Finite Sample Performance of Statistic Selection Procedures for Approximate Bayesian Computation

Yannick Max Schindler

Dissertation Supervisor:
Dennis Kristensen

Dissertation submitted in part-fulfilment of the MSc in Economics, University College London
September 2016

Declaration

I, Yannick Max Schindler, hereby declare that the work presented in this dissertation is my own original work. Where information has been derived from other sources, I confirm that this has been clearly and fully identified and acknowledged. No part of this dissertation contains material previously submitted to the examiners of this or any other university, or any material previously submitted for any other assessment.

Name: Yannick Max Schindler
Date: 14 September 2016

Classification

□ an empirical/econometric study
☑ the examination of a theoretical problem
□ a critical analysis of a policy issue
□ an analytical survey of empirical and/or theoretical literature

Dissertation Library

I give permission for a PDF electronic copy of this dissertation to be made available to future students by way of the Dissertation Library, should the Department of Economics choose to use it in this way.

Yes

Abstract

This simulation study assesses the performance of two statistic selection techniques for approximate Bayesian computation. Performance is first assessed in a benchmark linear regression setting and subsequently in the context of estimating a discrete time stochastic volatility process for which no likelihood function is known in closed form. A discussion of the sufficiency criteria from which each selection procedure is motivated is also provided. Lastly, this study makes explicit the relationship between one of the selection procedures, based on Kullback-Leibler divergence, and the general information theoretic framework which underpins it. The code used to generate the simulation results in this study is accessible at: https://github.com/stat-select-abc/statistic_selection_for_abc.git.

Acknowledgements

I express my gratitude to Dennis Kristensen for his immensely helpful guidance and supervision in writing this dissertation. I would also like to thank my parents for their unwavering support and encouragement over the years. Lastly, I thank Jorien for her invaluable edits on earlier drafts.

Contents

List of Figures

List of Tables

1 Introduction

The uncanny accuracy of Gordon Moore's forecast made in the 1960s, familiarly known as Moore's Law, and the associated exponential growth in computing power experienced since then has influenced many areas of research, with econometrics being no exception. In particular, this growth in computing has sparked a so-called simulation revolution for simulation-based Bayesian econometric techniques that rely, at their core, on the generation of pseudo-random numbers and varied simulated sampling techniques (Hoogerheide et al., 2009).

With its first formulations dating back to the early 18th century, Bayesian inference has long been a foundational concept in statistical theory as well as application. However, it was not until the 20th century that application of the Bayesian paradigm to econometrics emerged through the pioneering work of Zellner (1971) with an increase in simulation methods used for econometric inference developing over recent decades (Greenberg, 2008; Hoogerheide et al., 2009). Econometric techniques based on Bayesian principles lend themselves particularly well to such simulation-based techniques since a fundamental component required for Bayesian inference, namely the likelihood function, may not always be known in closed form. In such a setting it may nevertheless be possible to simulate data from the stochastic model in question, allowing for the application of simulation-based inference techniques such as approximate Bayesian computation (ABC).

Approximation Bayesian computation, introduced in more detail in the next section, is a method that compares observed data to simulated data to infer about the parameters of a model. Generally, estimates based on such techniques suffer from the curse of dimensionality necessitating the use of dimension reduction techniques and opening the door to a delicate balancing act between the reduction of dimensionality and the loss of information contained in simulated data. This trade-off has motivated literature on various methods for minimizing dimensionality constrained by restrictions on information loss. An overview of some works in this body of literature is provided in **Section 3**.

Creel and Kristensen (2015*b*) propose such dimension reduction techniques and a review of these techniques is given **Section 4**. Furthermore, the current study provides simulation results that compare the performance of these data reduction techniques in a benchmark setting proposed by Creel and Kristensen (2015*b*). Additionally, the performance of one of these dimension reduction techniques, namely one based on Kullback-

Leibler divergence (KLD), is assessed in the context of estimating the parameters of a discrete time stochastic volatility (SV) process. These simulation results, both from the benchmark case and the discrete time SV process, are given in **Section 6**.

2 Introduction to Approximate Bayesian Computation

2.1 Motivation

Bayesian statistical inference fundamentally rests on using Bayes' theorem as a learning device to update beliefs on the parameters of a model:

$$f(\theta|y) \propto f(y|\theta)\pi(\theta) \tag{1}$$

where θ is vector of parameters parameterizing the model $f(y|\theta)$ for y and $\pi(\theta)$ is a subjective prior density for θ. Such updating results in the construction of a posterior density function of parameters conditional on data generated from the model in question. It is this conditional density that represents the object of main focus in Bayesian statistical inference. For example, the posterior distribution can be used to calculate point estimates for the parameters of a model as:

$$\hat{\theta} = \int_{\Theta} \theta f(\theta|y) d\theta \tag{2}$$

or to provide information on the region of highest posterior density which is expressed as an interval on which, for example, 95% of the posterior density lies. Furthermore, the posterior density can be used as a model selection method by calculating Bayes factors (see e.g. Jeffreys, 1961 or Kass and Raftery, 1995).

In general, it is possible to express the posterior density in closed form given that both the prior density and the likelihood function of the model are known. As the prior density is a subjective choice of the researcher, it is always known. However, some models do not provide an analytically tractable likelihood function precluding a direct application of Bayes' theorem to express the posterior density in closed form. Approximate Bayesian computation is motivated by this inherent limitation of Bayesian statistical inference and expands the class of models to which Bayesian inference can be applied to include models with analytically intractable likelihood functions.

2.2 Simulating the posterior distribution

Approximate Bayesian computation techniques generate a sample from an approximated posterior density and subsequently use sample statistics to infer the properties of the population posterior density. Approximating a posterior distribution in this manner inherently relies on simulating data from a model and comparing simulated data to observed data. These methods date back to the 1980s (Rubin, 1984) and were first developed in the context of population genetics with the term "approximate Bayesian computation" originating from four papers at the turn of the millennium: Tavare et al. (1997), Pritchard et al. (1999), Marjoram et al. (2003), and Beaumont et al. (2002) (Beaumont, 2010). Approximate Bayesian computation techniques are relevant to any discipline in which models with unknown likelihood functions are used, including econometrics, and the popularity of such techniques should be buoyed by ever increasing computational resources available to researchers.

To generate samples from an approximated posterior distribution, ABC methods rely on different sampling methods including simple rejection sampling (e.g. Pritchard et al., 1999), importance sampling (e.g. Creel and Kristensen, 2015*a*), Markov Chain Monte Carlo (MCMC) (e.g. Marjoram et al., 2003), and sequential Monte Carlo (SMC) (e.g. Luciani et al., 2009). Although differing in mechanics, each of these sampling techniques rests on the aforementioned method of comparing simulated data to observed data. The current study utilizes rejection sampling, a sampling technique that is motivated by (1). The relationship in (1) illustrates that repeatedly drawing from the prior $\pi(\theta)$ and accepting draws with probability according to $f(y|\theta)$ generates a sample of draws drawn from a distribution proportional to the posterior density. In essence, this type of rejection sampling forms the numerical analogue of Bayesian updating with information about the posterior distribution being introduced via the comparison of simulated and observed data. The above steps are at the center of ABC-based inference and while drawing from the prior distribution is in practice a straightforward exercise, accepting draws with a probability induced by $f(y|\theta)$ requires more care.

Given the relationship presented by (1), a natural criteria for the acceptance of simulated draw θ^s from the prior is $y^s = y$ where y^s is data simulated from the model parametrized by θ^s and y is an n-dimensional vector of observed data. Put differently, a draw from the prior is accepted whenever the data simulated from it is identical to the observed data. This criterion is only of theoretical and expository interest as implementing such a rejection algorithm would be computationally inefficient given that in many applications the probability of simulating data identical to the observed data is negligible. The acceptance rate of the above accept-reject scheme is $\mathcal{P}(y) = \mathcal{P}(y|\theta)\mathcal{P}(\theta)$ inducing a negative binomial distribution for the number of simulations given that a sample of size m is accepted. This distribution has mean $m/\mathcal{P}(y)$ highlighting the computational effort required to generate adequately sized samples when $dim(y)$ is large and rendering such a sampling technique impractical to implement (Wilkinson and Tavare, 2008).

Algorithm 1 Accept-reject

```
Observe data $y$.
Draw $S$ parameter vectors $\theta^s$ from the prior $\pi(\theta)$.
for all $\theta^s$ do
    Simulate a dataset $y^s$ according to the model parameterized by $\theta^s$.
end for
Accept all $\theta^s$ for which the associated $y^s$ is such that $||y^s - y|| \leq \epsilon$.
```

A natural extension of the above acceptance criterion is to relax its restrictiveness by accepting draws whose respective simulated data are approximate, rather than identical, to the observed data. One approach to relax the condition $y^s = y$ involves defining a tolerance level ϵ as well as some metric $|| \cdot ||$ on the vector space $\mathbb{R}^n$ and accepting draw θ^s whenever $||y^s - y|| \leq \epsilon$ giving rise to **Algorithm 1** which is the classic accept-reject approach to ABC. With a non-zero tolerance ϵ, simulated draws are accepted with a probability that is approximately $f(\theta|y)$ with the approximation worsening as ϵ increases. The following limiting behavior illustrates the relationship between ϵ and the respective approximation of the posterior distribution $\hat{f}_\epsilon(\theta|y)$ using this type of rejection sampling (Wilkinson, 2013):

$$\lim_{\epsilon \to \infty} \hat{f}_\epsilon(\theta|y) = \pi(\theta) \qquad\qquad \lim_{\epsilon \to 0} \hat{f}_\epsilon(\theta|y) = f(\theta|y)$$

Dean et al. (2014) deliver theoretical results on the relationship between the magnitude of ϵ and the resulting bias in the context of an ABC maximum likelihood estimator. Furthermore, Fearnhead and Prangle (2012) provide a theoretical investigation of the relationship between the choice for ϵ and the quadratic loss of an ABC-based posterior mean estimate.

