Choosing Summary Statistics by Least Angle Regression for Approximate Bayesian Computation

Muhammad Faisal[*,1,2] , Andreas Futschik[3], Ijaz Hussain[4], and Mitwali Abd-el.Moemen[5]

[1]Faculty of Health Studies, University of Bradford, United Kingdom

[2]Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, United Kingdom

[3]Institute of Applied Statistics, JK University Linz, Austria

[4]Department of Statistics, Quaid-i-Azam University, Islamabad

[5]College of Law and Political Sciences, King Saud University Saudi Arabia.

*Corresponding Author

Email: m.faisal1@bradford.ac.uk

Tel: +44(1)274236129

Abstract

Bayesian statistical inference relies on the posterior distribution. Depending on the model, the posterior can be more or less difficult to derive. In recent years, there has been a lot of interest in complex settings where the likelihood is analytically intractable. In such situations, approximate Bayesian computation (ABC) provides an attractive way of carrying out Bayesian inference. For obtaining reliable posterior estimates however, it is important to keep the approximation errors small in ABC. The choice of an appropriate set of summary statistics plays a crucial role in this effort. Here, we report the development of a new algorithm that is based on least angle regression (LARS) for choosing summary statistics. In two population genetic examples, the performance of the new algorithm is better than a previously proposed approach that uses partial least squares.

Keywords: *Likelihood-free Methods, Least Angle Regression, Mutation, Population Genetics, Recombination.*

1 Introduction

In Bayesian statistics, the relevant information in data is summarized by the posterior distribution $f(\theta|D)$. The posterior is proportional to $f(\theta|D) \propto f(\theta)f(D|\theta)$, where $f(\theta)$ is prior distribution and $f(D|\theta)$ the likelihood. In many applications, the normalizing constant of $f(\theta|D)$ is computationally intractable. In such cases simulation based approaches such as MCMC are often used to sample from the posterior. Furthermore the numerical computation of the likelihood function $f(D|\theta)$ itself can sometimes be prohibitively expensive or even impossible. Such a situation frequently occurs for instance in population genetics, where the likelihood involves the summation over a huge number of potential genealogical trees.

Approximate Bayesian computation (ABC) methods provide an approximation to the posterior without the need to compute the likelihood explicitly. Instead, data are simulated from the model under various parameter values. For each simulated data set, a vector $S' = [s_1', s_2', \ldots, s_p']$ of summary statistics is computed. If S' is close to the summary vector $S = [s_1, s_2, \ldots, s_p]$ observed for the actual data, the parameter vector θ used to generate S' is added to an approximate posterior sample. In typical applications, no sufficient summary statistics are available. Thus the choice of summary statistics involves a trade-off between computational efficiency and speed: Relevant information may be lost when choosing too few summaries, but the computations become inefficient when too many are chosen. To illustrate this feature, we now introduce rejection sampling as the most basic version of ABC:

Algorithm 1: ABC-REJ-1 Algorithm

```
(1) Simulate a parameter vector $\theta$ from the chosen prior distribution $f(\theta)$.
(2) Simulate $D'$ from model $M$ with parameter $\theta$,
    and calculate the summary statistics $S'$ from $D'$.
(3) Calculate the distance $d(S', S)$ between $S'$ and $S$.
(4) Accept $\theta$, if $d(S', S) \leq \epsilon$.
(5) Go to step 1 until $N$ iterations have been carried out.
```

As an alternative to step (4), the values from the N iterations $\theta_1, \ldots, \theta_N$ can be sorted with respect to their (ascending) distances $d(S_i',\ S)$. Out of the sorted values $\theta_1^*, \ldots, \theta_N^*$, the subset $\theta_1^*, \ldots, \theta_r^*$ consisting of the r parameter values with smallest distances $d(S_i',\ S)$ is then taken as sample from the approximate posterior For details concerning the choice of r see e.g. Faisal et al. (2013). We summarize the resulting algorithm:

Algorithm 2: ABC-REJ-2 Algorithm

```
1. For $i = 1, \ldots, N$, repeat
    1.1. Simulate parameter $\theta$ from prior distribution $f(\theta)$
    1.2. Simulate $D'$ from model $M$ with parameter $\theta$ , and
    1.3. Calculate the summary statistics $S' = [s_1', \ldots, s_p']$
    1.4. Calculate the distance $d(S', S)$ where $S = [s_1, \ldots, s_p]$.
2. Sort $\theta_1, \ldots, \theta_N$ in ascending order with respect to their corresponding
   distances $d(S_i',\ S)$. For a pre-specified cut-off $r$, return the subset
   $\theta_1^*, \ldots, \theta_r^*$.
```

