Statistical Applications in Genetics and Molecular Biology

Volume 9, *Issue* 1 2010 *Article* 34

On Optimal Selection of Summary Statistics for Approximate Bayesian Computation

Matthew A. Nunes* David J. Balding†

*Lancaster University, m.nunes@lancs.ac.uk

†University College London, d.balding@ucl.ac.uk

On Optimal Selection of Summary Statistics for Approximate Bayesian Computation*

Matthew A. Nunes and David J. Balding

Abstract

How best to summarize large and complex datasets is a problem that arises in many areas of science. We approach it from the point of view of seeking data summaries that minimize the average squared error of the posterior distribution for a parameter of interest under approximate Bayesian computation (ABC). In ABC, simulation under the model replaces computation of the likelihood, which is convenient for many complex models. Simulated and observed datasets are usually compared using summary statistics, typically in practice chosen on the basis of the investigator's intuition and established practice in the field. We propose two algorithms for automated choice of efficient data summaries. Firstly, we motivate minimisation of the estimated entropy of the posterior approximation as a heuristic for the selection of summary statistics. Secondly, we propose a two-stage procedure: the minimum-entropy algorithm is used to identify simulated datasets close to that observed, and these are each successively regarded as observed datasets for which the mean root integrated squared error of the ABC posterior approximation is minimized over sets of summary statistics. In a simulation study, we both singly and jointly inferred the scaled mutation and recombination parameters from a population sample of DNA sequences. The computationally-fast minimum entropy algorithm showed a modest improvement over existing methods while our two-stage procedure showed substantial and highly-significant further improvement for both univariate and bivariate inferences. We found that the optimal set of summary statistics was highly dataset specific, suggesting that more generally there may be no globally-optimal choice, which argues for a new selection for each dataset even if the model and target of inference are unchanged.

KEYWORDS: data reduction, computational statistics, likelihood free inference, entropy, sufficiency

---

*We thank Mark Beaumont for helpful comments on a manuscript draft, as well as two anonymous referees. Research funded by the U.K. Engineering and Physical Sciences Research Council under their Mathematics Underpinning the Life Sciences initiative. Software for implementing the two algorithms described in the paper is available from MN.

1 Introduction

In a typical application of approximate Bayesian computation (ABC), a set of summary statistics $\mathbf{S}$ is computed from both observed and simulated datasets in order to define a between-dataset distance. For example, in the simplest form of ABC, called "rejection ABC" (Tavaré, Balding, Griffiths, and Donnelly, 1997, Beaumont, Zhang, and Balding, 2002), many parameter values (ϕ) are simulated from the prior distribution, each is used to simulate a dataset under the given statistical model, and $\mathbf{S}$ is computed for each dataset. The ϕ values that generate an $\mathbf{S}$ value closest to that observed, usually measured by Euclidean distance, are then treated as an approximate random sample from the posterior distribution for ϕ. The key advantage of such an approach is that complex models and high-dimensional datasets can be handled in a Bayesian framework without having to explicitly compute the likelihood. Researchers are thus free to specify very detailed models, involving many latent variables: the only limitation is that the model can readily be simulated. ABC methods have been applied to inferences of demographic history in population genetics (Fagundes, Ray, Beaumont, Neuenschwander, Salzano, Bonatto, and Excoffier, 2007, François, Blum, Jakobsson, and Rosenberg, 2008), to infectious disease models (Luciani, Sisson, Jiang, Francis, and Tanaka, 2009, McKinley, Cook, and Deardon, 2009) and to systems biology (Ratmann, Jørgensen, Hinkley, Stumpf, Richardson, and Wiuf, 2007, Toni, Welch, Strelkowa, Ipsen, and Stumpf, 2009). Software is becoming available to assist its implementation (Hickerson, Stahl, and Takebayashi, 2007, Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, and Estoup, 2008, Lopes, Balding, and Beaumont, 2009). In addition to the simple rejection approach outlined above, ABC approaches have been developed within Markov chain Monte Carlo (MCMC) (Marjoram, Molitor, Plagnol, and Tavaré, 2003) and sequential Monte Carlo algorithms (Sisson, Fan, and Tanaka, 2007, Beaumont, Cornuet, Marin, and Robert, 2009).

In practice, all these approaches rely on a good choice of $\mathbf{S}$ to extract from the dataset most of its information about ϕ. Current practice is to rely on expert opinion and established practice in the field for the choice of $\mathbf{S}$, with little or no quantitative justification, and no tailoring of the choice to the observed dataset. Ideally $\mathbf{S}$ should be sufficient for ϕ, but this is almost never achievable in practice. Joyce and Marjoram (2008) introduced a notion of approximate sufficiency (AS), and used it to choose $\mathbf{S}$. Briefly, if $\mathbf{S}$ is sufficient then the posterior distribution for ϕ will be unaffected by replacing $\mathbf{S}$ with $\mathbf{S}' = \mathbf{S} \cup \{X\}$, where X is an additional summary statistic. Accordingly, Joyce and Marjoram (2008) propose an algorithm in which a potential summary statistic

X is chosen randomly from a pool Ω, and is accepted into $\mathbf{S}$ if the change in the corresponding posterior density approximation exceeds a threshold. The AS approach was the first principled approach to choosing summary statistics for ABC. Its limitations include dependence of the final value of $\mathbf{S}$ on the order in which statistics are tested for inclusion, and on the choice of threshold for accepting a new summary statistic.

