On Selection of Statistics for Approximate Bayesian Computing (or the Method of Simulated Moments)*

Michael Creel† *and Dennis Kristensen*‡

Universitat Autònoma de Barcelona, Barcelona Graduate School of Economics, MOVE

University College London, Institute of Fiscal Studies, CREATES

April, 2015

Abstract

A cross validation method for selection of statistics for Approximate Bayesian Computing, and for related estimation methods such as the Method of Simulated Moments, is presented. The method uses simulated annealing to minimize the cross validation criterion over a combinatorial search space that may contain an extremely large number of elements. A first simple example, for which optimal statistics are known from theory, shows that the method is able to select these optimal statistics out of a large set of candidate statistics. A second example of selection of statistics for a stochastic volatility model illustrates the method in a more complex case. Code to replicate the results, or to use the method for other applications, is provided.

Keywords: Approximate Bayesian Computation; likelihood-free methods; selection of statistics; method of simulated moments

1 Introduction

Bayesian analysis centers attention on the posterior density, $f(\theta|y)$, where y is the sample, that is, a realization of a random vector Y, and θ is the parameter vector. As is well known, the posterior is proportional to the product of the likelihood of the sample, $f(y|\theta)$, and the prior, $\pi(\theta)$, so a Bayesian analysis requires the likelihood function. Classical statisticians often use the maximum likelihood (ML) estimator, because of its desirable property

*Accompanying computer code is provided through links in the online version of the paper. The entire set of files is available at http://www.runmycode.org/companion/view/1116.

†(corresponding author) Department of Economics and Economic History, Universitat Autònoma de Barcelona, Bellaterra (Barcelona) 08193, Spain. email: michael.creel@uab.es. Tel.: +34 935811696. FAX: +34 935812012.

‡Department of Economics, University College London, Gower Street, London WC1E 6BT, United Kingdom. email: d.kristensen@ucl.ac.uk.

of asymptotic efficiency. For both approaches, the likelihood function is needed. However, there are many areas of research involving complex models where computation of the likelihood function is not possible, or is extremely expensive, effectively ruling out ordinary Bayesian methods and ML estimation.

Much research in the last several decades, both classical and Bayesian, has been focused on providing alternatives. In the classical framework, and among other approaches, the Generalized Method of Moments (GMM; Hansen, 1982) is based on a vector of moment conditions $m(\theta, y)$ that are assumed to have expectation equal to zero under the true model. A common way to define moments first defines a vector of statistics Z, a mapping operating on Y, $Z = Z(Y)$, with sample realization $z = Z(y)$. Let $E_\theta[Z]$ be the expectation of Z under the model and $m(\theta, y) = z - E_\theta[Z]$ be the moment restrictions implied by the model; by construction, these moments have expectation zero when $\theta = \theta_0$, where θ_0 is the true parameter value. Approximate Bayesian Computing (ABC) is a set of methods that attempts to use Bayesian ideas when the likelihood function of the sample is not available, but the model is simulable. An important part of this literature (e.g., Beaumont et al. 2009; Fearnhead and Prangle, 2012; Blum et al. 2013) focuses on the posterior of the parameter conditional on a statistic, $f(\theta|z)$. In both areas, use of a statistic offers the simple advantage of feasibility, as well as computational advantages, due to the reduction of the dimensionality of the problem from $\dim Y$ to $\dim Z$. The cost is a possible loss of statistical efficiency, compared to what would obtain were the likelihood function available.

It is well known that the asymptotic distributions of GMM estimators depend crucially on the specific Z that is chosen. We know that adding statistics to a baseline set will not cause the asymptotic variance of the GMM estimator to increase, which might cause one to think that selection of statistics is not important, and that when in doubt, additional statistics should be included. However, use of weakly informative moment conditions can lead to very poor small sample performance of the GMM estimator (e.g., Tauchen, 1986; Stock, Wright and Yogo, 2002). In some cases, the model can give strong guidance regarding what moments to use for estimation, but in others, the choice is not clear, and a researcher may need to select which moments to use from a set that may include weakly informative and uninformative moments. These issues are also relevant for ABC with the posterior distribution $f(\theta|z)$ in general being sensitive to the choice of the summary statistics. From a frequentist point of view, ABC is first order asymptotic equivalent to the efficient GMM estimator based on $m(\theta, y) = z - E_\theta[Z]$ (see Creel and Kristensen, 2013 and Forneron and Ng, 2015) and so the above cited results for moment selection also applies to ABC. Gao and Hong (2015) show how ABC-type ideas can be employed to compute Bayesian versions of GMM estimators in a general setting.

Thus, an important part of implementing GMM and ABC is the choice of the statistics Z. Systematic methods for selection of statistics have received a good amount of attention in both the ABC and GMM literatures. Within the ABC literature, Blum et al. (2013) provide a review of this work, with some new results based on use of regular-

ization methods. The methods that have been used fall into three classes: best subset selection; projection methods that define new statistics by combining statistics in some way to reduce dimension; and regularization techniques such as ridge regression.

This paper adds to this literature by proposing a simple cross validation method for selecting statistics from a possibly large collection of candidate statistics, denoted W, assuming that the model is simulable. The method lies in the best subset selection classification of Blum et al. (2013): By searching over the candidate statistics using the simulated annealing algorithm, it provides a systematic method of dealing with the problem of a potentially large set of candidate statistics, while limiting computational cost. One of the examples presented below has 56 candidate statistics, so the number of possible combinations to search over, 2^{56}, is more than 70,000 million millions. Specifically, selection of statistics is investigated based on the idea of minimization of an integrated version of Bayes expected loss, $\int L\left(E\left[\Theta|Z=z\right]-\theta\right)f(z|\theta)\pi(\theta)dzd\theta$ for some loss function L, with respect to all subsets of statistics, Z, of W. Here, $E\left[\Theta|Z=z\right]$ is the posterior mean, $f(z|\theta)$ is the density of the statistic conditional on the parameter, and $\pi(\theta)$ is the prior. The integrated Bayes expected loss cannot be computed analytically, but can easily be approximated through simulations as demonstrated in the following. The use of Bayes expected loss in a decision theoretic framework is quite standard, but there one usually conditions on the particular outcome z that has been observed. Here, on the other hand, we choose to integrate over all possible outcomes with weights assigned by their likelihoods; this is meant to reflect that we are interested in identifying a universal ("global") set of sufficient statistics that work well for the given model irrespectively of the particular outcome observed. Moreover, this integrated version of Bayes expected loss proves to be simpler to estimate compared to the conditional version.

The above selection rule aims at minimizing the loss of the posterior mean. If the interest lies in the whole posterior distribution, and not just its mean, we show how our method is easily adjusted so as to minimize the integrated Kullback-Leibler distance of the posterior density for a given set of summary statistics w.r.t. z. This version of our selection procedure chooses the set of summary statistics that maximizes the information content over the whole distribution and so is similar in spirit to the one of Barnes et al. (2012). However, the procedure of Barnes et al. (2012) only performs a partial search over the set of candidate statistics, while our algorithm does a full search.

From a frequentist point of view, $E\left[\Theta|Z=z\right]$ can be interpreted as a point estimator of the true data-generating value; using of the posterior mean as an estimator is defensible by the first order asymptotic equivalence of the posterior mean and the efficient GMM estimator (Chernozhukov and Hong, 2003; Creel and Kristensen, 2013). Our procedure can therefore be thought of as minimizing Bayes Risk, which again is a well-known object in decision theory, of this estimator. Thus, our procedure also applies to simulation-based GMM estimators, including the method of simulated moments (McFadden, 1989), indirect inference (Smith, 1993; Gourieroux et al., 1993), and the efficient method of moments (Gallant and Tauchen, 1996). The specific procedure that we propose for computing the

integrated Bayes expected loss computation requires that the model of interest is simulable. This, in general, requires a fully specified model. Thus, the implementation of our procedure would have to be modified in order to use it for selection of moments in GMM estimation of models that are only partially specified.