One of the characteristics of the above detailed approach to simulating a posterior distribution is that, conditional on belonging to the set $\{y^s \,|\, \epsilon \geq ||y^s - y||\}$, all associated parameter values θ^s are accepted with certainty and consequently with equal probability. Put differently, there is no mechanism to emphasize, within the pairs (θ^s, y^s) that have been accepted, the pairs with smaller $||y^s - y||$ over the pairs with greater $||y^s - y||$. This rectangular acceptance region, expressible as $\mathbb{1}(y = y^s)$ evaluating to unity if and only if $y = y^s$, can be modified to feature smooth weighting instead as proposed by Beaumont et al. (2002). Didelot et al. (2011) provide the theoretical motivation for this smooth weighting approach by expressing the likelihood function evaluated at the observed dataset $f(y|\theta)$ as a degenerate integral:

$$f(y|\theta) = \int_Y \mathbb{1}(y = \tilde{y}) f(y|\theta)\, d\tilde{y} \tag{3}$$

Alternatively, the indicator function in (3) can be seen as a density function with singular mass. Replacing this density having singular mass at $y = \tilde{y}$ with a kernel function centered on y gives the following approximation:

$$\hat{f}(y|\theta) = \int_Y K_h(y - \tilde{y}) f(y|\theta)\, d\tilde{y} \tag{4}$$

where $K_h(\cdot)$ is a kernel function with bandwidth $h > 0$. The above integral shows that the expected value of $K_h(y - y^s)$ over the distribution $f(y|\theta)$ is equal to $\hat{f}(y|\theta)$ allowing for Monte Carlo integration. The resulting ABC rejection algorithm, due to Pritchard et al. (1999) is then **Algorithm 2**. Initial variations of this approach to generating a posterior sample were proposed in the population genetics literature by Tavare et al. in 1997, Fu and Li in 1997, and Weiss and von Haeseler in 1998 (Didelot et al., 2011).

Algorithm 2 Smoothed accept-reject

```
Draw $S$ parameter vectors $\theta^s$ from the prior $\pi(\theta)$.
for all $\theta^s$ do
    Simulate a dataset $y^s$ according to the model parameterized by $\theta^s$.
    Accept $\theta^s$ with probability proportional to $K_h(y - y^s)$.
end for
```

Even with the above concession of accepting draws from the prior distribution with a probability that is approximate to $f(\theta|y)$, the number of simulations required to generate an adequately sized posterior sample may be prohibitive for high dimensional data y. In terms of ABC estimation of the posterior mean, this characteristic of the sampling method translates into a drastically slower rate of convergence to the population parameter (Blum and Francois, 2010). To address this particular manifestation of the curse of dimensionality, summary statistics can be used to drastically reduce the dimension of the full dataset. The selection of appropriate statistics for this type of data compression is the main focus of this study.

2.3 Statistic selection for approximate Bayesian computation

The core motivation of mapping the full data vector to a lower dimensional space through the use of summary statistics is to achieve feasibility as well as computational advantages in evaluating ABC-type estimators (Creel and Kristensen, 2015*b*). However, dimension reduction comes at a cost: possible loss of information and the consequent deterioration of the approximation of the posterior distribution as well as an inflation of credible intervals (Blum et al., 2013; Sunnaker et al., 2013). This property of ABC-type estimators can be viewed in the context of generalized method of moments (GMM) estimation. As Creel and Kristensen (2013) establish the first order equivalence of an ABC-type estimator that approximates (2) based on summary statistics Z_n and the efficient GMM estimator using Z_n as moment conditions, it is worthwhile to note the relationship between the choice of moment conditions and the asymptotic variance of GMM estimators. While adding moment conditions to a baseline set cannot increase the asymptotic variance of GMM estimators, the small sample performance of such estimators may suffer from the addition of weakly or uninformative moment conditions (Stock et al., 2002; Tauchen, 1986).

Exacerbating the deterioration of the small sample performance of the SBIL estimator is the bias and error of the estimator introduced alone by simulation. Creel and Kristensen (2015*b*) give this bias and variance from simulation as being of order $O(h^2)$ and $O(1/Sh^{dim(W)})$ respectively when a second order kernel is employed where $h > 0$ is the bandwidth of the kernel, S is the number of simulated draws, and W is the set of summary statistics used in the estimation. Consequently, as Creel and Kristensen (2015*b*) remark, with an optimal bandwidth choice of order $h = O(S^{-1/(4+dim(W))})$ the error rate due to simulations is of order $O(S^{-2/(4+dim(W))})$, further illustrating the deterioration in performance of the SBIL estimator when uninformative statistics are used. This sensitivity to the choice of summary statistics Z_n provides the impetus for considerate frameworks by which to select summary statistics for ABC-type estimation. The following section provides an overview of recent contributions made to the development of such frameworks.

3 Literature Review

Blum et al. (2013) provide a comprehensive review of the recent literature on statistic selection for ABC and propose a taxonomy for the different approaches proposed in the literature. This classification consists of three non-mutually exclusive categories: best subset selection, projection techniques, and regularization methods. As the statistic selection methods assessed and modified in this study fall into the best subset selection class of methods, this section is dedicated to providing a brief review of recent techniques proposed in this category. A review of the primary motivation of the current study, namely the two ABC statistic selection technique developed by Creel and Kristensen (2015*b*) is given in **Section 4**. The methods belonging to the best subset selection category proposed in the recent literature aim at finding an optimal balance between two common sources of error in ABC-based methods: 1) the error arising from loss of information through data reduction (e.g. by using summary statistics) and 2) the error that arises from the inefficiency of high dimensional kernel smoothers (Blum et al., 2013). As a result, a fundamental theme in the literature reviewed below is the identification of a subset of statistics from a candidate set that satisfies some notion of minimal sufficiency (Blum et al., 2013).

3.1 Joyce and Marjoram (2008)

In spirit of the central theme discussed above, Joyce and Marjoram (2008) base their subset selection method on a proposed concept of “approximately sufficient”. The authors define this concept as a set of statistics $\{S_1, S_2, \dots, S_{k-1}\}$ being ε-sufficient relative to a statistic X if:

$$\sup_\theta \log f(X|S_1, S_2, \dots, S_{k-1}, \theta) - \inf_\theta \log f(X|S_1, S_2, \dots, S_{k-1}, \theta) \leq \varepsilon \tag{5}$$

where θ is the parameter vector of interest and ε is a positive real number. Joyce and Marjoram (2008) propose sequentially adding statistics to a set of summary statistics $\{S_1, S_2, \dots, S_{k-1}\}$ by scoring a statistic S_k considered for addition as:

$$\delta_k = \sup_\theta \log f(S_k|S_1, S_2, \dots, S_{k-1}, \theta) - \inf_\theta \log f(S_k|S_1, S_2, \dots, S_{k-1}, \theta) \tag{6}$$

and stopping to add statistics once δ_k drops below a certain threshold. In practice, this concept is implemented through the use of a ratio of posterior distributions:

$$R_k(\theta) = \frac{f(\theta|S_1, S_2, \dots, S_k)}{f(\theta|S_1, S_2, \dots, S_{k-1})} \tag{7}$$

which, if above a θ-specific threshold value, dictates the addition of s_k to the set $\{S_1, S_2, \dots, S_{k-1}\}$. In such a case, an above-threshold $R_k(\theta)$ is considered as evidence that S_k contains enough information relative to $\{S_1, S_2, \dots, S_{k-1}\}$ to warrant the increase in dimensionality (Blum et al., 2013). If a subset of statistics is reached for which no remaining statistic from the candidate set merits inclusion, the reached subset is considered to be ε-sufficient relative to the remaining candidate statistics and is chosen as the optimal subset.

The authors propose assessing candidate statistics at random which may lead to the inclusion of a statistic after a less informative statistic has already been added to the set. Consequently, the resulting subset from the procedure may not be optimal in that its cardinality is unnecessarily large. To safeguard against the inclusion of relatively uninformative statistics simply due to the randomized order in which they were assessed, Joyce and Marjoram (2008) propose an additional step performed after each addition of a statistic. This second step considers each statistic S_i included in the updated set S_A by assessing its information content relative to the modified subset $S_A \setminus S_i$ and removing S_i if is not included according to the same criterion and methodology described above.

The use of the above-specified posterior ratio as a central criterion for assessing the information content of a statistic rests on the assumption that changes in the posterior density $f(\theta|S_1, S_2, \ldots, S_k)$ relative to $f(\theta|S_1, S_2, \ldots, S_{k-1})$ are an indication that S_k is informative about θ (Blum et al., 2013). It is however possible that an uninformative statistic S_k is added according to the posterior ratio criterion if S_k is evaluated on data drawn from a low density region of the likelihood function which sufficiently alters the numerator to generate what is, in essence, a false positive (Blum et al., 2013; Sisson et al., 2009).

The above disadvantage aside, another feature of the method proposed by Joyce and Marjoram (2008) that severely limits its applicability is, as pointed out by Blum et al. (2013), the computational burden required to evaluate $R_k(\theta)$ for each θ. This limitation restricts the applicability of the proposed method essentially to settings in which $dim(\theta) = 1$ as is the case in the example provided by the authors (Blum et al., 2013).

3.2 Nunes and Balding (2010)

Another approach to statistic selection for ABC is proposed by Nunes and Balding (2010) whose suggested technique has, at least in part, parallels to the method proposed by Creel and Kristensen (2015b). Nunes and Balding (2010) propose a two stage procedure that utilizes minimization of entropy in the posterior distribution and subsequent simulated cross validation as a method for the selection of statistics for ABC. The authors utilize the k-nearest neighbor estimator of entropy defined as:

$$\hat{H} = \log[\frac{\pi^{p/2}}{\Gamma(p/2+1)}] - (k) + \log n + \frac{p}{n}\sum_{i=1}^{n} \log R_{i,k} \tag{8}$$

where p is the dimension of the parameter space, $R_{i,k}$ is the Euclidean distance between the i-th simulated parameter and its k-th nearest neighbor, n is the number of simulations, and $\psi(x) = \Gamma'(x)/\Gamma(x)$ denotes the digamma function. The estimated entropy $\hat{H}$ is minimized over all possible subsets in the combinatorial space spanned by the candidate set where $\hat{H}$ is computed from a posterior distribution approximated by rejection sampling. A second stage then assesses the performance of the selected subset of statistics $\mathcal{Z}$, with associated random vector $Z = \mathcal{Z}(Y)$ and sample realization $z = \mathcal{Z}(y)$, by first identifying a set of J simulated parameter and reduced data pairs (z^j, θ^j) that are close (as determined by Euclidean distance) to the observed summary statistic vector $z(y)$.

Subsequently, the root sum of squared errors (RSSE) between the simulated parameter values (treated as observed) and parameter values drawn from an approximated posterior distribution is calculated. The subset of statistics that minimizes the mean RSSE averaged across the J simulated pairs (z^j, θ^j) is chosen as the optimal subset of statistics.