Wegmann, Leuenberger, and Excoffier (2009) explored the use of partial least squares (PLS) regression applied to a pool Ω of statistics that have (optionally) been Box-Cox transformed (Box and Cox, 1964). In this approach, the chosen $\mathbf{S}$ are the leading k PLS components, which are orthogonal linear combinations of elements of Ω, chosen to be maximally correlated with ϕ (see for example Tenenhaus, 1998, Boulesteix and Strimmer, 2007). To choose the optimal value of k, Wegmann et al. (2009) suggest a leave-one-out cross-validation criterion, implemented in the `pls` R software package (Mevik and Wehrens, 2007). For computational reasons, in practice the PLS components are computed using a subsample of the simulated values, and the resulting PLS loadings are applied to all simulated datasets. Below, we follow Wegmann et al. (2009) and use a subsample of size 10^4, and we also use the `ABCtoolbox` software (Wegmann, Leuenberger, Neuenschwander, and Excoffier, 2010) for the PLS implementation. Neural networks can also be used for data reduction instead of PLS (Blum and François, 2010).

Here, we first propose for use in ABC inference the set of summary statistics $\mathbf{S}_{ME} \subseteq \Omega$ that minimizes the entropy of the resulting approximate posterior distribution. This criterion is easy to apply, and performs reasonably well. For superior performance, at additional computational cost, we further propose a two-stage procedure. Ideally we would wish to choose $\mathbf{S}$ that satisfies an optimality criterion, such as minimising the square root of the sum of squared errors (RSSE) of the resulting posterior approximation. In practice, we can't compute the RSSE since it implies knowledge of the ϕ underlying the observed dataset, which is our target of inference. However, the ϕ values are known for the simulated datasets. In Stage 2, we seek $\mathbf{S}_2$ that minimises the mean RSSE (MRSSE) over simulated datasets that were selected in Stage 1 to be closest to that observed, in terms of Euclidean distance between $\mathbf{S}_{ME}$ values.

We evaluate the performance of our ME and two-stage approaches, relative to the AS and PLS methods, in estimating the scaled mutation and recombination rates for a population sample of DNA sequences.

DOI: 10.2202/1544-6115.1576

2 Methods

2.1 Minimum Entropy (ME) approach

The entropy of a probability distribution is a measure of information (Shannon and Weaver, 1948). High entropy corresponds to low information and vice-versa. A sharply-peaked distribution with thin tails has low entropy. For example, the entropy of a uniform distribution on the interval $(0, a)$ is $\log(a)$, reflecting low entropy (high information) when a is small. For a unimodal distribution, entropy is related to variance (Ebrahimi, Maasoumib, and Soofi, 1999) but there is no general relationship: the entropy of the distribution that is uniform over the pair of intervals $(0, a)$ and $(c, c+a)$, for $c \geq a$, does not depend on c whereas the variance does. Minimisation of entropy (Shannon and Weaver, 1948) has been applied as a parameter selection technique in many scientific fields (MacKay, 2003, Cover and Thomas, 2006). Since we seek to extract maximal information about the parameter ϕ from the data, we propose minimising the estimated entropy of the posterior approximation for ϕ as a heuristic to select $\mathbf{S}$ for ABC inference.

The entropy of a distribution is equivalent to its Kullback-Leibler distance from the uniform distribution (Kullback and Leibler, 1951). We also considered the Kullback-Leibler distance of the posterior approximation from the prior, but found that this could give a good score to a "noise" statistic in the case of an informative prior. We preferred minimising entropy rather than variance as an optimality criterion because we expect better handling of multimodal posteriors (see also Ebrahimi et al., 1999). Moreover, variance, unlike entropy, is often a property of direct interest and if we focussed on minimising variance our resulting variance estimates could be downwards biased for our first algorithm. Although this problem may not be completely eliminated for the ME algorithm, little if any effect is expected for the two-stage algorithm since the selection process uses simulated datasets rather than that observed.