It can be shown (see Marjoram et al., 2003) that Algorithm 1 generates a sample from $f(\theta | d(S, S') \leq \varepsilon)$. Besides summary statistics S, this approach also requires the selection of a suitable distance metric d as well as a choice for the acceptance cut-off ϵ. Notice that small values of ϵ lead to a sample close to the posterior $f(\theta|S)$, but for the price of a low acceptance rate. For larger ϵ, the acceptance rate gets higher, but the distribution of the sample obtained will deviate further from the actual posterior. In particular, as $\epsilon \to \infty$, observations from the prior are generated, and as $\epsilon \to 0$ observations from the posterior density $f(\theta|S)$. Acceptance rates can be very low for Algorithm 1 as candidate parameter vectors θ are generated from the prior $f(\theta)$, which can be diffuse with respect to the posterior. Algorithm 2 faces an analogous challenge.

ABC estimates can usually be improved by adjusting the i^{th} accepted parameter value θ_i to correct for the (small) discrepancy between the observed summary statistic S and its corresponding simulated summary statistic S'. For this purpose, (Beaumont et al., 2002) proposed a regression adjustment. Blum and François (2009) suggest a more general method for mean and variance adjustments using feed-forward neural networks.

Several other flavours of ABC methods are available that aim for improving the computational efficiency. They include ABC with Markov chain Monte Carlo (Marjoram et al., 2003), ABC with sequential Monte Carlo (Sisson et al., 2007), and ABC with population Monte Carlo (Beaumont et al., 2009). For a review on ABC methods see Marjoram and Tavaré (2006) as well as Csilléry et al. (2010).

All these methods depend on a good choice of summary statistics for the parameter of interest θ (Nunes and Balding, 2010). With complex models, such as those commonly considered in population genetics, sufficient summary statistics usually cannot be found (Marjoram et al., 2003). Therefore several alternative approaches have been proposed, such as approximate sufficiency (Joyce and Marjoram, 2008), maximum entropy (Nunes and Balding, 2010), averaged results of neural networks (Blum and Tran, 2010), partial least squares (Wegmann et al., 2010), and a semi-automatic approach (Fearnhead and Prangle, 2012). Blum et al. (2012) review and compare the performance of these methods with further ones (AIC and BIC, and Ridge regression).

Wegmann et al. (2010) suggest partial least squares (PLS) regression together with leave-one-out cross-validation to choose a good set of summaries. An implementation is available in "*pls*" package of R (Mevik and Wehrens, 2007).

We will compare our proposed algorithm with PLS using the root sum of square error (RSSE) and the mean of RSSE (MRSSE) as performance measures: More specifically, we consider

$$RSSE = \left(\frac{1}{r}\sum_{i=1}^{N} I_i \|\theta_i - \theta\|^2\right)^{\frac{1}{2}}$$

with r being the number of accepted observations and N the number of simulations. If the pair $(\theta_i \,, S_i)$ is accepted, we define $I_i = 1$, otherwise, $I_i = 0$. As an estimate of E($RSSE$) we consider the following average over q generated pseudo observed data sets:

$$MRSSE = \frac{1}{q}\sum_{j=1}^{q} RSSE(j),$$

In section 2, we propose a new algorithm for choosing summary statistics that is based on least angle regression (LARS) We will illustrate our approach with two examples from population genetics in section 3. Our first example is simpler involving 7 candidate summary statistics and 2 unknown parameters. The second example is more complicated with 32 available summary statistics and 4 unknown population genetic parameters. Finally, we discuss our findings in section 4.

2 Proposed Method

Our proposed approach for choosing summary statistics relies on regressing each parameter of interest onto all possible summary statistics. For selecting suitable summary statistics, we use least angle regression (LARS) (see Efron et al., 2004) together with cross validation (CV) for estimating the prediction error. First we introduce these two methods and afterwards we will establish how they can be used to extract informative summary statistics.

