A comparative review of dimension reduction methods in approximate Bayesian computation

M. G. B. Blum[*][†], M. A. Nunes[‡], D. Prangle[‡] and S. A. Sisson[§]

Abstract

Approximate Bayesian computation (ABC) methods make use of comparisons between simulated and observed summary statistics to overcome the problem of computationally intractable likelihood functions. As the practical implementation of ABC requires computations based on vectors of summary statistics, rather than full datasets, a central question is how to derive low dimensional summary statistics from the observed data with minimal loss of information. In this article we provide a comprehensive review and comparison of the performance of the principal methods of dimension reduction proposed in the ABC literature. The methods are split into three non-mutually exclusive classes consisting of best subset selection methods, projection techniques and regularisation. In addition, we introduce two new methods of dimension reduction. The first is a best subset selection method based on Akaike and Bayesian information criteria, and the second uses ridge regression as a regularisation procedure. We illustrate the performance of these dimension reduction techniques through the analysis of three challenging models and datasets.

Keywords: Approximate Bayesian computation; dimension reduction; likelihood-free inference; regularisation; variable selection.

[*] Corresponding Author: Email: michael.blum@imag.fr

[†] Université Joseph Fourier, Centre National de la Recherche Scientifique, Laboratoire TIMC-IMAG UMR 5525, Grenoble, F-38041, France.

[‡] Mathematics and Statistics Department, Fylde College, Lancaster University, Lancaster LA1 4YF, U.K.

[§] School of Mathematics and Statistics, University of New South Wales, Sydney 2052, Australia.

1 Introduction

Bayesian inference is typically focused on the posterior distribution $p(\theta|y_{obs}) \propto p(y_{obs}|\theta)p(\theta)$ of a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^q$, $q \geq 1$, representing the updating of one's prior beliefs, $p(\theta)$, through the likelihood (model) function, $p(y_{obs}|\theta)$, having observed data $y_{obs} \in \mathcal{Y}$. The term *approximate Bayesian computation* (ABC) refers to a family of models and algorithms that aim to draw samples from an approximate posterior distribution when the likelihood, $p(y_{obs}|\theta)$, is unavailable or computationally intractable, but where it feasible to quickly generate data from the model, $y \sim p(\cdot|\theta)$. ABC is rapidly becoming a popular tool for the analysis of complex statistical models in an increasing number and breadth of research areas. See e.g. Lopes and Beaumont (2009), Bertorelle et al. (2010), Beaumont (2010), Csilléry et al. (2010) and Sisson and Fan (2011) for a partial overview of the application of ABC methods.

ABC introduces two principal approximations to the posterior distribution. Firstly, the posterior distribution of the full dataset, $p(\theta|y_{obs})$, is approximated by $p(\theta|s_{obs}) \propto p(s_{obs}|\theta)p(\theta)$, where $s_{obs} = S(y_{obs})$ is a vector of summary statistics of lower dimension than the data y_{obs}. In this manner, $p(\theta|s_{obs}) \approx p(\theta|y_{obs})$ is a good approximation if s_{obs} is highly informative for the model parameters, and $p(\theta|s_{obs}) = p(\theta|y_{obs})$ if s_{obs} is sufficient. As $p(s_{obs}|\theta)$ is also likely to be computationally intractable if $p(y_{obs}|\theta)$ is computationally intractable, a second approximation is constructed as $p_{ABC}(\theta|s_{obs}) = \int p(\theta, s|s_{obs})ds$, with

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

where $K_\epsilon(\|u\|) = K(\|u\|/\epsilon)/\epsilon$ is a standard smoothing kernel with scale parameter $\epsilon > 0$. As a result of (1), approximating the target $p(\theta|s_{obs})$ by $p_{ABC}(\theta|s_{obs})$ can be shown to be a good approximation if the kernel scale parameter, ϵ, is small enough, following standard kernel density estimation arguments (e.g. Blum 2010a).

In combination, both approximations allow for practical methods of sampling from $p_{ABC}(\theta|s_{obs})$

that avoid explicit evaluation of the intractable likelihood function, $p(y_{obs}|\theta)$. A simple rejection-sampling algorithm to achieve this was proposed by Pritchard et al. (1999) (see also Marjoram et al. 2003), which produces draws from $p(\theta, s|s_{obs})$. In general terms, an importance-sampling version of this algorithm proceeds as follows:

1. Draw a candidate parameter vector from the prior, $\theta' \sim p(\theta)$;

2. Draw summary statistics from the model $s' \sim p(s|\theta')$;

3. Assign to (θ', s') a weight, w', that is proportional to $K_\epsilon(\|s' - s_{obs}\|)$.

Here, the sampling distribution for (θ', s') is the prior predictive distribution, $p(s|\theta)p(\theta)$, and the target distribution is $p(\theta, s|s_{obs})$. Using equation (1), it is then straightforward to compute the importance weight for the pair (θ', s'). The weight is proportional to $p(\theta', s'|s_{obs})/[p(s'|\theta')p(\theta')] = K_\epsilon(\|s' - s_{obs}\|)$, which is free of intractable likelihood terms, $p(s'|\theta')$. The manner by which the intractable likelihoods cancel between sampling and target distributions forms the basis for the majority of ABC algorithms.

Clearly, both ABC approximations to the posterior distribution help to avoid the computational intractability of the original problem. The first approximation allows the kernel weighting of the second approximation, $K_\epsilon(\|s - s_{obs}\|)$, to be performed on a lower dimension than that of the original data, y_{obs}. Kernel smoothing is known to suffer from the curse of dimensionality (e.g. Blum 2010a), and so keeping $\dim(s) \leq \dim(y)$ as small as possible helps to improve algorithmic efficiency. The second approximation (1) allows the sampler weights (or acceptance probabilities, if one considers rejection-based samplers, such as Markov chain Monte Carlo) to be free of intractable likelihood terms.

In practice, however, there is typically a tradeoff between the two approximations: If the dimension of s is large so that the first approximation, $p(\theta|s_{obs}) \approx p(\theta|y_{obs})$ is good, the second approximation may then be poor due to the inefficiency of kernel smoothing in large dimensions. Conversely, if the dimension of s is small, while the second approximation (1) will be good (with a small kernel scale parameter, ϵ), any loss of information in the mapping $s_{obs} = S(y_{obs})$ means that the first approximation may be poor. Naturally, a low-dimensional and near-sufficient statistic, s, would provide a near-optimal and balanced choice.

For a given set of summary statistics, much work has been done on deriving more efficient sampling algorithms to reduce the effect of the second approximation by allowing a smaller value for the kernel scale parameter, ϵ, which in turn improves the approximation $p_{ABC}(\theta|s_{obs}) \approx p(\theta|s_{obs})$. The greater the algorithmic efficiency, the smaller the scale parameter that can be achieved for a given computational burden. These algorithms include Markov chain Monte Carlo (Marjoram et al. 2003; Bortot et al. 2007) and sequential Monte Carlo techniques (Sisson et al. 2007; Toni et al. 2009; Beaumont et al. 2009; Drovandi and Pettitt 2011; Peters et al. 2012; Del Moral et al. 2012). By contrast, the regression-based methods described in Section 2.1 do not aim at reducing the scale parameter ϵ but rather explicitly account for the imperfect match between observed and simulated summary statistics (Beaumont et al. 2002; Blum and François 2010),

Achieving a good tradeoff between the two approximations revolves around the identification of a set of summary statistics, s, which are both low-dimensional and highly informative for θ. A number of methods, primarily based on dimension reduction ideas, have been proposed to achieve this (Joyce and Marjoram 2008; Wegmann et al. 2009; Nunes and Balding 2010; Blum and François 2010; Blum 2010b; Fearnhead and Prangle 2012). The choice of summary statistics is one of the most important aspects of a statistical analysis using ABC methods (along with the choice of algorithm). Poor specification of s can have a large and detrimental impact on both ABC model approximations.

In this article we provide the first detailed review and comparison of the performance of the current methods of dimension reduction for summary statistics within the ABC framework. We characterise these methods into three non-mutually exclusive classes: (i) best subset selection, (ii) projection techniques and (iii) regularisation approaches. As part of this analysis, we introduce two additional novel techniques for dimension reduction within ABC. The first adopts the ideas of Akaike and Bayesian information criteria to the ABC framework, whereas the second makes use of ridge regression as a regularisation procedure for ABC. The dimension reduction methods are compared through the analysis of three challenging models and datasets. These involve the analysis of a coalescent model with recombination (Joyce and Marjoram 2008), an evaluation of the evolutionary fitness cost of mutation in drug-resistant tuberculosis (Luciani et al. 2009), and an assessment of the number and size-distribution of particle inclusions in the production of clean steels (Bortot et al. 2007).

The layout of this article is as follows: In Section 2 we classify and review the existing methods of summary statistic dimension reduction in ABC, and in Section 3, we outline our two additional novel methods. A comparative analysis of the performance of each of these methods is provided in Section 4. We conclude with a discussion.

2 Classification of ABC dimension reduction methods

In a typical ABC analysis, an initial collection of statistics $s^\top = (s_1, \ldots, s_p)$ is chosen by the modeller, the elements of which have the potential to be informative for the model parameters, $\theta^\top = (\theta_1, \ldots, \theta_q)$. Choice of these initial statistics is highly problem specific, and the number of candidate statistics, p, often considerably outnumbers the number of model parameters, q i.e. $p >> q$ (e.g. Bortot et al. 2007; Allingham et al. 2009; Luciani et al. 2009). For example, Bortot et al. (2007) and Allingham et al. (2009) use the ordered observations $S(y) = (s_{(1)}, \ldots, s_{(p)})$ so that there is no loss of information at this stage. The analysis then proceeds by either using all p statistics in full, or by attempting to reduce their dimension while minimising information loss. Note that the most suitable set of summary statistics for an analysis may be dataset dependent, as the information content of summary statistics may vary within the parameter space, Θ (an exception is when sufficient statistics are known). As such, any analysis should also consider establishing potentially different summary statistics when re-implementing any model with a different dataset.

Methods of summary statistics dimension reduction for ABC can be broadly classified into three non-mutually exclusive classes. The first class of methods follows a *best subset selection* approach. Here, candidate subsets are evaluated and ranked according to various information-based criteria, such as measures of sufficiency (Joyce and Marjoram 2008) or the entropy of the posterior distribution (Nunes and Balding 2010). In this article we contribute additional criteria for this process derived from Akaike and Bayesian information criteria arguments. From these criteria, the highest ranking subset (or alternatively, a subset consisting of those summary statistics which demonstrate clear importance) is then chosen for the final analysis.

The second class of methods can be considered as *projection techniques*. Here, the dimension of $(s_1, \ldots, s_p)$ is reduced by considering linear or non-linear combinations of the summary statistics. These methods make use of a regression layer within the ABC framework, whereby the response variable, θ, is regressed by the (possibly transformed) predictor variables, s, (Beaumont et al. 2002; Blum and François 2010). These projection methods include partial least squares regression (Wegmann et al. 2009), feed-forward neural networks (Blum and François 2010) and regression guided by minimum expected posterior loss considerations (Fearnhead and Prangle 2012).

In this article we introduce a third class of methods for dimension reduction in ABC, based on *regularisation techniques*. Using ridge regression, we also make use of the regression layer between the parameter θ and the summary statistics, s. However, rather than explicitly considering selection of summary statistics, we propose to approach this implicitly, by shrinking the regression coefficients towards zero so that uninformative summary statistics have the weakest contribution in the regression equation.

In the remainder of this Section we discuss each of these methods in more detail. We first describe the ideas behind ABC regression adjustment strategies (Beaumont et al. 2002; Blum and François 2010), as many of the dimension reduction techniques build on this framework.