There are many sample-based estimators of entropy (Beirlant, Dudewicz, Györfi, and Van der Meulen, 1997) that can be applied to the output of an ABC algorithm. Some of these involve either kernel-based (Dmitriev and Tarasenko, 1974, Ahmad and Lin, 1976, Hall, 1987, Hall and Morton, 1993) or histogram-based density estimators (Hall and Morton, 1993, Györfi and van der Muelen, 1987, Scott, 1992), while other methods are based on nearest-neighbor distances (Cressie, 1976, Vasicek, 1976, Tsybakov and Van Der Meulen, 1996, Singh, Misra, Fedorowicz, and Demchuk, 2003). Since it is suitable for multivariate samples, here we adopt the unbiased *kth nearest neighbor estimator* of

entropy (Singh et al., 2003):

$$\hat{H} = \log\left[\frac{\pi^{p/2}}{\Gamma(p/2+1)}\right] - \psi(k) + \log n + \frac{p}{n}\sum_{i=1}^{n}\log R_{i,k}, \quad (1)$$

where p denotes the dimension of ϕ and $R_{i,k}$ is the Euclidean distance from ϕ_i to its kth nearest neighbor in the posterior sample, while $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. The summation in (1) is small if the data are clustered together so that the majority of kth nearest neighbour distances are small. Below, we take $k = 4$ as suggested by Singh et al. (2003) for its error properties.

Given simulated parameter values ϕ_i, $i = 1, \ldots, n_{sim}$, and corresponding simulated datasets, we can now define our ME algorithm:

Rejection-ABC algorithm: Compute $\mathbf{S} \subseteq \Omega$ for the ith simulated dataset (denoted $\mathbf{S}_i$) and accept the ϕ_i corresponding to the n_{acc} smallest values of $||\mathbf{S}_i - \mathbf{S}^*||$, where $||\cdot||$ denotes Euclidean distance and $\mathbf{S}^*$ is the value of $\mathbf{S}$ computed at the observed dataset.

ME algorithm: For every $\mathbf{S} \subseteq \Omega$, perform rejection-ABC and compute $\hat{H}$ from the n_{acc} accepted values. $\mathbf{S}_{ME}$ is the value of $\mathbf{S}$ that minimises $\hat{H}$, and the corresponding n_{acc} values of ϕ_i provide the approximation to the posterior distribution for ϕ.

For large Ω, rather than consider all $\mathbf{S} \subseteq \Omega$ it may be necessary to restrict attention for example to $\{\mathbf{S} \subset \Omega : |\mathbf{S}| < k\}$, or to some other class of subsets specified by the investigator. Alternatively, instead of an exhaustive search over some subset of Ω, an iterative updating algorithm for $\mathbf{S}$ such as that of Joyce and Marjoram (2008) could be implemented with an acceptance criterion based on entropy.

2.2 Two-stage procedure

As noted above, if we treat a simulated dataset as if it were observed, we can use the known ϕ to assess the performance of $\mathbf{S}$ in ABC inference, using any preferred measure of average error of a distribution for ϕ. For example, we could compute the RSSE using the n_{acc} accepted values of ϕ_i:

$$\text{RSSE} = \left(\frac{1}{n_{acc}}\sum_{i=1}^{n_{sim}} I_i \|\phi_i - \phi\|^2\right)^{1/2}, \quad (2)$$

where $I_i = 1$ if the pair $(\phi_i, \mathbf{S}_i)$ is accepted in the rejection-ABC algorithm, otherwise $I_i = 0$. For multivariate ϕ the value of RSSE depends on the scale

DOI: 10.2202/1544-6115.1576

of each component. In the bivariate inferences below θ and ϕ are on similar scales; in other settings it may be appropriate to initially standardise each component.

The $\mathbf{S}$ that minimises (2) will vary over datasets, and so we minimise the RSSE averaged over n_{obs} simulated datasets close to that observed. In practice the definition of "close" requires a good choice of $\mathbf{S}$, and this is the problem we have set out to solve. To overcome this circularity we first employ the ME algorithm to find $\mathbf{S}_{ME}$ and use it to identify the n_{obs} datasets.

Again assuming we have (ϕ_i,S_i), $i = 1,\ldots,n_{sim}$, our two-stage algorithm can be summarized as:

Stage 1: Implement the ME algorithm to identify $\mathbf{S}_{ME}$. Find the n_{obs} simulated datasets that minimize $||\mathbf{S}_{ME,i} - \mathbf{S}^*_{ME}||$, where $\mathbf{S}_{ME,i}$ and $\mathbf{S}^*_{ME}$ denote the values of $\mathbf{S}_{ME}$ computed from the ith simulated and the observed datasets, respectively.

Stage 2: For each $\mathbf{S} \subseteq \Omega$, perform rejection-ABC and evaluate (2) for each of the n_{obs} datasets; denote the jth value RSSE(j). Hence identify the $\mathbf{S}_2 \subseteq \Omega$ that minimizes

$$\text{MRSSE} = \frac{1}{n_{obs}} \sum_{j=1}^{n_{obs}} \text{RSSE}(j). \tag{3}$$

We consider all $\mathbf{S} \subseteq \Omega$ in both stages, but it would be more computationally efficient, and will usually give the same answer, to consider in the second stage only the $\mathbf{S}$ that performed at least moderately well in the first stage.