Subsequently, we use our method together with the Algorithm 2 (ABC-REJ-2). Since a good choice of summary statistics is important for other variants of ABC as well, our algorithm should be useful also with other versions of approximate Bayesian computation.

Algorithm 3: Least Angle Regression (LARS)

```
1. Standardize the predictors to have mean zero and unit norm and start
   with the residual vector $\phi = \theta, \quad \hat{\beta}_p = 0, \forall p$
2. Find the predictor $s_j$ most correlated with $\phi$.
3. Increase $\hat{\beta}_j$ in the direction of the sign of $corr(\phi, s_j)$ until some other
   competitor $s_k$ has as much correlation with the current residual as does $s_j$
4. Update $\phi$, and move $(\hat{\beta}_j, \hat{\beta}_k)$ in the joint least squares direction for the
   regression of $\phi$ on $(s_j, s_k)$, until some other competitor $s_l$ has as much
   correlation with the current residual.
5. Continue in this way until all $p$ predictors have been entered. Stop when
   $corr(\phi, s_j) = 0 \ \forall \ j$ that is, the OLS solution.
```

Least angle regression (LARS) may be viewed as a less greedy alternative to traditional forward selection. At each step, the predictor most correlated with the residuals is included into the model. This process continues until all predictors are in the model. It can be shown that the classical least squares solution is reached at this termination point (see Cohen, 2006). Notice that LARS can produce the least absolute shrinkage and selection operator (LASSO) solution after an additional step.

A further motivation for using LARS is that the algorithm is computationally fast. In population genetics, there is often a large set of potential summary statistics for each parameter. Sophisticated methods available in the literature are often computationally very demanding in such a context.

The cross-validation (CV) procedure is used for model selection, i.e. to find which solution to retain in the infinite number of solutions provided by the LARS algorithm. It is probably the simplest and most widely used method for estimating the expected prediction error $Err = E\left[L\left(\theta,\ \hat{f}(S)\right)\right]$, where $L(.)$ is the loss function and $\hat{f}(S)$ is the fitted regression model. Leave-one-out cross validation (LOOCV) is a common variant of cross validation, where we leave out the i^{th} observations and estimate the fitted regression model on the rest of the data. A computationally faster alternative is k-fold cross-validation (CV) where the data are partitioned into k subsets. In each of the k steps one specific subset is left out when fitting the function, and is used for validation instead. Here we use 10-fold cross-validation for estimating the prediction error.

The risk $\hat{R}_{CV}$ with any type of crodd validation is given as

$$\hat{R}_{CV} = \frac{1}{N}\sum_{i=1}^{N}\left(\theta_i - f_i(S_i)\right)^2,$$

with f_i denoting the estimate where the respective subset containing observation (S_i,θ_i) has been omitted.

From a computational point of view, it can also be advantageous not to carry out a cross-validation step at each iteration. One way of achieving this, is to choose a moderate number of instances m, at which cross-validation steps are carried out. To spread these instances out evenly, consider the L_1 norm w of the coefficient vector for the full least squares solution. Setting $x_j = j/m$ $(1 \leq j \leq m)$, a cross validation step is carried out each time the coefficient vector reaches one of the levels x_j^*. This strategy is available as an option within the R package *LARS* (Hastie and Efron, 2013).

We implemented our approach using the following algorithm for choosing summary statistics:

```
Algorithm 4: Choosing summary statistics for ABC
1. Take the sorted parameter values $\theta_1^*, \dots, \theta_N^*$, and the corresponding simulated summary statistics $S_i = [s'_{1i}, \dots, s'_{pi}]$ (1≤i≤N) from Algorithm 2.
2. Let $\theta^* := [\theta_1^*, \dots, \theta_r^*]$, where $r > p$ is a user defined cutoff.
3. Apply LARS (Algorithm 3) on the following multiple linear regression model $f(\theta^* | S') = \alpha + \beta_1 s'_1 + \beta_2 s'_2 + \cdots + \beta_p s'_p + \phi$, with residuals $\phi$
4. Define $x_j := \frac{j}{m}, 1 \le j \le m$, where $m$ is a user defined number of points at which cross validation (CV) is carried out;
5. Compute the CV prediction error at $x_j$;
   $\hat{R}_{CV}(x_k) = \frac{1}{r}\sum_{k=1}^{r}\left(\theta_k^* - \hat{f}_{k,x_j}(\theta^* | S_k')\right)^2$
   At the proportion $x_j$ of the full model, $\hat{f}_{k,x_j}(\theta^* | S')$ is the predicted value for θ when the $k^{th}$ observation is not used for fitting the model.
   Define $\hat{R}_{CV}^* := min_j\,[\hat{R}_{CV}(x_j)]$, and calculate the cutoff
   $x_j^* = \arg min_j [\hat{R}_{CV}(x_j)]$
6. At the cutoff $x_j^*$, if $|\hat{\beta}_p(x_j^*)| > 0$ , then select $s'_p$ as a summary statistic, otherwise reject $s'_p$.
```

