using a coalescent model using the Jukes-Cantor model of DNA substitution and incorporating the gene conversion recombination model of Wiuf and Hein (2000) with exponential demographic growth, as simulation of this scenario is straightforward using existing tools (detailed below). However, simulation of a large dataset is prohibitively slow so we used a random subsample of 100 isolates. Coalescent theory suggests that such a sample size captures much of the information of the full sample (Nordborg, 2004), and simulation based checks on informativeness are detailed in the supplementary material (Section 6). The selected isolates were confirmed to be *C. jejuni* using the PubMLST database and through a phylogeny analysis of these isolates and a representative *C. coli* sequence. Three models were considered, with equal prior weights: *Model 1* no growth; *Model 2* growth for 50 years (since the expansion of the New Zealand poultry industry); *Model 3* growth for 170 years (since the introduction of European livestock, primarily from Australia and the UK).

Each model has three biological parameters: a recombination rate, mean track length (i.e., length of recombining DNA segment) and mutation rate. Models 2 and 3 also have a growth parameter. To aid interpretability we parameterised this as the relative increase in the effective population size during the period of growth. Prior information on parameters is summarised in Table 3. Mutation and recombination rates are given per kilobase per $2N_e g$ years, where N_e is the effective population size and g the generation length in years. Wilson et al. (2009) estimated the mean time to coalescence, $N_e g$, at 209 with an interval estimate of [155,288]. To simplify our model, we fix $N_e g = 209$. We expect that variations of $N_e g$ within the quoted interval will not affect the detection of growth. Mean recombination length is in kilobase units. The relative growth parameter is unitless as it is a ratio of effective population sizes.

Growth priors are based on demographics of the principal host; poultry for model 2 and sheep/cattle for model 3. Rough estimates of host growth rates are used, based on the data of French et al. (2014), with variance increased to account for uncertainty of the link between bacterial and host demographics. Biological

Table 3 Details of the parameter priors used in Section 6. Priors are assumed to be the product of a log normal prior for each individual parameter. The point estimates are geometric means. The recombination length prior was truncated below 1 base pair, and the recombination rate above $25 \text{ kb}^{-1}(2N_e g)^{-1}$ to avoid excessively slow simulations. (All estimated posteriors for recombination rate were well below this – see Figure 2 of supplementary material.)

| Parameter | Units | Model | Point estimate | 95% CI | Log normal | |
|--------------------|-------------------------------|-------|----------------|--------------|------------|------|
| | | | | | Mean | Sd |
| Mutation rate | $\text{kb}^{-1}(2N_e g)^{-1}$ | All | 13.7 | [8.1, 23.2] | 2.62 | 0.27 |
| Recombination rate | $\text{kb}^{-1}(2N_e g)^{-1}$ | All | 1.31 | [0.03, 51.5] | 0.27 | 1.87 |
| Mean track length | kb | All | 4.52 | [0.1, 209.9] | 1.51 | 1.96 |
| Relative growth | – | 2 | 4.06 | [1.5, 10.8] | 1.40 | 0.50 |
| Relative growth | – | 3 | 33.1 | [2.9, 383.8] | 3.50 | 1.25 |

parameter priors are based on analysis of other *C. jejuni* data in Wilson et al. (2009). This assumed a no growth model, so these priors may not be appropriate for models 2 and 3. Sensitivity analysis detailed in the supplementary material (Section 7) also considers a much less informative biological prior.

6.2 Methods

Data sets were simulated using *ms* (Hudson, 2002) and *seq-gen* (Rambaut and Grassly, 1997). Genetic summaries required were calculated using R, which was also used to code the inference algorithms.

We implemented semi-automatic ABC (Algorithm 4) as follows. First a pilot analysis was performed using the ABC SMC algorithm of Toni and Stumpf (2010) (see supplementary material Section 1) with 1000 particles. This targeted log-transformed parameters, as on the original scale the target distribution is roughly log-normal and hard for the algorithm to explore. The summary statistics were a set of 15 genetic summaries (these, and other summaries used below, are listed in Section 2 of the supplementary material). The distance metric was Equation (5), Euclidean distance between normalised summary statistics, with standard deviations estimated from 100 datasets simulated from the prior predictive distribution. These simulations were also used to choose an initial ABC threshold: the median of the distances between these datasets and the observations. In the following SMC iterations, the threshold was the median of distances for accepted particles in the preceding step. The algorithm terminated at the first available point after 2×10^4 total simulated data sets. That is, it also performs the further simulations necessary to finish the current threshold. Alternative ABC SMC algorithms are discussed in Section 7.

