Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 26

Approximately Sufficient Statistics and Bayesian Computation

Paul Joyce, *University of Idaho*
Paul Marjoram, *USC*

Recommended Citation:
Joyce, Paul and Marjoram, Paul (2008) "Approximately Sufficient Statistics and Bayesian Computation," *Statistical Applications in Genetics and Molecular Biology*: Vol. 7: Iss. 1, Article 26.
DOI: 10.2202/1544-6115.1389

Approximately Sufficient Statistics and Bayesian Computation

Paul Joyce and Paul Marjoram

Abstract

The analysis of high-dimensional data sets is often forced to rely upon well-chosen summary statistics. A systematic approach to choosing such statistics, which is based upon a sound theoretical framework, is currently lacking. In this paper we develop a sequential scheme for scoring statistics according to whether their inclusion in the analysis will substantially improve the quality of inference. Our method can be applied to high-dimensional data sets for which exact likelihood equations are not possible. We illustrate the potential of our approach with a series of examples drawn from genetics. In summary, in a context in which well-chosen summary statistics are of high importance, we attempt to put the `well' into `chosen.'

Author Notes: Joyce was supported by grants from the NIH (P20 RR16448, NIH R01 GM076040-01) and the NSF (NSF-DEB-0515738); Marjoram was funded by NIH grants GM069890. We would like to thank Simon Tavare, Peter Calabrese and the reviewers for helpful comments.

1 Introduction

We are in the midst of an era in which the size of data sets is growing at an increasingly rapid pace. While more data is always, in principle, helpful, there are several problems associated with the increasingly high dimensionality of data. Motivated by the growing number of such data sets appearing in genetics and genomics, in this paper we focus on issues that arise with these large data sets when model-based analysis methods are used.

In general, we collect data $\mathcal{D}^*$ and wish to make inference about a parameter, or set of parameters θ. The $*$ notation is used to distinguish between the given data set and the simulated data sets that we will exploit below. If we do not use the $*$ superscript we are referring to simulated data. A variety of methods exist for inference in this context, such as rejection algorithms (Ripley, 1982), Markov chain Monte Carlo [MCMC] methods (*e.g.*, the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970)), and Importance Sampling (Ripley, 1982). When taking a Bayesian perspective, inference regarding θ typically proceeds via calculation of the posterior distribution $P(\theta \mid \mathcal{D}^*) = P(\mathcal{D}^* \mid \theta)P(\theta)/P(\mathcal{D}^*)$. In some contexts calculation of the term $P(\mathcal{D}^* \mid \theta)$ is problematic, either because the sheer size of the data makes the calculation computationally intractable, or because calculation is impossible when using realistic models for how the data arise. These problems have motivated a drive to more approximate methods, in particular the field of approximate Bayesian computation [ABC] (*e.g.*, Beaumont et al., 2002; Marjoram et al., 2003; Sisson et al., 2007). For a general, low-level discussion of this progression in a biological context see Marjoram and Tavaré (2006). In this paper we focus on the use of rejection algorithms.

1.1 Rejection Algorithms

Assuming the existence of a model M to explain the generation of the data, and a prior $\pi(\cdot)$ for the parameter(s) θ, the aim of a rejection method is to produce observations from the posterior distribution $f(\theta \mid \mathcal{D}^*)$. In its simplest form, when calculation of $f(\mathcal{D}^* \mid \theta)$ is intractable, the algorithm takes the following form:

R1 Sample θ from $\pi(\cdot)$.

R2 Simulate data $\mathcal{D}$ from the model M using parameters θ.

R3 Accept θ if $\mathcal{D} = \mathcal{D}^*$, and go to **R1**.

The accepted observations have the required posterior distribution.

However, for high-dimensional data the probability of observing $\mathcal{D} = \mathcal{D}^*$ is often extremely small. This drove a move towards methods that directly approximate

the likelihood of the full data $\mathcal{D}^*$. As the complexity of data sets has continued to grow these methods have themselves become intractable, forcing one to consider summaries of the data. The choice of which summaries to use is frequently not obvious, motivating a need for methods to help determine which summary statistics carry information useful for inference. It is this problem we focus on here. These issues are related to the concept of "nearly sufficient" statistics (Le Cam, 1964; Abril, 1994; Cabrera and Yohai, 1999).