In our simulations, we observed an improved performance of the above algorithm when modifying step 5 using the one standard error rule (‘1 SE rule’) as a stopping cut-off (see Breiman et al., 1984; Hastie et al., 2009): This slightly more parsimonious strategy calculates the smallest cutoff x_o such that

$$\hat{R}_{CV}(x_o) \le \hat{R}_{CV}^* + SE[\hat{R}_{CV}^*].$$

In the following section we consider two examples and evaluate the performance of our proposed method and compare it in particular to PLS, another computationally fast method.

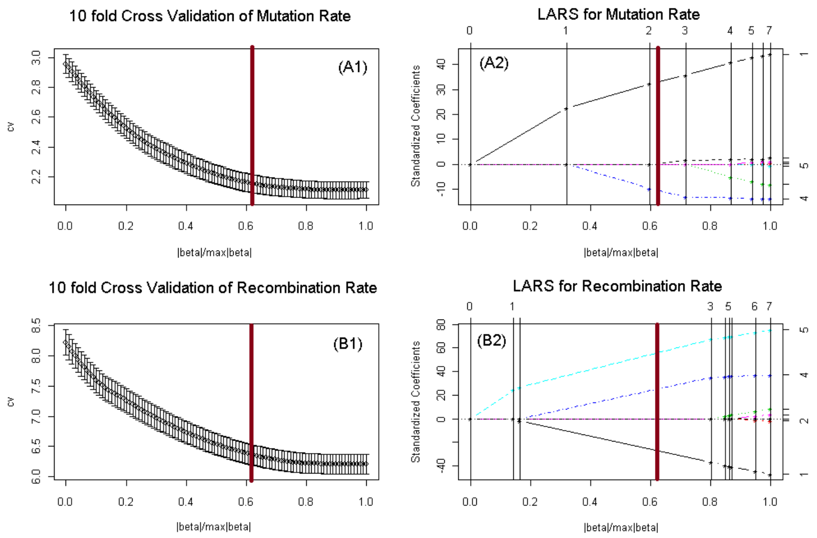
3 Simulation Results

3.1 Example 1: Estimation of the Mutation and Recombination Rates

The setup of our simulation study is similar to studies done previously (see Joyce and Marjoram, 2008; Nunes and Balding, 2010). The parameters are the scaled mutation and recombination rates, θ and ρ respectively. Each simulated data set consists of 50 haplotypes generated by using the ms software (Hudson, 2002) under the standard coalescent infinite-sites (IS) model (Nordborg, 2007). We chose the prior distribution for the scaled mutation rate as $\theta \sim U(2, 10)$, and $\rho \sim U(0, 10)$ for the scaled recombination rate. We computed seven summary statistics (see the appendix for details on the summary statistics). To carry out ABC, we used the R packages "abctools" (see Nunes and Balding, 2010) and "abc" (see Csilléry et al., 2012). Further parameters were chosen as follows: the number of ABC simulation runs $N = 10^6$, and the number of observed data sets $q = 10^2$. Furthermore, we used 1% as our acceptance cutoff ($r = 0.01 * N = 10000$) and the Euclidean distance for our metric $d(.)$. To carry out least angle regression, the R package *LARS* (Hastie and Efron, 2013) has been used.

We now discuss the accuracy of the resulting estimates of the mutation and the recombination rate.