2.3 Regression adjustment

ABC algorithms can usually be improved by adjusting the ith accepted parameter value ϕ_i to correct for the (small) discrepancy between its corresponding summary statistic $\mathbf{S}_i$ and the observed value $\mathbf{S}^*$. Fitting the homoscedastic regression model

$$\phi_i = \alpha + (\mathbf{S}_i - \mathbf{S}^*)^T\beta + \varepsilon_i$$

Beaumont et al. (2002) replaced ϕ_i with

$$\phi_i' = \hat{\alpha} + \hat{\varepsilon}_i = \phi_i + (\mathbf{S}^* - \mathbf{S}_i)^T\hat{\beta}, \tag{4}$$

where $(\hat{\alpha},\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\phi}$ is the weighted least squares estimator of (α,β) and $\boldsymbol{X}$ is the design matrix with ith row equal to (1, $\mathbf{S}_i - \mathbf{S}^*$). The

weight matrix $\boldsymbol{W}$ was taken to be

$$\boldsymbol{W}_{ij} = \begin{cases} K(\|\mathbf{S}_i - \mathbf{S}^*\|), & i = j \\ 0 & \text{otherwise.} \end{cases}$$

with K the Epanechnikov kernel

$$K_\delta(t) = \begin{cases} 3(1-(t/\delta)^2)/(4\delta), & t \le \delta \\ 0 & t > \delta. \end{cases}$$

Beaumont et al. (2002) also applied the weights W_{ii} to the ϕ_i' in posterior approximations, but we have not used this reweighting below.

It may also be useful to adjust for systematic changes in the variance of ϕ_i' as $\mathbf{S}_i$ deviates from $\mathbf{S}^*$, using a locally log-linear regression for the squared residuals from the mean adjustment:

$$\log(\hat{\varepsilon}_i^{\,2}) = \alpha' + (\mathbf{S}_i - \mathbf{S}^*)^T\beta' + \epsilon_i,$$

and estimating (α', β') using the same weighted least squares approach (Fan and Yao, 1998, Yu and Jones, 2004). Then we obtain the adjusted parameter values

$$\phi_i'' = \hat{\alpha} + \hat{\varepsilon}_i \frac{\hat{\sigma}(\mathbf{S}^*)}{\hat{\sigma}(\mathbf{S}_i)} = \hat{\alpha} + \hat{\varepsilon}_i \exp\{(\mathbf{S}^* - \mathbf{S}_i)^T\hat{\beta}'/2\}. \quad (5)$$

Feed-forward neural networks have also been proposed to make both mean and variance adjustments in the ABC setting (Blum and François, 2010).

2.4 Simulation study: estimation of the scaled mutation and recombination rates

Our simulation study is similar to that of Joyce and Marjoram (2008) except that we consider joint inference of the two parameters in addition to inference of each parameter separately. The parameters are the scaled mutation and recombination rates, θ and ρ, and datasets consist of 50 haplotypes generated using the `ms` software (Hudson, 2002) under the standard coalescent model, with infinite-sites mutation (Nordborg, 2007). The (θ, ρ) values were simulated from the uniform distribution on $(2, 10) \times (0, 10)$. This distribution was also adopted as the prior for inference. Although it is unrealistic in practice for the prior to be exactly the distribution from which the parameters were drawn, this should not bias comparisons of methods.

Univariate inferences for θ and ρ using the ME and two-stage algorithms were implemented with $n_{sim} = 10^6$, $n_{acc} = 10^4$ and $n_{obs} = 10^2$. These are

DOI: 10.2202/1544-6115.1576

compared with both the PLS and AS methods, which have only been proposed for univariate inferences. All four algorithms were repeated, implementing each of the regression adjustments (4) and (5). The pool of summary statistics Ω is shown in Table 1.

We were able to modify PLS, but not AS, for bivariate inference comparisons. Using cross-validation we found three and five to be the optimal numbers of PLS components for univariate inferences about θ and ρ, respectively, and the results shown below are for these numbers of selected components. For bivariate inferences, assessing predictive accuracy under cross-validation is more difficult and instead we observed that three components minimized the MRSSE of the resulting posterior approximation over the n_{obs} datasets, and we reported the results for three components below. In practice, RSSE isn't available because it requires knowledge of θ and/or ρ, and its use here should bias results in favour of the PLS method.

3 Results

Table 1 shows the numbers of datasets for which each $C \in \Omega$ was included in the optimal set by the AS, ME and two-stage methods in univariate, unadjusted ABC inference about the scaled mutation parameter, θ and the scaled recombination parameter, ρ. The results for AS are taken from Table 5 of Joyce and Marjoram (2008); these authors used a uniform (0,10) prior for each of θ and ρ while the other parameter was fixed. The "statistic" C_2 does not depend on the data and so should never be included in the chosen $\mathbf{S}$. No sufficient statistics are available for either θ or ρ, but C_1 (number of segregating sites) is highly informative about θ (Hudson, 1990, Nordborg, 2007), while C_5 (number of distinct haplotypes) is informative about ρ, and so we expect these to be often included in the respective optimal sets.