2.1 Regression adjustment in ABC

Standard ABC methods suffer from the curse of dimensionality in that the rate of convergence of posterior expectations with respect to $p_{ABC}(\theta|s_{obs})$ (such as the Nadaraya-Watson estimator of the posterior mean) decreases dramatically as the dimension of the summary statistics, p, increases (Blum 2010a). ABC regression adjustment (Beaumont et al. 2002) aims to avoid this by explicitly modelling the discrepancy between s and s_{obs}. When describing regression adjustment methods, for notational simplicity and clarity of exposition, we assume that the parameter of interest, θ, is univariate (i.e. $q = 1$). Regression adjustment methods may be readily applied to multivariate θ, by using a different regression equation for each parameter, $\theta_1, \ldots, \theta_q$, separately.

The simplest model for this is a homoscedastic regression in the region of s_{obs}, so that

$$\theta^i = m(s^i) + e^i,$$

where $(\theta^i, s^i) \sim p(s|\theta)p(\theta)$ are $i = 1, \ldots, n$ draws from the prior predictive distribution, $m(s^i) = \mathbb{E}[\theta|s = s^i]$ is the mean function, and the e^i are zero-mean random variates with common variance. To estimate the conditional mean $m(\cdot)$, Beaumont et al. (2002) assumed a linear model

$$m(s^i) = \alpha + \beta^\top s^i \tag{2}$$

in the neighborhood of s_{obs}. An estimate of the mean function, $\hat{m}(\cdot)$, is obtained by minimizing the weighted least squares criterion $\sum_{i=1}^n w^i \|m(s^i) - \theta^i\|^2$ where $w^i = K_\epsilon(\|s^i - s_{obs}\|)$.

A weighted sample from the posterior distribution, $p_{ABC}(\theta|s_{obs})$ is then obtained by the adjustment

$$\theta^{*i} = \hat{m}(s_{obs}) + (\theta^i - \hat{m}(s^i)) \tag{3}$$

for $i = 1, \ldots, n$. In the above, the kernel scale parameter ϵ controls the bias-variance tradeoff: Increasing ϵ reduces variance by increasing the effective sample size—the number of accepted simulations when using a uniform kernel K—but increases bias arising from departures from a linear mean function $m(\cdot)$ and homoscedastic error structure (Blum 2010a).

Blum and François (2010) proposed the more flexible, heteroscedastic model

$$\theta^i = m(s^i) + \sigma(s^i)e^i, \tag{4}$$

where $\sigma^2(s^i) = \mathbb{V}[\theta|s = s^i]$ denotes the conditional variance. This variance is estimated using a second regression model for the log of the squared residuals i.e. $\log(\theta^i - \hat{m}(s^i))^2 = \log\sigma^2(s^i) + \eta^i$, where the η^i are independent, zero-mean variates with common variance. The equivalent adjustment to (3) is then given by

$$\theta^{*i} = \hat{m}(s_{obs}) + \left[\theta^i - \hat{m}(s^i)\right] \frac{\hat{\sigma}(s_{obs})}{\hat{\sigma}(s^i)}, \tag{5}$$

where $\hat{\sigma}(s)$ denotes the estimate of $\sigma(s)$. The kernel scale parameter, ϵ, plays the same role as for the homoscedastic model, except with more flexibility on deviations from homoscedasticity. Nott et al. (2011) have demonstrated that regression adjustment ABC algorithms produce samples, $\{\theta^{*i}\}$, for which first- and second-order moment summaries approximate adjusted expectation and variance for a Bayes linear analysis. We do not describe here an alternative regression adjustment method where the summary statistics are rather considered as the dependent variables and the parameters as the independent variables of the regression (Leuenberger and Wegmann 2010).

2.2 Best subset selection methods

Best subset selection methods are conceptually simple, but are cumbersome to manage for large numbers of potential summary statistics, $s = (s_1, \ldots, s_p)$. Exhaustive enumeration of the $2^p - 1$ possible combinations of summary statistics is practically infeasible beyond a moderate value of p. This is especially true of Markov chain Monte Carlo or sequential Monte Carlo based analyses, which require one sampler implementation per combination. As a result, stochastic or deterministic (greedy) search procedures, such as forward or backward selection, are required to implement them.

2.2.1 A sufficiency criterion

The first principled approach to dimension reduction in ABC was the ε-sufficiency concept proposed by Joyce and Marjoram (2008), which was used to determine whether to include an additional summary statistic, s_k, to a model already containing statistics $s_1, \ldots, s_{k-1}$. Here, noting that the difference between the log likelihoods of $p(s_1, \ldots, s_k|\theta)$ and $p(s_1, \ldots, s_{k-1}|\theta)$ is $\log p(s_k|s_1, \ldots, s_{k-1}, \theta)$, Joyce and Marjoram (2008) defined the set of statistics $s_1, \ldots, s_{k-1}$ to be ε-sufficient relative to s_k if

$$\delta_k = \sup_\theta \log p(s_k|s_1, \ldots, s_{k-1}, \theta) - \inf_\theta \log p(s_k|s_1, \ldots, s_{k-1}, \theta) \leq \varepsilon. \tag{6}$$

Accordingly, if an estimate of δ_k (i.e. the "score" of s_k relative to $s_1, \ldots, s_{k-1}$) is greater than ε, then there is enough additional information content in s_k to justify including it in the model. In practice, Joyce and Marjoram (2008) implement a conceptually equivalent assessment, whereby s_k is added to the model if the ratio of posteriors

$$R_k(\theta) = \frac{p_{ABC}(\theta|s_1, \ldots, s_{k-1}, s_k)}{p_{ABC}(\theta|s_1, \ldots, s_{k-1})}$$

differs from one by more than some threshold value $T(\theta)$ for any value of θ. As such, a statistic s_k will be added to the model if the resulting posterior changes sufficiently at any point. The threshold, $T(\theta)$, is user-specified, with one particular choice described in Section 5 of Joyce and Marjoram (2008).

This procedure can be implemented within any stepwise search algorithm, each of which have various pros and cons. Following the definition (6), the resulting optimal subset of summary statistics is then ε-sufficient relative to each one of the remaining summary statistics. Here ε intuitively represents an acceptable error in determining whether s_k contains further useful information in addition to $s_1, \ldots, s_k$. This quantity is also user-specified, and so the final optimal choice of summary statistics will depend on the chosen value.

Sensitivity to the choice of ε aside, this approach may be criticised in that it assumes that every change to the posterior obtained by adding a statistic, s_k, is beneficial. It is conceivable that attempting to include a completely non-informative statistic where the observed statistic is unlikely to have been generated under the model, will result in a sufficiently modified posterior as measured by ε, but one which is more biased away from the true posterior $p(\theta|y_{obs})$ than without including s_k. A toy example illustrating this was given by Sisson and Fan (2011).

A further criticism is that the amount of computation required to evaluate $R_k(\theta)$ for all θ, and on multiple occasions, is considerable, especially for large q. In practice, Joyce and Marjoram (2008) considered θ to be univariate, and approximated continuous θ over a discrete grid in order to keep computational overheads to acceptable levels. As such, this method appears largely restricted to dimension reduction for univariate parameters ($q = 1$).

2.2.2 An entropy criterion

Nunes and Balding (2010) propose the entropy of a distribution as a heuristic to measure the informativeness of candidate combinations of summary statistics. Since entropy measures information and a lack of randomness (Shannon and Weaver 1948), the authors propose minimising the entropy of the approximate posterior, $p_{ABC}(\theta|s_{obs})$, over subsets of the summary statistics, s, as a proxy for determining maximal information about a parameter of interest. High entropy results from a diffuse posterior sample, whereas low entropy is obtained from a posterior which is more precise in nature.

Nunes and Balding (2010) estimate entropy using the unbiased k-th nearest neighbour estimator of Singh et al. (2003). For a weighted posterior sample, $(w^1, \theta^1), \ldots, (w^n, \theta^n)$, where $\sum_i w^i = 1$, this estimator can be written as

$$\hat{E} = \log\left[\frac{\pi^{q/2}}{\Gamma(q/2{+}1)}\right] - \psi(k) + \log n + q\sum_{i=1}^{n} w^i \log \hat{C}_i^{-1}(k/(n-1)), \tag{7}$$

where $q = \dim(\theta)$, $\psi(x) = \Gamma'(x)/\Gamma(x)$ denotes the digamma function, and where $\hat{C}_i(\cdot)$ denotes the empirical distribution function of the Euclidean distance from θ^i to the remainder of the weighted posterior sample i.e. of the weighted samples $\{(\tilde{w}^j, \|\theta^i - \theta^j\|)\}_{j\neq i}$, where $\tilde{w}^j = w^j / \sum_{j\neq i} w^j$. Following Singh et al. (2003), the original work of Nunes and Balding (2010) used $k = 4$, and was based on an equally weighted posterior sample (i.e. with $w^i = 1/n, i = 1, \ldots, n$), so that $\hat{C}_i^{-1}(k/(n-1))$ denotes the Euclidean distance from θ^i to its k-th closest neighbour in the posterior sample $\{\theta^1, \ldots, \theta^{i-1}, \theta^{i+1}, \ldots, \theta^n\}$.

While minimum entropy could in itself be used to evaluate the informativeness of a vector of summary statistics for θ (although see the criticism of entropy below), Nunes and Balding (2010) propose a second stage to their analysis, which aims to assess the performance of a candidate set of summary statistics using a measure of posterior error. For example, when the true parameter vector, θ_{true}, is known, the authors suggest the root sum of squared errors (RSSE), given by

$$\text{RSSE} = \left(\sum_{i=1}^{n} w^i \|\theta^i - \theta_{true}\|^2\right)^{1/2}, \tag{8}$$

where the measure $\|\theta^i - \theta_{true}\|$ compares the components of θ on a suitable scale (and so some component-wise standardisation may be required). Naturally, the true parameter value, θ_{true}, is unknown in practice. However, if the simulated summary statistics from the samples (θ^i, s^i) are treated as observed data, it is clear that $\theta_{true} = \theta^i$ for the posterior $p_{ABC}(\theta|s^i)$. As such, the RSSE can be easily computed with a leave-one-out technique.

As the subset of summary statistics that minimises (8) will likely vary over observed datasets, s^i, Nunes and Balding (2010) propose minimising the average RSSE over some number of simulated datasets which are close to the observed, s_{obs}. To avoid circularity, Nunes and Balding (2010) define these "close" datasets to be the $j = 1, \ldots, n^*$ simulated datasets, $\{s^j\}$, that minimise $\|s^j_{ME} - s_{ME}\|$, where s^j_{ME} and s_{ME} are the vectors of minimum entropy summary statistics computed via (7) from s^j and the observed summary statistics, s_{obs}, respectively. That is, the quantity

$$\overline{\text{RSSE}} = \frac{1}{n^*} \sum_{j=1}^{n^*} \text{RSSE}_j, \tag{9}$$

is minimised (over subsets of summary statistics), where RSSE_j corresponds to (8) using the simulated dataset s^j.

This approach is intuitive, and is attractive because the second stage directly measures error in the posterior with respect to a known truth, θ_{true}, which is not typically considered in other ABC dimension reduction approaches, albeit at the extra computational expense of a two-stage procedure. A weakness of the first stage however, is the assumption that addition of an informative statistic will reduce the entropy of the resulting posterior distribution. An example of when this does not occur is when the posterior distribution is diffuse with respect to the prior – for instance, if an overly precise prior is located in the distributional tails of the posterior (e.g. Jeremiah et al. 2011). In this case, attempting to include an informative additional statistic, s_k, can result in a distribution that is more diffuse than with s_k excluded. As such, the entropic approach is therefore mostly suited to models with relatively diffuse prior distributions. Another potential criticism of the first stage is that minimising the entropy does not necessarily provide the *minimal* subset of sufficient statistics. This provides an argument for considering the mutual information between θ and s, rather than the entropy (Barnes et al. 2012; see also Filippi et al. 2012). However, it is clear that the overall approach of Nunes and Balding (2010) could easily be implemented with alternative first-stage selection criteria.