Figure 1: Choosing summary statistics for mutation and recombination rate by using LARS



For Figure 1, the number of iteration is $N = 10^4$, and $m = 100$. This figure consists of four plots (A1, A2, B1, B2). In all these plots, solid vertical lines indicate the model complexity selected by the algorithm. For comparison purposes, the x-*axis* is normalized in the same way for all plots (range of coefficients 0 - 1). The plots A1 and B1 display the 10-fold cross validation prediction error both for the mutation and recombination rate. The plots A2 and B2 show at which stages the predictors enter the model. In plot A2, summary statistics s_1' and s_4' have been entered before the cutoff, and therefore will be used for estimating the mutation rate. Similarly, for estimating the recombination rate, s_1', s_4', and s_5' have been chosen by the algorithm in this particular example (see plot B2).

The summary statistic s_2' has been chosen by generating independent uniform random numbers. As s_2' and the responses are independent it makes sense that s_2' is included in neither set of summary statistics. As the summary statistics s_1'

(number of segregating sites) provides important information on θ (Hudson, 1990; Nordborg, 2007) and s_5' (number of distinct haplotype) important information on ρ, it is natural that they are included in the optimal sets of summary statistics (Nunes and Balding, 2010).

Table 1: Performance of PLS, and LARS methods, by MRSSE

PAR	s_1'	s_2'	s_3'	s_4'	s_5'	s_6'	s_7'	**All6**	**PLS**	**LARS**
θ	1.75	3.27	2.26	3.15	2.33	2.89	2.45	1.89	1.85	**1.75**
ρ	3.93	3.95	3.93	3.92	3.83	3.84	3.88	3.60	3.56	**3.46**

In Table 1, the performance of LARS is compared to that of other approaches in terms of the MRSSE. Additionally, the first seven columns ($s_1' - s_7'$) state the performance when only a single summary statistic is used; column eight (All6) shows the MRSSE when all summary statistics except the uninformative statistic s_2' are used together. The last two columns show the results for LARS and PLS. From Table 1 we can conclude that the sets of summary statistics selected by LARS produce—on average—the most accurate estimates.

Table 2: MRSSE with All6, PLS, and LARS for different choices of the acceptance cutoff, both with and without regression adjustment.

Acceptance	**Regression**	**Mutation (θ)**			**Recombination (ρ)**		
Cutoff (r)	**Adjustment**	**All6**	**PLS**	**LARS**	**All6**	**PLS**	**LARS**
1000	**No Adj**	1.804	1.786	1.743	3.480	3.525	3.342
	Mean	1.723	1.763	1.738	3.294	3.510	3.291
	Mean + Var	1.689	1.755	1.738	3.200	3.501	3.261
5000	No Adj	1.858	1.824	1.751	3.563	3.545	3.425
	Mean	1.737	1.771	1.743	3.317	3.518	3.314
	Mean + Var	1.701	1.750	1.740	3.209	3.487	3.240
10000	**No Adj**	1.890	1.849	1.754	3.604	3.559	3.464
	Mean	1.744	1.776	1.743	3.326	3.524	3.320
	Mean + Var	1.701	1.747	1.738	3.212	3.484	3.230
20000	No Adj	1.931	1.892	1.766	3.647	3.579	3.521
	Mean	1.752	1.786	1.747	3.330	3.530	3.327
	Mean + Var	1.701	1.745	1.737	3.218	3.478	3.220
30000	**No Adj**	1.959	1.925	1.776	3.675	3.593	3.561
	Mean	1.757	1.793	1.750	3.333	3.535	3.332
	Mean + Var	1.701	1.741	1.737	3.222	3.475	3.215
40000	No Adj	1.983	1.955	1.786	3.694	3.605	3.591
	Mean	1.762	1.799	1.753	3.335	3.538	3.336
	Mean + Var	1.701	1.739	1.737	3.226	3.473	3.225
50000	**No Adj**	2.004	1.981	1.795	3.709	3.614	3.614
	Mean	1.766	1.805	1.756	3.335	3.540	3.338
	Mean + Var	1.701	1.737	1.736	3.228	3.470	3.214
100000	No Adj	2.087	2.089	1.839	3.759	3.649	3.693
	Mean	1.781	1.827	1.769	3.341	3.550	3.346
	Mean + Var	1.698	1.737	1.733	3.247	3.466	3.222

In Table 2, both methods (PLS and LARS) are compared for different values of the acceptance cut off. Regression adjustment is also considered. With regression adjustment, the choice of the acceptance cut off becomes less important. This is since the adjustment applies corrections to the parameter points that increase with the distance measured in terms of the summary statistics. In general regression adjustment leads to an improved performance, both with LARS and PLS. Though smaller, there is still a slight advantage visible when using LARS instead of PLS. Also, in our example mean plus variance adjustment (Blum and François, 2009) leads to slightly better results than just mean adjustment (Beaumont et al., 2002).