In comparison to existing methods developed in the ABC literature, our proposal has the advantage that it provides a universal search over all possible candidate sets of statistics. In contrast, most competing methods either rely on fairly complex step-up procedures, where one statistic is added at a time, or on the construction of approximately optimal statistics. These methods tend to be more complex to implement and require much computational effort to achieve a solution that is stable across repeated runs. In contrast, our method is relatively simple to implement and tends to converge quite quickly. We investigate the performance of the proposed method through two examples, and in both cases our procedure performs well. The first example takes the form of a linear regression model, where the optimal statistics are known, and thus it provides a good testing ground for our procedure. We show that our method can identify these statistics quite rapidly and does so robustly across many simulated samples. This simple test problem should be of independent interest since it might be used by proponents of other methods to evaluate performance and to allow comparison of computational demands. The second example is more complex. The model is a continuous-time jump-diffusion model with latent stochastic volatility and jump intensity, with 10 parameters. In this case, the optimal statistics are unknown, so we cannot directly validate whether our procedure is identifying the correct statistics from the initial pool, which contains 56 statistics. We can however compare the performance of the resulting posterior mean, based on the selected statistics, relative to alternative selection methods (such as including all statistics). We find that our proposed selection method performs very well in terms of this metric.

In the remaining sections, a particular ABC estimator is first reviewed, and then our proposal for selection of statistics is presented and discussed in relation to existing summary statistic selection methods and nonparametric variable selection; we demonstrate the performance of the proposed algorithm through two examples, and finally, conclusions are offered. Software to replicate the results is available at http://www.runmycode.org/companion/view/1116. The software is written in the Julia programming language, which is a free, high level, high performance language that runs on all popular operating systems. This language may be unfamiliar to many readers, but it has a syntax very similar to widely used languages, and it is very easy to install and begin using in little time.

2 Review of the SBIL ABC Estimator

Creel and Kristensen (2013) define several indirect likelihood estimators, one of which, the simulated Bayesian indirect likelihood (SBIL) estimator, is an ABC estimator. The SBIL estimator is a specific version of ABC that is focused on point estimation, and which

uses particular methods. It is very similar to the ABC estimator proposed by Beaumont, Zhang and Balding, 2002. Here the name SBIL is used simply to avoid confusion with other versions of ABC estimators, such as versions which do not use a statistic, rejection-based ABC, ABC via Markov chain Monte Carlo, etc. The name also serves to highlight the relationship of the estimator to other indirect likelihood estimators, which are not ABC estimators.

Suppose we have a fully specified model indexed by a parameter $\theta \in \mathbb{R}^k$. Given a random sample of, say n, observations, y, which is a realization of the random vector Y, generated at the unknown true parameter value θ_0, define a random vector of statistics $Z = Z(Y)$, and let $z = Z(y)$ be the realized sample value of the statistic. Let Θ be the random parameter vector, in the Bayesian context. Creel and Kristensen, 2013 (henceforth, CK13) propose a Bayesian indirect likelihood (BIL) estimator

$$\hat{\theta}_{BIL} = E[\Theta|Z = z] = \int_{\Theta} \theta f(\theta|z)\, d\theta, \tag{2.1}$$

where, for some prior density $\pi(\theta)$ on the parameter space Θ, $f(\theta|z)$ is the posterior distribution given by

$$f(\theta|z) = \frac{f(z,\theta)}{f(z)} = \frac{f(z|\theta)\pi(\theta)}{\int_{\Theta} f(z|\theta)\pi(\theta)\, d\theta}.$$

This same concept is found in Fearnhead and Prangle (2012, Theorem 3). This is very much like the widely used Bayesian posterior mean, except that the likelihood is formulated in terms of the density of the statistic, $f(z|\theta)$, rather than the full sample. Advantages of the BIL estimator over GMM are the avoidance of optimization, avoidance of the need to compute the efficient weight matrix (see Hansen, 1982), and higher order efficiency relative to the GMM estimator that uses the same statistic and the associated optimal weight matrix (CK13).

Computation of the BIL estimator requires $f(z|\theta)$, which is normally not known. Just as the simulated method of moments may be required when GMM is not possible, simulation and nonparametric regression may be used to compute a simulated BIL (SBIL) estimator. This is implemented as follows (using standard notation from the literature on nonparametric regression):

Algorithm 1. *Basic SBIL algorithm (Beaumont, Zhang and Balding, 2002).*

1. Make independent and identically distributed draws θ^s, $s = 1, ..., S$, from the prior density $\pi(\theta)$

2. For each draw, generate a realized sample $y(\theta^s)$ from the model at this parameter value, and then compute the corresponding statistic $z^s = Z(y(\theta^s))$, $s = 1, ..., S$.

3. Given the i.i.d. draws (θ^s, z^s), $s = 1, ...S$, the SBIL estimator is

$$\hat{\theta}_{SBIL} = \hat{E}_S(\theta|z) = \frac{\sum_{s=1}^{S} \theta^s K_h(z^s - z)}{\sum_{s=1}^{S} K_h(z^s - z)},$$

where $K_h(z) = K(z/h)/h^{\dim z}$, $K(z) \geq 0$ *is a kernel function, and* $h > 0$ *is a common bandwidth parameter.*

The computation of the conditional mean in Step 3 is implemented using a kernel regression (or smoother) estimator with a global bandwidth, but other nonparametric regression estimators can be used in its place. For example, it could be computed using a nearest neighbor estimator, where the bandwidth is adaptive, so that the kernel places nonzero weight only on the k closest neighbors to z. In practice, this is what we do. One could also employ local linear estimators as advocated by Gao and Hong (2015).

Several factors affect the SBIL estimator. Among these are the prior, the kernel, the bandwidth, and the statistics chosen to form Z. Choice of kernels and bandwidths has been addressed in detail in the literature on nonparametric regression. Choice of the kernel is known to be relatively unimportant, if the bandwidth is selected appropriately (Marron and Nolan, 1989). In this work, we use simple K nearest neighbors regression, with the number of neighbors K selected using a rule of thumb, $K = S^{0.25}$, rounded down to the nearest integer. We do not enter further into these particular details, as they are not in the scope of this paper and the details are given in the code that accompanies the paper.

Regarding the prior, the basic SBIL algorithm outlined above uses simple sampling from the prior. When the prior and the posterior are not similar, direct sampling from the prior can be computationally inefficient, because many draws of the parameter θ^s lead to simulated statistics z^s that are often so far away from the observed statistic, z, so that the associated parameter draw is rarely retained as one of the neighbors that affects the estimated value. Recognizing this problem, methods of computing estimators using likelihood-free Markov chain Monte Carlo, sequential Monte Carlo, and importance sampling have been studied in some detail in the ABC literature (among others, see Marjoram *et al.*, 2003; Sisson, Fan and Tanaka, 2007; Beaumont *et al.* 2009; Del Moral, Doucet and Jasra, 2012). As noted in Creel and Kristensen (2011), one way of improving computational efficiency is to adapt the basic SBIL algorithm to use importance sampling as follows:

Algorithm 2. *Importance Sampling SBIL algorithm.*

1. *Make i.i.d. draws* θ^s, $s = 1, ..., S$, *from an importance sampling density* $g(\theta|z)$
2. *For each draw, generate a sample and compute the corresponding statistic,* z^s.
3. *Given the i.i.d. draws* (θ^s, z^s), $s = 1, ...S$, *we can obtain the SBIL estimator using*

$$\hat{\theta}_{SBIL} = \hat{E}_S[\Theta|Z = z] = \frac{\sum_{s=1}^{S} w_s \theta^s K_h(z^s - z)}{\sum_{s=1}^{S} K_h(z^s - z)},$$

where $w_s = \pi(\theta^s)/g(\theta^s|z)$ *are the importance sampling weights.*

If the importance sampling density $g(\theta|z)$ is close to the posterior density, then this algorithm will allow accurate computation of the estimator using many fewer draws (S)

than would be needed if Algorithm 1 were used. This is a straightforward application of importance sampling, and no claims to originality are made.