2.2.3 AIC and BIC criteria

Information criteria based on Akaike and Bayesian information are natural best subset selection techniques for summary statistic dimension reduction in ABC analyses. We introduce and develop these criteria in Section 3.1.

2.3 Projection techniques

Selecting a best subset of summary statistics from $s = (s_1, \ldots, s_p)$ suffers from the problem that it may require several statistics to provide the same information content as a single, highly informative statistic that was not specified in the initial set, s. To avoid this, projection techniques aim to combine the elements of s through linear or non-linear transformations, in order to construct a potentially much lower dimensional set of highly informative statistics.

One of the main advantages of projection techniques is that, unlike best subset selection methods, they scale well with increasing numbers of summary statistics. They can handle large numbers of possibly uninformative summary statistics, in addition to accounting for high levels of inter-dependence and multicollinearity. A minor disadvantage of projection techniques is that the final sets of projected summary statistics typically (but not universally) lack interpretability. In addition, most projection methods require the specification of a hyperparameter that governs the number of projections to perform.

2.3.1 Partial least squares regression

Partial least squares regression seeks the orthogonal linear combinations of the explanatory variables which have high variance and high correlation with the response variable (e.g. Boulesteix and Strimmer 2007; Vinzi et al. 2010; Abdi and Williams 2010). Wegmann et al. (2009) proposed the use of partial least squares regression for dimension reduction in ABC, where the explanatory variables are the suitably (e.g. Box-Cox) transformed summary statistics, s, and the response variables is the parameter vector, θ.

The output of a partial least squares analysis is the set of k orthogonal components of the regression design matrix

$$X = \begin{pmatrix} 1 & s_1^1 & \cdots & s_p^1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & s_1^n & \cdots & s_p^n \end{pmatrix} \tag{10}$$

that are optimally correlated (in a specific sense) with θ. Here, s_j^i denotes the j^{th} component of the i^{th} simulated summary statistic, s^i. To choose the appropriate number of orthogonal components, Wegmann et al. (2009) examine the root mean square error of θ for each value of k, as estimated by a leave-one-out cross validation strategy. For a fixed number of components, k, this corresponds to

$$\text{RMSE}_k = \left(\frac{1}{n} \sum_{i=1}^{n} \| \hat{m}_k^{-i}(s^i) - \theta^i \|^2 \right)^{1/2}, \tag{11}$$

where $\hat{m}_k^{-i}(s)$ denotes the mean response of the partial least squares regression, estimated without the i-th simulated summary statistic, s^i (e.g. Mevik and Cederkvist 2004). The optimal number of components is then chosen by inspection of the RMSE_k values, based on minimum gradient change arguments (e.g. Mevik and Wehrens 2007).

A potential disadvantage of partial least squares regression, as performed by Wegmann et al. (2009), is that it aims to infer a global linear relationship between θ and s based on draws from the prior predictive distribution, $p(s|\theta)p(\theta)$. This may differ from the relationship observed in the region around $s = s_{obs}$, and as such may produce unsuitable orthogonal components as a result. A workaround for this would be to follow Fearnhead and Prangle (2012) (see Section 2.3.3) and elicit the relationship between θ and s based on samples from a truncated prior $(\theta^i, s^i) \sim p(s|\theta)p(\theta)I(\theta \in \Theta^R)$, where $\Theta^R \subset \Theta$ restricts the samples, θ^i, to regions of significant posterior density. One simple way to identify such a region is through a pilot ABC analysis (Fearnhead and Prangle 2012).

2.3.2 Neural networks

In the regression setting, feed-forward neural networks can be considered as a non-linear generalisation of the partial least squares regression technique described above. Blum and François (2010) proposed the neural network as a machine learning approach to dimension reduction by estimating the conditional mean and variance functions, $m(\cdot)$ and $\sigma^2(\cdot)$ in the non-linear, heteroscedastic regression adjustment model (4) – see Section 2.1.

The neural network reduces the dimension of the summary statistics to $H < p$, using H hidden units in the network, $z_1, \ldots, z_H$, defined as

$$z_j = h\left(\sum_{k=1}^{p} \omega_{jk}^{(1)} s_k + \omega_{j0}^{(1)}\right), \tag{12}$$

for $j = 1, \ldots, H$. The $\omega_{jk}^{(1)}$ terms are the weights of the first layer of the neural network, and $h(\cdot)$ is a non-linear function, typically the logistic function. The reduced and non-linearly transformed summary statistics of the hidden units, z_j, are then combined through the regression function of the neural network

$$m(s) = g\left(\sum_{j=1}^{H} \omega_j^{(2)} z_j + \omega_0^{(2)}\right), \tag{13}$$

where $\omega_j^{(2)}$ denotes the weights of the second layer of the neural network and $g(\cdot)$ is a link function. A similar neural network is used to model $\log \sigma^2(s)$ (e.g. Nix and Weigend 1995), with the possibility of allowing for a different number of hidden units to estimate heteroscedasticity in the regression adjustment compared to that in the mean function $m(s)$.

Rather than dynamically determining the number of hidden units H, Blum and François (2010) propose to specify a fixed value, such as $H = q$ where $q = \dim(\theta)$ is the number of parameters to infer. The weights of the neural network are then obtained by minimising the regularised least-squares criterion

$$\sum_{i=1}^{n} w^i \|m(s^i) - \theta^i\|^2 + \lambda\|\omega\|^2,$$

where ω is the vector of all weights in the neural network model for $m(s)$, $w^i = K_\epsilon(\|s^i - s_{obs}\|)$ is the weight of the sample $(\theta^i, s^i) \sim p(s|\theta)p(\theta)$, and $\lambda > 0$ denotes the regularisation parameter (termed the weight-decay parameter for neural networks). The idea of regularisation is to shrink the weights towards zero so that only informative summary statistics contribute in the model (12) and (13) for $m(s)$. Following the estimation of $m(s)$, a similar regularisation criterion is used to estimate $\log \sigma^2(s)$. Both mean and variance functions can then be used in the regression adjustment of equation (5).

2.3.3 Minimum expected posterior loss

Fearnhead and Prangle (2012) proposed a decision-theoretic dimension reduction method with a slightly different aim to previous dimension reduction approaches. Here, rather than constructing appropriate summary statistics to ensure that $p_{ABC}(\theta|s_{obs}) \approx p(\theta|y_{obs})$ is a good approximation, $p_{ABC}(\theta|s_{obs})$ is alternatively required to be a good approximation in terms of the accuracy of specified functions of the model parameters. In particular, assuming that interest is in point estimates of the model parameters, if θ_{true} denotes the true parameter value and $\hat{\theta}$ an estimate, then Fearnhead and Prangle (2012) propose to choose those summary statistics that minimise the quadratic loss

$$L(\theta_{true}, \hat{\theta}) = (\theta_{true} - \hat{\theta})^\top A(\theta_{true} - \hat{\theta}),$$

for some $p \times p$ positive-definite matrix A. This loss is minimised for $s_{obs} = E_{p(\theta|y_{obs})}(\theta)$, the true posterior mean.

To estimate $E_{p(\theta|y)}(\theta)$, Fearnhead and Prangle (2012) propose least-squares regression models for the $k = 1, \ldots, q$ model parameters, $(\theta_1, \ldots, \theta_q)$, given by

$$\theta_k^i = E_{p(\theta|y)}(\theta_k) + \eta_k^i = \alpha_k + \beta_k^\top f(s^i) + \eta_k^i \tag{14}$$

where $(\theta^i, s^i) \sim p(s|\theta)p(\theta)$ are draws from the prior predictive distribution, α_k and β_k are unknown regression parameters to be estimated, and η_k^i denotes a zero-mean noise process. Here $f(s)$ is a vector of potentially non-linear transformations of the data (i.e. of the original summary statistics). For example, in one application, Fearnhead and Prangle (2012) use the polynomial basis functions $f(s) = (s, s^2, s^3, s^4)$; that is, a vector of length $4p$, where $p = \dim(s)$ is the number of elements in s, consisting of the first four powers of each element of s. The choice of $f(s)$ can be based on standard diagnostics of regression fit, such as BIC. If the prior $\pi(\theta)$ is diffuse with respect to the posterior, then one may estimate the regression model (14) based on samples from a truncated prior $(\theta^i, s^i) \sim p(s|\theta)p(\theta)I(\theta \in \Theta^R)$, where $\Theta^R \subset \Theta$ restricts the samples, θ^i, to regions of significant posterior density (e.g. via a pilot ABC analysis). Clearly, more sophisticated alternatives to least-squares regression may be used.

After fitting equation (14), the new, single summary statistic for the parameter θ_k is $\hat{\beta}_k^\top f(s)$, where $\hat{\beta}_k$ denotes the least squares estimate of β_k. The resulting q-dimensional vector of new summary statistics is then used in a standard ABC analysis. Fearnhead and Prangle (2012) show that these new statistics can lead to posterior inferences that considerably outperform inferences based on the original statistics, s. Nott et al. (2012) demonstrate that these summary statistics can be viewed as Bayes linear estimates of the posterior mean.

2.4 Regularisation approaches

Regularisation approaches aim to reduce overfitting in a model by penalising model complexity. A simple example where overfitting can occur in ABC is the standard regression adjustment (Beaumont et al. 2002; Section 2.1), where there is a risk of over adjusting the parameters, θ^i, in the direction of uninformative summary statistics via (3). Regularisation is used as part of the estimation of the neural network weights in the projection technique proposed by Blum and François (2010) (see Section 2.3.2). As such, the regression adjustment of Beaumont et al. (2002) is a procedure that could greatly benefit from the inclusion of regularisation techniques. We introduce the ridge regression adjustment to ABC in Section 3.2.

2.5 Other methods

There are a number of alternative approaches to dimension reduction for ABC, including methods that aim to circumvent the dimensionality issue, that we do not include in our comparative analysis (Section 4). Drovandi et al. (2011) proposed to adopt ideas from indirect inference (e.g. Heggland and Frigessi 2004) as a means to identify summary statistics for an ABC analysis. This involves specification of a model $\tilde{p}(\cdot|\tilde{\theta})$ which is similar to $p(\cdot|\theta)$, but which is computationally tractable. The idea is that estimates of $\tilde{\theta}$ under $\tilde{p}(\cdot|\tilde{\theta})$, such as maximum likelihood estimates or posterior means, are likely to be informative about θ if $p(\cdot|\theta)$ and $\tilde{p}(\cdot|\tilde{\theta})$ are sufficiently similar. This approach can be considered similar in spirit to that of Fearnhead and Prangle (2012) which uses estimated posterior means under a pilot ABC analysis (see Section 2.3.3). Blum (2010b) proposed a Bayesian criterion related to the BIC (see Section 3.1) as a best subset selection procedure. The idea is to implement a Bayesian analysis of the standard regression adjustment model (3). The criterion, called the *evidence* approximation, seeks the best subset of summary statistics to regress the parameter θ. In comparison to the BIC, the evidence criterion is attractive because it contains no approximation in its derivation. However, the downside is that its computation requires the tuning of the Bayesian linear regression hyperparameters. Additionally, Jung and Marjoram (2011) developed a genetic algorithm that weights the summary statistics so that individual statistics do not contribute equally to the comparisons between observations and simulations. The aim is that the uninformative summary statistics should ideally have negligible weights.