Thus, supposing the existence of a set of summary statistics $\mathcal{S} = \{S_1, \ldots, S_n\}$, and letting $\mathcal{S}^*$ denote the value taken by these statistics on the observed data $\mathcal{D}^*$ while $\mathcal{S}$ denotes the value taken on simulated data $\mathcal{D}$, [R3] in the above algorithm is replaced with the following:

R3' Accept θ if $\mathcal{S} = \mathcal{S}^*$, and go to **R1**.

The accepted observations now represent independent samples from $f(\theta \,|\, \mathcal{S} = \mathcal{S}^*)$. In situations in which a sufficient statistic exists (and is contained in $\mathcal{S}$) this distribution is equal to $f(\theta \mid \mathcal{D}^*)$. However, it will often be the case that a sufficient statistic does not exist, in which case the resulting distribution represents an approximation to $f(\theta \mid \mathcal{D}^*)$. Note that the closeness of the approximation is in general unknown and depends upon the choice of statistics in $\mathcal{S}$. A further complexity is that, even when using summary statistics, the probability of observing $\mathcal{S} = \mathcal{S}^*$ often remains small. In such a setting, it is common to define a distance metric $d(S, S^*)$ to measure the distance between S and S^* (with low distances corresponding to S being relatively similar to S^*). Step **R3'** is then replaced with:

R3" Accept θ if $d(\mathcal{S}, \mathcal{S}^*) < E$, and go to **R1**.

where E is an arbitrary, small, constant. The accepted observations now represent independent samples from $f(\theta \mid d(\mathcal{S}, \mathcal{S}^*) \lessdot E)$, but again the degree of approximation between this distribution and $f(\theta \mid \mathcal{D}^*)$ is, in general, unknown. Despite these issues, these algorithms have been widely used (*e.g.*, Plagnol and Tavaré, 2004; Tavaré et al., 1997; Fu and Li, 1997; Innan et al., 2005). There is, however, a pressing need for theory to guide the choice of summary statistics.

Of course, *in principle* maximum information is gained by using all summary statistics, since, in the worst case scenario that a statistic adds no information regarding θ, including that statistic has no effect on the posterior for θ. However, that observation is a theoretical observation (in some sense, corresponding to the use of an infinite number of iterations in the rejection algorithm). In practice, the addition of a new statistic to those considered within a rejection method will result in fewer iterations being accepted, which leads to the addition of stochastic noise to the empirical estimator of the posterior distribution that is calculated from the set of

DOI: 10.2202/1544-6115.1389

accepted θs (unless the statistic is completely correlated with an already included statistic - in which case it adds no information anyway). Thus, there is a trade-off between the information added by a new statistic and the corresponding additional stochastic noise that is caused by the fact that fewer iterations will now be accepted. This means that a sensible strategy will be to only include an additional statistic if it alters the posterior distribution in a meaningful way.

Thus, the purpose of this paper is to develop a conceptual framework with which one can begin to assess the information content of a list of summary statistics. The goal is to develop a procedure to score summary statistics in such a way that the score provides a natural assessment of the utility of the summary, with the requirement that the score is computable for problems where exact likelihoods calculations are not possible.

2 Theoretical Results

We focus on the following situation. Suppose we have a list of summary statistics $S_1, S_2, \cdots, S_{k-1}$ and a candidate summary S_k. The question we wish to address is the following. If we add the statistic S_k to our existing list, will it substantially improve the quality of inference, or is the added information associated with S_k so low as to suggest that it can safely be ignored? We begin by considering the log-likelihood of the summary statistics:

$$\begin{aligned} \ln P(S_1, S_2, \cdots, S_k|\theta) &= \ln P(S_1|\theta) + \ln P(S_2|S_1, \theta) + \\ &\quad \cdots + \ln P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta). \end{aligned}$$

Note that $\ln P(S_1, S_2, \cdots, S_k|\theta)$ differs from $\ln P(S_1, S_2, \cdots, S_{k-1}|\theta)$ only by the term

$$\ln P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta).$$

If $S_1, \cdots, S_{k-1}$ were sufficient then $\ln P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)$ would not depend on θ and thus would not contribute any information to any inference of θ. The term would drop out of the log-likelihoood calculation and under the Bayesian perspective $P(\theta|S_1, S_2 \cdots, S_k)$ would equal $P(\theta|S_1, S_2 \cdots, S_{k-1})$ because $P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)$ would equal $P(S_k|S_1, S_2, \cdots, S_{k-1})$ and would appear in both the numerator and denominator of the Bayes calculation and would cancel. Motivated by these properties of sufficient statistics we will define a concept of approximately sufficient.