To fit summary statistics, 2×10^4 datasets were simulated using the training regions. Model choice summaries were fitted as described in Section 4.1 and summaries for parameter inference by linear regression (detailed shortly). For all regressions the vector of covariates $f(\cdot)$ consisted of 3 cubic B-spline bases for each of 125 genetic summaries, giving a total of 375 covariates, and a constant term. Spline transformations were included to capture non-linear effects. Due to the large number of covariates, L_1 penalised versions of logistic and linear regression were used, using the “glmnet” R package (Friedman et al., 2010) with the tuning parameter chosen by cross-validation. Cross-validation estimates of fitting error were used to investigate which genetic summaries were most informative and to validate many of our modelling and tuning choices (details in Section 6 of the supplementary material).

Exploratory analysis showed that for each parameter a single estimator could perform reasonably well under all models (supplementary material Section 6). To keep $\text{dim}(S)$ small, our S is the concatenation of such estimators with model choice statistics. A single hypercube training region was used for all models: the product of the parameter ranges from the entire pilot output, regardless of model. The motivation was based on pilot run results. This placed little weight on some models (see Table 6 below) making estimation of their training regions difficult. However marginal parameter posteriors were similar for all models (see Figure 2 of the supplementary material) allowing pooling of information. The regression responses were log-transformed

parameters, supported by exploratory analysis of Box-Cox transformations. The resulting predictors were exponentiated to use in S . Regressions for biological parameters were fitted using the simulations from all models, while those for the demographic parameter used simulations from the growth models only.

The final S vector used in the main ABC analysis consisted of four parameter estimators and three statistics for model choice. The analysis used the distance metric (5) with $\hat{\sigma}_i$ estimated from the training data (n.b. These differ from earlier $\hat{\sigma}_i$ values as a different S is used.) The analysis used the same SMC ABC algorithm as the pilot run, again with 1000 particles and targeting log-transformed parameters. The initial threshold was the median of distances to the observed summary statistics calculated from the training data, with subsequent thresholds chosen as in the pilot run. The algorithm terminated after the iteration which reached 4×10^4 simulated data sets.

To investigate performance we also analysed simulated datasets, where the results can be compared to the known generating models. As in Section 5, we simulated 100 data sets split roughly evenly between the three models with parameters drawn from the priors, and performed ABC analyses of each. To make the computational cost manageable we used ABC rejection sampling (Algorithm 2) once S was selected, and reused the same (\mathcal{M}, θ, x) simulations in each analysis. Each analysis used $N=4 \times 10^6$ total simulations and accepted 100. As before normalised Euclidean distance (5) was used. Semi-automatic ABC was implemented without truncation steps (i.e. Algorithm 3), as they would be hard to carry out while reusing simulations, and in any case the pilot output for real data is not particularly concentrated (see Figure 2 of the supplementary material). The same simulations were used for both the regression fitting and ABC stage. The simulation study compared (i) semi-automatic S as described above, giving $\dim(S)=7$ (ii) only the model choice components, giving $\dim(S)=3$ (iii) S as in the pilot analysis. A further simulation study using the method of Marin et al. (2013) is detailed in the supplementary material (Section 8).

6.3 Results

6.3.1 Simulation study

Table 4 shows confusion matrices for two choices of S . In both cases, model 1 (no growth) is well identified, but it is harder to differentiate between models 2 and 3. More detailed results, taking into account the magnitude of probability estimates, are shown in Table 5 which gives mean entropic loss (defined by Equation (6) above) results under all choices of S . One dataset is omitted as an outlier here; it was simulated under model 1 but under all analyses the estimated model 1 probability was 0.02 or less, and if not omitted Monte Carlo error in this estimate would dominate the results. The loss figures confirm a modest overall improvement for semi-automatic ABC which is maintained when parameter inference summaries are added. Other features of interest in Table 5 are that the pilot choice of S is slightly better at identifying model 2, and post-processing only improves results under the lowest dimensional choice of S .

A simulation study using the approach of Marin et al. (2013) (details in Section 8 of the supplementary material) suggested that both the pilot and the semi-automatic ABC summary statistics are suitable to detect differences between the models in the asymptotic case of highly informative data.