Finally, a number of recent ABC modelling approaches have attempted to find ways of accurately handling the full vector of initial statistics, s, (or the full dataset, $s = S(y) = y$) thereby avoiding the need to perform dimension reduction. Bonassi et al. (2011) propose fitting a $(p + q)$-dimensional mixture of Gaussian distributions to the sample $(\theta^i, s^i) \sim p(s|\theta)p(\theta)$, $i = 1, \ldots, n$, and then find the distribution of $\theta|s_{obs}$ by conditioning on observing $s = s_{obs}$. This approach potentially requires a large number of mixture components to accurately model the joint density when $(p + q)$ is large. Fan et al. (2012) suggest using an approximation to $p(s|\theta)$ by approximating each marginal likelihood function, $p(s_i|\theta)$, using a mixtures of experts model, where the weights, mean, and variance of each mixture component is allowed to depend on θ, and then inducing dependence between these marginals using a mixture of multivariate Gaussian distributions. This approach requires continuous summary statistics for the mixture regression, and is practically useful for moderate p (i.e. hundreds of summary statistics). Writing $y = (y_1, \ldots, y_p)$, Barthelmé and Chopin (2011) propose to factorise the likelihood as $p(y|\theta) = \prod_i p(y_i|y_{1:i-1}, \theta)$ and construct an ABC approximation of each component in turn (i.e. $p_{ABC}(y_i|y_{1:i-1}) = \int K_\epsilon(\|y_i - y_{obs,i}\|)p(y_i|y_{1:i-1}, \theta)dy_i$) with computation performed using an expectation-propagation algorithm (Minka 2001). This approach, while potentially fast and accurate, assumes that conditional simulation of $y_i \sim p(y_i|y_{1:i-1}, \theta)$ is available for $i = 1, \ldots, n$, and so is not suitable for all models and analyses. Last, Jasra et al. (2012) exploit the structure of hidden Markov models to perform an iterative sequence of ABC analyses, each using only a single data point in each analysis, and Nakagome et al. (2012) propose a novel approach to post-processing ABC importance sampling output whose convergence rate is claimed to avoid the curse of dimensionality.

3 New dimension reduction methods

In this Section, we introduce two new dimension reduction criteria for ABC methods. The first is a best subset selection procedure deriving from AIC and BIC criteria, constructed under implementation of the local linear model of equation (2) (Beaumont et al. 2002). A similar idea was proposed and tested for a Gaussian model by Sedki and Pudlo (2012). The second is a modification to the fitting of (2) by considering ridge regression instead of least-squares regression. Both of these methods are now implemented in the freely available `R` package `abc` (Csilléry et al. 2012).

3.1 AIC and BIC criteria

Akaike information criterion (AIC) and Bayesian information criterion (BIC) provide a measure of the relative goodness of fit of a statistical model. Each can be expressed as the sum of the maximized log-likelihood that measures the fit of the model to the data, and a penalty for model complexity (Akaike 1974; Schwarz 1978). While evaluation of $\log p(y_{obs}|\hat{\theta}_{mle})$ or $\log p(s_{obs}|\hat{\theta}_{mle})$ is unavailable in the ABC framework, and determination of the maximum likelihood estimator, $\hat{\theta}_{mle}$, is challenging, a simple and tractable likelihood function is available though the local-linear regression model of equation (2) (Section 2.1).

Specifically, we consider the local linear regression model equation (2) of Beaumont et al. (2002) for each parameter $\theta_1, \ldots, \theta_q$, and assume independent Gaussian errors, $e^j \sim N(0, \sigma_j^2)$, for $j = 1, \ldots, q$. Then the AIC becomes

$$\mathrm{AIC} = \tilde{n} \log \prod_{j=1}^{q} \hat{\sigma}_j^2 + 2d, \tag{15}$$

where $d = q(p+1)$ is the number of estimated regression parameters, and $\tilde{n}$ is the *effective* number of simulations used in the local-linear regression model, which we define as $\tilde{n} = \sum_{i=1}^{n} I(w^i > 0)$ when the kernel K_ϵ has compact support. Alternative definitions of the effective number of simulations, such as $c\sum_{i=1}^{n} w^i$ for some $c > 0$, can be on an arbitrary scale, since the least-squares regression solution is insensitive to the scale of the weights. For any fixed value of c, the value of $c\sum_{i=1}^{n} w^i$ will decrease as $p = \dim(s)$ increases so that it will artificially favour larger numbers of (even uninformative) summary statistics. Our definition of $\tilde{n}$ guarantees that the AIC scores are comparable for different subsets of summary statistics. A downside is that this definition of $\tilde{n}$ is only suitable for kernels, K_ϵ, with a compact support.

In equation (15), $\hat{\sigma}_j^2$ is defined as the weighted mean of squared residuals for the regression of θ_j, and is given by

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} w^i [\theta_j^i - \hat{m}_j(s^i)]^2}{\sum_{i=1}^{n} w^i}.$$

where θ_j^i is the j^{th} component of θ^i, and $\hat{m}_j(s)$ denotes the estimate of the mean function $m_j(s) = \mathbb{E}[\theta_j|s]$. For small sample sizes, the corrected AIC, the so-called AICc, is given by replacing d in (15) by $d(d+1)/(\tilde{n} - d - 1)$ (Hurvich and Tsai 1989). In the same manner the BIC can be defined as

$$\text{BIC} = \tilde{n} \log \prod_{j=1}^{q} \hat{\sigma}_j^2 + d \log \tilde{n}. \tag{16}$$

Alternative penalty terms involving the hat matrix of the regression could also be used in the above (e.g. Hurvich et al. 1998; Irizarry 2001; Konishi et al. 2004).

It is instructive to note that in using the linear regression adjustment (3), the above information criteria may be expressed as

$$\text{xIC} = \tilde{n} \log \prod_{j=1}^{q} \text{Var}(\theta_j^*) + \text{penalty term},$$

where θ_j^* is the j^{th} element of the regression adjusted vector $\theta^* = (\theta_1^*, \ldots, \theta_q^*)$. As such, up to the penalty terms, both AIC and BIC seek the combination of summary statistics that minimizes the product of the marginal variances of the adjusted posterior sample. Similarly to the entropy criterion of Nunes and Balding (2010) (see Section 2.2.2), these information criterion will select those summary statistics that maximise the precision of the posterior distribution, $p_{ABC}(\theta|s_{obs})$. However, unlike Nunes and Balding (2010), this precision is traded off by a penalty for model complexity.

A rationale for the construction of AIC and BIC in this manner is that the summary statistics that should be included within an ABC analysis are those which are good predictors of θ. However, an obvious requirement for AIC or BIC to identify an informative statistic is that the statistic varies (with θ) within the local range of the regression model. If a statistic is informative outside of this range, but uninformative within it, it will not be identified as informative under these criteria.

3.2 Regularisation via ridge regression

As described in Section 2.1, the local-linear regression adjustment of Beaumont et al. (2002) fits the linear model

$$\theta^i = \alpha + \beta^\top s^i + e^i$$

based on the prior predictive samples $(\theta^i, s^i) \sim p(s|\theta)p(\theta)$, and with regression weights given by $w^i = K_\epsilon(\|s^i - s_{obs}\|)$. (As before, we describe the case where θ is univariate for notational simplicity and clarity of exposition, but the approach outlined below can be readily implemented for each component of a multivariate θ.) However, in fitting the model by minimising the weighted least squares criteria, $\sum_{i=1}^n w^i \|\alpha - \beta^\top s^i - \theta^i\|^2$, there is a risk of over-adjustment by adjusting the parameter values via (3) in the direction of uninformative summary statistics.

To avoid this, implicit dimension reduction within the regression framework can be performed by alternatively minimising the regularised weighted sum of squares (Hoerl and Kennard 1970)

$$\sum_{i=1}^n w^i \|\theta^i - (\alpha + \beta^\top s^i)\|^2 + \lambda \|\beta\|^2, \tag{17}$$

with regularisation parameter $\lambda > 0$. As with the regularisation component within the neural network model of Blum and François (2010) (Section 2.3.2), with ridge regression the risk of over-adjustment is reduced because the regression coefficients, β, are shrunk towards zero by imposing a penalty on their magnitudes. Note that while we consider ridge regression here, a number of alternative regularisation procedures could be implemented, such as the Lasso method.

An additional advantage of ridge regression is that standard least squares estimates, $(\hat{\alpha}_{LS}, \hat{\beta}_{LS})^\top = (X^\top W X)^{-1} X^\top W \Theta$, are not guaranteed to have a unique solution. Here X is a $n \times (p+1)$ design matrix given in equation (10), $\Theta = (\theta^1, \ldots, \theta^n)$ is the $n \times 1$ column vector of sampled θ^i, and $W = \text{diag}(w^1, \ldots, w^n)$ is an $n \times n$ diagonal matrix of weights. The lack of a unique solution can arise through multicolinearity of the summary statistics, which can result in singularity of the matrix $X^\top WX$. In contrast, minimisation of the regularised weighted sum of squares (17) has always a unique solution, provided that $\lambda > 0$. This solution is given by $(\hat{\alpha}_{\text{ridge}}, \hat{\beta}_{\text{ridge}})^\top = (X^\top WX + \lambda I_p)^{-1} X^\top W\Theta$, where I_p denotes the $p \times p$ identity matrix. There are several approaches for dealing with the regularisation parameter λ, including cross-validation and generalised cross-validation to identify an optimal value of λ (Golub et al. 1979), as well as averaging the regularised estimates $(\hat{\alpha}_{\text{ridge}}, \hat{\beta}_{\text{ridge}})^\top$ obtained for different values of λ (Taniguchi and Tresp 1997).

4 A comparative analysis

We now provide a comparative analysis of the previously described methods of dimension reduction within the context of three previously studied analyses in the ABC literature. Specifically, this includes the analysis of a coalescent model with recombination (Joyce and Marjoram 2008), an evaluation of the evolutionary fitness cost of mutation in drug-resistant tuberculosis (Luciani et al. 2009), and an assessment of the number and size-distribution of particle inclusions in the production of clean steels (Bortot et al. 2007).

Each analysis is based on $n = 1,000,000$ simulations where the parameter θ is drawn from the prior distribution $p(\theta)$. The performance of each method is measured through the $\overline{\text{RSSE}}$ criterion (9) following Nunes and Balding (2010), based on the same randomly selected subset of $n^* = 100$ samples $(\theta^i, s^i) = (\theta_{true}, s_{obs})$ as 'observed' datasets. When evaluating the RSSE error measure of equation (8), we give a weight $w^i = 1$ for the accepted simulations and a weight of 0 otherwise. As the value of the RSSE (8) depends on the scale of each parameter, we standardise the parameters in each example by dividing the parameter values by the standard deviation obtained from the $n = 1,000,000$ simulations (with the exception of the first example, where the parameters are on similar scales). For comparative ease, and to provide a performance baseline for each example, all $\overline{\text{RSSE}}$ results are presented as relative to the $\overline{\text{RSSE}}$ obtained when using the maximal vector of summary statistics and no regression adjustment. In this manner, a relative $\overline{\text{RSSE}}$ of $x/-x$ denotes an $x\%$ worsening/improvement over the baseline score.

Within each ABC analysis, we use Euclidean distance within an Epanechnikov kernel $K_\epsilon(\|s - s_{obs}\|)$. The Euclidean distances are computed after standardizing the summary statistics with a robust estimate of the standard deviation (the mean absolute deviation). The kernel scale parameter, ϵ, is determined as the value at which exactly 1% of the simulations, (θ^i, s^i), have non-zero weight. This yields exactly $\tilde{n} = 10,000$ simulations that form the final sample from each posterior. To perform the method of Fearnhead and Prangle (2012), a randomly chosen 10% of the n simulations were used to fit the regression model that determines the choice of summary statistics, with the remaining 90% used for the ABC analysis. The final ABC sample size $\tilde{n} = 10,000$ was kept equal to the other methods by slightly adjusting the scale parameter, ϵ. In addition for the method of Fearnhead and Prangle (2012), following exploratory analyses, the regression model (14) was fitted using $f(s) = (s, s^2, s^3, s^4)$ for Examples 1 and 2 (as described in Section 2.3.3) and using $f(s) = (\log s, [\log s]^2, [\log s]^3, [\log s]^4)$ for Example 3 resulting in always $4 \times p$ independent variables in the regression model of equation (14).