One weakness in the above-described method is that the first stage relies on the assumption that adding an informative statistic to a set of statistics will reduce the entropy of the respective posterior distribution which is not a general result and does not, for example, hold when the posterior is relatively diffuse compared to the prior (Blum et al., 2013). Additionally, this two stage procedure is, as pointed out by Creel and Kristensen (2015*b*), computationally intractable for large sets of candidate statistics. Recognizing this shortcoming of the method, Nunes and Balding (2010) suggest either limiting the cardinality of the sets to search over (this choice being left to the researcher) or employing an iteratively updating search that adds and removes candidate statistics to a subset, as done by Joyce and Marjoram (2008).

3.3 Barnes et al. (2012)

The selection procedures proposed by Joyce and Marjoram (2008) and Nunes and Balding (2010) are special cases of the information theoretic framework upon which Barnes et al. (2012) develop their proposal for a selection algorithm. To motivate their subset selection algorithm, Barnes et al. (2012) rely on the following expression for mutual information:

$$I(\Theta; Y) \equiv \int_{\Theta} \int_{\mathcal{Y}} f(\theta, y) \log \frac{f(\theta, y)}{f(\theta) f(y)} \, d\theta \, dy$$

and define a set of summary statistics $\mathcal{W}$ to be sufficient for Θ if and only if $I(\Theta; Y)$ is equal to $I(\Theta; W)$ where $W = \mathcal{W}(Y)$ and Y is distributed according to $\theta \in \Theta$. To identify the subset $\mathcal{Z} \subseteq \mathcal{W}$ with minimum cardinality, such that $\mathcal{Z}$ remains sufficient for Θ according to the definition given above, the authors note that the mutual information between Θ and W conditional on Z can be expressed as:

$$I(\Theta; W|Z) = I(\Theta; W) - I(\Theta; Z) \tag{9}$$

where $\mathcal{W}$ is assumed to be a sufficient set of summary statistics for Θ. As a consequence of (9), Barnes et al. (2012) endeavor to find the subset $\mathcal{Z}$ such that $I(\Theta; W|Z) = 0$ or, put differently, such that $I(\Theta; W) = I(\Theta; Z)$.

With this objective formally established, Barnes et al. (2012) further note that given a subset $\mathcal{T} \subseteq \mathcal{Z} \subseteq \mathcal{W}$, the mutual information between Θ and W conditional on $T = \mathcal{T}(Y)$ can be expressed as:

$$I(\Theta; W|Z) = I(\Theta; W|T) - I(\Theta; Z|T)$$

from which it follows that $I(\Theta; W|Z)$ is always less than $I(\Theta; W|T)$ and to construct $\mathcal{Z}$ such that $I(\Theta; W|Z) = 0$, it is sufficient to add summary statistics $S_k \in \mathcal{W}$ incrementally until the condition holds. Barnes et al. (2012) further remark that an increase in efficiency can be gained by noting that:

$$I(\Theta; W|S_1, \ldots, S_k) = I(\Theta; W|S_1, \ldots, S_{k-1}) - I(\Theta; S_1, \ldots, S_k|S_1, \ldots, S_{k-1})$$

and adding a candidate statistic to the subset being constructed if, at a given increment $k > 1$, the statistic under consideration for addition $\tilde{S}$ is such that:

$$S_k = \underset{\tilde{S} \in S \setminus \{S_1, \ldots, S_{k-1}\}}{\text{argmax}} \quad I(\Theta; S_1, \ldots, S_{k-1}, \tilde{S})|S_1, \ldots, S_{k-1})$$

Put differently, at each increment, the statistic added to the subset is the one that maximizes the difference between $I(\Theta; W|S_1, \ldots, S_k)$ and $I(\Theta; W|S_1, \ldots, S_{k-1})$ where $I(\Theta; W|S_1, \ldots, S_k)$ is always less than or equal to $I(\Theta; W|S_1, \ldots, S_{k-1})$. To motivate a selection procedure from this framework, $I(\Theta; W|Z)$ is expressed as the expectation of a Kullback-Leibler divergence from $f(\Theta|\mathcal{Z}(Y))$ to $f(\Theta|\mathcal{W}(Y))$ with respect to the distribution of the data. The resulting sufficiency criterion:

$$I(\Theta; W|Z) = E_{\sim W}[KL(f(\Theta|W)||f(\Theta|Z))] = 0 \tag{10}$$

makes explicit that a subset of statistics Z satisfying (10) is such that:

$$KL(f(\Theta|w)||f(\Theta|z)) = 0$$

for every realization w of W. This feature of the subset follows from the non-negativity of $KL(f(\Theta|w)||f(\Theta|z))$ for all w. As such, the sufficiency criterion $I(\Theta;W|Z) = 0$ is achieved for subsets of summary statistics for which $f(\theta|w) = f(\theta|z)$ for all θ and all w. However, in practice, the researcher is often presented with a particular realization of Y, namely y^*, which relaxes the sufficiency criterion to apply only locally at y^* rather than across the entire distribution of Y. This relaxed sufficiency criterion then becomes $f(\Theta|w^*) = f(\Theta|z^*)$ and is achieved if and only if:

$$KL(f(\Theta|\mathcal{W}(y^*))||f(\Theta|\mathcal{Z}(y^*))) = 0 \tag{11}$$

Consequently, the practical implementation of the above described method when the researcher has observed y^* consists of adding the following candidate statistic at each increment $k > 1$:

$$S_k^* = \underset{\tilde{S}^* \in S^* \setminus \{S_1^*, \ldots, S_{k-1}^*\}}{\text{argmax}} \; KL(\hat{f}(\Theta|S_1^*, \ldots, S_{k-1}^*, \tilde{S}^*)||\hat{f}(\Theta|S_1^*, \ldots, S_{k-1}^*)$$

until, in theory:

$$KL(\hat{f}(\Theta|S_1^*, \ldots, S_{k-1}^*, S_k^*)||\hat{f}(\Theta|S_1^*, \ldots, S_{k-1}^*) = 0 \tag{12}$$

Since the left hand side in (12) is approximated, it may not reach zero even when $f(\Theta|S_1^*, \ldots, S_{k-1}^*, S_k^*) = f(\Theta|S_1^*, \ldots, S_{k-1}^*)$ and (11) is satisfied. As such, the researcher must define a criterion whereby the selection algorithm stops as soon as:

$$KL(\hat{f}(\Theta|S_1^*, \ldots, S_{k-1}^*, S_k^*)||\hat{f}(\Theta|S_1^*, \ldots, S_{k-1}^*) \leq \delta$$

where $\delta > 0$ is a tolerance level chosen by the researcher.

One drawback of the selection algorithm proposed by Barnes et al. (2012) is the associated high computational burden when the cardinality of the candidate set is large. Furthermore, the stopping criterion requires the researcher to make a choice for δ which may not be straightforward and may involve additional computation. Irrespective of these characteristics however, the information theoretic framework delivered by Barnes et al. (2012) serves as a useful generalization of the two selection methods proposed by Joyce and Marjoram (2008) and Nunes and Balding (2010). This generalization offers a glimpse into the underlying mechanics of these two selection procedures as well as the KLD-based selection procedure discussed in this study.

3.4 Summarizing remarks

Parallel to the common theme in the recent literature to first formally define and then systematically identify a minimally sufficient subset of statistics is a common limiting factor manifested in computational expense. Prohibitive computational expense restricts the applicability of the above reviewed selection methods in settings with many candidate statistics (e.g. Barnes et al., 2012; Nunes and Balding, 2010) or in settings with multiple parameters (e.g. Joyce and Marjoram, 2008). The two main methods proposed by Creel and Kristensen (2015*b*) address these shortcomings and remain computationally tractable in settings with multivariate parameters and/or large candidate sets. Furthermore, the methods proposed by Creel and Kristensen (2015*b*) allow for varying degrees of sufficiency with one method aimed preserving information about the posterior mean and the other motivated by a sufficiency criterion that encompasses the entire posterior distribution. The relative performance of these selection methods is investigated in the current study in **Section 6**. Additionally, the selection methods provided by Creel and Kristensen (2015*b*) are readily modified to select the subset of summary statistics that is optimal with respect to an observed sample rather than the entire sample space. The current study assesses the performance of these sample-specific analogues in **Section 6** as well.

4 Motivation: A review of Creel and Kristensen (2015*b*)

4.1 Expected Bayesian loss (EBL)

As was the case with the methods summarized in the literature review of this study, the procedures proposed by Creel and Kristensen (2015*b*) belong to the best subset selection class of methods aiming to select an optimal subset from a candidate set of statistics. To give form to the notion of "optimality" in the context of selecting statistics for ABC, Creel and Kristensen (2015*b*) provide the following constrained minimization:

$$\delta_0 = \operatorname*{argmin}_{\delta \in \Delta} \sum_{p=1}^{P} \delta_p \,\text{subject to}\, E[\Theta|W] = E[\Theta|Z(\delta)] \tag{13}$$

where W is the candidate set of statistics, $P = dim(W)$, δ is a $P \times 1$ vector of zeros and ones with a one indicating that the corresponding summary statistic is included in the selected subset of W, $\Delta = \{0,1\}^P$, and $Z(\delta) \subseteq W$ is the subset of the candidate set dictated by δ. The constraint in (13):

$$E[\Theta|W] = E[\Theta|Z(\delta)] \tag{14}$$

forms the sufficiency criterion from which Creel and Kristensen (2015*b*) develop the following statistic selection procedure. The solution to (13) is interpretable as the subset of statistics with the smallest cardinality for which no information loss on the posterior mean occurs.