From Table 1, we see that the AS method performed well relative to these expectations for unadjusted univariate inference of θ and ρ. ME performed slightly less well, but the two-stage algorithm had a perfect score: it never included C_2 in the optimal set, while always including C_1 and C_5 for inferences about, respectively, θ and ρ.

Table 2 includes the main result of the paper: that the set of summary statistics selected by our novel two-stage algorithm led to substantially better ABC inference in the simulation study than any of the other methods of selecting summary statistics considered here. The two-stage algorithm gives the lowest MRSSE in each row of Table 2, and its improvement over the PLS and AS algorithms is statistically significant (Table 3). ME also performed

Table 1: The pool of summary statistics Ω considered for summarising datasets of DNA sequence haplotypes in the simulation study. For each statistic, we show the number of observed datasets (out of 100) for which it was included in the optimal set in univariate, unadjusted ABC inference by the methods described in the text.

		Selected for θ (%)			Selected for ρ (%)		
Statistic	Description	AS	ME	2-stage	AS	ME	2-stage
C_1	no. of segregating sites	75	67	100	73	67	97
C_2	Uniform[0,25] random variable	4	3	0	2	5	0
C_3	mean no. of differences over all pairs of haplotypes	27	54	25	52	30	19
C_4	25*(mean r^2 across pairs separated by $< 10\%$ of the simulated genomic region)	56	35	50	35	59	78
C_5	no. of distinct haplotypes	43	19	20	78	73	100
C_6	frequency of the most common haplotype	36	20	1	11	23	2
C_7	no. of singleton haplotypes	16	14	5	16	31	5

Table 2: MRSSE for ABC inference in the simulation study for each of the seven summary statistics taken alone (columns 1 – 7), all six summary statistics other than C_2 (column 8), and four methods of choosing $\mathbf{S}$ described in the text (columns 9 – 12). Bold indicates the lowest value in each row. In each section of the table, first row: no regression adjustment; second row: mean adjustment (4); third row: mean and variance adjustments (5).

		C_1	C_2	C_3	C_4	C_5	C_6	C_7	All 6	PLS	AS	ME	2-stage
θ	no adj.	1.75	3.27	2.26	3.15	2.33	2.89	2.45	1.87	1.83	1.86	1.80	**1.70**
	mean	1.75	3.27	2.26	3.14	2.33	2.89	2.45	1.74	1.78	1.76	1.74	**1.68**
	mean+var.	1.75	3.27	2.26	3.14	2.33	2.89	2.45	1.70	1.75	1.76	1.70	**1.67**
ρ	no adj.	3.93	3.95	3.93	3.92	3.83	3.84	3.88	3.59	3.91	3.68	3.54	**3.44**
	mean	3.92	3.95	3.93	3.92	3.83	3.84	3.89	3.33	3.37	3.83	3.56	**3.31**
	mean+var.	3.92	3.95	3.93	3.92	3.83	3.83	3.88	3.21	3.35	3.60	3.27	**3.17**
(θ, ρ)	no adj.	4.36	5.19	4.62	5.10	4.55	4.89	4.65	4.81	4.15	-	4.03	**3.97**
	mean	4.36	5.19	4.62	5.10	4.56	4.89	4.65	4.83	4.08	-	4.06	**3.81**
	mean+var.	4.36	5.19	4.62	5.10	4.55	4.89	4.65	4.75	4.03	-	3.71	**3.66**

better than PLS, AS and the best-performing single summary statistic, but the differences between these are usually not significant.

As expected, C_1 was the best individual statistic for univariate inference about θ, and also for the bivariate inference (Table 2). For ρ, all the statistics performed almost equally poorly: no single summary statistic can capture much information about this parameter, but using all six statistics (other than C_2) was almost as effective as the two-stage algorithm after regression adjustment. Reassuringly, the noise statistic C_2 performed worst in each inference, and generated MRSSE values similar to sampling from the prior distribution. The statistic C_4 performs poorly on its own for all inferences, although it is included in several near-optimal combinations of summary

DOI: 10.2202/1544-6115.1576

Table 3: Difference in MRSSE (ΔMRSSE) for the pair of methods indicated in the column heading, together with its standard error and p-value (one-sided t_{99} test).