When using neural networks or ridge regression to estimate the conditional mean and variance, $m(s)$ and $\sigma^2(s)$, we take the pointwise median of the estimated functions obtained with the regularisation parameters $\lambda = 10^{-3}, 10^{-2}$ and 10^{-1}. These values of λ assume that the summary statistics and the parameters have been standardized before fitting the regression function (Ripley 1994). However, because the optimisation procedure for neural networks (the `R` function `nnet`) only finds local optima, in this case we take the pointwise median of ten estimated functions, with each optimisation initialised from a different random starting point, and randomly choosing the regularisation parameter with equal probability from the above values (see Taniguchi and Tresp 1997).

4.1 Example 1: A coalescent analysis

This model was previously considered by Joyce and Marjoram (2008) and Nunes and Balding (2010), each while proposing their respective ABC dimension reduction strategies (see Sections 2.2.1 and 2.2.2). The analysis focuses on joint estimation of the scaled mutation rate, $\tilde{\theta}$, and the scaled recombination rate, ρ, in a coalescent model with recombination (Nordborg 2007). Under this model, 5,001 basepair DNA sequences for 50 individuals are generated from the coalescent model, with recombination, under the infinite-sites mutation model, using the software `ms` (Hudson 2002). The initial summary statistics, $s = (s_1, \dots, s_6)$, are the number of segregating sites (s_1), the pairwise mean number of nucleotidic differences (s_2), the mean R^2 across pairs separated by $< 10\%$ of the simulated genomic regions (s_3), the number of distinct haplotypes (s_4), the frequency of the most common haplotype (s_5), and the number of singleton haplotypes (s_6).

We first examine the performance of ABC without using dimension reduction techniques. For different parameter combinations, $\tilde{\theta}, \rho$ and $(\tilde{\theta}, \rho)$, we compute the relative $\overline{\text{RSSE}}$ obtained with a single optimal summary statistic, and the relative $\overline{\text{RSSE}}$ obtained when using all six population genetic statistics (s_1–s_6) (Table 1). When estimating $\tilde{\theta}$ only, we find that using only the number of segregating sites (s_1) provides lower relative $\overline{\text{RSSE}}$ than when including all 6 summary statistics even when performing regression adjustment. For all other parameter combinations, using a single statistic produce substantially worse than the rejection algorithm with all summary statistics. For all inferences (i.e. of $\tilde{\theta}$, ρ and $(\tilde{\theta}, \rho)$ jointly), regression adjustments generally improve the inference when using all six summary statistics, which is consistent with previous results (Nunes and Balding 2010). The only exception is when jointly estimating $(\tilde{\theta}, \rho)$, where homoscedastic linear adjustment neither decreases nor increases the error obtained with the pure rejection algorithm.

Next, we investigate the performance of each dimension reduction technique. Table 2 and Figure 1 shows the relative $\overline{\mathrm{RSSE}}$ obtained under each dimension reduction method for each parameter combination and under heteroscedastic regression adjustment. For all three examples, more complete tables that contain the results obtained with no regression adjustment and homoscedastic adjustment can be found in the supplementary information to this article.

The performance achieved with AIC, AICc, or BIC is comparable to (i.e. the same or slightly better than) the result obtained when including all six population genetic statistics. When using the ε-sufficiency criterion, we find that the performance is improved for the inference on $\tilde{\theta}$ only. The only best subset selection method for dimension reduction that substantially and uniformly improves the performance of ABC posterior estimates is the entropy-based approach. For the projection techniques, all methods (partial least squares, neural nets, and minimum expected posterior loss) outperform the adjustment method based on all six population genetics statistics, with a large performance advantage for partial least squares when estimating $(\tilde{\theta}, \rho)$ jointly. By contrast, ridge regression provides no improvement over the standard regression adjustment (the "All" column).

Based on these results, a loose performance ranking of the dimension reduction methods can be obtained by computing, for each method, the mean (relative) $\overline{\mathrm{RSSE}}$ over all parameter combinations $\tilde{\theta}$, ρ, and $(\tilde{\theta}, \rho)$ using the heteroscedastic adjustment. The worst performers were ridge regression and the ε-sufficiency criterion (with a mean relative $\overline{\mathrm{RSSE}}$ of -3%). These are followed by the standard regression adjustment with all summary statistics (-5%) and the AIC/BIC, neural nets and the posterior loss method (-6%). The best performing methods are partial least squares (-10%), and the two-stage entropy based procedure (-16%).

4.2 Example 2: The fitness cost of drug resistant tuberculosis

We now consider an example of Markov processes for epidemiological modeling. If a pathogen, such as Mycobacterium tuberculosis, mutates to gain an evolutionary advantage, such as antibiotic resistance, it is biologically plausible that this mutation will come at a cost to the pathogen's relative fitness. Based on a stochastic model to describe the transmission and evolutionary dynamics of Mycobacterium tuberculosis, and based on incidence and genotypic data of the IS6110 marker, Luciani et al. (2009) estimated the posterior distribution of the pathogen's transmission cost and relative fitness. The model contained $q = 4$ free parameters: the transmission rate, α, the transmission cost of drug resistant strains, c, the rate of evolution of resistance, ρ, and the mutation rate of the IS6110 marker, μ.

Luciani et al. (2009) summarised information generated from the stochastic model through $p = 11$ summary statistics. These statistics were expertly elicited as quantities that were expected to be informative for one or more model parameters, and included the number of distinct genotypes in the sample, gene diversity for sensitive and resistant cases, the proportion of resistant cases and measures of the degree of clustering of genotypes etc. It is considered likely that there is dependence, and potentially replicate information within these statistics.

As before, we examine the relative performance of the statistics without using dimension reduction techniques. Table 1 shows that for the univariate analysis of c, ρ, or μ, performing rejection sampling ABC with a single, well chosen summary statistic, can provide an improved performance over a similar analysis using all 11 summary statistics, under any form of regression adjustment. In particular, the proportion of isolates that are drug resistant is the individual statistic which is most informative to estimate c (with a relative $\overline{\text{RSSE}}$ of -7%) and ρ (-9%). For the marker mutation rate, μ, the most informative statistic is the number of distinct genotypes, with a relative $\overline{\text{RSSE}}$ of -14%. Conversely, an analysis using all summary statistics with a regression adjustment offers the best inferential performance for α alone, or for (α, c, ρ, μ). These results provide support for recent arguments in favour of "marginal regression adjustments," (Nott et al. 2011) whereby the univariate marginal distributions of a full multivariate ABC analysis are replaced by separately estimated marginal distributions using only statistics relevant for each parameter. Here, more precisely estimated margins can improve the accuracy of the multivariate posterior sample, beyond the initial analysis.

The performance results of each dimension reduction method are shown in Table 2 and Figure 1. In contrast with the previous example, here the use of the AIC/BIC criteria can substantially decrease posterior errors. For example, compared to the linear adjustment of all 11 parameters, which produces a mean relative $\overline{\text{RSSE}}$ between -3% and -8% depending on the parameter (Table 2), using the AIC/BIC criteria results in a relative $\overline{\text{RSSE}}$ of between -15% and -19%. The ε-sufficiency criterion produces more equivocal results, however, as the error is sometimes increased with respect to baseline performance (e.g. $+6\%$ when estimating α with homoscedastic adjustment) and sometimes reduced (e.g. -8% for c, ρ and θ with heteroscedastic adjustment). As with the previous example, the entropy criterion provides a clear improvement to the ABC posterior, and this improvement is almost comparable to that produced by AIC/BIC. Finally, the projection and regularisation methods mostly all provide comparable and substantive improvements compared to the baseline error, with only partial least squares producing more equivocal results (e.g. $+1\%$ when estimating ρ).

Based on these results, the loose performance ranking of the dimension reduction methods determines the worst performers to be the standard least-squares regression adjustment (with a mean relative $\overline{\text{RSSE}}$ of -5%), the ε-sufficiency approach (-6%) and partial least squares (-8%). These are followed by ridge regression (-11%), neural networks and the posterior loss method (-12%). The best performing methods for this analysis are the two-stage entropy-based procedure (-15%) and the AIC/BIC criteria (-17%).

In this example, it is interesting to compare the performance of the standard linear regression adjustment of all 11 summary statistics (mean relative $\overline{\text{RSSE}}$ of -5%) with that of the ridge regression equivalent (mean relative $\overline{\text{RSSE}}$ of -11%). The increase in performance with ridge regression may be attributed to its more robust handling of multicolinearity of the summary statistics than under the standard regression adjustment. To see this, Figure 2 illustrates the relationship between the relative $\overline{\text{RSSE}}$ (again, relative to using all summary statistics and no regression adjustment), and the condition number of the matrix $X^\top WX$, for both the standard regression (top panel) and ridge regression (bottom panel) adjustments based on inference for (α, c, ρ, μ). The condition number of $X^\top WX$ is given by $\kappa = \sqrt{\lambda_{\max}/\lambda_{\min}}$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $X^\top WX$. Extremely large condition numbers are evidence for multicolinearity.

Figure 2 demonstrates that for large values of the condition number (e.g. for $\kappa > 10^8$), the least-squares-based regression adjustment clearly performs very poorly. The region of $\kappa > 10^8$ corresponds to almost 5% of all simulations, and for these cases the relative error is hugely increased (w.r.t. rejection) to anywhere between 5% and 200%. In contrast, for ridge-regression, the relative errors corresponding to $\kappa > 10^8$ are not larger than the errors obtained for non-extreme condition numbers. This analysis clearly illustrates that, unlike ridge regression, the standard least-squares regression adjustment can perform particularly poorly when there is multicolinearity between the summary statistics.

In terms of the original analysis of Luciani et al. (2009) which used all eleven summary statistics with no regression adjustment (although with a very low value for ϵ), the above results indicate that a more efficient analysis may have been achieved by using a suitable dimension reduction technique.

4.3 Example 3: Quality control in the production of clean steels

Our final example concerns the statistical modelling of extreme values. In the production of clean steels, the occurrence of microscopic imperfections (termed *inclusions*) is unavoidable. The strength of a clean steel block is largely dependent on the size of the largest inclusion.

Bortot et al. (2007) considered an extreme value twist on the standard stereological problem (e.g. Baddeley and Jensen 2004), whereby inference is required on the size and number of 3-dimensional inclusions, based on data obtained from those inclusions that intersect with a 2-dimensional slice. The model assumes a Poisson point process of inclusion locations with rate parameter $\tau > 0$, and that the distribution of inclusion size exceedances above a measurement threshold of 5μm are drawn from a generalised Pareto distribution with scale and shape parameters $\sigma > 0$ and ξ, following standard extreme value theory arguments (e.g. Coles 2001).

The observed data consist of 112 cross-sectional inclusion diameters measured above 5μm. The summary statistics thereby comprise 112 equally spaced quantiles of the cross-sectional diameters, in addition to the number of inclusions observed, yielding $p = 113$ summary statistics in total. The ordering of the summary statistics creates strong dependences between them, a fact which can be exploited by dimension reduction techniques. Bortot et al. (2007) considered two models based on spherical or ellipsoidal shaped inclusions. Our analysis here focuses on the ellipsoidal model.

By construction, the large number (2^{113}) of possible combinations of summary statistics means that best subset selection methods are strictly not practicable for this analysis, unless the number of summary statistics is reduced further a priori. In order to facilitate at least some comparison with the other dimension reduction approaches, for best subset selection methods *only*, we consider six candidate subsets. Each subset consists of the number of observed inclusions in addition to 5, 10, 20, 50, 75 or 112 empirical quantiles of the inclusion size exceedances (the latter corresponds to the complete set of summary statistics). Due to the extreme value nature of this analysis, the parameter estimates are likely to be more sensitive to the precise values of the larger quantiles. As such, rather than using equally spaced quantiles, we use a scheme which favours quantiles closer to the maximum inclusion and we always include the maximum inclusion.