As $f(\Theta, W)$ is not known in closed form, $E[\Theta|W]$ and $E[\Theta|Z(\delta)]$ cannot be evaluated analytically and must instead be approximated numerically. To translate (13) into a numerically evaluable analogue, Creel and Kristensen (2015*b*) propose a decision theoretic framework consisting of action $\hat{E_S}[\Theta|Z(\delta) = z(\delta)]$ and decision $\delta \in \Delta$ where $\hat{E_S}[\Theta|Z(\delta) = z(\delta)]$ is the simulated Bayesian indirect likelihood (SBIL) estimator:

$$\hat{\theta}_{SBIL} = \hat{E}_S[\theta|z_n] = \frac{\sum_{s=1}^{S} \theta^s K_h(z_n^s - z_n)}{\sum_{s=1}^{S} K_h(z_n^s - z_n)} \tag{15}$$

for which $z_n = Z(y^*)$ is a set of statistics calculated on observed data and (θ^s, z_n^s) are simulated pairs consisting of a parameter vector drawn from a subjective prior $\pi(\theta)$ and $z_n^s = Z(y^s)$ where y^s is data simulated from a model parameterized by respective θ^s. Details of the SBIL estimator are given in **Appendix 1**. The resulting expected Bayesian loss is then formulated as:

$$\mathscr{B}_S(\delta|w) \equiv E\Big[L\big(\Theta - \hat{E}_S[\Theta|Z(\delta)]\big)\Big|Z(\delta) = z(\delta)\Big] \tag{16}$$

$$= \int_{\mathbb{R}^k} L\Big(\theta - \hat{E}_S[\Theta|Z(\delta) = z(\delta)]\Big) f(\theta|z(\delta))\, d\theta \tag{17}$$

where $L(\cdot)$ is a loss function chosen by the researcher and $k = dim(\theta)$. Minimizing the expected Bayesian loss in (17) would provide an estimate for the optimal subset of statistics according to the sample dependent optimality condition (14), namely $E[\Theta|W = w] = E[\Theta|Z(\delta) = z(\delta)]$.

Creel and Kristensen (2015*b*) propose expanding (17) to identify the $Z(\delta)$ that is optimal for the model of interest irrespective of the observed data. Integrating (17) with respect to all realizable W gives the following integrated loss function:

$$\mathscr{L}_S(\delta) \equiv \int_W \mathscr{B}_S(\delta|w) f(w)\, dw \tag{18}$$

$$= \int_{Z(\delta)} \int_{\mathbb{R}^k} L\Big(\theta - \hat{E}_S[\Theta|Z(\delta) = z(\delta)]\Big) f(\theta, z(\delta))\, d\theta\, dz(\delta) \tag{19}$$

which provides an estimate for δ_0 that is optimal, in the sense defined in (14), irrespective of which data have been observed. From a practical perspective, the estimate generated by (19) differs from (17) in that (17) may be specific to the sample that has been observed whereas (19) is not and can be applied in any context for which the same model is used.

Algorithm 3 Selection procedure (Creel and Kristensen, 2015*b*)

```
1: Draw $R$ parameter vectors $\theta^r$ from the prior.
2: For each $\theta^r$ simulate data $Y(\theta^r)$.
3: For each $\theta^r$ compute candidate statistics $w(Y(\theta^r))$.
4: Minimize (20) with respect to $\delta$ to obtain an estimate $\hat{\delta}$ characterizing a subset of the candidate set.
```

Since the integrals constituting (19) cannot be evaluated analytically, Creel and Kristensen (2015*b*) employ Monte Carlo integration to approximate the above integrated Bayesian loss criterion with:

$$\hat{\mathscr{L}_S}(\delta) = \frac{1}{R} \sum_{r=1}^{R} L\Big(\theta^r - \hat{E}_S[\Theta|Z(\delta) = z^r(\delta)]\Big) \tag{20}$$

where the θ^r are drawn from the prior $\pi(\theta)$ and $z^r(\delta)$ is a subset of summary statistics calculated from data simulated according to the respective θ^r. The above expression reveals the cross validation inherent in the proposed method whereby the performance of R estimates, each based on S simulations, is assessed.

The selection algorithm Creel and Kristensen (2015*b*) propose based on the above criterion is then **Algorithm 3**. Step 4 involves the minimization of a non-differentiable objective function over a potentially very large, $2^{dim(W)}$, discrete choice set necessitating the use of a numerical optimization method. Creel and Kristensen (2015*b*), as well as the present study, use simulated annealing, a stochastic optimization algorithm, to minimize the objective function given in Step 4. Simulated annealing, covered in more detail in **Section 4** section of this study, is capable of escaping local minima and is a popular choice for discrete optimization problems (Henderson et al., 2003). The authors implement the above algorithm and resulting simulating annealing procedure in Julia, a high performance programming language.

4.2 Kullback-Leibler divergence (KLD)

The expected Bayesian loss criterion arose from the objective of selecting statistics for which no information loss on the posterior mean occurs. However, Creel and Kristensen (2015*b*) note that it is as well interesting to strengthen the restriction (14) to the following:

$$f(\Theta|W) = f(\Theta|Z(\delta)) \tag{21}$$

which restricts $Z(\delta)$ to have a stronger property, namely sufficiency relative to the candidate set. With the condition in (21), the restriction is no longer on information loss for the first moment of the posterior distribution but for higher moments as well. Revisiting the framework provided by Barnes et al. (2012), the following remark motivates the information theoretic foundation with which to identify subsets of the candidate set that are sufficient with respect to the candidate set:

$$\begin{aligned}
I(\Theta;W|Z(\delta)) &= \int_\Theta \int_Z \int_W f(\theta,w,z(\delta)) \log \frac{f(\theta,w|z(\delta))}{f(\theta|z)f(w|z(\delta))}\, dw\, dz\, d\theta \\
&= \int_\Theta \int_W f(w)f(\theta|w) \log \frac{f(\theta|w)}{f(\theta|z(\delta))}\, dw\, d\theta \\
&= \int_W f(w) KL\big(f(\Theta|w)||f(\Theta|z(\delta))\big)\, dw \\
&= E_{\sim W}\Big[KL\big(f(\Theta|W)||f(\Theta|Z(\delta))\big)\Big]
\end{aligned} \tag{22}$$

The above mutual information conditional on $Z(\delta)$ expressed in terms of the expected value of the Kullback-Leibler divergence from $f(\Theta|Z)$ to $f(\Theta|W)$ provides the theoretical foundation of an additional selection procedure proposed by Creel and Kristensen (2015*b*). Rewriting 22 gives:

$$\begin{aligned}
E_{\sim W}[KL(f(\Theta|W)||f(\Theta|Z))] &= \int_\Theta \int_W f(w) f(\theta|w) \log \frac{f(\theta|w)}{f(\theta|z)}\, dw\, d\theta \\
&= \int_\Theta \int_W \log f(\theta|w) f(\theta, w)\, dw\, d\theta \\
&\quad - \int_\Theta \int_W \log f(\theta|z) f(\theta, w)\, dw\, d\theta \qquad (23)
\end{aligned}$$

which is the integrated Kullback-Leibler divergence criterion that Creel and Kristensen (2015*b*) propose to use as a basis for a statistic selection criterion. Viewing the above criterion from an information theoretic perspective illustrates its foundation in the general framework provided by Barnes et al. (2012).

Creel and Kristensen (2015*b*) note that only the second term in (23) depends on δ, motivating the following estimate for δ_0:

$$\hat{\delta} = \underset{\delta \in \Delta}{\operatorname{argmax}} \int_\Theta \int_W \log f(\theta|z(\delta)) f(\theta, w) dw\, d\theta \qquad (24)$$

which effectively minimizes $E_{\sim W}[KL(f(\Theta|W)||f(\Theta|Z))]$ with respect to δ. In (24) the authors use a kernel density estimator for $f(\theta|z(\delta))$:

$$\hat{f}(\theta|z) = \frac{\sum_{s=1}^{S} K_h(\theta^s - \theta) K_h(z^s(\delta) - z(\delta))}{\sum_{s=1}^{S} K_h(z^s(\delta) - z(\delta))} \qquad (25)$$

and apply Monte Carlo integration for a numerical analogue of (24) expressed as:

$$\hat{\delta} = \underset{\delta \in \Delta}{\operatorname{argmax}} \sum_{r=1}^{R} \log \hat{f}(\theta^r | z^r(\delta)) \qquad (26)$$

For large samples, the selection algorithm based on the integrated expected Bayesian loss (19) and the selection algorithm based on the integrated Kullback-Leibler divergence from $f(\Theta|Z)$ to $f(\Theta|W)$ (23) are asymptotically equivalent under specific regularity conditions (Creel and Kristensen, 2015*b*). However, their small sample behavior may be dissimilar. Part of the current study examines the relative performance of these selection procedures in the context of the benchmark case provided by Creel and Kristensen (2015*b*), namely estimating a linear regression model, as well as in the context of estimating a discrete time SV process.

Barnes et al. (2012) as well as Creel and Kristensen (2015*b*) allude to a non-integrated version of (24). Recalling the relaxed sufficiency criterion (11) that applies at a particular realization of Y rather than the entire distribution of Y motivates the non-integrated, conditional Kullback-Leibler divergence from $f(\Theta|z^*(\delta))$ to $f(\Theta|w^*)$:

$$KL(\delta|w) = \int_\Theta \log f(\theta|w)f(\theta|w)d\theta - \int_\Theta \log f(\theta|z(\delta))f(\theta|w)d\theta \tag{27}$$

The Kullback-Leibler based selection criterion (23) proposed by Creel and Kristensen (2015*b*) is (27) integrated over the entire parameter space. Both Barnes et al. (2012) and Creel and Kristensen (2015*b*) remark that (27) is only of theoretical interest, as evaluating (27) involves an approximation of $f(\theta|w)$ which can be computationally prohibitive in cases where $dim(w)$ is large.

4.3 A benchmark case

Creel and Kristensen (2015*b*) provide a diagnostic and benchmarking case with which to assess the performance of the proposed EBL-based statistic selection procedure. This benchmark case comprises of running the selection procedure using a model for which a likelihood function is available and for which optimally informative statistics are available from theory. The authors suggest the linear regression model as such a benchmarking case and assess their proposed selection procedure by constructing a candidate set that includes statistics that are, from theory, optimally informative, statistics that are weakly informative, and statistics that are not informative at all for inference about the model's parameters. In the context of the benchmark linear regression case, the candidate set is constructed from the following three regression equations:

$$y_i = \alpha_L + \sum_{j=1}^{4} \beta_j^L x_{ij} + \sigma_L u_i \tag{28}$$

$$y_i = \alpha_Q + \sum_{j=1}^{4} \beta_j^Q x_{ij} + \sum_{j=1}^{4} \gamma_j^Q x_{ij}^2 + \sigma_Q u_i \tag{29}$$

$$y_i = \alpha_C + \sum_{j=1}^{4} \beta_j^C x_{ij} + \sum_{j=1}^{4} \gamma_j^C x_{ij}^2 + \sum_{j=1}^{4} \delta_j^C x_{ij}^3 + \sigma_C u_i \tag{30}$$

where the data used in the statistic selection procedure are generated from (28). The set of candidate statistics is then constructed from OLS estimates for each of the regression equations above in addition to five noise statistics as detailed in **Table 1**. To identify the set of statistics for which the associated SBIL estimation of (28) is asymptotically efficient, Creel and Kristensen (2015*b*) note that treating the OLS estimators for the parameters of (28) as moment conditions $m(\theta) = \theta - \hat{\theta}_{OLS}$ where $\theta = (\alpha_L, \beta_0^L, \ldots, \beta_4^L, \sigma_L)$ induces an exactly identified GMM estimator with criterion $s(\theta) = m(\theta)'m(\theta)$. This criterion is minimized at $\theta = \hat{\theta}_{OLS}$ and as such the just identified GMM estimate is the maximum likelihood estimate (Creel and Kristensen, 2015*b*). As a result, the just identified GMM estimator is efficient for θ and since Creel and Kristensen (2013) establish the first order equivalence of the SBIL estimator and the efficient GMM estimator, the SBIL estimator employing these statistics is asymptotically efficient as a point estimator (Creel and Kristensen, 2015*b*). Consequently, the OLS estimates for the parameters of (28) are considered optimal statistics for the SBIL estimator and the descriptions in **Table 1** refer to optimality in the sense just described.