Table 4 Confusion matrices from ABC analyses of 100 simulated datasets from the *C. jejuni* application with two choices of S . These are contingency tables where the rows represents the true model, and the columns the model with the largest estimated probability.

| Pilot ($\dim(S)=15$) | Predicted | | | Semi-automatic ($\dim(S)=7$) | Predicted | | | | |
|---------------------------|-----------|----|----|-----------------------------------|-----------|----|---|----|----|
| | 1 | 2 | 3 | | 1 | 2 | 3 | | |
| Correct | 1 | 31 | 2 | 1 | 1 | 32 | 0 | 2 | |
| | 2 | 3 | 16 | 14 | Correct | 2 | 2 | 17 | 14 |
| | 3 | 1 | 12 | 20 | | 3 | 1 | 11 | 21 |

Table 5 Mean entropic losses from ABC analyses of 100 simulated datasets from the *C. jejuni* application, omitting one outlying dataset (see main text). The columns show results for test datasets simulated from particular models. Results following regression post-processing as in Beaumont (2008) are shown in brackets. The two semi-automatic ABC rows are for model choice only summaries ($\dim(S)=3$) and model choice plus parameter inference summaries ($\dim(S)=7$).

| Summary statistics | Model 1 | Model 2 | Model 3 | All models |
|----------------------------|---------------|---------------|---------------|---------------|
| Pilot | 0.408 (1.58) | 0.776 (0.857) | 0.725 (1.05) | 0.636 (1.16) |
| Semi-automatic (MC only) | 0.247 (0.192) | 0.816 (0.877) | 0.740 (0.690) | 0.601 (0.586) |
| Semi-automatic (MC and PI) | 0.329 (0.674) | 0.798 (0.794) | 0.686 (0.686) | 0.604 (0.718) |

6.3.2 Real data

Table 6 summarises the model choice results for the pilot and main analyses, including the effect of regression post-processing. They agree in putting the majority of the weight on model 1, the no growth model. Effective sample sizes (Liu, 1996) show that Monte Carlo error is approximately equal to that of a moderately large independent sample (n.b. the largest achievable ESS here is 1000, the number of particles). The supplementary material (Section 7) details sensitivity analyses which vary the parameter priors and the subsample of isolates used as observations. With the exception of some pilot analyses, the weight placed on model 1 remains in the range 80–100%. Also, the simulation results above confirm that the analyses can reliably distinguish the no growth model from the others.

Parameter inference results are less informative, and are detailed in the supplementary material (Section 3). The main finding is a low estimate of recombination rate. Also, the regression and ABC results were also used to find which genetic summaries were particularly informative, and to show that some aspects of the data fitted poorly under any model. These results are given in the supplementary material (Sections 4 and 5), and can inform future modelling and analyses.

6.4 Conclusions

Our main finding is support for a model with no change in the effective population size of *C. jejuni*. This is surprising over a period where its ecological niche has greatly increased. Analysis in the supplementary material (Section 4) shows some features of the data are poorly fitted under all models, suggesting that more detailed demographic structure is necessary to fit the data well. One potential modification is subpopulation structure amongst the hosts, which might reveal that only some support growing *C. jejuni* populations. Parameter inference findings are discussed in the supplementary material (Section 3).

Simulation results (Table 5) show that semi-automatic ABC choice of S gives a modest quantitative improvement over the pilot choice. We argue that our method also produces an improvement in confidence over an ad-hoc S . This is discussed in the next section.

Table 6 Estimated posterior probabilities and effective sample sizes from ABC analyses on *C. jejuni* data. Some rows do not total 1 due to rounding.

| Analysis | ESS | Post-processed | Model 1 | Model 2 | Model 3 |
|----------|-----|----------------|---------|---------|---------|
| Pilot | 348 | No | 0.86 | 0.11 | 0.04 |
| | | Yes | 1.00 | 0.00 | 0.00 |
| Main | 600 | No | 0.96 | 0.03 | 0.01 |
| | | Yes | 0.92 | 0.03 | 0.05 |

7 Discussion

This paper proposes methodology to generate summary statistics for ABC model choice. These estimate a low dimensional vector of summary statistics that we prove to be sufficient for model choice. The statistics are fitted using logistic regression on a set of simulated data. We argue for using simulated data from models truncated to the most likely parameter values, which necessitates model choice taking place between these truncated models. Another contribution is to provide a method to use these results to estimate the model choice results of interest. An important motivation for a regression based approach is that it scales well when there are a large number of potential summary statistics and searching all subsets of them is impractical and costly.