Definition A set of statistics $S_1, S_2, \cdots, S_{k-1}$ are $\epsilon-$sufficient relative to a statistic X if

$$\sup_{\theta} \ln P(X|S_1, S_2, \cdots, S_{k-1}, \theta) - \inf_{\theta} \ln P(X|S_1, S_2, \cdots, S_{k-1}, \theta) \leq \epsilon$$

Note that if X represents the entire data and ϵ is zero then the above reduces to the definition of sufficiency. We are interested in the case where $X = S_k$. We now wish to consider sequentially adding new summary statistics to a list of summaries and scoring each new summary as follows.

Definition The score of S_k relative to $S_1, S_2, \cdots, S_{k-1}$ is defined as follows.

$$\delta_k = \sup_{\theta} \ln P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta) - \inf_{\theta} \ln P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta). \quad (1)$$

Once the score drops below a certain threshold we will stop adding new statistics. The result below shows how the score of the proposed summary is related to the posterior distributions of interest.

Result 1 If δ_k is the score of a statistic S_k relative to $S_1, S_2, \cdots, S_{k-1}$, and $\pi(\theta)$ is the prior distribution on θ, and the odds-ratio $R_k(\theta)$ is defined to be

$$R_k(\theta) = \frac{P(\theta|S_1, S_2, \cdots, S_k)}{P(\theta|S_1, S_2, \cdots, S_{k-1})} \quad (2)$$

then

$$e^{-\delta_k} \leq R_k(\theta) \leq e^{\delta_k}$$

Result 1 follows from the fact that

$$P(\theta|S_1, S_2, \cdots, S_k) = \frac{P(S_1, S_2, \cdots, S_{k-1}|\theta)P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)\pi(\theta)}{\int P(S_1, S_2, \cdots, S_{k-1}|\theta)P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)\pi(\theta)d\theta}$$

$$\leq \frac{P(S_1, S_2, \cdots, S_{k-1}|\theta)\pi(\theta)}{\int P(S_1, S_2, \cdots, S_{k-1}|\theta)\pi(\theta)d\theta} \frac{\sup_\theta \left(P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)\right)}{\inf_\theta \left(P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)\right)}$$

$$= P(\theta|S_1, S_2, \cdots, S_{k-1}) \frac{\sup_\theta \left(P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)\right)}{\inf_\theta \left(P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)\right)}$$

$$= P(\theta|S_1, S_2, \cdots, S_{k-1})e^{\delta_k},$$

DOI: 10.2202/1544-6115.1389

and similarly

$$P(\theta|S_1, S_2, \cdots, S_k) \geq P(\theta|S_1, S_2, \cdots, S_{k-1}) \frac{\inf_\theta \left(P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)\right)}{\sup_\theta \left(P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)\right)}$$

$$= P(\theta|S_1, S_2, \cdots, S_{k-1}) e^{-\delta_k} \blacksquare$$

Note that if we choose our threshold score small enough then $e^{\delta_k} \approx e^{-\delta_k} \approx 1$ and there is little difference between the posterior distribution for θ given $S_1, S_2, \cdots, S_k$ and the posterior for θ given $S_1, S_2, \cdots, S_{k-1}$. So the odds-ratio $R_k(\theta)$ defined by (2) is close to 1. In fact, the next result shows that the scoring function can be defined in terms of the odds-ratio $R_k(\theta)$.

Result 2 Let δ_k be the score of a statistic S_k relative to $S_1, S_2, \cdots, S_{k-1}$ defined by (1) and let $R_k(\theta)$ be the odds-ratio defined by (2) then

$$e^{\delta_k} = \frac{\sup_\theta R_k(\theta)}{\inf_\theta R_k(\theta)} \tag{3}$$

Result 2 follows by first noting that

$$\begin{aligned} P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta) = & \frac{P(\theta, S_k|S_1, \cdots, S_{k-1})}{P(\theta|S_1, \cdots, S_{k-1})} \\ & = \frac{P(\theta|S_1, \cdots, S_k) P(S_k|S_1, \cdots, S_{k-1})}{P(\theta|S_1, \cdots, S_{k-1})} \qquad (4) \\ & = R_k(\theta) P(S_k|S_1, \cdots, S_{k-1}). \end{aligned}$$