The following algorithm is a means of constructing an importance sampling density. The final importance sampling density $g(\theta|z)$ is simply a finite mixture density. The algorithm uses perturbation and a particular method of selection to determine the location parameters of the component densities. Because the algorithm has similarities to particle filtering methods, the means of the components of the mixture are referred to as "particles". The algorithm adaptively perturbs and selects the components in order to obtain a final density that has high mass in the region around the observed z.

Algorithm 3. *Construction of Importance Sampling density.*

1. Generate S_0 particles (draws) θ^s, $s = 1, 2, ..., S_0$ from the prior $\pi(\theta)$ and compute the associated statistics z^s for each particle. Set the iteration counter i to 0.

2. Set iteration counter i to $i + 1$.

3. Select the best S_1 current particles, based on proximity of z^s to z.

4. From the selected particles, randomly draw, with replacement, S_i particles.

5. Perturb each of the new particles by adding a mean zero random draw to one or more of the elements.

6. Generate a z^s for each of the S_i perturbed particles. Add the new particles to the current set of particles from Step 3, so that the set of current particles has $S_1 + S_i$ elements

7. Proceed to step 8 if the iteration counter is equal to a pre-established limit, otherwise, go to step 2.

8. Choose the best S particles from the $S_1 + S_i$ current particles, as in step 3.

9. Define the importance sampling density as the mixture of densities associated with each final particle: $g(\theta|z) = \frac{1}{S}\sum_{s=1}^{S} g_s(\theta|z;\theta^s)$.

To give a concrete example, if the particles are subjected to a mean zero multivariate normal perturbation in step 5, so that the density of a perturbation is $N(0, \Sigma)$, then the components of the final importance sampling density defined in Step 9 are $g_s(\theta|z;\theta^s) = N(\theta^s, \Sigma)$.

This algorithm has strong similarities to the sequential Monte Carlo (particle filtering) versions of ABC that have been proposed. For example, it is similar to the population Monte Carlo algorithm of Beaumont et al. (2009), to the replenishment algorithm of Drovandi and Pettitt (2011) and to Algorithm 4 of Marin et al., (2012). However, there are differences, too. The goal is simply to construct an importance sampling density, rather than an ABC posterior, so there is no declining tolerance level in the proposed algorithm. The actual ABC estimation is done using Algorithm 2, after the importance sampling density is constructed; it is not the product of the algorithm. Also, selection of particles is by nearest neighbors, rather than rejection. Previous experience with this algorithm (Creel and Kristensen, 2015) indicates that it is very effective in restricting attention to the portion of the parameter space that has non-negligible mass, which allows one to

specify a relatively uninformative initial prior, because the algorithm quickly focuses attention on the region which can generate auxiliary statistics that are close to the observed z. The only drawback to setting a very uninformative prior is that S_0 in Step 1 may need to be increased to compensate for excessive dispersion of the initial statistics, z^s. Note that steps 1, 4, 5 and 6 are "embarrassingly parallelizable" and step 3 can also be parallelized with some communications overhead. This means that the algorithm can easily be programmed to take advantage of available computational resources, such as multiple cores or a cluster of computers. In the application of Creel and Kristensen (2015), S_1 of step 3 was set to retain the best 20% of the particles. A careful exploration of the effect of this and other tuning details, such as the perturbation method, remains a topic for further study.

This discussion makes clear the nature of the estimator. The BIL estimator, $E[\Theta|z]$ (which cannot actually be computed) is a posterior mean, conditional on a statistic, rather than the full sample. The BIL estimator has the same asymptotic normal distribution as the optimal GMM estimator that uses the same statistic (CK13, see also Chernozhukov and Hong, 2003). Thus, the relationship between the BIL estimator and the ordinary posterior mean $E[\Theta|y]$ based on the full sample is essentially the same as the relationship between the GMM estimator and the maximum likelihood estimator: the first is in general not fully efficient, while the second is. The relationship between the SBIL estimator, which can be computed using importance sampling and nonparametric regression, and the BIL estimator, is like that between an ordinary Bayesian posterior mean computed using Markov chain Monte Carlo or some other computational technique, and the desired true posterior mean: the first is a numeric approximation of the second, which can be made as precise as needed by means of additional computational resources. Our argument for using the SBIL estimator is one of computational convenience and statistical performance. In terms of convenience, the SBIL estimator can be reliably computed using the above outlined algorithms, which are amenable to parallel computing techniques. There is no need for minimization or computation of the efficient weight matrix, as is the case for the GMM estimator. The remaining question is statistical performance, and this is fundamentally related to the main topic of this paper, the choice of which statistics to include in Z. Six factors are essential to the performance of the SBIL estimator: the prior, the number of simulations S, the importance sampling density, the kernel, the bandwidth, and the statistics chosen to form Z. The first five of these have been addressed in this section. We now turn to the last.

3 Selection of Auxiliary Statistics

In most implementations of the ABC estimator, a large number of candidate statistics are available, say, W. In such scenarios it may be tempting to simply include all of them in the computation of the estimator. However, recall that the simulated version of the ABC estimator takes the form of a nonparametric regression estimator, and such esti-

mators are known to suffer from the so-called *curse-of-dimensionality*: For a given set of simulations of size S and bandwidth choice h, the bias and variance of $\hat{E}_S[\Theta|W=w]$ due to simulations is of order $O\left(h^2\right)$ and $O(1/(Sh^{\dim(W)}))$, respectively, assuming that a so-called second order kernel is employed, c.f. Creel and Kristensen (2013). In particular, the optimal bandwidth choice should be of order $h = O\left(S^{-1/(4+\dim(W)}\right)$ yielding an error rate due to simulations of $O(S^{-2/(4+\dim(W))})$. Thus, the performance of the ABC estimator deteriorates as the number of summary statistics, $\dim(W)$, increases. For a good performance, we would therefore ideally like to only use the "relevant" portion of the set of candidate statistics W, where by relevant we mean summary statistics from W that are informative about θ_0, while leaving out the irrelevant ones. This is not special to the particular ABC estimator considered here. The same issue arises in the approximate Bayesian computation of the posterior density given W. There, the best rate one can hope for is $O(S^{-2/(4+\dim(\theta)+\dim(W)})$.

To formalize this idea, let δ be a $\dim(W) \times 1$ vector of zeros and ones, where a zero indicates that the corresponding auxiliary statistic is not used in the computation of the ABC estimator, and a one indicates that it is, and let $Z(\delta)$ be the corresponding vector of selected statistics. Let $\Delta = \{0,1\}^{\dim(W)}$ be the set of all possible values of δ. This set has $2^{\dim(W)}$ elements, a number which can be very large when a number of statistics are under consideration. We then define δ_0 as the minimal set of statistics for which there is no predictive loss,

$$\delta_0 = \arg\min_{\delta\in\Delta} \sum_{k=1}^{\dim(W)} \delta_k \text{ subject to } E[\Theta|W] = E[\Theta|Z(\delta)].$$

Note that the set of permissible δ's are non-empty since $\delta = (1,....,1)$ satisfies the constraint. The corresponding set of summary statistics $Z_0 = Z(\delta_0) \subseteq W$ is the minimal set of statistics for which

$$E[\Theta|W] = E[\Theta|Z_0]. \tag{3.1}$$

This is a weaker property compared to the concept of sufficient statistics. For Z_0 to be a sufficient statistic we would need that $f(\theta|z_0) = f(\theta|y)$ which in turn would imply that $E[\Theta|Y] = E[\Theta|Z_0]$. In contrast, the criterion we use for selecting relevant statistics only considers the impact of the summary statistic on the posterior mean and only ask it to be sufficient relative to the initial set of candidate statistics W, not the full sample. One could strengthen the above objective and require that Z_0 satisfies $f(\theta|z_0) = f(\theta|w)$, where $f(\theta|w)$ is the posterior of $\theta|W = w$. In the next section, we explain how our proposed method can be modified so as to select statistics that satisfy this property.