Table 1: Candidate statistics for linear regression benchmark case (Creel and Kristensen (2015*b*))

Statistics	Count	Information content
$\hat{\alpha}_L, \hat{\boldsymbol{\beta}}^{\boldsymbol{L}}, \hat{\sigma}_L, \hat{\sigma}_Q, \hat{\sigma}_C$	8	Optimally informative.
$\hat{\alpha}_Q, \hat{\alpha}_C, \hat{\boldsymbol{\beta}}^{\boldsymbol{Q}}, \boldsymbol{\beta}^{\boldsymbol{C}}, \hat{\boldsymbol{\gamma}}^{\boldsymbol{Q}}, \hat{\boldsymbol{\gamma}}^{\boldsymbol{C}}, \hat{\boldsymbol{\delta}}^{\boldsymbol{C}}$	22	Relevant yet suboptimally informative.
$\boldsymbol{\varepsilon}$	5	Uninformative.
Total:	35	

Already for the benchmark case, the candidate set of cardinality 35 generates a search space of over 34 billion combinations. For the benchmark case, Creel and Kristensen (2015*b*) use as a loss function a weighted mean absolute deviation (MAD) of the SBIL estimator:

$$L(\theta) = \frac{1}{dim(\theta)} \sum_{i=1}^{dim(\theta)} \frac{1}{\sigma_i} |\theta_i - \hat{\theta}_i| \tag{31}$$

where the weight associated with each absolute deviation is the inverse of the respective parameter's prior standard deviation and $\hat{\theta}_i$ is the SBIL estimate of the i-th element of θ. The expected Bayesian loss criterion (20) is then minimized 100 times using a simulated annealing procedure. Running multiple iterations of the selection procedure allows for insight into both the performance of the selection procedure when one conservative run is performed as well as the performance that may be expected from the procedure when a single smaller run is conducted.

The results from the benchmark linear regression case with a sample size of $n = 30$ are promising with the six optimally informative statistics, $\hat{\alpha}_L, \hat{\beta}_{1L}, \hat{\beta}_{2L}, \hat{\beta}_{3L}, \hat{\beta}_{4L}, \hat{\sigma}_L$, picked in the best run of the battery of 100 and, when viewing each of the 100 runs separately, picked with the frequencies reported in **Table 2**. Encouragingly, the uninformative noise statistics where only picked twice. For more detailed results, see Creel and Kristensen (2015*b*).

Table 2: Selection frequencies for optimally informative statistics (Creel and Kristensen, 2015*b*)

Statistic	Selected (pct)
$\hat{\alpha}_L$	100%
$\hat{\beta}_1^L$	100%
$\hat{\beta}_2^L$	97%
$\hat{\beta}_3^L$	98%
$\hat{\beta}_4^L$	100%
$\hat{\sigma}_L$	64%
$\hat{\sigma}_Q$	66%
$\hat{\sigma}_C$	55%

5 Method

5.1 Aims of the current study

The current study provides simulation results for the two selection procedures proposed by Creel and Kristensen (2015*b*) as well as their non-integrated counterparts. These simulation results aim to 1) provide further evidence of the efficacy of the statistic selection procedures proposed by Creel and Kristensen (2015*b*), 2) compare the small sample performance of the EBL-based and KLD-based selection procedures, and 3) compare the performance of the integrated and non-integrated versions of the selection procedures.

5.2 Implementation and assessment of selection procedures

The four selection algorithms assessed in the current study comprise of the EBL and KLD algorithms proposed by Creel and Kristensen (2015*b*) and their respective non-integrated versions. The integrated EBL and KLD based selection procedures proposed by Creel and Kristensen (2015*b*) are given by (20) and (26) respectively and Creel and Kristensen (2015*b*) provide a third estimator:

$$\hat{\delta}_{EBL}|w = \underset{\delta}{\operatorname{argmin}} \frac{\sum_{r=1}^{R} L(\theta^r - \hat{E}_S[\theta|z(\delta)])K_h(z^r(\delta) - z(\delta))}{\sum_{r=1}^{R} K_h(z^r(\delta) - z(\delta))} \tag{32}$$

which is the minimum of an approximation of the non-integrated expected Bayesian loss criteria given in (17). The estimator above is written as conditional on the observed instance of W to highlight its dependence on the observed sample, a dependence that does not apply to the estimators based on integrated criteria. For a four-way comparison of the two EBL and KLD-based estimators, each having an integrated and a non-integrated version, a non-integrated KLD-based estimator is proposed by this study as:

$$\hat{\delta}_{KLD}|w = \underset{\delta}{\operatorname{argmax}} \int_{\Theta} \log \hat{f}(\theta|z(\delta))\hat{f}(\theta|w)\, d\theta \tag{33}$$

where the integral can be approximated using Monte Carlo integration:

$$\int_{\Theta} \log \hat{f}(\theta|z(\delta))\hat{f}(\theta|w)\, d\theta \approx \frac{\sum_{r=1}^{R} \log \hat{f}(\theta|z(\delta))K_h(w^r - w)}{\sum_{r=1}^{R} K_h(w^r - w)} = \hat{E}_{\theta|w}[\log \hat{f}(\theta|z(\delta))] \tag{34}$$

and:

$$\hat{f}(\theta|z(\delta)) = \frac{\sum_{s=1}^{S} K_h(\theta^s - \theta^r)K_h(z^s(\delta) - z(\delta))}{\sum_{s=1}^{S} K_h(z^s(\delta) - z(\delta))} \tag{35}$$

The estimator in (33) arises out of the minimization of the non-integrated KLD criteria given in (27). As the first term in (27) is independent of δ, the minimization of (27) with respect to δ reduces to a maximization of the second term.

Each of the four estimators involves an optimization over a discrete and finite search space $\Delta = \{0, 1\}^{dim(W)}$ of size $2^{dim(W)}$ and the present study adopts the same combinatorial optimization algorithm utilized by Creel and Kristensen (2015*b*), namely simulated annealing, to find optima. The characterizing features of this stochastic search algorithm are twofold. First, simulated annealing allows for so-called hill-climbing moves in which moves to points in the search space are accepted even if they result in an increase of the cost function. Such hill-climbing moves allow the algorithm to escape local minima which can be of crucial importance for non-exhaustive searches over spaces that have local minima. Second, the tolerance for accepting moves to points that increase the cost function decreases on each move, effectively constricting the probability of allowing a hill-climbing move as the algorithm progresses (Henderson et al., 2003).

The programming language used to implement the above detailed selection procedures is Python. Fortunately, many of the steps involved in the selection procedures readily lend themselves to being performed in parallel without any communication overhead and the current study exploits these so-called “pleasantly parallel” steps by utilizing Python’s multiprocessing package. The current study also uses a Python implementation of the simulated annealing algorithm made available by Richard J. Wagner and Matthew T. Perry with a URL to the source code made available in the bibliography. From a hardware perspective, the selection procedures discussed in this study were run on a virtual machine based on Xeon E5-2666 v3 (Haswell) processors with 32 virtual CPUs. Compute times for each procedure are given, along with results, in **Table 4**.

5.3 Linear regression (benchmark case)

As suggested and done by Creel and Kristensen (2015*b*), this study uses a benchmark case in which optimally informative statistics are known from theory to assess the performance of the four selection algorithms. This benchmark case comprises of the selection of statistics for ABC estimation of the parameters of a linear regression model. The implementation of the four selection algorithms applied to this benchmark case departs slightly from the implementation of Creel and Kristensen (2015*b*) in that the present study reduces the scope of the selection problem. This reduction in scope is due to slower performance achieved by this study's Python implementation compared to that achieved in Julia by Creel and Kristensen (2015*b*). Reduction in scope is achieved by generating data according to (36) and forming a candidate set of statistics from the counterparts of the linear model and two auxiliary regressions (28), (29), and (30) with two regressors per order of regressors as opposed to four.

$$y_i = \beta_0 + \beta x_{1i} + \beta_2 x_{2i} + \sigma u_i \tag{36}$$

Rather than using a kernel function in the posterior mean estimation $\hat{E}_S[\Theta|z^r(\delta)]$ for the EBL-based selection algorithms, this study follows the estimator adopted by Creel and Kristensen (2015*b*) and uses a k-nearest neighbors (k-nn) analogue of the SBIL estimator (15) motivated by **Algorithm 1**:

$$\tilde{E}_S[\Theta|z(\delta)] = \frac{\sum_{s=1}^{S} \theta^s \mathbb{1}_k||z^s - z||}{\sum_{s=1}^{S} \mathbb{1}_k||z^s - z||} \tag{37}$$

where $\mathbb{1}_k||\cdot||$ is an indicator function evaluating to 1 if and only if $||z^s - z||$ is less than the k-th smallest distance of the set of distances $\{||z^s - z|| \mid s \in S\}$. In the current study, k is set to be $S^{1/4}$. Furthermore, all simulated statistics z^s and parameters θ^s are standardized to give equal weighting to each summary statistic in the evaluating of $||z^s - z||$ and equal weighting to each parameter in the evaluation of $L(\theta - \tilde{E}_S[\Theta|z(\delta)])$. For choice of loss function, the current study adopts the choice made by Creel and Kristensen (2015*b*) in utilizing the mean absolute error specified in (31). Furthermore, prior distributions for the four parameters of (36) are chosen to be $\beta_i \sim U(-2, 2), \forall i \in [0, 2]$ and $\sigma \sim U(0, 1)$.