		AS vs. 2-stage		PLS vs. 2-stage	
		ΔMRSSE (s.e.)	p-value	ΔMRSSE (s.e.)	p-value
θ	no adjustment	0.151 (0.024)	$< 10^{-8}$	0.130 (0.020)	$< 10^{-8}$
	mean	0.078 (0.029)	0.0042	0.099 (0.020)	$< 10^{-5}$
	mean+variance	0.097 (0.024)	$< 10^{-4}$	0.086 (0.023)	0.0002
ρ	no adjustment	0.239 (0.048)	$< 10^{-5}$	0.471 (0.070)	$< 10^{-9}$
	mean	0.525 (0.090)	$< 10^{-7}$	0.059 (0.031)	0.0300
	mean+variance	0.426 (0.081)	$< 10^{-6}$	0.181 (0.042)	$< 10^{-4}$
(θ, ρ)	no adjustment			0.185 (0.053)	0.0004
	mean			0.270 (0.065)	$< 10^{-4}$
	mean+variance			0.369 (0.071)	$< 10^{-6}$

statistics (Table 4); this suggests that it conveys little information about θ or ρ, but that information is apparently not captured by the other statistics.

The mean regression adjustment usually increases the efficacy of ABC inference, and the variance adjustment typically offers a small further improvement (Table 2). The regression adjustments have little effect for the seven single-statistic inferences ($|\mathbf{S}| = 1$); the tolerance we use is small (acceptance rate = 1%) relative to the range of tolerances for which mean regression adjustment was found to be useful by Beaumont et al. (2002). A fixed acceptance rate corresponds to an increasing tolerance in each dimension of $\mathbf{S}$ as $|\mathbf{S}|$ increases, and a 1% acceptance rate does allow for regression adjustments to convey an advantage when $|\mathbf{S}| > 1$. Choosing all six summary statistics is not efficacious without regression adjustment, but since $|\mathbf{S}| = 6$ there is a substantial gain from each of the regression adjustments for the univariate inferences, though much less so for the bivariate inference.

Table 4 shows the best-performing $\mathbf{S} \subseteq \Omega$ for unadjusted ABC inferences. It is striking that there is rarely a clear "winner": typically no one $\mathbf{S}$ is near-optimal for a majority of datasets and thus the optimality of $\mathbf{S}$ in extracting information from the data about the parameter holds at best only locally. Our study included a wide range of values for θ and ρ, and so an $\mathbf{S}$ that performs well in, say, the high-θ, low-ρ domain may be inferior when θ is low and ρ is high. The superiority of the two-stage algorithm that we have observed may lie in the fact that it optimizes $\mathbf{S}$ locally to the observed dataset. While there is no universally best $\mathbf{S}$, the set $\{C_1, C_4, C_5\}$ is the best performing set for inferences about ρ and (θ, ρ).

The two-stage procedure is computationally the most expensive, requiring about 10 hours of CPU time on a standard desktop computer for an

Table 4: Top three sets of summary statistics, measured as the frequency (number of datasets out of 100) for which unadjusted ABC inference using that set achieved within 1% of the optimal value of: true RSSE (which requires knowledge of the parameter that is the target of inference), estimated entropy, and MRSSE (over the 100 nearest datasets, as used in the two-stage algorithm).

	true RSSE	frequency	est. entropy	frequency	MRSSE	frequency
θ	$\{C_1, C_3\}$	26	$\{C_1, C_3\}$	31	$\{C_1, C_4\}$	60
	$\{C_1, C_4\}$	23	$\{C_1, C_4\}$	29	$\{C_1, C_3\}$	34
	$\{C_1, C_5\}$	12	$\{C_1, C_3, C_4\}$	17	$\{C_1, C_4, C_5\}$	33
ρ	$\{C_1, C_4, C_5\}$	34	$\{C_1, C_4, C_5\}$	33	$\{C_1, C_4, C_5\}$	73
	$\{C_1, C_5\}$	17	$\{C_1, C_5\}$	23	$\{C_1, C_3, C_4, C_5\}$	40
	$\{C_1, C_3, C_4, C_5\}$	15	$\{C_3, C_4, C_5\}$	16	$\{C_1, C_5\}$	23
(θ, ρ)	$\{C_1, C_4, C_5\}$	29	$\{C_1, C_5\}$	41	$\{C_1, C_4, C_5\}$	76
	$\{C_1, C_3, C_5\}$	17	$\{C_1, C_4, C_5\}$	40	$\{C_1, C_3, C_4, C_5\}$	44
	$\{C_1, C_5\}$	16	$\{C_1, C_3, C_5\}$	27	$\{C_1, C_5\}$	39

exhaustive search over all 127 sets of summary statistics for one dataset. The corresponding computational time for the ME, AS, and PLS methods are 6 minutes, 3 minutes and 2 minutes, respectively (including the cross-validation for PLS). These timings do not include the $\sim$ 11 hrs required for the initial 10^6 simulations of parameter value and corresponding dataset, which are common to all methods. (This step, as well as both the ME calculations and the two-stage analysis, are readily parallelized). We repeated the two-stage analysis for θ using $n_{obs} = 50$ datasets closest to each observed dataset (previously $n_{obs} = 100$) and this led to only very slightly worse performance while halving the compute time. To further reduce computing time, the search in Stage 2 can be limited to sets of summary statistics that performed well in Stage 1.