The goal is then to identify δ_0 (or correspondingly, Z_0) satisfying eq. (3.1). Evidently, given the model is fully parametric, Z_0 could in principle be obtained given knowledge of $f(\theta, W) = f(W|\theta)\pi(\theta)$. Unfortunately, $f(\theta, W)$ is not known on closed form; we can only simulate values from this distribution. Thus, based on a finite set of draws (θ^s, w^s), $s = 1, ...S$, we wish to develop a statistical procedure for choosing "an estimate" of Z_0.

For this purpose, we introduce the *Bayesian expected loss* of the action $\hat{E}_S\left[\Theta|Z\left(\delta\right)=z\left(\delta\right)\right]$ associated with a given decision $\delta \in \Delta$,

$$\begin{aligned}\mathcal{B}_S\left(\delta|w\right) &:= E\left[L\left(\Theta-\hat{E}_S\left[\Theta|Z\left(\delta\right)\right]\right)|Z\left(\delta\right)=z\left(\delta\right)\right] \\ &= \int_{\mathbb{R}^k} L\left(\theta-\hat{E}_S\left[\theta|Z\left(\delta\right)=z\left(\delta\right)\right]\right) f\left(\theta|z\left(\delta\right)\right) d\theta,\end{aligned} \tag{3.2}$$

where $L:\mathbb{R}^k \mapsto \mathbb{R}_+$ is some loss function chosen by the researcher. A natural choice of L is the (weighted) L_2 loss, $L_\Sigma\left(\theta\right)=\theta'\Sigma\theta$ for some positive definite weighting matrix $\Sigma \in \mathbb{R}^{k\times k}$, but alternative risk measures are allowed for, such as (weighted) L_1 loss. Given data w, we could then seek to minimize this w.r.t. δ. The above criterion reflects that $E\left[\Theta|Z\left(\delta\right)=z\left(\delta\right)\right]$ is approximated (estimated) by $\hat{E}_S\left[\Theta|Z\left(\delta\right)=z\left(\delta\right)\right]$, and takes into account the additional errors that the simulation draws generate. As pointed out earlier, these additional errors are a major reason for trying to reduce the number of summary statistics. As is well-known, $\mathcal{B}_S\left(\delta|w\right)$ can also be given a frequentist interpretation in terms of *Bayes risk* of the simulated estimator $\hat{E}_S\left[\Theta|Z\left(\delta\right)\right]$ associated with a particular choice of summary statistics.

The δ that minimizes the Bayesian expected loss would be dependent on the particular observation w of W in our sample. Here, we aim at finding a universal (global) set of statistics that works well for the model of interest, independently of the sample, and so modify the above expected loss function by integrating over all possible outcomes of W w.r.t. its marginal distribution $f\left(w\right)$, yielding the following *integrated loss function* $\mathcal{L}_S\left(\delta\right)$,

$$\mathcal{L}_S\left(\delta\right) := \int \mathcal{B}_S\left(\delta|w\right) f\left(w\right) dw = \int\int L\left(\theta-\hat{E}_S\left[\theta|Z\left(\delta\right)=z\left(\delta\right)\right]\right) f\left(\theta,z\left(\delta\right)\right) d\theta dz\left(\delta\right). \tag{3.3}$$

We will then use this for selecting statistics. In Section 4, we discuss how the procedure described below can be modified if one would rather select statistics according to $\mathcal{B}_S\left(\delta|w\right)$.

In principle, we would now minimize $\mathcal{L}_S\left(\delta\right)$ to obtain an estimate of δ_0. However, the integrals appearing in eq. (3.3) cannot be computed analytically, but Monte Carlo integration may be used: Let $\tilde{\theta}^r$ be a draw from the prior and $\tilde{z}^r\left(\delta\right)$ be the associated simulated draw from $Z\left(\delta\right)$ generated at the parameter value $\tilde{\theta}^r$, $r=1,...,R$. Note that $\left(\tilde{\theta}^r,\tilde{z}^r\left(\delta\right)\right)$, $r=1,...,R$, are drawn independently of the draws used to compute $\hat{E}_S\left[\Theta|Z\left(\delta\right)=z\left(\delta\right)\right]$; to emphasize this feature, we here use ˜ for the "test sample" of size R in order to differentiate it from the initial set of draws of size S. We then obtain the following simulated version of $\mathcal{L}_S\left(\delta\right)$,

$$\hat{\mathcal{L}}_S\left(\delta\right) = \frac{1}{R}\sum_{r=1}^{R} L\left(\tilde{\theta}^r-\hat{E}_S\left[\Theta|Z\left(\delta\right)=\tilde{z}^r\left(\delta\right)\right]\right).$$

One can think of this procedure as a "split sample method", where we have a pool of $S+R$ draws from the target distribution, and we then set aside the last R draws for

evaluation of $\hat{E}_S[\Theta|Z(\delta) = z(\delta)]$. This criterion is designed to do variable selection while safe guarding against over-fitting, which in our case corresponds to including irrelevant summary statistics. Instead of the simple sample splitting method used in $\hat{\mathcal{L}}_S(\delta)$, one could employ more advanced cross-validation method to evaluate the "out-of-sample" performance of $\hat{E}_S[\Theta|Z(\delta) = \tilde{z}^r(\delta)]$; we refer to Stone (1974) and the discussion in the following section for more on this issue.

Finally, one can introduce a penalty term $p(\delta)$ to encourage a more parsimonious selection. One example that will be used in the following is

$$p(\delta) = \left(1 + a \sum_{i=1}^{\dim W_n} \delta_i\right), \tag{3.4}$$

where $a > 0$ is a penalty weight, but other choices are possible. This penalty is increasing in the number of nonzero elements in δ. Multiplying the penalty onto the simulated loss function, so that the penalty increases the basic loss by a percentage, we arrive at our recommended decision rule:

$$\hat{\delta} = \arg\min_{\delta \in \Delta} CV(\delta|p),$$

where

$$CV(\delta|p) = \frac{p(\delta)}{R} \sum_{r=1}^{R} L\left(\tilde{\theta}^r - \hat{E}_S[\Theta|Z(\delta) = \tilde{z}^r(\delta)]\right). \tag{3.5}$$

In normal usage, cross validation by itself already enforces a parsimonious selection of statistics, so the penalty term $p(\delta)$ can be neutralized by setting a in equation (3.4) to zero. An example of use of the penalty term is given below in Section 5.2, when selection of statistics is targeted to a single parameter, in an effort to reduce bias.

The method is simple to implement in practice:

Algorithm 4. *Selection of statistics.*

1. draw R θ^r from the prior $\pi(\theta)$.

2. for each θ^r, generate the realized candidate statistics w^r.

3. minimize the criterion in equation (3.5) with respect to δ (recall that z^r is determined by δ, given w^r) to obtain $\hat{\delta}$, which defines the selected statistics.

Note that Steps 1 and 2 of Algorithm 1 need be performed only once, and then the result may be used for all of the iterations in Step 3 of Algorithm 4. A user only needs to supply a set of draws from the prior and the associated candidate statistics. This is a trivial requirement if the user is engaged in estimation using any ABC method. The algorithm does the rest.