The relative $\overline{\text{RSSE}}$ obtained under each dimension reduction method is shown in Table 2 and Figure 1. In comparison to an analysis using all 113 summary statistics and regression adjustment (the "All" column), the best subset selection approaches do not in general offer any improvement. While the entropy based method provides a slight improvement, the relative $\overline{\text{RSSE}}$ under the ε-sufficiency criterion is substantially worse (along with partial least squares). Of course, these results are limited to the few subsets of statistics considered, and it is possible that alternative subsets could perform substantially better. However, it is computationally untenable to evaluate this possibility based on exhaustive enumeration of all subsets.

When using neural networks to perform the regression adjustment based on computing the pointwise median of the $m(s)$ and $\sigma^2(s)$ estimates, obtained using varying regularisation parameter values (see the introduction to Section 4), the relative performance is quite poor (left hand side $\overline{\text{RSSE}}$ values in Table 2). The mean relative $\overline{\text{RSSE}}$ is -13% for neural networks, compared to -40% for heteroscedastic least-squares regression. As an alternative approach, rather than averaging over the regularization parameter λ, we rather choose the value of $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ that minimises the leave-one-out error of θ (equation (11)). This approach considerably improves the performance of the network (right hand side $\overline{\text{RSSE}}$ values in Table 2) with the mean relative $\overline{\text{RSSE}}$ improving to the same level as for heteroscedastic regression. Adopting the same procedure to determine the regularisation parameter within ridge regression, there is also a mean gain in performance from -39% to -42%, although the joint parameter inference on (τ, σ, ξ) actually performs worse under this alternative approach. The variability in these results highlights the importance of making an optimal choice of the regularisation parameter for an ABC analysis.

The minimum expected posterior loss approach performs particularly well here. This approach has also been shown to perform well in a similar analysis; that of performing inference using quantiles of a large number of independent draws from the (intractable)

g-and-k distribution (Fearnhead and Prangle 2012).

The loose performance ranking of each of the dimension reduction methods finds that the worst performers are the ε-sufficiency criterion (with a mean relative $\overline{\text{RSSE}}$ of -16%) and partial least squares (-19%). Neural networks and AIC/BIC perform just as well as standard least squares regression (-40%), ridge regression slightly outperforms standard regression (-42%) and the entropy based approach is a further slight improvement at -44%. The clear winner in this example is the posterior loss approach with a mean relative $\overline{\text{RSSE}}$ of -58%.

5 Discussion

The process of dimension reduction is a critical and influential part of any ABC analysis. In this article we have provided a comparative review of the major dimension reduction approaches (and introduced two new ones) in order to provide some guidance to the practitioner in choosing the most appropriate technique for their own analysis. A summary of the qualitative features of each dimension reduction method is shown in Table 3, and a comparison of the relative performances of each method for each example, is illustrated in Figure 3. As with each individual example, we may compute an overall performance ranking of the dimension reduction methods, by averaging the mean relative $\overline{\text{RSSE}}$ values over the examples. Performing worse, on average, than a standard least squares regression adjustment with no dimension reduction (with an overall mean relative $\overline{\text{RSSE}}$ of -17%), is the ε-sufficiency technique (-8%) and partial least squares (-12%). Performing better, on average, than standard least squares regression is ridge regression and neural networks (-19%) and AIC/BIC (-21%). In this study, the top performers, on average, were the entropy-based procedure and the minimum expected posterior loss approach, with an overall mean relative $\overline{\text{RSSE}}$ of -25%. It is worth emphasising that the potential gains in performing a regression adjustment alone (with all summary statistics and *no* dimension reduction) can be quite substantial. This suggests that regression adjustment should be an integral part of the majority of ABC analyses. Further gains in performance can then be obtained by combining regression adjustment with dimension reduction procedures, although in some cases (such as with the ε-sufficiency technique and partial least squares) performance can sometimes worsen.

While being ranked in the top three, a clear disadvantage of the entropy based procedure and the AIC/BIC criteria is the quantity of computation required. This primarily occurs as best subset selection procedures require evaluation of all 2^p potential models. For examples 1 and 2, a greedy algorithm was able to find the optimum solution in a reasonable time. This was not possible for example 3. Additionally in this latter case, for the subsets of summary statistics considered, the performance obtained by implementing computationally expensive methods of dimension reduction was barely an improvement over the computationally cheap, least squares regression adjustment. This raises the important point, that the benefits of performing potentially expensive forms of dimension reduction over, say, the simple linear regression adjustment, should be evaluated prior to their full implementation. We also note that the second stage of the entropy-based method (Section 2.2.2) targets minimisation of (9), the same error measure used in our comparative analysis. As such, this approach is likely to be numerically favoured in our results.

The top ranked (ex aequo) minimum expected posterior loss approach particularly outperforms other dimension reduction methods in the final example (the production of clean steels). In such analyses, with large numbers of summary statistics (here $p = 113$), non-linear methods such as neural networks may become overparametrised, and simpler alternatives such as least squares or ridge regression adjustment can work more effectively. This is naturally explained through the usual bias-variance tradeoff: more complex regression models such as neural networks reduce the bias of the estimate of $m(s)$ (and optionally $\sigma^2(s)$), but in doing so the variance of the estimate is increased. This effect can be especially acute for high-dimensional regression (Geman et al. 1992).

Our analyses indicate that the original least squares, linear regression adjustment (Beaumont et al. 2002) can sometimes perform quite well, despite its simplicity. However, the presence of multicolinearity between the summary statistics can cause severe performance degradation, compared to not performing the regression adjustment (see Figure 2). In such situations, regularisation procedures such as ridge regression (e.g. Example 2, and Figure 2) and projection techniques can be beneficial.

However, an important issue with regularisation procedures, such as neural networks and ridge regression, is the handling of the regularisation parameter, λ. The 'averaging' procedure that was used in the first two examples, proved quite suboptimal in the third, where a cross-validation procedure to select a single best parameter value produced much improved results. This problem can be particularly critical for neural networks with large numbers of summary statistics, p, as the number of network weights is much larger than p, and accordingly, massive shrinkage of the weights (i.e. large values of λ) is required to avoid overfitting.

The posterior loss approach produced the superior performance in the third example. In general, a strong performance of this method can be primarily attributed to two factors. Firstly, in the presence of large numbers of highly dependent summary statistics, the extra analysis stage in determining the most appropriate regression model (14) by choosing $f(s)$ through e.g. BIC diagnostics, affords the opportunity to reduce the complexity of the regression in a simple and relatively low-parameterised manner. This was not a primary contributor in example 3, however, as the regression (equation (14)) was directly performed on the full set of 113 statistics. Given the benefits of using regularisation methods in this setting, it is possible that a ridge regression model would allow a more robust estimate of the posterior mean (as a summary statistic) as part of this process. Secondly, the posterior loss approach determines the number of summary statistics to be equal to the number of posterior quantities of interest – in this case, $q = 3$ posterior parameter means. This small number of derived summary statistics naturally allows more precise posterior statements to be made, compared to dimension reduction methods that produce a much larger number of equally informative statistics. Of course, the dimension advantage here is strongly related to the number of parameters ($q = 3$) and summary statistics ($p = 113$) in this example. However, it is not fully clear how any current methods of dimension reduction for ABC would perform for substantially more challenging analyses with considerably higher numbers of parameters and summary statistics. This is because the curse of dimensionality in ABC (Blum 2010a) has tended to restrict existing applications of ABC methods to problems of moderate parameter dimension, although this may change in the future.

What is very apparent from this study, is that there is no single 'best' method of dimension reduction for ABC. For example, while the posterior loss and entropy based methods were the best performers for example 3, AIC and BIC were ranked first in the analysis of example 2, and partial least squares outperformed the posterior loss approach in example 1. A number of factors can affect the most suitable choice for any given analysis. As discussed above, these can include the number of initial summary statistics, the amount of dependence and multicolinearity within the statistics, the computational overheads of the dimension reduction method, the requirement to suitably determine hyperparameters, and sensitivity to potentially large numbers of uninformative statistics.

One important point to understand is that all of the ABC analyses of this review were performed using the rejection algorithm optionally followed by some form of regression adjustment. While alternative, potentially more efficient and accurate methods of ABC posterior simulation exist, such as Markov chain Monte Carlo or sequential Monte Carlo based samplers, the computational cost of separately implementing such an algorithm 2^p times (in the case of best subset selection methods) means that such dimension reduction methods can become rapidly untenable, even for small p. The price of the benefit of using the more computationally practical, fixed large number of samples, is that decisions on the dimension reduction of the summary statistics will be made on potentially worse estimates of the posterior than those available under superior sampling algorithms. As such, the final derived summary statistics may in fact not be those which are most appropriate for subsequent use in e.g. Markov chain Monte Carlo or sequential Monte Carlo based algorithms.

However, this price is arguably a necessity. It is practically important to evaluate the performance of any dimension reduction procedure in a given analysis. Here we used a criterion (the $\overline{\mathrm{RSSE}}$ of equation (9)) that is based on a leave-one-out procedure. When using a fixed, large number of samples, evaluation of such a performance diagnostic is entirely practicable, as no further model simulations are required. This idea is also relevant to methods of dimension reduction for model selection (Barnes et al. 2012; Estoup et al. 2012) where a misclassification rate based on a leave-one-out procedure can serve as a comparative criterion.

Acknowledgments

SA Sisson is supported by the Australian Research Council through the Discovery Project Scheme (DP1092805). MGB Blum is supported by the French National Research Agency (DATGEN project, ANR-2010-JCJC-1607-01).

References

Abdi, H. and L. J. Williams (2010). Partial least square regression, projection on latent structure regression. Wiley Interdisciplinary Reviews: Computational Statistics 2, 433–459.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723.

Allingham, D., R. A. King, and K. L. Mengersen (2009). Bayesian estimation of quantile distributions. *Statistics and Computing* *19*, 189–201.

Baddeley, A. and E. B. V. Jensen (2004). *Stereology for Statisticians*. Boca Eaton, FL: Chapman and Hall/CRC.

Barnes, C., S. Filippi, M. P. H. Stumpf, and T. Thorne (2012). Considerate approaches to achieving sufficiency for ABC model selection. *Statistics and Computing*, in press.

Barthelmé, S. and N. Chopin (2011). Expectation-propagation for summary-less, likelihood-free inference. *Arxiv*, http://arxiv.org/abs/1107.5959.

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

Beaumont, M. A., J.-M. Marin, J.-M. Cornuet, and C. P. Robert (2009). Adaptivity for ABC algorithms: the ABC-PMC scheme. *Biometrika* *96*, 983–990.

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

Bertorelle, G., A. Benazzo, and S. Mona (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Molecular Ecology* *19*, 2609–2625.

Blum, M. G. B. (2010a). Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association* *105*(491), 1178–1187.

Blum, M. G. B. (2010b). Choosing the summary statistics and the acceptance rate in approximate Bayesian computation. In G. Saporta and Y. Lechevallier (Eds.), *COMPSTAT 2010: Proceedings in Computational Statistics*, pp. 47–56. Springer, Physica Verlag.

Blum, M. G. B. and O. François (2010). Non-linear regression models for approximate

Bayesian computation. *Statistics and Computing* *20*, 63–73.

Bonassi, F., L. You, and M. West (2011). Bayesian learning from marginal data in bionetwork models. *Statistical Applications in Genetics and Molecular Biology* *10*(1), 49.

Bortot, P., S. G. Coles, and S. A. Sisson (2007). Inference for stereological extremes. *Journal of the American Statistical Association* *102*, 84–92.