3.2 Example 2: Estimation of Mutation, Recombination, Migration and Time Parameters.

This example is on population genetic inference under a model that includes demography: two subpopulations that split in the past with migration occurring between them. We consider the estimation of four parameters; the mutation rate θ, the recombination rate ρ, the migration rate θ_m, and the time η_c at which sub-population 2 and sub-population 1 have split. The *ms* (Hudson, 2002) software is again used to generate data sets that consist of 50 haplotypes. The prior distributions for the parameters were chosen as $\theta \sim U(0, 10)$, $\rho \sim U(0, 10)$, $\theta_m \sim U(0, 0.4)$, and $\eta_c \sim U(0.5, 0.9)$. Twenty-nine summary statistics have been calculated using msABC (see Pavlidis et al., 2010), and three uniform random variables (see appendix) unrelated to the parameters are added to this set of summary statistics. We compare PLS with LARS using $N = 10^6$ simulation runs, $r = 500$ accepted observations, and $q = 10^2$ different data sets . Thus we tried ABC with $N = 10^6$ runs on each of $q = 10^2$ data sets. As before, we used the Euclidean distance as our metric $d(.)$.

Table 3: Comparison of PLS and LARS methods, by MRSSE.

Summary statistics	$\boldsymbol{\theta}$	$\boldsymbol{\rho}$	$\boldsymbol{\theta_m}$	$\boldsymbol{\eta_c}$
s'_1	1.875	3.479	0.148	0.151
s'_2	1.893	3.480	0.149	0.152
s'_3	1.528	3.488	0.153	0.152
s'_4	2.025	3.484	0.148	0.151
s'_5	2.058	3.456	0.149	0.151
s'_6	1.733	3.468	0.153	0.148
s'_7	1.876	3.479	0.148	0.151
s'_8	1.894	3.480	0.149	0.152
s'_9	1.528	3.488	0.153	0.152
s'_{10}	3.023	3.480	0.152	0.152
s'_{11}	2.961	3.485	0.152	0.153
s'_{12}	3.113	3.470	0.153	0.149
s'_{13}	2.959	3.398	0.151	0.153
s'_{14}	2.951	3.418	0.151	0.152
s'_{15}	3.006	3.446	0.152	0.151
s'_{16}	3.167	3.514	0.148	0.154
s'_{17}	2.296	3.443	0.132	0.155
s'_{18}	2.213	3.563	0.145	0.151
s'_{19}	3.006	3.507	0.145	0.155
s'_{20}	3.167	3.514	0.148	0.154
s'_{21}	3.077	3.483	0.151	0.153
s'_{22}	3.122	3.525	0.153	0.153
s'_{23}	3.196	3.515	0.152	0.153
s'_{24}	2.089	3.229	0.151	0.152
s'_{25}	2.307	3.296	0.151	0.152
s'_{26}	2.187	3.301	0.151	0.152
s'_{27}	2.353	3.354	0.151	0.152
s'_{28}	1.899	3.202	0.152	0.152
s'_{29}	2.084	3.289	0.152	0.152
s'_{30}	3.168	3.502	0.152	0.152
s'_{31}	3.168	3.502	0.152	0.152
s'_{32}	3.174	3.516	0.152	0.153
All 29	1.579	3.060	0.134	0.152
PLS	1.595	3.119	0.132	0.153
LARS	**1.536**	**3.042**	**0.129**	**0.149**

In Table 3, we present the estimates for the error (MRSSE) when estimating the four model parameters. Here, both PLS and LARS select from 32 individual summary statistics ($s'_1 - s'_{32}$) separately for each parameter. We also consider the use of all 29 informative summary statistics. Notice that the other three summary statistics (s'_{30}, s'_{31}, s'_{32}) have been chosen as random numbers, unrelated to the actual data. In Table 3, bold indicates the lowest value in each column. LARS produced slightly better results than PLS.