Minimization of $CV(\delta)$ in equation (3.5) presents us with a non-differentiable minimization problem over a discrete choice set that potentially (depending on the number of candidate statistics) contains an extremely large number of possible combinations. We propose to resolve this issue by employing simulated annealing (Černý, 1985; Goffe et al.,

1994), a global optimization method that performs a stochastic search over the domain of the objective function. Simulated annealing is an algorithm that is well known in many areas, including economics, where applications include Wu and Wang (1998), for the computation of economic equilibria, and Örkcü (2013), for selection of regressors in linear regression models. It is related to the Metropolis-Hastings algorithm, which should be familiar to users of Markov Chain Monte Carlo method in that movements that do not improve the objective function may be accepted, which allows the algorithm to escape from local minima, but the probability of accepting such upward movements declines as the algorithm progresses. The rate at which this probability declines is known as the cooling schedule of the algorithm. A slow cooling rate helps to ensure that the algorithm does not converge to a local minimum, but it also leads to a longer run time.

Our version of the algorithm is loosely based on the Fortran code provided by Goffe et al. (1994), with adaptation to a discrete choice set. It is written in the Julia programming language (http://julialang.org), which is a high level, high performance scientific programming language, with syntax not unlike that of Matlab, but performance that is similar to that of C. We obtain an approximately optimal set of statistics by applying simulated annealing to approximately minimize the cross validation score over the discrete choice set. The simulated annealing algorithm can theoretically find the actual minimizer, if the cooling schedule is slow enough (Černý, 1985), but this may entail a very high computational demand. It is possible to contract the search space faster than is allowed for convergence to the actual minimizer, or one may simply limit the number of objective function evaluations. If this is done, the actual minimizing value of δ may not be found, depending on how smooth and regular is the objective function, but at least a systematic means of search is used, and by dedicating enough computational power, the final solution should be close to being optimal. It is easy to do parallel runs of the minimization, to more thoroughly explore the δ space.

4 Discussion

We here discuss our proposal, relative to existing methods found in the ABC literature, and to variable selection methods developed in the literature on nonparametric regression. As mentioned earlier, existing methods for selection of summary statistics can broadly be categorized into three groups: best subset selection, projection methods and regularization techniques.

Among papers that use projection, Fearnhead and Prangle (2012) explore using a training set of simulated observations, regressing parameters drawn from the prior on statistics computed using the simulated data, and then using the fitted values (a weighted index of the initial statistics that is an approximation of the posterior mean of the parameter) as a single statistic for ABC estimation, for each parameter in turn. This approach reduces to one (per parameter) the dimension of the initial large set of statistics. This may lead to an excessive loss of information, in that there may exist no single linear index that

captures the information in the large body of initial statistics. A further limitation is the use of linear regression, which may perform poorly if the true relationship between the parameter and the informative statistics is nonlinear. Fearnhead and Prangle (2012) address this by using methods to endogenously choose the training set from a region of reasonably high posterior mass. In contrast, our method works with a consistent estimator of the posterior mean and so avoids any biases due to the use of linear regression techniques.

As an example of papers that select a subset of statistics, Joyce and Marjoram (2008) propose an algorithm to determine whether or not a new statistic should be added to a previously selected set of statistics, based on approximations of the posterior odds ratio, computed using rejection-based ABC. If the new statistic is added, then removal of the previously selected statistics is contemplated. This proceeds until all statistics have been considered. Also proposing a subset selection strategy, Nunes and Balding (2010) perform rejection-based ABC for all subsets of the candidate statistics, and select statistics to minimize the entropy of the resulting ABC posterior. This may be followed by a second step that again searches over all subsets of the candidate statistics to minimize a criterion that depends in part on the statistics that were selected in the first step. The first step of their procedure may not work well in certain contexts (see Blum, et al., 2013), and this could undermine the effectiveness of the second step. Their second step is quite similar to what is proposed in the present paper. However, the present paper proposes a systematic and computationally efficient method to do the minimization, rather than searching over all possible solutions, which may not be possible to do when a large number of candidates are under consideration. A final example of papers that use subset selection is Barnes et al. (2012). This paper proposes a stepwise method of introducing statistics until the Kullback-Leibler divergence between ABC posterior distributions is below a tolerance. This method is interesting, and appealing from the Bayesian perspective, in that the full posterior influences the selection of statistics, compared to the method proposed here, which uses only the posterior mean. In comparison to these methods, we conduct a full search over the set of possible candidate statistics without relying on step-wise procedures. Moreover, our method is computationally very simple as Algorithm 4 reveals. In particular, our procedure does not involve any additional tuning parameters (besides, optionally, the penalization weight). Below, we discuss how the information theoretic criterion developed in Barnes, et al. (2012) in principle could be combined with the ideas of this paper to yield an alternative version of our cross-validation method.

In the GMM literature, attention has also been given to the choice of moment conditions (which corresponds to the summary statistics in our setting). Notable examples include Andrews and Lu (2001), Cheng and Liao (2012) and Hall et al. (2012) who proposed methods for selecting moment conditions out of a potentially large set of possible moment conditions using penalization methods. All these papers focus on the case of a partially specified model which complicates the selection algorithms and their analysis considerably, which tends to rely on asymptotic arguments. In contrast, since we have a

fully specifed model, more information about the relative importance of a given statistic is available, and an exact analysis can be made.

Our proposed cross-validation method is also related to variable selection methods in nonparametric regression. By definition, we know that the ABC estimator related to the set of ideal set of statistics Z_0 satisfies

$$\Theta = m\left(Z_0\right) + \epsilon, \quad m\left(z_0\right) := E\left[\Theta | Z_0 = z_0\right], \tag{4.1}$$

where $E\left[\epsilon | Z_0\right] = 0$. In the context of this regression, one can interpret the problem as one of variable selection: If we add more statistics (regressors) to the above regression, there will be no improvement in the fit, while at the same time the nonparametric estimator of m will deteriorate due to the curse-of-dimensionality. In this context, we would like to choose δ so as to minimize the distance between the optimal predictor, $m\left(z_0\right)$ and $\hat{E}_S\left[\theta | Z\left(\delta\right) = z\right]$ while taking into account the simulations. This could be done by minimizing

$$\begin{aligned} \int E\left[L_\Sigma\left(m\left(z_0\right) - \hat{E}_S\left[\Theta | Z\left(\delta\right) = z\left(\delta\right)\right]\right)\right] f\left(z\left(\delta\right)\right) dz \quad &= \\ \int E\left[L_\Sigma\left(\theta - \hat{E}_S\left[\Theta | Z\left(\delta\right) = z(\delta)\right]\right)\right] f\left(z(\delta)\right) dz \quad &+ E\left[L_\Sigma\left(\epsilon\right)\right], \end{aligned}$$

where, as before, $f\left(z(\delta)\right)$ is the density of $Z\left(\delta\right)$, while $E\left[\cdot\right]$ denotes expectations w.r.t. the simulations, and $L_\Sigma\left(\theta\right) = \theta'\Sigma\theta$ is the weighted L_2 -loss. Note that the first term on the right hand side is simply $\mathcal{L}_S\left(\delta\right)$ while the second one is independent of δ. Thus, one can interpret our procedure as one of variable selection in a nonparametric regression framework. In the literature on nonparametric regression, $\mathcal{L}_S\left(\delta\right)$ is usually estimated by

$$\tilde{CV}\left(\delta\right) := \frac{1}{S}\sum_{s=1}^{S} L_\Sigma\left(\theta^s - \hat{E}_{-s}\left[\Theta | Z\left(\delta\right) = z^s(\delta)\right]\right),$$

where $\hat{E}_{-s}\left[\Theta | Z\left(\delta\right) = z(\delta)\right]$ is the so-called leave-on-out version of $\hat{E}_S\left[\Theta | Z\left(\delta\right) = z(\delta)\right]$. Very often penalty terms are added to $\tilde{CV}\left(\delta\right)$ in order to further regularize the selection procedure; see e.g., Härdle et al. (1998, p. 88). Thus, the main difference between the above alternative cross-validation criterion and the one proposed in the previous section are that while we use sample splitting, $\tilde{CV}\left(\delta\right)$ estimates the integrated loss using the same simulated sample which is used to estimate $E\left[\Theta | Z\left(\delta\right) = z(\delta)\right]$. The two measures $CV\left(\delta | p\right)$ (with $p = 1$) and $\tilde{CV}\left(\delta\right)$ are asymptotically equivalent, but in practice, with not too large R, $CV\left(\delta\right)$ is faster to implement relative to $\tilde{CV}\left(\delta\right)$ since they involve RS and S^2 computations, respectively.