It therefore follows that

$$e^{\delta_k} = \frac{\sup_\theta P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)}{\inf_\theta P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)} = \frac{\sup_\theta R_k(\theta)}{\inf_\theta R_k(\theta)} \blacksquare \tag{5}$$

Remark Note that $P(S_k|S_1, \cdots, S_{k-1})$ depends on the prior distribution, since an integral involving the prior is required to calculate this quantity. Therefore, while $P(S_k|S_1, S_2, \cdots, S_{k-1}, \theta)$ does not depend on any prior, the odds ratio $R_k(\theta)$ will be influenced by the prior, even though equation (5) shows that $\frac{\sup_\theta R_k(\theta)}{\inf_\theta R_k(\theta)}$ will not depend on the prior.

An ABC algorithm for approximating δ_k

We now consider a fixed data set D^* and a set of summary statistics that take observed value $S_1^*, S_2^* \cdots, S_k^*$. The algorithm for scoring a summary statistic S_k^* given $S_1^*, S_2^*, \cdots, S_{k-1}^*$ is quite simple. First generate a large enough number of data sets so that one can reasonably approximate the posterior $P(\theta|S_1^*, S_2^*, \cdots, S_k^*)$ using the standard ABC rejection algorithm. Note that this single simulation can be used to approximate all of the posteriors for θ given any subset of the k summary statistics under consideration. Since the algorithm will produce a finite set of accepted θ's, we can estimate δ_k by

$$\delta_k = \max_j \ln R_k(\theta_j) - \min_j \ln R_k(\theta_j) = \max_{j,l} |\ln R_k(\theta_j) - \ln R_k(\theta_l)|$$

Since

$$\hat{\delta}_k \leq 2 \max_i |\ln R_k^*(\theta_i)|$$

then if the estimate of the score $\hat{\delta}_k$ exceeds some threshold then the absolute value of the log of the odds-ratio $|\ln R_k^*(\theta_i)|$ will exceed half that threshold. Therefore we can base our decision whether or not to accept a statistic S_k based on whether or not the score departs significantly from the null expectation ($\delta_k = 0$) or the absolute odds-ratio $|R_k^*(\cdot)|$ departs significantly from 1.

3 Examples

In this section we give several example applications of the above ideas. The general schema is to generate a set of 100 data sets (the 'observed' data) and then use our algorithm to decide which of a family of test statistics should be used for inference on each of those data sets. In each case, we then assess the accuracy of the final choice of statistics by calculating the error e_i to be the difference between the mean of the posterior distribution for θ for data set i and the value of θ that was used to generate that data. We report the mean of e_i^2 over the 100 data sets.

We simulate a set, $\mathbf{D}$, of 5 million data sets which we will use in order to choose which statistics should be used for each observed data set. In principle, for each observed data set, we would like to proceed by adding randomly chosen statistics, one-by-one, and determining whether δ_k, as defined in (1), exceeds some threshold T after each addition. In practise, we actually use a conceptually equivalent statement derived from equation (5), and, after the addition of each new statistic, determine whether the ratio of posteriors

$$R_k^*(\theta) = \frac{P(\theta|S_1^*, S_2^*, \cdots, S_{k-1}^*, S_k^*,)}{P(\theta|S_1^*, S_2^*, \cdots, S_{k-1}^*)} \tag{6}$$

DOI: 10.2202/1544-6115.1389

differs from 1 by more than some threshold value $T(\theta)$ for any value of θ. We defer details of this to the appendix.

Since we choose to attempt to add a randomly chosen statistic at each iteration of the algorithm it is entirely possible that a more informative statistic might be added after a statistic which is less informative has already been included. This would result in a final set of statistics that was non-optimal. To help avoid this we implemented an additional step in which, after the addition of any statistic, we attempt to remove each of the other already accepted statistics. In this context, suppose we are currently attempting to remove statistic S_i from the set of statistics S_A that are currently included. Conceptually speaking, we proceed as if the set of currently included statistics were $S_A \backslash S_i$, and try to add statistic S_i. If S_i is not added in this scenario, we drop it from the set S_A.

We now give three example applications.

3.1 Example 1: The Ewens Sampling Formula

We begin with a "proof of principle" example using the Ewens Sampling formula [ESF]. This formula was introduced in (Ewens, 1972) to describe the distribution of allelic types in a genetic sample drawn from the so-called infinite sites model (in which every new mutation results in a unique, new type) under a number of standard assumptions, such as neutrality. See (Ewens, 1972) for a more complete discussion. The ESF has the appealing property that the number of types is a sufficient statistic for the mutation rate (see, for example, Joyce, 1998). Thus we use it as an elementary example application of our algorithm, in which the correct answer is known: if the algorithm performs well it will select the number of types as the only statistic to use when estimating the mutation rate.