4 Discussion

For implementing ABC reliably, an appropriate choice of summary statistics is crucial. We propose a new approach for this purpose that uses least angle regression (LARS) in combination with cross validation. It is computationally fast, and related to LASSO which is a popular approach for selecting sparse sets of coefficients for a large set of potential variables. We compared our approach to partial least squares (PLS, Wegmann et al., 2010), another computationally fast method for choosing summary statistics. In our simulations, least angle regression performed slightly better than PLS.

Several other methods are available, such as approximate sufficiency (Joyce and Marjoram, 2008), maximum entropy (Nunes and Balding, 2010), avarages over neural networks (Blum and Tran, 2010), a semi-automatic approach (Fearnhead and Prangle, 2012). These methods tend to be computationally more expensive, making them less attractive when the goal is to choose from a large set of candidate summary statistics (say greater than 10).

Applications where large sets of potential summary statistics often occur is population genetics (up to a few 100 for instance when allele frequency spectra are involved). Thus we illustrated our approach in the context of two population genetic examples with different levels of complexity.

A limitation of our approach may be that we consider only one parameter a time as response. This seems appropriate when aiming for marginal posteriors, but does not permit to investigate the joint distribution of several parameters.

However, any version of ABC will suffer from the curse of dimensionality at least when trying to explore high dimensional joint distributions of several parameters.

Furthermore, this study also demonstrates that mean and variance regression adjustment can help to make ABC less sensitive with respect to the choice of an acceptance cutoff (see Table 2). While we assumed a linear relationship between parameter and summary statistics, it would be interesting to explore also nonlinear relationships.

Appendix

List of Summary Statistics for Example 1

Statistic	**Description**
s_1'	No. of segregating sites
s_2'	Uniform [0,25] random variable
s_3'	Mean no. of differences over all pairs of haplotypes
s_4'	25*(mean r^2 across pairs separated by <10% of the simulated genomic region)
s_5'	No. of distinct haplotypes
s_6'	Frequency of the most common haplotype
s_7'	No. of singleton haplotypes

List of Summary Statistics for Example 2

Statistic	**Description**
s_1'	number of segregating sites for sub-population 1
s_2'	number of segregating sites for sub-population 2
s_3'	number of segregating sites for total sample
s_4'	Tajima’s π pi for sub-population 1
s_5'	Tajima’s π for sub-population 2
s_6'	Tajima’s π for total sample
s_7'	Watterson’s estimator for sub-population 1
s_8'	Watterson’s estimator for sub-population 2
s_9'	Watterson’s estimator for total sample
s_{10}'	Tajima's D for sub-population 1
s_{11}'	Tajima's D for sub-population 2

s'_{12}	Tajima's D for total sample
s'_{13}	the Zns for sub-population 1
s'_{14}	the Zns for sub-population 2
s'_{15}	the Zns for total sample
s'_{16}	the Fst (total sample, hbk calculation)
s'_{17}	the percentage of shared polymorphisms between sub-populations 1 and 2
s'_{18}	the percentage of private polymorphisms between sub-populations 1 and 2
s'_{19}	the percentage of fixed difference polymorphisms between sub-populations 1 and 2
s'_{20}	the Fst between sub-populations 1 and 2
s'_{21}	H in sub-population 1
s'_{22}	H in sub-population 2
s'_{23}	H in total sample
s'_{24}	the number of haplotypes in sub-population 1
s'_{25}	the heterozygosity of haplotypes in sub-population 1
s'_{26}	the number of haplotypes in sub-population 2
s'_{27}	the heterozygosity of haplotypes in sub-population 2
s'_{28}	the number of haplotypes in the total sample
s'_{29}	the Heterozygosity of haplotypes in the total sample
s'_{30}	Uniform [0,1] random variable
s'_{31}	Uniform [0,10] random variable
s'_{32}	Uniform [0,25] random variable

For a further description of these summary statistics see (Pavlidis et al., 2010).

Acknowledgment

Muhammad Faisal gratefully acknowledges the financial support by Higher Education Commission (HEC), of Pakistan during the entire duration of PhD. Mitwali Abd-el.Moemen is thankful to the College Deanship of Scientific Research, King Saud University, Riyadh Saudi Arabia for funding research through research group project RGP-VPP-280.