Cross-validation methods are mostly used for bandwidth selection (see, e.g., Härdle et al., 1998), but can also be used for variable selection as advocated by Hall et al. (2004) in the context of conditional density estimation. For variable selection, all candidate statistics, W, are included in the nonparametric kernel regression, but with individual band-

widths $h_1, ..., h_{\dim W}$ so that

$$\hat{E}\left[\Theta|W=w\right] = \frac{\sum_{s=1}^{S} \theta^s K_H\left(w^s - w\right)}{\sum_{s=1}^{S} K_H\left(w^s - w\right)},$$

where $K_H\left(w^s - w\right) = |H|^{-1} K\left(H^{-1}\left(w^s - w\right)\right)$, and H is the diagonal matrix with diagonal elements $h_1, ..., h_{\dim W}$. By choosing these bandwidths to minimize $\tilde{CV}\left(h_1, ..., h_{\dim W}\right) = \frac{1}{S}\sum_{s=1}^{S} L_\Sigma\left(\theta^s - \hat{E}_{-s}\left[\Theta|W=w^s\right]\right)$, where $\hat{E}_{-s}\left[\Theta|W=w^s\right]$ is the leave-one-out estimator, bandwidth selection and variable selection is done jointly with irrelevant statistics being removed implicitly by being assigned (very) large bandwidths, making them vanish in practice; see Hall et al. (2004). Comparing this variable selection method with the one proposed here, one can think of our procedure as a computational shortcut to the variable selection procedure using multiple bandwidths: When $\dim W$ is large, the joint minimization problem of $\tilde{CV}\left(h_1, ..., h_{\dim W}\right)$ w.r.t. $h_1, ..., h_{\dim W}$ is quite challenging and very time consuming, with local minima very often being present. In contrast, by using a single bandwidth for all variables and using δ to select statistics, our variable selection problem is of a lower dimension which can be solved more quickly and is numerically more stable.

The selection rule developed in the previous section chooses the summary statistics so as to minimize the (integrated) Bayes loss of the posterior mean. In many cases, the interest lies in the whole of the posterior distribution, not just its mean, however. Fortunately, it is straightforward to modify our selection rule to handle this case by simply replacing the Bayes loss as objective function by the Kullback-Leibler information criterion (KLIC) of the posterior distribution conditional on data. The (conditional) KLIC of the posterior distribution, corresponding to a given choice $Z\left(\delta\right)$, relative to the full set of candidate summaries is given by

$$KL\left(\delta|w\right) = \int \log f\left(\theta|w\right) f\left(\theta|w\right) d\theta - \int \log f\left(\theta|z\left(\delta\right)\right) f\left(\theta|w\right) d\theta.$$

One can now minimize $KL\left(\delta|w\right)$ to obtain the minimal set of summary statistics for which $f\left(\theta|z_0\right) = f\left(\theta|w\right)$, where $Z_0 = Z\left(\delta_0\right)$ and $\delta_0 = \min_\delta KL\left(\delta|w\right)$. Thus, compared to the selection criterion proposed in the previous section, this KLIC selection rule satisfies the stronger property of Z_0 being a sufficient statistic relative to the candidate set of statistics W. The use of KLIC for selecting summary statistics was already proposed by Barnes et al. (2012); however, while they use a step-up procedure for identifying δ_0, where one statistic is added at a time until there is no further improvement in terms of KLIC, we here do a full search over all possible sets of candidates.

As already noted by Barnes et al. (2012), KLIC is only of theoretical interest since its computation involves approximating $f\left(\theta|w\right)$ which is a high-dimensional problem when $\dim W$ is large. To resolve this issue, we proceed as in the previous section and introduce

the integrated (or expected) version of this measure,

$$\begin{aligned} KL(\delta) &= \int KL(\delta|w) f(w) dw \\ &= \int\int \log f(\theta|w) f(\theta|w) f(w) d\theta dw - \int\int \log f(\theta|z(\delta)) f(\theta|w) f(w) d\theta dw \\ &= \int\int \log f(\theta|w) f(\theta) d\theta dw - \int\int \log f(\theta|z(\delta)) f(\theta, z(\delta)) d\theta dz(\delta). \end{aligned}$$

Note that the first term in the expression of $KL(\delta)$ is independent of δ, while the second term can be estimated by $\frac{1}{R}\sum_{r=1}^{R} \log \hat{f}_S(\theta^r|z^r(\delta))$ where $\hat{f}_S(\theta|z(\delta))$ is the kernel density estimator of the posterior density,

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

Note again that we use sample splitting to safeguard against over-fitting (i.e., including irrelevant summary statistics). With this simulated version of the posterior density, we then propose the following KLIC selection rule,

$$\hat{\delta}_{KL} = \arg\max_{\delta \in \Delta} \frac{1}{R} \sum_{r=1}^{R} \log \hat{f}_S(\theta^r|z^r(\delta)).$$

As with the Bayes expected loss, one could add a penalty term to the above objective function so as to further regularize the problem.

In large samples, the KLIC and the Bayes expected loss versions of our selection method will be asymptotically equivalent under certain regularity conditions. Specifically, if the summary statistics satisfy a central limit theorem, the posterior distribution is well-approximated by a Gaussian distribution centered around the posterior mean (see Creel and Kristensen, 2013),

$$\log f(\theta|z) \propto (\theta - E[\Theta|Z = z])' \Omega^{-1}(\theta) (\theta - E[\Theta|Z = z]),$$

where $\Omega(\theta)$ is the asymptotic variance of $E[\Theta|Z = z]$. This implies that the two cross-validation methods using KLIC and Bayes expected loss (using a weighted L_2 loss function), respectively, will in many cases be asymptotically equivalent, as sample sizes increase.

If the objective is to identify the optimal set of statistics for the given sample at hand, our proposed method is easily adjusted to handle this: Using the "test sample" (θ^r, z^r), $r = 1, ..., R$, to estimate the posterior density $f(\theta|z(\delta))$ and then substitute this into the Bayes expected loss $\mathcal{B}_S(\delta|z(\delta))$ defined in eq. (3.2), we obtain

$$\begin{aligned} \hat{\mathcal{B}}_S(\delta|w) &= \int_{\mathbb{R}^k} L\left(\theta - \hat{E}_S[\theta|Z(\delta) = z(\delta)]\right) \hat{f}_R(\theta|z(\delta)) d\theta \\ &= \frac{\sum_{r=1}^{R} L\left(\tilde{\theta}^r - \hat{E}_S[\theta|Z(\delta) = z(\delta)]\right) K_h(\tilde{z}^r(\delta) - z(\delta))}{\sum_{r=1}^{R} K_h(\tilde{z}^r(\delta) - z(\delta))}. \end{aligned}$$

One could now minimize this criterion w.r.t. δ using simulated annealing to obtain an estimate of the optimal set of summary statistics conditional on the observed sample. The precise algorithm is an obvious adjustment of Algorithm 4. One disadvantage of this procedure is that $\hat{\mathcal{B}}_S(\delta|w)$ is more sensitive to the bandwidth h compared to $\hat{\mathcal{L}}_S(\delta)$ - one would in fact probably wish to use a different bandwidth for the test sample of size R. To reduce the effect of h, one could potentially proceed in two steps: First, minimize $CV(\delta)$, and then do a second stage of selection of statistics by minimizing $\hat{\mathcal{B}}_S(\delta|w)$ with initial starting value chosen as the estimated obtained in the first stage. This idea is similar to Nunes and Balding's (2010) second stage procedure, which has a criterion that depends on the accepted draws, and to Fearnhead and Prangle's (2012) suggestion to use an endogenously chosen target set, given z, to focus on regions of non-negligible posterior density.