A battery of 100 simulated annealing runs is conducted to asses the performance of the four selection procedures in the benchmark case. Reporting the results from 100 runs provides some insight into the performance of the algorithms when less time or computational resources are dedicated to the optimization. **Table 4** displays results for both the integrated EBL and KLD-based algorithms with asterisks denoting the selection made at the optimum achieved in the 100 runs. This "best run" selection can be interpreted as the result of a single, more expansive, run of the simulated annealing procedure with the caveat that it features a cooling schedule not appropriate for a search of this size (Creel and Kristensen, 2015*b*). The frequencies at which statistics were selected across the 100 runs illustrate the type of performance that could be expected when the selection procedure is applied with significantly fewer moves across the search space.

5.4 Discrete time stochastic volatility

While the linear regression model provides a useful sandbox in which to test the efficacy of selection algorithms, the impetus for developing such algorithms is to utilize them in the estimation of parameters for more complex, potentially non-linear, models for which no likelihood is known in closed form. To investigate the integrated KLD selection algorithm's performance in identifying optimally informative subsets of summary statistics for such models, the current study utilizes this algorithm to estimate the parameters of a discrete time SV process. In the assessment of the selection algorithm's performance, the specification of the SV process used in this study departs from the prevailing standard specification in that the current study fixes the variance of the white noise random variable in the latent process (39) to be unity:

$$y_t = \mu_y + \phi y_{t-1} + \exp(\sigma_t/2)\epsilon_t \qquad \epsilon_t \sim N(0,1) \tag{38}$$

$$\sigma_t = \mu_\sigma + \varphi(\sigma_{t-1} - \mu_\sigma) + \varepsilon_t \qquad \varepsilon_t \sim N(0,1) \tag{39}$$

with associated likelihood function:

$$f(y|\theta) = \int_\sigma f(y|\theta,\sigma)f(\sigma|\theta)\,d\sigma \tag{40}$$

where $\theta = (\mu_y, \mu_\sigma, \phi, \varphi)$ is a vector of parameters, $y = (y_T, y_{T-1}, \ldots, y_0)$ is a vector of the observed variables, and $\sigma = (\sigma_T, \sigma_{T-1}, \ldots, \sigma_0)$ is a vector of the latent variables. The integral in (40) is a T-dimensional integral involving latent variables and is analytically intractable precluding the use of standard maximum likelihood estimation techniques (Li et al., 2011).

In applying statistic selection methods to a continuous time jump-diffusion SV model, Creel and Kristensen (2015*b*) utilize auxiliary models to capture features of the SV process and consequently use statistics motivated by these auxiliary models to construct a set of candidate statistics. The current study adopts this approach as well and the auxiliary models chosen for the discrete time SV process are:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-1}^2 + \alpha_3 y_{t-2} + \alpha_4 y_{t-2}^2 + u_t \tag{41}$$

$$y_t^2 = \lambda_0 + \lambda_1 y_{t-1} + \lambda_2 y_{t-1}^2 + \lambda_3 y_{t-2} + \lambda_4 y_{t-2}^2 + w_t \tag{42}$$

The candidate set of summary statistics for the discrete time SV model then comprises of the 22 statistics detailed in **Table 3**. Simulated data is generated from (38) with $T = 2,000$. As was done in the linear regression case, all summary statistics and parameter draws are standardized according to sample means and variances to place equal weighting on each summary statistic in the computation of norms as well as on each parameter when evaluating the loss function.

Table 3: Candidate set of summary statistics for SV process

Statistic	Description
$(\hat{\mu}_y,\, \hat{\mu}_{y^2})$	Mean of y_t and y_t^2
$(\hat{\sigma}_y,\, \hat{\sigma}_{y^2})$	Standard deviation of y_t and y_t^2
$(\hat{\gamma}_{1,y},\, \hat{\gamma}_{1,y^2})$	Skewness of y_t and y_t^2
$(\hat{\gamma}_{2,y},\, \hat{\gamma}_{2,y^2})$	Kurtosis of y_t and y_t^2
$\hat{\rho}_{y,y^2}$	Correlation between y_t and y_t^2
$\hat{\boldsymbol{\alpha}}$	Vector of coefficients of (41)
$\hat{\boldsymbol{\lambda}}$	Vector of coefficients of (42)
$(SE_{\hat{\alpha}}, SE_{\hat{\gamma}})$	Standard error of (41) and (42)
$\hat{\rho}_{u,w}$	Cross equation correlation of residuals from (41) and (42)

The same estimator from the linear regression case (37) and the same loss function (31) are used in this assessment as well. To simulate data according to the discrete time SV model (38), priors for the parameters are chosen to be $\mu_y \sim U(-6, 6)$, $\mu_\sigma \sim U(0, 3)$, $\phi \sim U(0, 0.7)$ and $\varphi \sim U(0, 0.7)$. To assess how well the selection algorithm performs, this study simulates a single benchmark time series from known parameters and then assesses the performance of the ABC estimator (37) in estimating these parameters using the algorithm's selection of summary statistics. This assessment comprises of computing the root mean squared error (RMSE) and estimating the bias of the ABC estimator based on the selected subset as well as a random subset selection and the entire candidate set. Results from this assessment are provided in **Table 7**.

Table 4: Performance of integrated selection procedures (linear regression)

Statistic	Integrated EBL $n = 30$		Integrated KLD $n = 30$	
	Selection frequency (%)	Selected in best run	Selection frequency (%)	Selected in best run
$\hat{\alpha}_L$	96	*	100	*
$\hat{\beta}_{1L}$	91	*	99	*
$\hat{\beta}_{2L}$	96	*	96	*
$\hat{\sigma}_L$	88		78	*
$\hat{\alpha}_Q$	34		19	
$\hat{\beta}_{1Q}$	57		45	
$\hat{\beta}_{2Q}$	54	*	41	
$\hat{\gamma}_{1Q}$	11		2	
$\hat{\gamma}_{2Q}$	16		3	
$\hat{\sigma}_Q$	72	*	63	
$\hat{\alpha}_C$	29		12	
$\hat{\beta}_{1C}$	14		6	
$\hat{\beta}_{2C}$	21		4	
$\hat{\gamma}_{1C}$	11		2	
$\hat{\gamma}_{2C}$	10		1	
$\hat{\delta}_{1C}$	6		3	
$\hat{\delta}_{2C}$	6		1	
$\hat{\sigma}_C$	65	*	46	*
ε_1	1		1	
ε_2	2		0	
ε_3	2		0	
ε_4	1		0	
ε_5	0		0	
Average number of statistics chosen: 7.8			Average number of statistics chosen: 7.0	
Total time to run: 8 hours and 30 minutes			Total time to run: 3 hours and 45 minutes	
$R = 300$			$R = 300$	
$S = 3,000$			$S = 3,000$	

6 Simulation Results

6.1 Linear regression (benchmark case)

The results displayed in **Table 4** for the integrated EBL-based algorithm report a decent performance of the selection algorithm to select informative summary statistics from a candidate set. With regards to the optimally informative statistics described in **Table 1**, the estimates for the variance of the linear model's error term were picked in roughly equal proportion while the estimates for the coefficients of the correctly specified model were picked over 90% of the time. Comparing best runs reveals a selection that is slightly less accurate than the optimal selection achieved by the implementation in Creel and

Kristensen (2015*b*). The best run in the current study selected six statistics to estimate four parameters and while correctly choosing the estimates for the coefficients of the correctly specified model it also selected a relevant, yet suboptimal, estimate for a coefficient appearing in the quadratic specification. Furthermore, the best run in the current study selected the optimally informative estimate for the variance of the model's error term twice. Nevertheless, when viewed as 100 separate runs, the EBL-based selection algorithm selected the estimates for $\gamma_L, \gamma_C, \delta_C$ with much smaller frequency than the more strongly informative estimates for $\beta_L, \beta_Q, \beta_C$.

Overall, these simulation results supplement the evidence already provided by Creel and Kristensen (2015*b*) that the EBL-based selection algorithm is able to discern informative statistics from weakly and uninformative statistics in a candidate set. The selection of six statistics for four parameters and in particular the selection of close substitutes $\hat{\sigma}_Q$ and $\hat{\sigma}_C$, which provide information about the same parameter, suggests that a penalty function placing weight on a more parsimonious selection may provide improvements in performance at these simulation sizes. The results provided by Creel and Kristensen (2015*b*) show that with a larger simulation size, the EBL-based selection procedure performs very well even without such a penalty function.

The integrated KLD-based selection algorithm delivered a more promising performance with the optimally informative estimates for the coefficients of the correctly specified model all being picked in nearly each run. Furthermore, uninformative statistics and weakly informative statistics were picked with far smaller frequencies than was the case in the EBL-based selection procedure, suggesting that, in this linear model case, the KLD-based selection algorithm is better able to discriminate between informative and weakly informative statistics in a candidate set. The KLD-based selection procedure did however select the optimally informative estimate for the variance of the error term twice, a result that again highlights the potential benefits from employing an appropriate penalty function at these simulation sizes and search specifications. The low frequencies with which the KLD-based selection algorithm selected weakly informative statistics suggest that it has a good chance of performing well even when the selection procedure conducts a radically faster, and less thorough, search through the search space. In other words, a single run, requiring only 1 percent of the time taken to produce the results in **Table 4** provided a high probability of selecting optimal and strongly informative statistics from the candidate set. Particularly striking is the difference in compute time required

Table 5: Performance of non-integrated selection procedures (linear regression)

Statistic	Non-integrated EBL $n = 30$		Non-integrated KLD $n = 30$	
	Selection frequency (%)	Selected in best run	Selection frequency (%)	Selected in best run
$\hat{\alpha}_L$	94	*	93	*
$\hat{\beta}_{1L}$	95	*	85	*
$\hat{\beta}_{2L}$	89	*	90	*
$\hat{\sigma}_L$	61	*	1	
$\hat{\alpha}_Q$	55		51	*
$\hat{\beta}_{1Q}$	84		69	
$\hat{\beta}_{2Q}$	78	*	70	
$\hat{\gamma}_{1Q}$	34	*	1	
$\hat{\gamma}_{2Q}$	32		0	
$\hat{\sigma}_Q$	59		2	
$\hat{\alpha}_C$	62	*	42	
$\hat{\beta}_{1C}$	46	*	12	
$\hat{\beta}_{2C}$	49		16	
$\hat{\gamma}_{1C}$	34		0	
$\hat{\gamma}_{2C}$	35		2	
$\hat{\delta}_{1C}$	30		2	
$\hat{\delta}_{2C}$	30		1	
$\hat{\sigma}_C$	59		0	
ε_1	1		6	
ε_2	0		3	
ε_3	0		1	
ε_4	0		5	
ε_5	1		1	
Average number of statistics chosen: 9.9			Average number of statistics chosen: 5.6	
Total time to run: 6 hours and 15 minutes			Total time to run: 5 hours and 44 minutes	
$R = 30,000$			$R = 3,000$	
$S = 30,000$			$S = 30,000$	

for both algorithms with the KLD-based selection algorithm finishing in less than half the time than the EBL-based algorithm for identical simulation sizes, cooling schedule, and number of moves made by the simulated annealing procedure. Overall, this comparison provides at least some evidence that the KLD-based selection procedure, based on a stricter sufficiency criterion, may provide better finite sample performance than its EBL-based counterpart. Somewhat counter-intuitively, the KLD-based selection procedure chose, on average across the 100 runs, more parsimoniously than the EBL-based algorithm even though the sufficiency criterion it is based on is stricter and encompasses more information about the posterior than the EBL criterion.