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

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer-Verlag.

Csillėry, K., M. G. B. Blum, O. Gaggiotti, and O. François (2010). Approximate Bayesian computation in practice. *Trends in Ecology & Evolution* *25*, 410–418.

Csillėry, K., O. François, and M. G. B. Blum (2012). abc: An R package for approximate Bayesian computation. *Methods in Ecology and Evolution* *3*, 475–479.

Del Moral, P., A. Doucet, and A. Jasra (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing* *22*, 1009–1020.

Drovandi, C. C. and A. N. Pettitt (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics* *67*, 225–233.

Drovandi, C. C., A. N. Pettitt, and M. J. Faddy (2011). Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C* *60*, 317–338.

Estoup, A., E. Lombaert, J.-M. Marin, T. Guillemaud, P. Pudlo, C. Robert, and J.-M. Cornuet (2012). Estimation of demo-genetic model probabilities with approximate bayesian computation using linear discriminant analysis on summary statistics.

Molecular Ecology Resources in press.

Fan, Y., D. J. Nott, and S. A. Sisson (2012). Regression density estimation ABC. In preparation.

Fearnhead, P. and D. Prangle (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic ABC (with discussion). Journal of the Royal Statistical Society: Series B 74, 419–474.

Filippi, S., C. P. Barnes, and M. P. H. Stumpf (2012). Contribution to the discussion of Fearnhead and Prangle (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. Journal of the Royal Statistical Society: Series B 74, 459–460.

Geman, S., E. Bienenstock, and R. Doursat (1992). Neural networks and the bias/variance dilemma. Neural computation 4(1), 1–58.

Golub, G., M. Heath, and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223.

Heggland, K. and A. Frigessi (2004). Estimating functions in indirect inference. Journal of the Royal Statistical Society: Series B 66, 447–462.

Hoerl, A. and R. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

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

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60(2), 271–293.

Hurvich, C. M. and C.-L. Tsai (1989, 06). Regression and time series model selection in small samples. Biometrika 76(2), 297–307.

Irizarry, R. A. (2001). Information and posterior probability criteria for model selection in local likelihood estimation. Journal of the American Statistical Association 96, 303–315.

Jasra, A., S. Singh, J. Martin, and E. McCoy (2012). Filtering via approximate Bayesian computation. Statistics and Computing in press.

Jeremiah, E., S. A. Sisson, L. Marshall, R. Mehrotra, and A. Sharma (2011). Bayesian calibration and uncertainty analysis for hydrological models: A comparison of adaptive-Metropolis and sequential Monte Carlo samplers. Water Resources Research 47, W07547, 13pp.

Joyce, P. and P. Marjoram (2008). Approximately sufficient statistics and Bayesian computation. Statistical Applications in Genetics and Molecular Biology 7. Article 26.

Jung, H. and P. Marjoram (2011). Choice of summary statistic weights in approximate bayesian computation. Statistical Applications in Genetics and Molecular Biology 10.

Konishi, S., T. Ando, and S. Imoto (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. Biometrika 91, 27–43.

Leuenberger, C. and D. Wegmann (2010). Bayesian computation and model selection without likelihoods. Genetics 184, 243–252.

Lopes, J. S. and M. A. Beaumont (2009). ABC: A useful Bayesian tool for the analysis of population data. Infection, Genetics and Evolution 10, 826–833.

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*. Proceedings of the National Academy of Sciences of the United States of America 106, 14711–14715.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavare (2003). Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America 100, 15324–15328.

Mevik, B.-H. and H. R. Cederkvist (2004). Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). Journal of Chemometrics 18(9), 422–429.

Mevik, B.-H. and R. Wehrens (2007). The pls package: Principal component and partial least squares regression in R. Journal of Statistical Software 18(2), 1–24.

Minka, T. (2001). Expectation propagation for approximate Bayesian inference. Proceedings of Uncertainty in Artificial Intelligence 17, 362–369.

Nakagome, S., K. Fukumizu, and S. Mano (2012). Kernel approximate Bayesian computation for population genetic inferences. Arxiv, http://arxiv.org/abs/1205.3246.

Nix, D. A. and A. S. Weigend (1995). Learning local error bars for nonlinear regression. In G. Tesauo, D. Touretzky, and T. Leen (Eds.), Advances in Neural Information Processing Systems 7 (NIPS‘94), pp. 489–496. MIT Press, Cambridge.

Nordborg, M. (2007). Coalescent theory. In D. J. Balding, M. J. Bishop, and C. Cannings (Eds.), Handbook of Statistical Genetics (Third ed.)., pp. 179–208. Wiley: Chichester.

Nott, D. J., Y. Fan, L. Marshall, and S. A. Sisson (2011). Approximate Bayesian computation and Bayes linear analysis: Towards high-dimensional approximate Bayesian computation. Arxiv, http://arxiv.org/abs/1112.4755.

Nott, D. J., Y. Fan, and S. A. Sisson (2012). Contribution to the discussion of Fearnhead and Prangle (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. Journal of the Royal Statistical Society: Series B 74, 466.

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

Peters, G. W., Y. Fan, and S. A. Sisson (2012). On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. Statistics and Computing, in press.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution 16, 1791–1798.

Ripley, B. (1994). Neural networks and related methods for classification. Journal of the Royal Statistical Society. Series B (Methodological) 56, 409–456.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Sedki, M. A. and P. Pudlo (2012). Contribution to the discussion of Fearnhead and Prangle (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. Journal of the Royal Statistical Society: Series B 74, 466–467.

Shannon, C. E. and W. Weaver (1948). A mathematical theory of communication. Bell System Technical Journal 27, 379–423.

Singh, H., N. H. V. Misra, A. Fedorowicz, and E. Demchuk (2003). Nearest neighbor estimates of entropy. American Journal of Mathematical and Management Sciences 23, 301–321.

Sisson, S. A. and Y. Fan (2011). Likelihood-free Markov chain Monte Carlo. In S. P. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), Handbook of Markov Chain Monte Carlo, pp. 319–341. Chapman and Hall/CRC Press.

Sisson, S. A., Y. Fan, and M. Tanaka (2007). Sequential Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America 104, 1760–1765. Errata (2009), 106, 16889.

Taniguchi, M. and V. Tresp (1997). Averaging regularized estimators. Neural Computation 9(5), 1163–1178.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of The Royal Society Interface 6, 187–202.

Vinzi, V., W. W. Chin, J. Henseler, and H. Wang (Eds.) (2010). Handbook of Partial Least Squares. Springer.

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

		One optimal statistic	All summary statistics		
		(no adj.)	no adj.	homo adj.	hetero adj.
Example 1	$\tilde{\theta}$	**-7** (s_1)	0	-3	-3
	ρ	9 (s_5)	0	**-5**	-4
	$(\tilde{\theta}, \rho)$	7 (s_1)	0	0	**-7**
Example 2	α	6	0	**-3**	**-3**
	c	**-7**	0	-5	-5
	ρ	**-9**	0	-8	-8
	μ	**-14**	0	-5	-6
	(α, c, ρ, μ)	5	0	**-4**	**-4**

Table 1: Relative $\overline{\text{RSSE}}$ for Examples 1 and 2. The leftmost column shows the minimal $\overline{\text{RSSE}}$ when considering only one summary statistic (with no regression adjustment). Rightmost columns show relative $\overline{\text{RSSE}}$ using all summary statistics under no, homoscedastic and heteroscedastic regression adjustment. All $\overline{\text{RSSE}}$ are relative to the $\overline{\text{RSSE}}$ obtained when using no regression adjustment with all summary statistics. The score of the best method in each analysis (row) is emphasised in boldface.

		Best subset selection					Projection techniques			Regularisation
	All	BIC	AIC	AICc	ε-suff	Ent	PLS	NNet[1]	Loss	Ridge[1]
$\tilde{\theta}$	-3	-5	-5	-5	-6	**-11**	-6	-4	-7	1
ρ	-4	-6	-6	-6	0	**-12**	-7	-8	-7	-3
$(\tilde{\theta}, \rho)$	-7	-7	-7	-7	–	**-24**	-16	-7	-6	-6
α	-3	-15	-15	-15	0	**-17**	-13	-15	**-17**	-15
c	-5	**-15**	**-15**	**-15**	-8	**-15**	-8	-12	-9	-9
ρ	-8	**-16**	**-16**	**-16**	-8	**-16**	1	-12	-9	-10
μ	-6	**-18**	**-18**	**-18**	-8	-13	-10	-13	-12	-12
(α, c, ρ, μ)	-4	**-19**	**-19**	**-19**	–	-13	-10	-9	-12	-11
τ	-49	-47	-47	-48	-19	-52	-22	-20/-42	**-75**	-48/-48
σ	-45	-46	-47	-46	-15	-50	-15	-21/-37	**-56**	-43/-43
ξ	-27	-29	-29	-28	-13	-32	-28	-7/-41	-41	-26/**-44**
(τ,σ,ξ)	-39	-39	-40	-39	–	-42	-11	-4/-38	**-60**	-39/-32

[1] For the third Example, the first value is found by integrating out the regularisation parameter whereas the second one is found by choosing an optimal regularisation parameter with cross-validation. In Examples 1 and 2, integration over the regularisation parameter is performed.

Table 2: Relative $\overline{\text{RSSE}}$ for Examples 1-3 for different parameter combinations using each method of dimension reduction, and under heteroscedastic regression adjustment. Columns denote no dimension reduction (All), BIC, AIC, AICc, the ε-sufficiency criterion (ε-suff), the two-stage entropy procedure (Ent), partial least squares (PLS), neural networks (NNet), minimum expected posterior loss (Loss) and ridge regression (Ridge). All $\overline{\text{RSSE}}$ are relative to the $\overline{\text{RSSE}}$ obtained when using no regression adjustment with all summary statistics. The score of the best method in each analysis (row) is emphasised in boldface.

Class	Method	Hyper-parameter	Choice of hyper-parameter	Computational burden
Best subset selection	AIC/BIC	None	–	Substantial/greedy alg.
	ε-suff	$T(\theta)$	User choice	Substantial/greedy alg.
	Ent	None	–	Substantial/greedy alg.
Projection techniques	PLS	Number of PLS components, k	Cross-validation	Weak
	NNet	Regularisation parameter, λ	Integration or cross-validation	Moderate (optimization algorithm)
	Loss	Choice of basis functions	BIC	Weak (closed-form solution)
Regularisation	Ridge	Regularisation parameter, λ	Integration or cross-validation	Weak (closed-form solution)

Table 3: Summary of the main features of the different methods of dimension reduction for ABC.

Figure 1: Relative $\overline{\text{RSSE}}$ for the different methods of dimension reduction in the three Examples. All $\overline{\text{RSSE}}$ are relative to the $\overline{\text{RSSE}}$ obtained when using no regression adjustment with all summary statistics. Methods of dimension reduction include no dimension reduction (All), AIC/BIC, the ε-sufficiency criterion (ε-suff), the two-stage entropy procedure (Ent), partial least squares (PLS), neural networks (NNet), minimum expected posterior loss (Loss) and ridge regression (Ridge). The crosses correspond to situations for which there is no result available.