5 Examples

5.1 A simple example

This section presents a simple example that illustrates the proposed methods using Bayes expected loss as criterion. The data generated by a simple linear regression model with normally distributed errors

$$y_i = \alpha + \sum_{j=1}^{4} \beta_j x_{ij} + \sigma u_i, \tag{5.1}$$

$i = 1, 2, ..., n$. The x_{ij} are all independently drawn from the standard normal distribution, as is the error u_i. We investigate selection for two sample sizes: n =30 and n = 100 observations. The parameter vector is $\theta = (\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \sigma)$. The priors for the α and β parameters are independent $U(-2, 2)$ densities, and the prior for σ is $U(0, 5)$. To be clear, the regressors are random, so they vary across samples, but for every sample, they are observed without error. The model is simply the classical linear regression model with Gaussian errors, with the exception that the regressors are not fixed across repeated samples. The model details are apparent in the accompanying code "make_simdata.jl".

`[insert online link to "make_simdata.jl" accompanying text file about here]`

For this linear model, the ordinary least squares (OLS) estimator of the α and the β parameters is the maximum likelihood estimator, and this vector, along with the associated estimator of σ, provide auxiliary statistics for the SBIL estimator such that it is asymptotically efficient as a point estimator. This is because (i) these statistics cause the just identified efficient GMM estimator to coincide with the ML estimator, which is asymptotically efficient[1], and (ii) the efficient GMM and SBIL estimators are first order asymptotically equivalent (CK13). Thus, the SBIL estimator that uses these statistics is fully asymptotically efficient as a point estimator, and these statistics are optimal in this

[1] To be more clear, the GMM estimator uses the moment conditions $m(\theta) = \theta - \hat{\theta}$, where $\hat{\theta}$ are the OLS estimates of the parameters. The GMM criterion to be minimized is $s(\theta) = [m(\theta)]' m(\theta)$. The solution is clearly to set the GMM estimate to the OLS estimate, which is the ML estimate.

sense. We do not claim that these statistics lead to an optimal posterior density, in any sense. It is of interest to include optimal statistics in the pool of candidates to see whether or not the selection procedure can select them.

Together with the "optimal" statistics, we added the OLS regression coefficients and estimators of σ from two incorrectly specified models:

$$y_i = \alpha + \sum_{j=1}^{4} \beta_j x_{ij} + \sum_{j=1}^{4} \gamma_j x_{ij}^2 + \sigma u_i \tag{5.2}$$

and

$$y_i = \alpha + \sum_{j=1}^{4} \beta_j x_{ij} + \sum_{j=1}^{4} \gamma_j x_{ij}^2 + \sum_{j=1}^{4} \delta_j x_{ij}^3 + \sigma u_i. \tag{5.3}$$

These models include irrelevant regressors. The estimators of the α, β and σ parameters from the quadratic and cubic regressions are still consistent estimators of the associated parameters, though. However, the α and β estimators from these regressions are not efficient, due to the added irrelevant regressors. The σ estimators are asymptotically efficient in all three of the regressions, so we expect that selection between them will be more or less evenly spread. The estimated γ and δ parameters of the quadratic and cubic regressions will be partially informative for the β parameters, due to collinearity between the x regressors and the square and cubic terms. Thus, these are informative, but not optimal statistics. Finally, 5 statistics that are simply standard normal white noise are added to the set of candidates. These statistics are completely uninformative, and ideally, will never be chosen by the selection procedure. Thus, the entire set of candidate statistics includes:

- 6 optimal statistics from the estimation of equation 5.1
- 9 relevant but non-optimal statistics, and one optimal statistic, from the estimation of equation 5.2
- 13 relevant but non-optimal statistics, and one optimal statistic, from the estimation of equation 5.3
- 5 irrelevant (pure noise) statistics

for a total of 35 statistics. There are 5 optimal statistics for the α and β parameters, one for each element, coming from the linear regression. There are 3 alternative optimal statistics for σ, from each of the 3 regressions. The best possible result of a selection method would be to choose the α and β estimators from the linear regression model, and one of the σ estimators, from any of the three regressions models. The file "make_simdata.jl" which accompanies this paper allows one to generate replications of (θ^s, z^s).

The loss function L in equation (3.5) is chosen as the mean absolute error of the SBIL estimator, with scaling by the inverse of prior standard deviation of each parameter. That

is, $L(\theta) = \frac{1}{\dim\theta}\sum_{j=1}^{\dim\theta}\frac{1}{\sigma_j}\left|\theta_j\right|$. Here, σ_j is the sample standard deviation of the jth parameter in the S draws made in the first step of Algorithm 1. Scaling is important to make the cross validation criterion place similar weight on all parameters. In its absence, if some parameter had a much tighter prior than the others, it would have little effect on the chosen statistics. The penalty term $p(\delta)$ in equation (3.5) has no effect, because a in equation (3.4) is set to zero. The number of draws from the prior (R in equation 3.5) is set to $R = 1000$, and the number of simulations (S in algorithm 1) is set to $S = 10000$.

`[insert online links to "Selection.jl" and "Selection_mpi.jl" accompanying text files about here]`

The main selection algorithm is contained in the file "Selection.jl", which accompanies this paper. A version which uses the message passing interface (MPI) to run on a multicore computer or a cluster is in the file "Selection_mpi.jl". Instructions for how to do this are in the file "README", which accompanies the paper. The previously mentioned file "make_simdata.jl" can be used to generate the data sets used by these scripts. All of the code may be downloaded in one file from http://www.runmycode.org/companion/view/1116.

`[insert online link to README accompanying text file about here]`

The selection algorithm was run using a battery of 100 simulated annealing minimizations, for both sample sizes. We may think of this as investigation of 100 separate runs of the algorithm, or as investigation of a single conservative run, where the use of a battery of runs allows more rapid cooling than would otherwise be advisable. The results are summarized in Table 1. Interpreting the results for the overall battery (the columns headed with "selected in best run"), over the 100 replications, the replication that had the smallest criterion value selected the optimal (in the sense defined above) statistics that correspond to the estimated parameters using the linear model, for the sample size $n = 30$. For the sample size $n = 100$, the same statistics were selected, with the exception that the regression coefficient for β_3 from the quadratic model was selected, and the corresponding coefficient from the linear model was not. In both cases, only 6 statistics were selected to estimate the six parameters.

Next, we may interpret the 100 runs separately, to examine what results might be expected if the method is used with less care. For both sample sizes, the γ_i and δ_i statistics of the quadratic and cubic models were never selected in any of the runs, so they do not appear in the table. Two of the pure noise statistics were selected, one time each, for the sample size $n = 30$. Overall, this is an important result: weakly informative and non-informative statistics are almost never selected. However, the estimators of α,β and σ of the linear, quadratic and cubic models are all more strongly informative, and they may be fairly close substitutes for one another. For the sample size $n = 30$, the "optimal" statistics corresponding to the linear model were selected almost always, the β estimators of the quadratic model were selected with some frequency, and the β estimators of the cubic model were selected rarely. For this sample size, the quadratic and cubic models are certainly over-parameterized, so their estimators have fairly high variances, compared

Tab. 1: Selected statistics, linear regression example, 100 replications of selection procedure