In each application of this example we compare results from our algorithm to those obtained from a rejection method estimator that uses only the number of types, N_T, in the sample. Since N_T is sufficient for the mutation rate θ, this latter estimator represents the optimum performance that could be attained.

We simulate data sets of size 50 and attempt to estimate the mutation rate. For each data set the prior for the mutation parameter θ is assumed to be uniformly distributed on [0,10]. We then simulate $N = 5000000$ data sets to use when deciding whether to add statistics. For convenience, when calculating posterior distributions for θ we discretize the range of θ using 10, equally-spaced bins. We begin by presenting a case in which we consider two statistics:
S_1: the number of types in the sample, N_T;
S_2: p, where p is a random number that is uniformly distributed on [0,25].
The range of p is chosen so that both statistics have roughly the same degree of variation between data sets. Obviously, S_2 is designed to be completely uninformative.

Table 1: Example application: the ESF using statistics S_1 and S_2. The table shows the frequency with which statistics S_1 and S_2 are chosen to be used in this example, along with the mean square error of the resulting estimator over the 100 test data sets.

Statistic		Error	
S_1	S_2	baseline	algorithm
100	0	2.19	2.19

In Table 1, we show the results, presented as counts of the number of times each of the statistics was chosen to be used. We see that S_1 was always chosen to be used, while S_2 was never chosen. (We note that it is entirely possible for application of our algorithm to result in no statistics being chosen.) These results are as one would hope, since the number of types is sufficient for the mutation parameter in this model. Of course, stochastic noise will lead to non-optimal choices from time-to-time (but this does not occur in this example). The frequency of such occurrences can be reduced by increasing L, the number of data sets used to determine which statistics should be chosen.

We now consider the same example again, but this time we use S_1 and S_3, where S_3 is defined as $50H$, and H is the homozygosity of the sample. The factor of 50 is chosen so that the two statistics have comparable variances (see Discussion). Results are shown in Table 2. Here, it is the case that both the number of types and homozygosity contain signal regarding θ (Ewens, 1972). However, since the number of types is sufficient for θ it should generally be the only statistic chosen. This is indeed the case, there being only two exceptions. Despite these exceptions the mean square error of the estimators derived from our algorithm matches that resulting from the baseline algorithm which uses only S_1, indicating that on the two data sets on which S_3 was chosen the resulting estimator performed as well as one constructed from the sufficient statistic.

Finally, we consider an example in which we introduce two further statistics:
S_4=25*frequency of the commonest type (as a proportion);
S_5=Number of singleton types (i.e. types that have but one representative in the sample).
These two statistics both carry some information regarding θ (albeit less information than is carried by S_1). We allow the algorithm to consider any of $S_1, \ldots, S_5$. Results are shown in Table 3. Again the mean square error from our algorithm is essentially identical to that resulting from the algorithm that uses only the sufficient statistic. However, given the greater range of informative statistics, the algorithm

DOI: 10.2202/1544-6115.1389

Table 2: Example application: the ESF using statistics S_1 and S_3. The table shows the frequency with which statistics S_1 and S_3 are chosen to be used in this example, along with the mean square error of the resulting estimator over the 100 test data sets.

Statistic		Error	
S_1	S_3	baseline	algorithm
98	2	2.19	2.19

Table 3: Example application: the ESF using statistics S_1 through S_5. The table shows the frequency with which each statistic is chosen to be used in this example, along with the mean square error of the resulting estimator over the 100 test data sets.

Statistic					Error	
S_1	S_2	S_3	S_4	S_5	baseline	algorithm
91	1	5	4	6	2.19	2.19

sometimes includes other statistics in addition to, or in place of, S_1. We note that though the algorithm performs well, it would not be obvious from these results that S_1 was a sufficient statistic (although the results clearly indicate that this statistic is by far the most informative with respect to mutation rate).