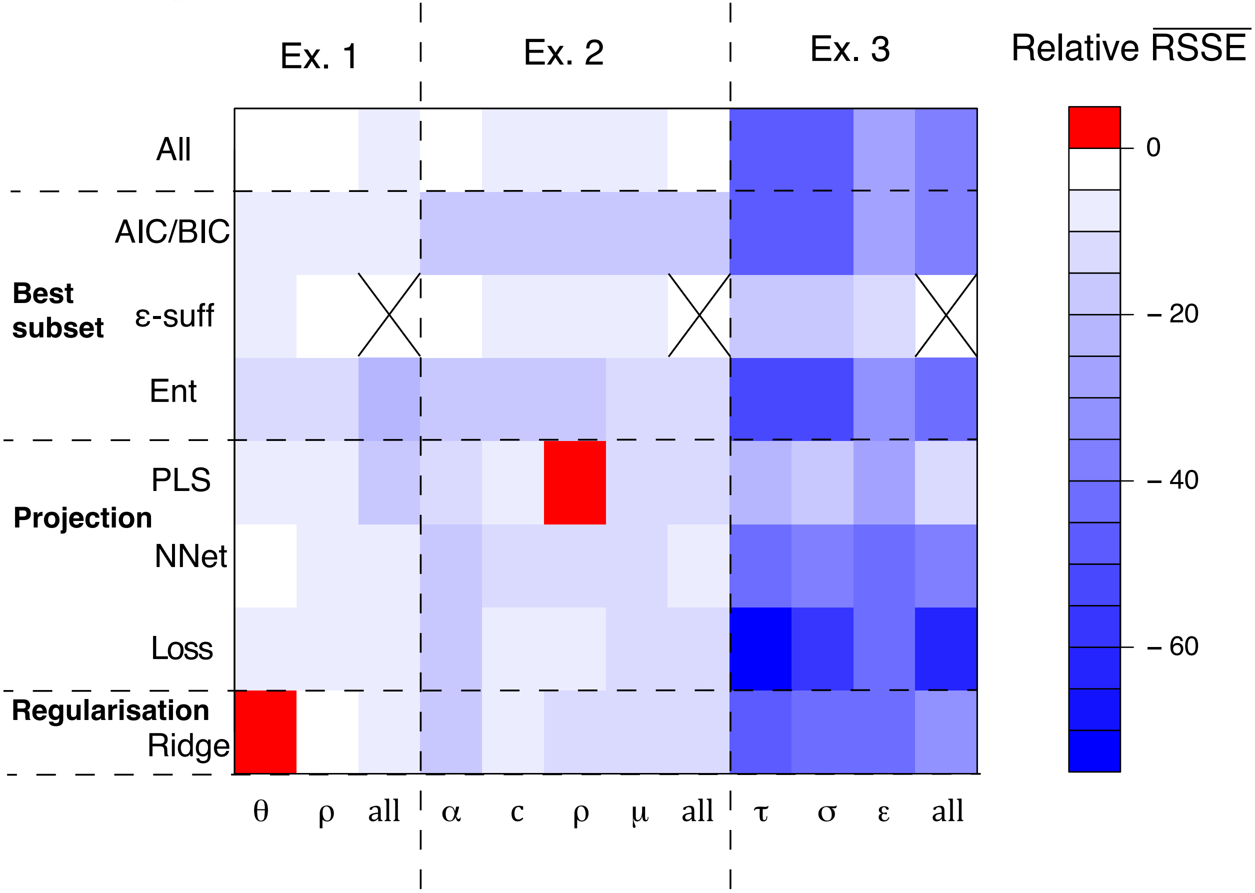


Figure 2: Scatterplots of relative RSSE versus the condition number of the matrix $X^\top WX$ for linear least-squares (top) and ridge (bottom) regression adjustments. Points are based on joint inference for (α, c, ρ, μ) in Example 2 using 1,000 randomly selected vectors of summary statistics, s^i, as "observed" data. When the minimum eigenvalue, λ_{min}, is zero, the (infinite) condition number is arbitrarily set to be 10^{25} for visual clarity (open circles on the scatterplot).

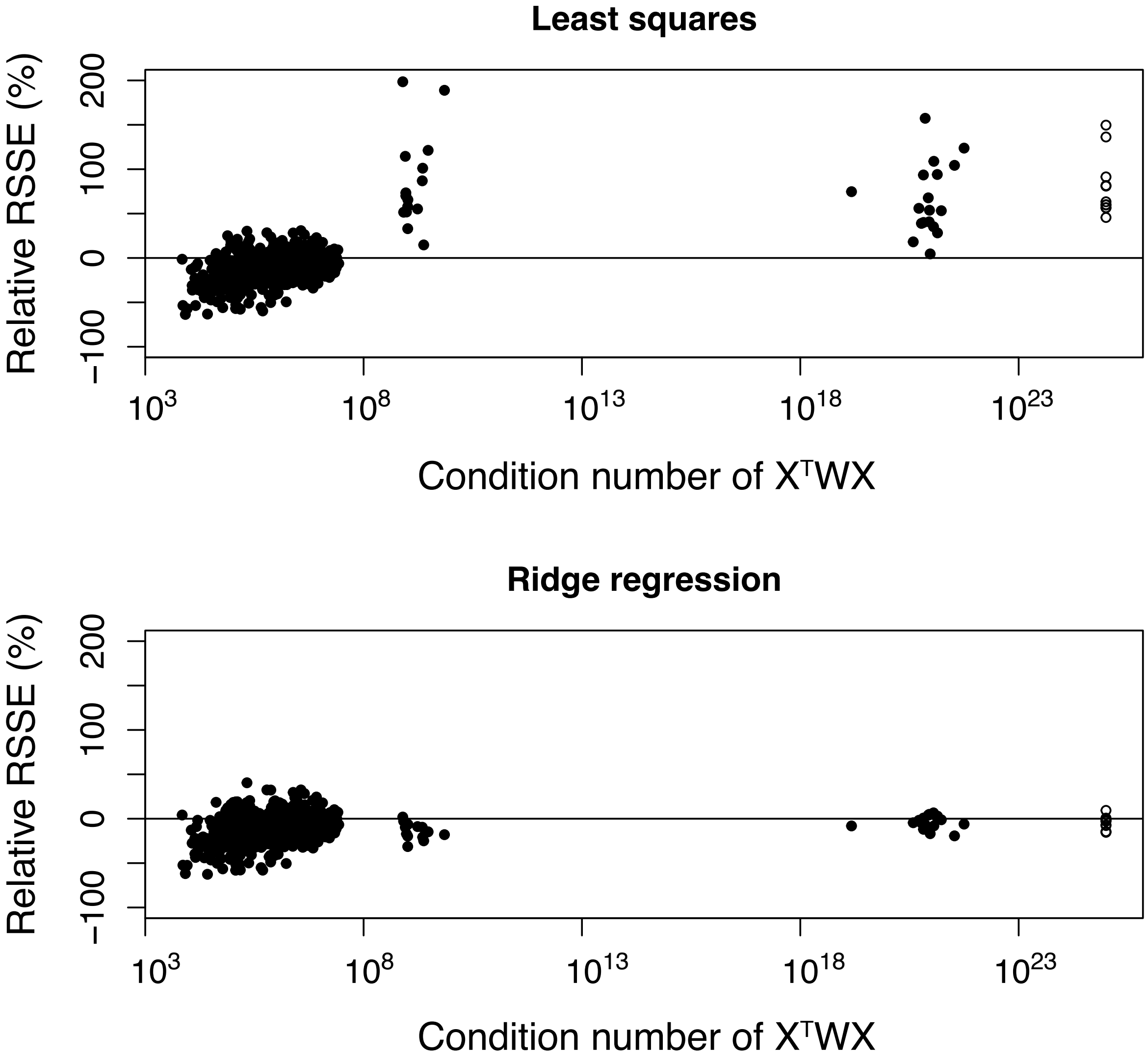
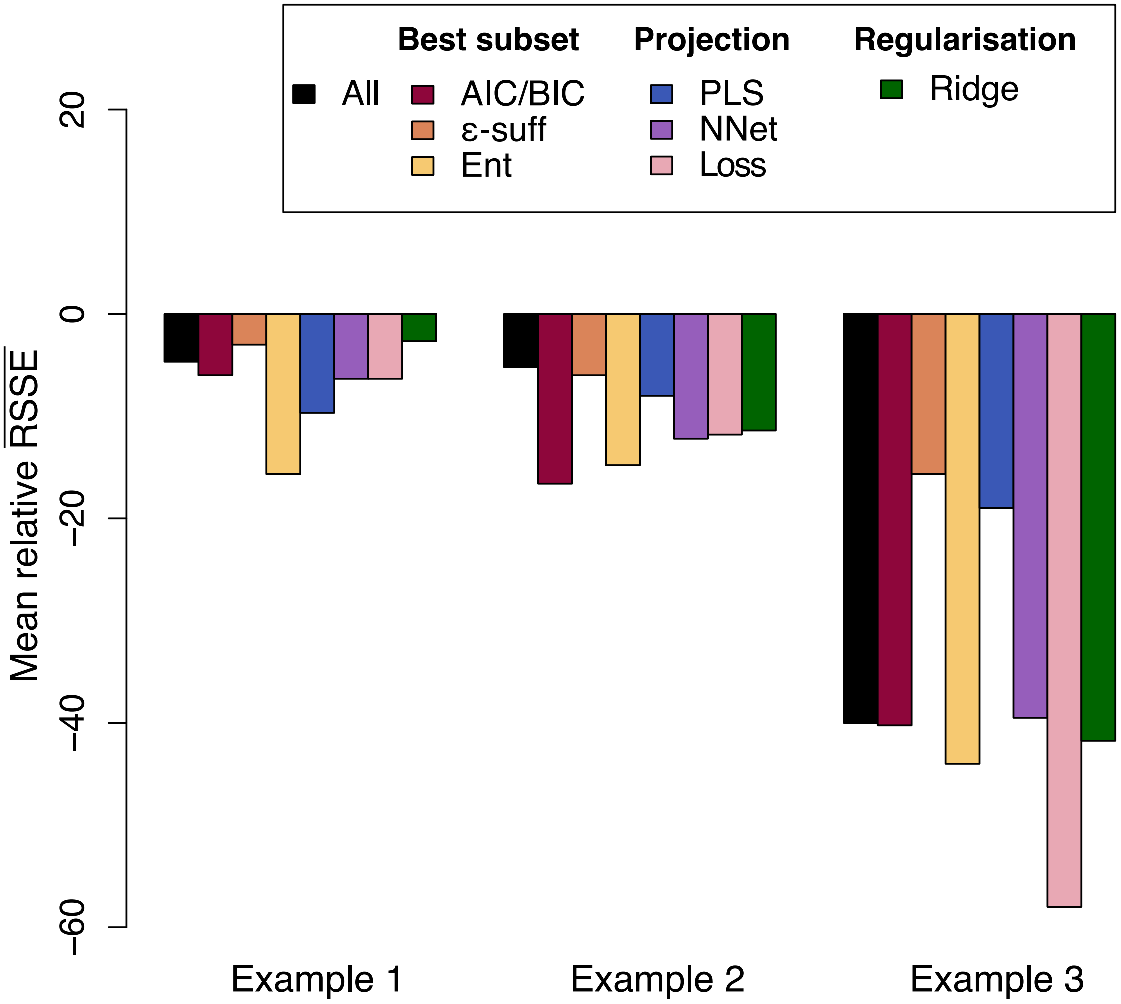


Figure 3: Mean relative $\overline{\text{RSSE}}$ values using each method of dimension reduction and for each example. Methods of dimension reduction include no dimension reduction (All), AIC/BIC, the ε-sufficiency criterion (ε-suff), the two-stage entropy procedure (Ent), partial least squares (PLS), neural networks (NNet), minimum expected posterior loss (Loss) and ridge regression (Ridge). For Examples 1 and 2, the results for ridge regression and neural networks estimate $m(s)$ and $\sigma^2(s)$ have been obtained by taking the pointwise median curve over varying values of the regularisation parameter; $\lambda = 10^{-3}, 10^{-2}$ and 10^{-1} (see introduction to Section 4). For Example 3, an optimal value of λ was chosen based on a cross-validation procedure (see Section 4.3).


Best subset
Projection
Regularisation
All
AIC/BIC
ε-suff
Ent
PLS
NNet
Loss
Ridge
Mean relative $\overline{\text{RSSE}}$
Example 1
Example 2
Example 3