4 Discussion

We have proposed two new algorithms that automate the choice of summary statistics in rejection-ABC inference. Both use the heuristic of minimising the estimated entropy of the resulting posterior distribution approximation, for which we adopted the unbiased 4th nearest neighbor estimator of entropy (Singh et al., 2003). For our two-stage algorithm this heuristic is only used in the first stage. In the second stage, selection is based on simulated datasets similar to that observed, and since the true (simulated) parameter value is known any measure of precision of a sample about a target value can be employed to identify the optimal posterior approximation – here we have adopted the square root of the sum of squared errors (RSSE).

DOI: 10.2202/1544-6115.1576

In our simulation study, based on estimating the scaled mutation and recombination rates from DNA sequence data, we found that the two-stage algorithm gave a large, and usually highly-significant, improvement over the other methods considered for both univariate and bivariate inferences. The minimum entropy (ME) algorithm also performed consistently better than the approximate-sufficiency (AS) (Joyce and Marjoram, 2008) and partial least-squares (PLS) (Wegmann et al., 2009) methods, which each performed about as well as the best single-statistic inference. The ME method conveys advantages in being easy to apply and interpret, and does not require any user-supplied quantities. The AS algorithm makes a series of pairwise comparisons of summary statistics, and the result can depend on the way the sequence is chosen and on an arbitrary threshold. The PLS algorithm requires optimisation of the number of components using cross-validation, and because the selected statistics are linear combinations of an original set of summary statistics, the selected statistics can lack interpretability.

We found that the best choice of summary statistics $\mathbf{S} \subseteq \Omega$ varied over datasets, which suggests that more generally there may be no universally optimal $\mathbf{S}$ for a given inference problem. Thus efficient ABC inference may require a choice of summary statistics specific to the observed dataset. It was, however, possible to identify some particular $\mathbf{S}$, for example the set $\{C_1, C_4, C_5\}$, that performed well overall in our simulation study. Choosing all summary statistics in Ω performed well here for univariate but not bivariate inferences. It is unlikely that this will provide a good approach even for univariate inferences when Ω is large.

In our simulation study we standardized each statistic but did not orthogonalize. Better inferences should be possible by considering different weightings of an orthogonal set of statistics, though a thorough exploration of the space of possible weightings would be computationally prohibitive. A more tractable route to improvement could be to consider weighted combinations of posterior approximations from several near-optimal subsets of Ω.

The statistical efficiency of our two-stage algorithm comes at a substantial computational cost. While this cost remains modest relative to the large simulation cost common to all rejection-ABC methods, and we have developed software to facilitate the implementation of our method (available on request), there is likely to be ample scope for reducing computational cost with little statistical loss. The exhaustive search over (a large subset of) Ω could be replaced with an iterative updating scheme for $\mathbf{S}$, such as that of Joyce and Marjoram (2008). The update decision could be based on the RSSE for a single dataset chosen with a probability that declines with distance from the observed dataset. In this case the updates would not terminate at an optimal

choice of **S**, but could be run for a large number of iterations, from which the most frequently-selected values of **S** could be selected and their corresponding posterior approximations averaged, or else used to assign weights to the individual statistics.

In closing, we note that while we have focussed on rejection-ABC, similar approaches should also be applicable to other ABC schemes.

References

Ahmad, I. and P. E. Lin (1976): "A nonparametric estimation of the entropy for absolutely continuous distributions," *IEEE Trans. Inf. Theory*, 22, 372–375.

Beaumont, M. A., J. Cornuet, J. Marin, and C. P. Robert (2009): "Adaptive approximate Bayesian computation," *Biometrika*, 96, 983–990.

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

Beirlant, J., E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen (1997): "Nonparametric entropy estimation: An overview," *Intern. J. Math. Stat. Sci.*, 6, 17–40.

Blum, M. G. B. and O. François (2010): "Non-linear regression models for approximate Bayesian computation," *Stat. Comput.*, 20, 63–73.

Boulesteix, A. L. and K. Strimmer (2007): "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," *Briefings Bioinf.*, 8, 32–44.

Box, G. E. P. and D. R. Cox (1964): "An analysis of transformations," *J. Roy. Stat. Soc. B*, 26, 211–252.

Cornuet, J., F. Santos, M. Beaumont, C. Robert, J. Marin, D. Balding, T. Guillemaud, and A. Estoup (2008): "Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computations," *Bioinformatics*, 24, 2713–2719.

Cover, T. M. and J. A. Thomas (2006): *Elements of information theory*, Wiley-Interscience.

Cressie, N. (1976): "On the logarithms of high-order spacings," *Biometrika*, 63, 343–355.

Dmitriev, Y. G. and F. P. Tarasenko (1974): "On the estimation of functionals of the probability density and its derivatives," *Theory of Prob. Appl.*, 18, 628–633.