3.2 Example 2: Coalescent Simulation - Estimation of Mutation Rate

We now move onto an example in which there is no sufficient statistic, but where there is a statistic that is known to be nearly sufficient: estimation of mutation rate in coalescent simulation. The coalescent, introduced by (Kingman, 1982c,a,b), is a widely-used model for the evolution of genetic material. For a general overview of the coalescent see (*e.g.,* Hudson, 1990; Nordborg, 2001). In this example we simulate samples of 50 haplotypes under the coalescent using the infinite sites model (where each mutation occurs at a unique position).

We begin with a simple simulation in which there is no recombination and where we attempt to estimate the scaled mutation rate (generally denoted by θ, and which varies continuously from 0 to 10 in data sets simulated for this example). Initially we consider just two statistics:

Table 4: Example application: Estimation of mutation rate in the coalescent, without recombination, using statistics C_1 and C_2. The table shows the frequency with which each statistic is chosen to be used in this example, along with the mean square error of the resulting estimator over the 100 test data sets.

Statistic		Error	
C_1	C_2	baseline	algorithm
100	2	1.77	1.77

C_1: the number of mutations in the data;
C_2: p, where p is a random number that is uniformly distributed on [0,25].
Here, C_1 is a highly informative, but not a sufficient statistic for θ. Consequently, in order to assess performance, in each application in this example we compare results from our algorithm to those obtained from a baseline rejection method estimator that uses only C_1. Results of this analysis are shown in Table 4. Again, we see that in this simple scenario the algorithm successfully chooses only the informative statistic in almost all cases.

We now make the example more interesting by adding five more statistics to the mix:
C_3: The mean number of pairwise differences between haplotypes;
C_4: 25*(The mean pairwise LD across all pairs of loci that are within a distance of 0.1 of each other in the sample - for convenience we re-scale the length of the region being simulated to be 1 unit.);
C_5: The number of haplotypes;
C_6: The frequency of the commonest haplotype (as an integer);
C_7: The number of singleton haplotypes.
Here, the number of pairwise differences between two haplotypes is simply the number of mutations that are present in one, but not both, of the haplotypes; and pairwise LD is measured as r^2.

Results of an analysis in which the algorithm is now able to choose between all seven statistics are presented in Table 5. We see that the algorithm chooses a combination of statistics that varies from one data set to another, but that the resulting estimate has a lower mean square error than that obtained from a rejection method that uses just the number of mutations.

DOI: 10.2202/1544-6115.1389

Table 5: Example application: Estimation of mutation rate in the coalescent, without recombination, using statistics C_1 through C_7. The table shows the frequency with which each statistic is chosen to be used in this example, along with the mean square error of the resulting estimator over the 100 test data sets.

Statistic							Error	
C_1	C_2	C_3	C_4	C_5	C_6	C_7	baseline	algorithm
75	4	27	56	43	18	16	1.77	1.59

Table 6: Example application: Estimation of recombination rate in the coalescent, using statistics C_1 through C_7. The table shows the frequency with which each statistic is chosen to be used in this example, along with the mean square error of the resulting estimator over the 100 test data sets.

Statistic							Error	
C_1	C_2	C_3	C_4	C_5	C_6	C_7	baseline	algorithm
73	2	52	35	78	11	16	7.41	6.96

3.3 Example 3: Coalescent Simulation - Estimation of Recombination Rate

We conclude with an example of a more complex situation: estimation of recombination rate in a coalescent setting. Here there is no known sufficient, or nearly sufficient statistic. However, many of the statistics $(C_1, \ldots, C_7)$ above are known to be informative regarding recombination rate (see, *e.g.*, Innan et al., 2005). We proceed as for example 2, but now set the mutation parameter θ equal to 5 for all simulations while allowing the scaled recombination rate parameter (commonly denoted by ρ) to be sampled uniformly at random from the interval $[0, 10]$. In order to assess performance, we compare the estimator constructed using our algorithm to a baseline estimate resulting from a rejection method that uses C_4, which is somewhat informative for ρ. Results are shown in Table 6. We see that the algorithm appears to choose statistics that we would expect to be informative regarding ρ but, once again, the exact set of statistics that is used varies across data sets. However, overall, the set of statistics that is chosen represents an improvement over the baseline estimator.

4 Discussion

This paper presents a first step towards solving a difficult but increasingly common problem: how to choose summary statistics in an ABC application. We presented an algorithm that, given a particular data set to analyze, will choose a set of summary statistics to use based upon the effect of inclusion of those statistics on an empirically calculated posterior distribution. The algorithm performs well and is able to distinguish useful statistics from noise. However, a number of issues remain to be explored. We discuss some of these below.