The methodology is applied to several simple examples from the literature and a substantive population genetic application, and in all cases improves performance compared to summary statistics previously used in the literature and other ad-hoc choices. The size of the improvement varies from modest to substantial, with results for some examples close to those of the true posterior. When the improvement is modest, as in the main application, we argue our method is still useful, as follows. Let S_0 denote ad-hoc summaries with which a first ABC analysis has been performed. There is potentially large approximation error, due to the curse of dimensionality if S_0 is high dimensional, or due to ignored information if S_0 is a low dimensional summary of high dimensional data. Verifying the result with our approach adds confidence that these issues are avoided, thus addressing the criticisms of ABC model choice outlined in Section 1.

The paper has concentrated on methodology for $M \leq 3$, following previous ABC literature. The method outlined in Section 4.1 for calculating summary statistics using logistic regression on each pair of models does generalise to $M > 3$. However this gives $\dim(S) = M(M-1)/2$, whereas Theorem 1 shows there are sufficient statistics of dimension $M-1$. Alternative regression methods can be used to give $\dim(S) = M-1$, for example estimating an appropriate subset of the Bayes factors or multinomial regression. For the case of $M=3$ we believe pairwise logistic regression offers some robustness: even if the logistic regression for one pair of models fits poorly, the others can still allow a good overall estimate of sufficient statistics. Further work is required to assess this and to investigate the case $M > 3$, where new methodological problems for ABC model choice may arise. We note that inference for discrete parameters is a natural application of ABC model choice where this case can easily occur.

It is often desirable to perform model choice and parameter inference using the same simulations. Our methodology focuses on producing S appropriate for model choice only. Section 6 contains an application-specific example of adding a small number of further summaries to S which are informative for parameter inference. General purpose methods to choose such low dimensional summaries would be useful. However, often each model may require separate summaries, so that a choice of S suitable for model choice and parameter inference would be high dimensional. An alternative strategy is to develop ABC methods in which comparisons of simulated and observed data do not always use the same summaries. A simple approach would be to perform separate rejection sampling analyses for model choice and for parameter inference under each model, and more efficient methods would be desirable.

There are numerous alternatives to logistic regression to fit summary statistics for model choice in our framework, such as linear discriminant analysis (Estoup et al., 2012), and a comparison of their performance, as well as the subset selection method of Barnes et al. (2012b), would be interesting. Other parts of our semi-automatic method could also be varied. For example, our choice of S is a vector of one-to-one transformations of Bayes factors, and other transformations may perform differently. Also, other methods could produce a more accurate training region, such as fitting a flexible model to the pilot output. There are similarities between our method and that of Nunes and Balding (2010), including a pilot analysis using rough summary statistics and searching for summaries which perform well locally rather than globally, and some hybrid of the two may be possible.

For simplicity we have used relatively simple ABC algorithms. However, much progress is being made in improving algorithmic efficiency, especially of ABC SMC (e.g., Drovandi and Pettitt, 2011; Del Moral et al., 2012). Our work is complementary to this and it could be used with many such improved algorithms. Indeed

ABC SMC algorithms can also be modified to incorporate semi-automatic ABC. For example, recall that in Section 5 the training data were reused as the simulations needed for ABC rejection sampling. As suggested by Barnes et al. (2012a), in ABC SMC they could be similarly reused for the first SMC iteration. Also, note one disadvantage of the ABC SMC algorithm we use is that the number of model simulations is random, making it difficult to compare results fairly. We are unaware of ABC SMC algorithms without this property, although that of Drovandi and Pettitt (2011) reduces the variability considerably.

Acknowledgements: The authors acknowledge the Marsden Fund project 08-MAU-099 (Cows, starlings and *Campylobacter* in New Zealand: unifying phylogeny, genealogy, and epidemiology to gain insight into pathogen evolution) for funding this project. This publication made use of the *Campylobacter* Multi Locus Sequence Typing website (<http://pubmlst.org/campylobacter/>) developed by Keith Jolley and sited at the University of Oxford (Jolley and Maiden 2010, BMC Bioinformatics, 11:595). The development of this site has been funded by the Wellcome Trust. The paper has benefited from many helpful suggestions of two anonymous reviewers.