The non-integrated KLD and EBL-based selection algorithms perform very poorly at the simulation sizes of $R = 300$ and $S = 3,000$. With larger values for R and S, the non-integrated versions show improved performance, albeit still noticeably worse than their integrated counterparts. Even with the increased simulation sizes detailed in **Table 5**, the EBL-based non-integrated procedure picked suboptimal statistics very frequently and its best run selection failed to adequately avoid suboptimal statistics as well. The KLD-based non-integrated procedure on the other hand recorded a much more parsimonious best run selection but did not pick any of the three estimates for σ in (36). In fact, the non-integrated KLD-based procedure very rarely selected $\hat{\sigma}$ suggesting that perhaps the inclusion of this statistic has little effect on the procedure's sufficiency criterion (27). As the non-integrated KLD-based algorithm involves the approximation of $\hat{f}(\theta|w)$, the very object whose dimensionality these selection algorithms aim to reduce, this deterioration in performance relative to the algorithm's integrated counterpart is not surprising.

6.2 Discrete time stochastic volatility

The best run selections and selection frequencies are somewhat similar across the integrated EBL and KLD selection algorithms for the SV process but do feature some noteworthy differences. The integrated KLD selection algorithm almost exclusively selected summary statistics from the auxiliary regression specified in (41) with the OLS estimate for the intercept $\hat{\alpha}_0$ and the first coefficient $\hat{\alpha}_1$ being selected in each run and the standard error of (41) being selected in 99 out of 100 runs. In contrast, the EBL-based selection algorithm, while also picking $\hat{\alpha}_0, \hat{\alpha}_1 \hat{SE}_{\hat{\alpha}}$ frequently, did select other statistics with large frequencies as well. In particular, the EBL-based algorithm deviated from the selections made by the KLD-based algorithm in that it frequently selected estimates for the mean and standard deviation of the distribution of y_t as well as estimates for the coefficients of (42). In fact, the EBL-based algorithm selected the estimate for the intercept of the second auxiliary regression, $\hat{\lambda}_0$, 45 out of 100 runs while the KLD-based algorithm never selected it. The frequent selection of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ by both algorithms is not surprising given that these statistics should capture information on μ_y and ϕ in (38), however it is rather unexpected that the KLD algorithm so sparsely selected any estimates derived from (42). A reasonable candidate for an informative statistic on φ in the latent process for volatility would be $\hat{\lambda}_2$ as this estimate should capture some of the relationship between σ_t and its one period lag. In contrast to the benchmark linear regression case,

Table 6: Performance of integrated selection procedures (stochastic volatility)

Statistic	Integrated EBL $n = 2000$		Integrated KLD $n = 2000$	
	Number of times selected	Selected in best run	Number of times selected	Selected in best run
$\hat{\mu}_y$	83	*	5	
$\hat{\sigma}_y$	53		19	
$\hat{\gamma}_{1,y}$	3		1	
$\hat{\gamma}_{2,y}$	10		5	
$\hat{\mu}_{y^2}$	25		1	
$\hat{\sigma}_{y^2}$	24		0	
$\hat{\gamma}_{1,y^2}$	12		14	
$\hat{\gamma}_{2,y^2}$	23		19	
$\hat{\rho}_{y,y^2}$	20		1	
$\hat{\alpha}_0$	100	*	100	*
$\hat{\alpha}_1$	100	*	100	*
$\hat{\alpha}_2$	0		0	
$\hat{\alpha}_3$	0		1	
$\hat{\alpha}_4$	0		0	
$\hat{SE}_{\hat{\boldsymbol{\alpha}}}$	85	*	99	*
$\hat{\lambda}_0$	45		0	
$\hat{\lambda}_1$	17		7	
$\hat{\lambda}_2$	19		0	
$\hat{\lambda}_3$	13		1	
$\hat{\lambda}_4$	1		0	
$\hat{SE}_{\hat{\boldsymbol{\lambda}}}$	21		1	
$\hat{\rho}_{u,w}$	7		4	
Average number of statistics chosen: 6.6			Average number of statistics chosen: 3.8	
Total time to run: 3 hours and 32 minutes			Total time to run: 5 hours and 3 minutes	
$R = 200$			$R = 200$	
$S = 2,000$			$S = 2,000$	

the integrated EBL-based algorithm recorded a faster compute time than the integrated KLD-based algorithm.

A more extensive, single run of the simulated annealing procedure with a much slower cooling schedule and using the KLD-based algorithm produced selections identical with those of the best KLD run in **Table 6**. This single run differed relative to the battery of runs only in that it dedicated 100 times more moves to each stage of the cooling schedule resulting in a much more expansive coverage of the search space. Encouragingly, the k-nn ABC estimator (37) based on the selected subset from this more thorough search showed substantially better performance, in a RMSE sense, than the same estimator based on randomly selected subsets as reported in **Table 7**. Naturally, the improvements over the estimator based on the entire candidate set are less pronounced since, as pointed out by Creel and Kristensen (2015*b*), the candidate set may have been chosen well. There are

Table 7: Performance of k-nn ABC estimator (stochastic volatility process)

Parameter	Value	Selection (KLD)		Random selection		Entire candidate set	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
μ_y	2	**0.0077**	**0.0683**	-0.7320	1.4597	-1.9813	1.9961
μ_σ	2	-0.6954	0.7009	-0.3762	0.4321	**-0.2693**	**0.2800**
φ	0.2	**0.0971**	**0.0996**	0.1070	0.1258	0.2402	0.2409
ϕ	0.2	**-0.0152**	**0.0163**	0.0772	0.0999	0.0281	0.0331

Note: Bolded figures denote the minimum RMSE and the minimum absolute bias for the corresponding parameter.

however caveats to this improved performance. Firstly, the estimator using the selected subset underperformed in its estimation of the mean of the latent process relative to the estimator based on the entire candidate set, suggesting that a statistic in the candidate set is informative about μ_σ but that the selection algorithm was unable to select it. Secondly, the improved performance of the estimator based on the selected subset compared to the randomly selected subset stems disproportionately from an improved RMSE for the $\hat{\mu}_y$ estimate, suggesting that the selected subset is particularly informative μ_y relative to the other parameters. As such, the candidate set may not include statistics informative about the features of the latent volatility process or, if such statistics are indeed present in the candidate set, the selection algorithm is not able to select them.

7 Concluding remarks

Avenues for future research are numerous. In particular, the effects of model misspecification on statistic selection and means to ensure robustness of statistic selection procedures in the face of model uncertainty are of interest. Ratmann et al. (2009) investigate methods by which to simultaneously perform model criticism and model inference in the context of ABC. However, Ratmann et al. (2009) use as a starting point a set of summary statistics assumed to be optimal and consequently do not dedicate any discussion to the effects of model uncertainty on statistic selection techniques. Furthermore, the use of a different divergence criterion other than the Kullback-Leibler divergence, such as the Kolmogorov-Smirnov statistic, could potentially provide improvements for selection algorithms that aim to preserve information about the entire posterior distribution.

Overall, this study delivers empirical evidence that the efficacy of both the EBL and KLD-based selection algorithms is maintained for reduced simulation sizes. Additionally, the comparison between the performance of both algorithms in the benchmark cases suggests that the KLD-based algorithm performs better at finite sample sizes and may be a more appropriate choice in practice when only limited computational resources can be dedicated to the optimization step. The best run of the KLD-based algorithm picked a near optimal subset for the benchmark case in less than half of the compute time of its EBL-based counterpart and using only a third of the simulation size used in Creel and Kristensen (2015*b*). Furthermore, the low frequency with which the KLD-based algorithm selected weakly informative statistics suggests that even a haphazard and quick application of the optimization will produce reliable results. Lastly, in the estimation of the discrete time SV process, the selected subset did generate improved ABC point estimates for all but one of the parameters, even if most gains were achieved in the estimation of the mean of the observable process. In culmination, the results provided in this study corroborate both the ease of implementation of the selection procedures proposed by Creel and Kristensen (2015*b*) as well as the limited computational expense required by these procedures to produce dependable results.

Bibliography

Barnes, C. P., Filippi, S., Stumpf, M. P. H. and Thorne, T. (2012), 'Considerate approaches to constructing summary statistics for ABC model selection', *Statistics and Computing* **22**(6), 1181–1197.

Beaumont, M. A. (2010), 'Approximate Bayesian Computation in Evolution and Ecology', *Annual Review of Ecology, Evolution, and Systematics* **41**(1), 379–406.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002), 'Approximate Bayesian computation in population genetics', *Genetics* **162**(4), 2025–2035.

Blum, M. G. B. and Francois, O. (2010), 'Non-linear regression models for Approximate Bayesian Computation', *Statistics and Computing* **20**(1), 63–73.

Blum, M. G. B., Nunes, M. A., Prangle, D. and Sisson, S. A. (2013), 'A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation', *Statistical Science* **28**(2), 189–208.

Creel, M. and Kristensen, D. (2013), 'Indirect Likelihood Inference', *UFAE and IAE Working Papers, Unitat de Fonaments de l'Anàlisi Econòmica (UAB) and Institut d'Anàlisi Econòmica (CSIC).* pp. 1–32.