One potential drawback to our approach follows from the observation that the order in which you add the statistics will matter. There is no way of knowing *a priori* which statistics hold the most information. You would like to be able to add statistics in decreasing order of information. Since this is likely to be unknown we might like to try all possible subsets of the statistics, but this is likely to be computationally intractable in most settings. Thus, in this paper we have implemented a scheme in which we attempt to add a randomly chosen statistic at each step, and then, if the statistic is added to the set of statistics used by the rejection method, we subsequently attempt to drop each other statistic. We admit this is an imperfect solution to the problem, but it is practical to implement and appears to work reasonably well. For example, when we took one of the coalescent data sets contained within the results presented in Table 5 and analyzed it 100 times, with the order in which we attempted to introduce statistics randomized for each analysis, we discovered that the algorithm always chose to use statistic C_1, the number of mutations. In each case, one other statistic was also chosen to be used, but the identity of that statistic varied across analyses. If one wished to improve this behavior, one might explore generalizations of our approach in which we attempt to add or remove a randomly chosen set of statistics at each iteration, or use a larger set of test data in order to reduce the level of stochastic noise (*i.e.* increase the value of N).

It is relevant to note that our algorithm decides which statistics should be included in a rejection method, but it does not address the issue of how each chosen statistic should be weighted. When using exact rejection this will not matter, since simulated data sets must exactly match the observed data for the chosen statistics. However, in many real applications, exact matching is either impossible (because of statistics that vary on a continuous scale), or computationally intractable (because of the extremely large number of simulations that will typically be needed in order to produce each exact match).

We illustrate the effect of varying the weight of a statistic using the scenario of Example 3. In Table 7 we show results for three analyses in which the weight of statistic C_4 is varied. We see that if the weight of C_4 is low the statistic is not used, but if the weight is increased the statistic becomes more informative. This

DOI: 10.2202/1544-6115.1389

Table 7: Frequency with which statistics C_1 through C_7 are chosen to be used in the coalescent example when estimating recombination rate and allowing the weight of C_4 to vary.

Weight of C_4	Statistic							Error
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	
1	83	1	60	0	85	9	9	7.21
10	77	0	51	27	81	15	15	6.98
25	73	2	52	35	78	11	16	6.96

demonstrates that the weight that is placed on a statistic (in the form of a constant by which the statistic is multiplied) can have a significant impact upon the information added by including the statistic. It is important to note that this is only an issue when approximate methods are being used (as opposed to a situation in which exact matching between S and S^* can be insisted upon), but this is likely to be the case in many real applications. It is straightforward to imagine a generalization of the approaches we give here, in which as well as considering the addition/removal of statistics we also consider altering the weight placed on each statistic. The algorithm would now proceed by considering a set of weights $W_1, \ldots, W_n$, with $W_i = 0$ corresponding to a statistic not being included in the estimator; with weights being altered from iteration to iteration. The overall logic of the approach would be the same, but the details would be somewhat more complex. We propose to investigate the feasibility of such an approach in future work. We also note the development of alternative methods for estimating the optimum weights for a set of statistics, such as the use of projection-pursuit methods (Peter Calabrese - personal communication).

Finally, we note that our scoring scheme does not, in principle, require that one use the rejection method. Ultimately, the score function depends only on the odds-ratio, which does not require the calculation of the constant of integration of the posterior, so it is also amenable to MCMC and other computationally intensive methods. We speculate that convergence of an MCMC algorithm would be much quicker if one only had to propose moves in 'summary statistics' space rather than the full high dimensional space of most computationally intensive problems. So the same trade-off between computational efficiency and loss of information that applies to ABC techniques would still apply.

5 Appendix

Here we give more details of the implementation of our algorithm. In actual implementations of such an algorithm we will, for practical purposes, need to discretize the set of possible θ values into a set, $\{\theta_1, \ldots, \theta_L\}$ say, in order to construct an empirical estimate of the posterior distributions used in (5). (It would be possible to implement a versions that instead applied a density estimation procedure, but this would make substantially greater computational demands than the method we describe here.) What results will be an estimate of the posterior, subject to stochastic noise; the degree of noise being a function of the number of accepted data sets before, and after, the addition of the new statistic. This leads to a stochastic estimate of the odds-ratio in (5). The intuition here is that if the observed value of

$$R_k^*(\theta_i) = \frac{P(\theta_i|S_1^*, S_2^*, \cdots, S_{k-1}^*, S_k^*,)}{P(\theta_i|S_1^*, S_2^*, \cdots, S_{k-1}^*)}$$

differs from 1 by more than would expected by chance, we consider the new statistic to be informative. Consequently, we use an intuitively reasonable definition of the threshold $T(\theta)$ which is defined in terms of the the probability of observing a deviation from 1 as large as that which was actually observed. We now explain the details of this procedure.