References

- Atkinson, I. A. and E. K. Cameron (1993): “Human influence on the terrestrial biota and biotic communities of New Zealand,” *Trends in Ecology & Evolution*, 8, 447–451.
- Barnes, C. P., S. Filippi, M. P. H. Stumpf and T. Thorne (2012a): “Considerate approaches to constructing summary statistics for ABC model selection,” *Statistics and Computing*, 22, 1181–1197.
- Barnes, C. P., S. Filippi and M. P. H. Stumpf (2012b): “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, 453.
- Beaumont, M. A. (2008): “Joint determination of topology, divergence time, and immigration in population trees,” In: C. Renfrew, S. Matsumura, and P. Forster, editors, *Simulation, Genetics and Human Prehistory*. McDonald Institute Monographs, pp. 134–154.
- Beaumont, M. A., W. Zhang and D. J. Balding (2002): “Approximate Bayesian computation in population genetics,” *Genetics*, 162, 2025–2035.
- Blum, M. G. B. (2010): “Approximate Bayesian computation: a nonparametric perspective,” *Journal of the American Statistical Association*, 105 (491), 1178–1187.
- Blum, M. G. B. and O. François (2010): “Non-linear regression models for approximate Bayesian computation,” *Statistics and Computing*, 20, 63–73.
- Blum, M. G. B., M. A. Nunes, D. Prangle and S. A. Sisson (2013): “A comparative review of dimension reduction methods in approximate Bayesian computation,” *Statistical Science*, 28, 189–208.
- Del Moral, P., A. Doucet and A. Jasra (2012): “An adaptive sequential Monte Carlo method for approximate Bayesian computation,” *Statistics and Computing*, 22 (5), 1009–1020.
- Didelot, X., R. G. Everitt, A. M. Johansen and D. J. Lawson (2011): “Likelihood-free estimation of model evidence,” *Bayesian Analysis* 6 (1), 49–76.
- Dingle, K. E., F. M. Colles, D. R. A. Wareing, M. C. J. Maiden, M. C. J. Ure, R. Maiden, A. J. Fox, F. E. Bolton, H. J. Bootsma, R. J. Willems, R. Urwin and M. C. Maiden (2001): “Multilocus sequence typing system for *Campylobacter jejuni*,” *Journal of Clinical Microbiology*, 39, 14–23.
- Drovandi, C. C. and A. N. Pettitt (2011): “Estimation of parameters for macroparasite population evolution using approximate Bayesian computation,” *Biometrics*, 67 (1), 225–233.
- Estoup, A., E. Lombaert, J.-M. Marin, T. Guillemaud, P. Pudlo, C. P. Robert and J. Cornuet (2012): “Estimation of demo-genetic model probabilities with approximate Bayesian computation using linear discriminant analysis on summary statistics,” *Molecular Ecology Resources*, 12 (5), 846–855.
- Fan, Y., D. J. Nott and S. A. Sisson (2013): Approximate Bayesian computation via regression density estimation. *Stat*, 2, 34–48.
- Fearnhead, P. and D. Prangle (2012): “Constructing summary statistics for approximate Bayesian computation: semi-automatic ABC,” *Journal of the Royal Statistical Society, Series B*, 74, 419–474.
- French, N., S. Yu, P. Biggs, B. Holland, P. Fearnhead, B. Binney, A. Fox, D. H. Grove-White, J. Leigh, W. Miller, P. Muellner and P. Carter (2014): “Evolution of *Campylobacter* species in New Zealand,” In S. Sheppard and G. Méric, editors, *Campylobacter Ecology and Evolution*. Caister Academic Press, Norfolk.
- Friedman, J., T. Hastie, and R. Tibshirani (2010): “Regularization paths for generalized linear models via coordinate descent,” *Journal of Statistical Software*, 33 (1).