Ebrahimi, N., E. Maasoumib, and E. Soofi (1999): "Ordering univariate distributions by entropy and variance," *J. Econometrics*, 90, 317–336.

DOI: 10.2202/1544-6115.1576

Fagundes, N. J. R., N. Ray, M. A. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto, and L. Excoffier (2007): "Statistical evaluation of alternative models of human evolution," *Proc Natl Acad Sci USA*, 104, 17614–17619.

Fan, J. and Q. Yao (1998): "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, 85, 645–660.

François, O., M. G. B. Blum, M. Jakobsson, and N. A. Rosenberg (2008): "Demographic history of European populations of Arabidopsis thaliana," *PLoS Genet.*, 4, e1000075.

Györfi, L. and E. C. van der Muelen (1987): "Density-free convergence properties of various estimators of entropy." *Comput. Stat. Data Anal.*, 5, 425–436.

Hall, P. (1987): "On Kullback-Leibler loss and density estimation," *Ann. Stat.*, 15, 1491–1519.

Hall, P. and S. C. Morton (1993): "On the estimation of entropy," *Ann. Inst. Stat. Math.*, 45, 69–88.

Hickerson, M., E. Stahl, and N. Takebayashi (2007): "msBayes: Pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation," *BMC Bioinformatics*, 8.

Hudson, R. R. (1990): "Gene genealogies and the coalescent process," in D. Futuyma and J. Antonovics, eds., *Oxford Surveys in Evolutionary Biology*, volume 7, Oxford University Press, 1–44.

Hudson, R. R. (2002): "Generating samples under a Wright-Fisher neutral model of genetic variation," *Bioinformatics*, 18, 337–338.

Joyce, P. and P. Marjoram (2008): "Approximately sufficient statistics and Bayesian computation," *Stat. Appl. Gen. Mol. Biol.*, 7, 1–16.

Kullback, S. and R. A. Leibler (1951): "On information and sufficiency," *Ann. Math. Stat.*, 22, 79–86.

Lopes, J., D. Balding, and M. A. Beaumont (2009): "PopABC: a program to infer historical demographic parameters." *Bioinformatics*, 25, 2747–2749.

Luciani, F., S. A. Sisson, H. Jiang, A. R. Francis, and M. M. Tanaka (2009): "The epidemiological fitness cost of drug resistance in Mycobacterium tuberculosis," *Proc Natl Acad Sci USA*, 106, 14711–14715.

MacKay, D. J. C. (2003): *Information theory, inference, and learning algorithms*, Cambridge University Press.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003): "Markov chain monte carlo without likelihoods," *Proc Natl Acad Sci USA*, 100, 15324–15328.

McKinley, T., A. R. Cook, and R. Deardon (2009): "Inference in epidemic models without likelihoods," *Intern. J. Biostat.*, 5, 24.

Mevik, B.-H. and R. Wehrens (2007): "The PLS package: principal component and partial least squares regression in R," *J. Stat. Soft.*, 18, 1–24.

Nordborg, M. (2007): "Coalescent theory," in *Handbook of Statistical Genetics*, Wiley: Chichester, 3rd edition, 179–208.

Ratmann, O., O. Jørgensen, T. Hinkley, M. P. H. Stumpf, S. Richardson, and C. Wiuf (2007): "Using likelihood free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum," *PLoS Comput. Biol.*, 3, 2266–2278.

Scott, D. W. (1992): *Multivariate density estimation: Theory, practice, and visualization*, Wiley-Interscience.

Shannon, C. E. and W. Weaver (1948): "A mathematical theory of communication," *Bell Syst. Tech. J.*, 27, 379–423.

Singh, H., V. Misra, N.and Hnizdo, A. Fedorowicz, and E. Demchuk (2003): "Nearest neighbor estimates of entropy," *Am. J. Math. Man. Sci.*, 23, 301–321.

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

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

Tenenhaus, M. (1998): *La régression PLS: théorie et pratique*, Editions Technip.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf (2009): "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," *J. Roy. Soc. Interface*, 6, 187–202.

Tsybakov, A. B. and E. C. Van Der Meulen (1996): "Root-n consistent estimators of entropy for densities with unbounded support," *Scand. J. Stat.*, 23, 75–83.

Vasicek, O. (1976): "A test for normality based on sample entropy," *J. Roy. Stat. Soc. B*, 38, 54–59.

Wegmann, D., C. Leuenberger, and L. Excoffier (2009): "Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood," *Genetics*, 182, 1207–1218.

Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier (2010): "ABCtoolbox: A versatile toolkit for approximate Bayesian computations," *BMC Bioinformatics*, 11, 116.

Yu, K. and M. C. Jones (2004): "Likelihood-based local linear estimation of the conditional variance function," *J. Am. Stat. Assoc.*, 99, 139–144.

DOI: 10.2202/1544-6115.1576