Suppose that we accept N_{k-1} of the total set of $N = 5000000$ data sets when performing a rejection method using $S_1, \ldots, S_{k-1}$, and that this results in an empirically estimated posterior for θ_i of $P_{k-1}(i)$ for each i. Furthermore, let N_k denote the number of accepted data sets after the addition of the k^{th} statistic (the statistic we are currently considering adding). Under the null hypothesis that statistic S_k adds no information to the posterior for θ, when we add S_k we are equally likely to accept any subset (of size N_k) of the N_{k-1} data sets that were accepted using statistics $\theta_1, \ldots, \theta_{k-1}$. Thus, we can, in a relatively straightforward manner, approximate the probability of observing any particular deviation from 1 when constructing the ratio

$$\frac{P(\theta_i|S_1^*, S_2^*, \cdots, S_{k-1}^*, S_k^*,)}{P(\theta_i|S_1^*, S_2^*, \cdots, S_{k-1}^*)}.$$

In particular, for each i, we proceed as follows:

Suppose that there were $N_{k-1}(i)$ accepted data sets with $\theta = \theta_i$ when considering statistics $S_1, \ldots, S_{k-1}$. Then, under the null, the expected number of acceptances after adding the k^{th} statistic is $N_{k-1}(i)N_k/N_{k-1}$. For computational convenience we treat each θ_i $(i = 1, \ldots, L)$ independently. Since $N_k(i)$ and $N_k(i')$ are negatively correlated for any $i \neq i'$, this means that the approach is conservative. From this we can calculate the standard deviation of the number of acceptances. We

DOI: 10.2202/1544-6115.1389

then define $T(i)$, the threshold for acceptable differences in the ratio (5) in terms of this standard deviation, allowing the ratio to differ from 1 by up to 4 standard deviations.

References

Abril, J. (1994). On the concept of approximate sufficiency. *Pak. J. Statist.*, 10:171–177.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035.

Cabrera, J. and Yohai, V. (1999). A new computational approach for Bayesian and robust Bayesian statistical analysis. Technical report, available at http://www.rci.rutgers.edu/ cabrera/.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popn. Biol.*, 3:87–112.

Fu, Y.-X. and Li, W.-H. (1997). Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.*, 14:195–199.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In Futuyma, D. and Antonovics, J., editors, *Oxford Surveys in Evolutionary Biology*, volume 7, pages 1–44.

Innan, H., Zhang, K., Marjoram, P., Tavaré, S., and Rosenberg, N. (2005). Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics*, 169:1763–1777.

Joyce, P. (1998). Partition structures and sufficient statistics. *Jour. of App. Prob.*, 35:622–632.

Kingman, J. F. C. (1982a). The coalescent. *Stoch. Proc. Applns.*, 13:235–248.

Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations. In Koch, G. and Spizzichino, F., editors, *Exchangeability in probability and statistics*, pages 97–112. North-Holland Publishing Company.

Kingman, J. F. C. (1982c). On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43.

Le Cam, L. (1964). Sufficiency and approximate sufficiency. *Ann Math Stat*, 35:1419–1455.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Nat. Acad. Sci.*, 100:15324–15328.

Marjoram, P. and Tavaré, S. (2006). Modern computational approaches for analysing molecular genetic variation data. *Nat. Rev. Genet.*, 7:759–770.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091.

Nordborg, M. (2001). Coalescent theory. In Balding, D. J., Bishop, M. J., and Cannings, C., editors, *Handbook of Statistical Genetics*, pages 179–208. John Wiley & Sons, Inc., New York.

Plagnol, V. and Tavaré, S. (2004). Approximate Bayesian computation and MCMC. In Niederreiter, H., editor, *Proceedings of the 5th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*. Springer Verlag.

Ripley, B. D. (1982). *Stochastic simulation*. John Wiley & Sons, Inc., New York.

Sisson, S. A., Fan, Y., and M., T. M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.*, 104:1760–1765.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times for molecular sequence data. *Genetics*, 145:505–518.

DOI: 10.2202/1544-6115.1389