Creel, M. and Kristensen, D. (2015*a*), 'ABC of SV: Limited information likelihood inference in stochastic volatility jump-diffusion models', *Journal of Empirical Finance* **31**, 85–108.

Creel, M. and Kristensen, D. (2015*b*), 'On selection of statistics for approximate Bayesian computing (or the method of simulated moments)', *Computational Statistics & Data Analysis* **100**, 1–29.

Dean, T. A., Singh, S. S., Jasra, A. and Peters, G. W. (2014), 'Parameter estimation for hidden markov models with intractable likelihoods', *Scandinavian Journal of Statistics* **41**(4), 970–987.

Didelot, X., Everitt, R. G., Johansen, A. M. and Lawson, D. J. (2011), 'Likelihood-free estimation of model evidence', *Bayesian Analysis* **6**(1), 49–76.

Fearnhead, P. and Prangle, D. (2012), ‘Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **74**(3), 419–474.

Fu, Y.-X. and Li, W.-H. (1997), ‘Estimating the age of the common ancestor of a sample of DNA sequences.’, *Molecular Biology and Evolution* **14**(2), 195–199.

Greenberg, E. (2008), *Introduction to Bayesian Econometrics*, Cambridge University Press, New York, NY.

Henderson, D., Jacobson, S. H. and Johnson, A. W. (2003), The theory and practice of simulated annealing, *in* ‘Handbook of Metaheuristics’, pp. 287–319.

Hoogerheide, L. F., van Dijk, H. K. and van Oest, R. D. (2009), Handbook of Computational Econometrics, *in* D. A. Belsley and E. J. Kontoghiorghes, eds, ‘Handbook of Computational Econometrics’, John Wiley and Sons, pp. 1–496.

Jeffreys, H. (1961), *Theory of Probability*, Vol. 2.

Joyce, P. and Marjoram, P. (2008), ‘Approximately sufficient statistics and bayesian computation.’, *Statistical applications in genetics and molecular biology* **7**(1), Article26.

Kass, R. R. E. and Raftery, A. E. A. (1995), ‘Bayes factors’, *Journal of the American Statistical Association ...* **90**(430), 773– 795.

Li, Y., Ni, Z.-X. and Lin, J.-G. (2011), ‘A Stochastic Simulation Approach to Model Selection for Stochastic Volatility Models’, *Communications in Statistics - Simulation and Computation* pp. 1043–1056.

Luciani, F., Sisson, S. A., Jiang, H., Francis, A. R. and Tanaka, M. M. (2009), ‘The epidemiological fitness cost of drug resistance in Mycobacterium tuberculosis’, *Proc Natl Acad Sci U S A* **106**(34), 14711–14715.

Marjoram, P., Molitor, J., Plagnol, V. and Tavare, S. (2003), ‘Markov chain Monte Carlo without likelihoods’, *Proc Natl Acad Sci U.S A* **100**(0027-8424), 15324–15328.

Nadaraya, E. a. (1964), ‘On Estimating Regression’, *Theory of Probability & Its Applications* **9**(1), 141–142.

Nunes, M. and Balding, D. J. (2010), ‘On optimal selection of summary statistics for approximate Bayesian computation.’, *Statistical applications in genetics and molecular biology* **9**(1), Article34.

Perry, M. T. and Wagner, R. J. (2014), ‘simanneal’. Accessed: 3 July 2016.
URL: *https://github.com/perrygeo/simanneal*

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999), ‘Population growth of human Y chromosomes: a study of Y chromosome microsatellites.’, *Molecular biology and evolution* **16**(12), 1791–8.

Ratmann, O., Andrieu, C., Wiuf, C. and Richardson, S. (2009), ‘Model criticism based on likelihood-free inference, with an application to protein network evolution.’, *Proceedings of the National Academy of Sciences of the United States of America* **106**(26), 10576–10581.

Rubin, D. B. (1984), ‘Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician’, *The Annals of Statistics* **12**(4), 1151–1172.

Sisson, S. A., Fan, Y., Tanaka, M. M., Rogers, A., Huang, Y., Njegic, B., Wayne, L., Gordon, M. S., Dabdub, D., Gerber, R. B. and Finlayson-pitts, B. J. (2009), ‘Correction for Sisson et al., Sequential Monte Carlo without likelihoods’, *Proceedings of the National Academy of Sciences* **106**(39), 16889–16889.

Stock, J. H., Wright, J. H. and Yogo, M. (2002), ‘A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments’, *Journal of Business & Economic Statistics* **20**(4), 518–529.

Sunnaker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M. and Dessimoz, C. (2013), ‘Approximate Bayesian Computation’, *PLoS Computational Biology* **9**(1).

Tauchen, G. (1986), ‘Statistical Properties of Generalized Method-of-Moments Estimators of Structural Parameters Obtained from Financial Market Data’, *Journal of Business and Economic Statistics* **4**(4), 397–416.

Tavare, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997), ‘Inferring coalescence times from DNA sequence data’.

Watson, G. S. (1964), ‘Smooth regression analysis’, *The Indian Journal of Statistics* **26**(4), 359–372.

Weiss, G. and von Haeseler, A. (1998), ‘Inference of population history using a likelihood approach.’, *Genetics* **149**(July), 1539–1546.

Wilkinson, R. D. (2013), ‘Approximate Bayesian computation (ABC) gives exact results under the assumption of model error’, *Statistical Applications in Genetics and Molecular Biology* **12**(2), 129–141.

Wilkinson, R. D. and Tavare, S. (2008), ‘Approximate Bayesian Computation : a simulation based approach to inference’. Accessed: 28 June 2016.
URL: *http://www0.cs.ucl.ac.uk/staff/C.Archambeau/AIS/Talks/rwilkinson_ ais08.pdf*

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York ; London : Wiley.

8 Appendix 1: The simulated Bayesian indirect likelihood (SBIL) estimator

Creel and Kristensen (2015b) develop one of their statistic selection criteria with an emphasis on a particular ABC-type point estimator. This estimator is the simulated analogue of an indirect likelihood estimator proposed by Creel and Kristensen (2013):

$$\hat{\theta}_{BIL} = \int_{\Theta} \theta f_n(\theta|z_n) d\theta = \int_{\Theta} \theta \frac{f_n(z_n|\theta)\pi(\theta)}{\int_{\Theta} f_n(z_n|\theta)\pi(\theta)d\theta} d\theta \tag{43}$$

where $\pi(\theta)$ is a prior, n is sample size, Z_n is a statistic calculated from the observed sample with sample realization z_n and $dim(Z_n) = q$, and θ is a vector of parameters of interest. The above estimator can be seen as the indirect inference version of the Bayesian posterior mean estimator as it relies on a statistic Z_n rather than the full observed dataset. For cases in which the likelihood function $f_n(Z_n|\Theta)$ is not known in closed form but data can nevertheless be simulated, non-parametric regression may be employed to approximate (43) with:

$$\hat{\theta}_{SBIL} = \hat{E}[\theta|z_n] = \frac{\sum_{s=1}^{S} \theta^s K_h(z_n^s - z_n)}{\sum_{s=1}^{S} K_h(z_n^s - z_n)} \tag{44}$$

akin to the Nadaraya-Watson kernel regression estimator (Nadaraya, 1964; Watson, 1964). In 44, the S parameter vectors θ^s are draws from the prior density $\pi(\theta)$, the z_n^s are statistics calculated from data simulated by the model parameterized by respective θ^s, and $K_h(\cdot)$ is a kernel function with bandwidth $h > 0$.

Beaumont et al. (2002) propose the estimator in (44) by extending standard rejection-based ABC estimation methods to feature local-linear regression smooth weighting. To motivate this extension, Beaumont et al. (2002) restrict the posterior distribution to have structure according to:

$$\theta^i = \alpha + (z^i - z)^T \beta + \varepsilon^i \tag{45}$$

where the error term is centered on zero and has constant variance. In (45), θ^i is a scalar but the steps that follow can also be generalized to multidimensional parameter vectors. No further distributional restrictions are imposed on the error and consequently the distribution of θ^i. When $z^i = z$, the θ^i variables are drawn from the posterior distribution with mean α. Consequently estimates for α are point estimates for the posterior mean. The linear model (45) gives the following ordinary least squares (OLS) estimates of α and β for S simulated pairs (θ^s, z^s):

$$(\hat{\alpha}, \hat{\beta})' = (X^{'}X)^{-1}X^{'}\theta \tag{46}$$

where θ is a $S \times 1$ vector containing the parameter draws and:

$$X = \begin{pmatrix} 1 & z_{11} - z_1 & \dots & z_{1q} - z_q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{S1} - z_1 & \dots & z_{Sq} - z_q \end{pmatrix} \tag{47}$$

with z_{ij} being the j-th component of the summary statistic vector for the i-th simulated draw. Given the linear conditional structure imposed on the posterior distribution by (45), the S random variables defined as:

$$\theta^s_* = \theta^s - (z^s - z)^{'}\hat{\beta} \tag{48}$$

form an approximation of the posterior distribution. To weaken the restriction imposed by (45), which may very well be implausible in most cases, Beaumont et al. (2002) add a smoothing feature to the OLS estimation of $(\hat{\alpha}, \hat{\beta})'$ so as to only impose the structure of (45) locally for small $||z^s - z||$ and to weaken this restriction as $||z^s - z||$ increases:

$$(\hat{\alpha}, \hat{\beta})' = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{s=1}^{S} \{\theta^s - \alpha - (z^s - z)^T \beta\}^2 K_h(z^s - z) \tag{49}$$

where K_h is chosen by Beaumont et al. (2002) to be the Epanechnikov kernel. The minimization above has the solution:

$$(\hat{\alpha}, \hat{\beta}) = (X^T W X)^{-1} X^T W \theta \tag{50}$$

corresponding to the weighted least squares (WLS) estimator with a diagonal weight matrix whose i-th diagonal element is $K_h(z^s - z)$**.** The estimator in 50 gives a point estimate for α as:

$$\hat{\alpha} = \frac{\sum_{s=1}^{S} \theta_*^s K_h(z^s - z)}{\sum_{s=1}^{S} K_h(z^s - z)} \tag{51}$$

Imposing a local-constant, as opposed to local-linear, structure whereby (45) becomes $\theta^s = \alpha + \varepsilon^s$ and (48) becomes $\theta_*^s = \theta^s$, gives the SBIL estimator from (44). It is this SBIL estimator (44) which forms the basis for one of the statistic selection procedures proposed by Creel and Kristensen (2015*b*) and consequently the current study.