- Grelaud, A., C. Robert, J.-M. Marin, F. Rodolphe and J. F. Taly (2009): "ABC likelihood-free methods for model choice in Gibbs random fields," *Bayesian Analysis*, 4 (2), 317–336.
- Hudson, R. R. (2002): "[Generating samples under a Wright-Fisher neutral model of genetic variation](#)," *Bioinformatics*, 18, 337–338.
- Humphrey, T., S. O'Brien and M. Madsen (2007): "Campylobacters as zoonotic pathogens: a food production perspective," *International Journal of Food Microbiology*, 117 (3), 237–57.
- Joyce, P. and P. Marjoram (2008): "Approximately sufficient statistics and Bayesian computation," *Statistical Applications in Genetics and Molecular Biology*, 7, 2008. Article 26.
- Kolmogorov, A. N. (1942): "Determination of centre of dispersion and measure of accuracy from a finite number of observations (in Russian)," *Izv. Akad. Nauk, USSR Ser. Mat.*, 6, 3–32.
- Liu, J. S. (1996): "Metropolisized independent sampling with comparisons to rejection sampling and importance sampling," *Statistics and Computing*, 6, 113–119.
- Marin, J.-M., N. Pillai, C. P. Robert and J. Rousseau (2013): "Relevant statistics for Bayesian model choice," Preprint. Available at <http://www.arxiv.org/abs/1110.4700>.
- Mullner, P., S. E. F. Spencer, D. J. Wilson, G. Jones, A. D. Noble, A. C. Midwinter, J. M. Collins-Emerson, P. Carter, S. Hathaway and N. P. French (2009): "Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach," *Infection, Genetics and Evolution* 9 (6), 1311–1319.
- Nordborg, M. (2004): "Coalescent theory," In: D.J. Balding, M. Bishop, C. Cannings (Eds.). *Handbook of statistical genetics*, Wiley-Interscience, volume 2, New York.
- Nunes, M. A. and D. J. Balding (2010): "On optimal selection of summary statistics for approximate Bayesian computation," *Statistical Applications in Genetics and Molecular Biology*, 9 (1), 2010.
- Rambaut, A. and N. C. Grassly (1997): "Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees," *Computer Applications in the Biosciences*, 13, 235–238.
- Rayner, G. D. and H. L. MacGillivray (2002): "Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions," *Statistics and Computing*, 12 (1), 57–75.
- Robert, C. P. (1996): "Intrinsic losses," *Theory and decision*, 40 (2), 191–214.
- Robert, C. P. (2014): Bayesian computational tools. *Annual Review of Statistics and Its Application*, 1, 16.1–16.25.
- Robert, C. P., J. M. Cornuet, J.-M. Marin and N. Pillai (2011): "Lack of confidence in approximate Bayesian computation model choice," *Proceedings of the National Academy of Sciences*, 108 (37), 15112–15117.
- Savill, M., A. Hudson, M. Devane, N. Garrett, B. Gilpin and A. Ball (2003): "Elucidation of potential transmission routes of *Campylobacter* in New Zealand," *Water Science and Technology*, 47 (3), 31–38.
- Sears, A., M. G. Baker, N. Wilson, J. Marshall, P. Muellner, D. M. Campbell, R. J. Lake and N. P. French (2011): "Marked campylobacteriosis decline after interventions aimed at poultry, New Zealand," *Emerging Infectious Diseases*, 17 (6), 1007–1015.
- Sjödin, P., A. E. Sjöstrand, M. Jakobsson and M. G. B. Blum (2012): "Resequencing data provide no evidence for a human bottleneck in Africa during the penultimate glacial period," *Molecular Biology and Evolution*, 29 (7), 1851–1860.
- Sousa, V. C., M. A. Beaumont, P. Fernandes, M. M. Coelho and L. Chikhi (2012): "Population divergence with or without admixture: selecting models using an ABC approach," *Heredity*, 108, 521–530.
- Toni, T. and M. P. H. Stumpf (2010): "Simulation-based model selection for dynamical systems in systems and population biology," *Bioinformatics*, 26 (1), 104–110.
- Wilson, D. J., E. Gabriel, A. J. H. Leatherbarrow, J. Cheesbrough, S. Gee, E. Bolton, A. Fox, C. A. Hart, P. J. Diggle and P. Fearnhead (2009): "Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*," *Molecular Biology and Evolution*, 26 (2), 385–397.
- Wiuf C. and J. Hein (2000): "The coalescent with gene conversion," *Genetics*, 155, 451–462.
- Yu, S., P. Fearnhead, B. R. Holland, P. Biggs, M. Maiden and N. P. French (2012): "Estimating the relative roles of recombination and point mutation in the generation of single locus variants in *Campylobacter jejuni* and *Campylobacter coli*," *Journal of Molecular Evolution*, 74 (5–6), 273–280.