References

1. Beaumont M. A.,Zhang W., Balding D. J., Approximate Bayesian Computation in Population Genetics, Genetics 162 (4) (2002) 2025-2035.
2. Beaumont M. A., Cornuet J.-M., Marin J.-M., Robert C. P., Adaptive approximate Bayesian computation, Biometrika 96 (4) (2009) 983-990,
3. Blum M. G. B., François O., Non-linear regression models for Approximate Bayesian Computation, Statistics and Computing 20 (1) (2009) 63-73.

4. Blum M. G. B., Tran V. C., HIV with contact tracing: a case study in approximate Bayesian computation. Biostatistics (Oxford, England) 11 (4) (2010) 644-660.
5. Blum M. G. B., Nunes M. A., Prangle D., Sisson S. A., A comparative review of dimension reduction methods in approximate Bayesian computation (2012) 1-48.
6. Breiman L., Friedman J., Stone J. C., Olshen R., Classication and regression trees, Wadsworth International Group, Wadsworth, 1st edn., (1984).
7. Cohen R., Introducing the GLMSELECT Procedure for Model Selection, in: Proceedings of the Thirty-First Annual SAS Users Group International Conference, 2006.
8. Csilléry K., Blum M. G. B., O. E. Gaggiotti, O. François, Approximate Bayesian Computation (ABC) in practice, Trends in Ecology & Evolution 25 (7) (2010) 410-418.
9. Csilléry K., François O., Blum M. G. B., abc: An R package for approximate Bayesian computation, Methods in Ecology and Evolution in press. .
10. Efron B., Hastie T., Johnstone L., Tibshirani R., Least Angle Regression, The Annals of Statistics 32 (2) (2004) 407-451.
11. Faisal M., Futschik A., and Hussain I. A new approach to choose acceptance cutoff for approximate Bayesian computation. Journal of Applied Statistics 40 (4) (2013) 862-269.
12. Fearnhead P.,Prangle D., Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74 (3) (2012) 419-474.
13. Hastie T.,Tibshirani R., Friedman H. J., The elements of statistical learning: data mining, inference, and prediction, Springer Science Business Media, LLC, New York, 2nd edn., 2009.
14. Hudson R. R., Gene genealogies and the coalescent process, Oxford Survey Evol. Biol. 7 (1) (1990) 1-44.
15. Hudson R. R., Generating samples under a Wright-Fisher neutral model of genetic variation., Bioinformatics (Oxford, England) 18 (2) (2002) 337-338.
16. Joyce P., Marjoram P., Approximately sufficient statistics and Bayesian computation., Statistical applications in genetics and molecular biology 7 (1) (2008) Article 26.
17. Kohavi K., A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in: International Joint Conference on Artificial Intelligence (1995) 1137-1145.
18. Marjoram P., Tavaré S., Modern computational approaches for analysing molecular genetic variation data., Nature reviews. Genetics 7 (10) (2006) 759-770.
19. Marjoram P., Molitor J., Plagnol V., Tavare S., Markov chain Monte Carlo without likelihoods., Proceedings of the National Academy of Sciences of the United States of America 100 (26) (2003) 15324-15328.
20. Mevik B. H., Wehrens R., The pls Package: Principal Component and Partial Least Squares Regression in R, Journal of Statistical Software 18 (2) (2012).
21. Nordborg M., Coalescent theory , in: Handbook of Statistical Genetics, Wiley: Chichester, 3rd edn., 179-208, 2007.
22. Nunes M. A., Balding D. J., On Optimal Selection of Summary Statistics for Approximate Bayesian Computation, Statistical Applications in Genetics and Molecular Biology 9 (1) (2010) Article 34.

23. Pavlidis P., Laurent S., Stephan W., msABC: a modification of Hudson's ms to facilitate multilocus ABC analysis, Molecular Ecology Resources 10 (4) (2010) 723-727.
24. Sisson S. A., Fan Y., Likelihood-Free MCMC, in: A. Gelman, S. Brooks, G. Jones, X.-L. Meng (Eds.), Handbook of Markov Chain Monte Carlo, chap. 12, Taylor & Francis US, (2010) 313-351.
25. Sisson S. A., Fan Y., Tanaka M. M., Sequential Monte Carlo without likelihoods., Proceedings of the National Academy of Sciences of the United States of America 104 (6) (2007) 1760-1765.
26. Wegmann D., Leuenberger C., Neuenschwander S., Exco-er L., ABCtoolbox: a versatile toolkit for approximate Bayesian computations., BMC bioinformatics 11 (1) (2010) 116.