Statistic	$n = 30$		$n = 100$	
	% times selected	selected in best run	% times selected	selected in best run
$\widehat{\alpha_L}$	100	*	98	*
$\widehat{\beta_{1L}}$	100	*	75	*
$\widehat{\beta_{2L}}$	97	*	76	*
$\widehat{\beta_{3L}}$	98	*	74	
$\widehat{\beta_{4L}}$	100	*	92	*
$\widehat{\sigma_L}$	64	*	52	*
$\widehat{\alpha_Q}$	3		10	
$\widehat{\beta_{1Q}}$	24		56	
$\widehat{\beta_{2Q}}$	23		47	
$\widehat{\beta_{3Q}}$	21		48	*
$\widehat{\beta_{4Q}}$	16		55	
$\widehat{\sigma_Q}$	66		56	
$\widehat{\alpha_C}$	0		13	
$\widehat{\beta_{1C}}$	2		6	
$\widehat{\beta_{2C}}$	2		10	
$\widehat{\beta_{3C}}$	1		9	
$\widehat{\beta_{4C}}$	1		8	
$\widehat{\sigma_C}$	55		54	
N_2	1		0	
N_5	1		0	

to the variances of the corresponding estimators of the linear model. The selection procedure is able to choose the estimators of the linear model, which does not suffer from over-parameterization. However, for the sample size $n = 100$, the over-parameterization problem is less severe, and we see that the α and β estimators of the quadratic and cubic models are selected more often, as they become closer substitutes for the corresponding estimators of the linear model. For the σ parameter, recall that the estimators of all three models have the same asymptotic distribution. This explains why the estimators from all three models are selected more or less evenly, for both sample sizes. Note that if we sum, over the linear, quadratic and cubic models, the number of times an estimator of a given parameter is selected, the result is greater than 100, which means that the selection procedure sometimes selects more than one statistic that is informative about a given parameter.

To summarize, this discussion of the 100 separate runs is indicative of the result one could obtain if the selection procedure is used without much care, seeking a quick selection, rather than a careful selection. We see that we still obtain quite good results: weakly informative and uninformative statistics are almost never selected. There is some difficulty in choosing between statistics that are close substitutes. However, when statistics are close substitutes, either one will do well for the final ABC estimation, so the choice between them is not too important. When the procedure is used with care, selecting the best of repeated runs, the problem of not distinguishing between statistics that are close substitutes disappears.

Regarding the time needed to perform selection, a single run using a single core of a Macbook Air notebook computer, setting $S = 10000$ and $R = 1000$, takes less than 13 minutes. The "Selection_mpi.jl" version of the code uses MPI for parallel execution, allowing a number of runs to be done simultaneously, which helps a great deal if one wants to use a battery of runs. However, if one does not have MPI installed, it is also possible to improve selection using a single run, by making some small changes to Selection.jl. One may use a slower cooling schedule (set rt closer to 1), or use a larger "in sample" data set (S) and/or a larger "out of sample" data set (R). Such changes will improve the chances of arriving to the global minimum, or close to it, at the cost of increasing execution time. In this simple example, the number of candidate statistics is 35, so a method that uses an exhaustive search over all possible combinations of statistics, such as that of Nunes and Balding (2010), would require evaluating 2^{35} combinations, which is infeasible. The proposed method can deal with a large number of candidate statistics while using a reasonable amount of computational time.

5.2 Selection of statistics for a jump-diffusion stochastic volatility model

Creel and Kristensen (2015; henceforth CK15) use ABC methods to estimate a continuous time stochastic volatility model that has non-constant drift, leverage, and jumps with

dynamic jump intensity. In this model, the true log price, p_t, solves

$$dp_t = (\mu_0 + \mu_1(h_t - \alpha)/\sigma)\, dt + \sqrt{\exp(h_t)} dW_{1,t} + J_t dN_t. \tag{5.4}$$

where h_t is log volatility, J_t is jump size, and N_t is a Poisson process with time-varying jump intensity λ_t. Log-volatility is specified to follow

$$dh_t = h_t + \kappa(\alpha - h_t)dt + \sigma\left(\rho dW_{1,t} + \sqrt{1-\rho^2} dW_{2,t}\right), \tag{5.5}$$

where $W_{1,t}$ and $W_{2,t}$ are two independent standard Brownian motions. Jump sizes, conditional on the occurrence of a jump, are independent and conditionally normally distributed: $J_t \sim N(\mu_J, \sigma_J^2)$. Finally, the jump intensity process λ_t is modeled as

$$\lambda_t = 1(\lambda_t^* > 0) \text{ where } \lambda_t^* = \lambda_0 + \lambda_0\lambda_1(h_t - \alpha)/\sigma.$$

CK15 discusses the model and interprets the parameters. One of the results was that measurement error in p_t is not an important factor for the data set that was studied. For this reason, here, we use the model that does not include measurement error. We collect the 10 parameters of the model in the vector $\theta = (\mu_0, \mu_1, \alpha, \kappa, \sigma, \rho, \lambda_0, \lambda_1, \mu_J, \sigma_J)$.

The data series available for computing statistics include daily returns, r, realized volatility measured using 5 minute intervals (RV) , and realized bipower variation (BV), which is a measure of volatility that is robust to jumps. The candidate statistics are functions of these three variables. The variable $IJ = RV - BV$ is constructed as indicator of jump activity, as is discussed by Andersen, Bollerslev and Diebold (2007). Another variable, $r2$, is identical to r, except that returns are set to zero in periods when IJ_t is greater than 2.5 of its own standard deviations. This is an attempt to create a returns series that is net of jumps. The candidate statistics are: the mean of the jump indicator (statistic 1); the correlation of the jump indicator with the average of its 10 most recent lags (statistic 2); the mean, standard deviation, skew and kurtosis of the variables $r, r2, RV, BV, IJ$ (statistics 3-22); the correlations between the same 5 variables (statistics 23-32), and regression coefficients, standard errors and R^2 from several auxiliary models (statistics 33-53), along with cross equation correlations of the residuals of regressions 2, 3 and 4 (statistics 54-56).

The four auxiliary regressions [2] are:

$$\textbf{Aux. reg. } 1\text{: } \log BV_t = \alpha_1 \frac{\log BV_{t-1} - \log BV_{t-2}}{2} + \alpha_2 \frac{\log BV_{t-3} - \log BV_{t-4}}{2} + \epsilon_{1t}$$

$$\textbf{Aux. reg. } 2\text{: } IJ_t = \alpha_1 \log BV_{t-1} + \alpha_2 \left(\frac{\sum_{j=1}^{10} IJ_{t-j}}{10}\right) + \epsilon_{2t}$$

[2] All variables in the auxiliary regressions are standardized and and normalized before estimation of parameters.

Aux. reg. 3: $\log BV_t = \alpha_1 \log BV_{t-1} + \alpha_2 r2_{t-1} + \alpha_3 \left(\log BV_{t-1}\right)^2 + \alpha_4 \left(\log r2_{t-1}\right)^2 + \epsilon_{3t}$

Aux. reg. 4: $\log r2_t = \alpha_1 r2_{t-1} + \alpha_2 \log BV_{t-1} + \alpha_3 \left(r2_{t-1}\right)^2 + \alpha_4 \left(\log BV_{t-1}\right)^2 + \epsilon_{4t}.$

These auxiliary statistics are somewhat different than those used in CK15. In particular, the EGARCH model used in that paper was the most costly statistic to compute, but it contributed no statistics to those that were selected, so it has been dropped. There have been some other minor adjustments to the auxiliary regressions, too, but overall, the statistics in the candidate set are very similar to those of the previous paper.

The prior is different from that used in CK15, because the uniform prior of CK15, and the broad bounds chosen for its support, are probably unrealistically uninformative, and because CK15 already presents results using selected statistics, for the original prior. For the purpose of selecting informative statistics, it is probably of more interest to investigate how the procedure performs when draws are made from a region of non-negligible posterior mass. Thus, the prior used here is one that makes use of the previous estimation results for the S&P 500 series, during the period 2008-2011, to focus on such a region. Nevertheless, we seek to select statistics which will perform well even if market conditions change. Specifically, the prior is that the parameters have independent normal distributions with means given by the estimates for the S&P 500 series during the period 2008-2011, found in Table 3 of CK15, and standard deviations equal to two times the estimated standard errors, also reported in the same table. However, the lower and upper bounds for parameters given in Table 1 of CK15 continue to be enforced, through rejection sampling. Thus, the prior is not exactly unbiased, because the parameter bounds are occasionally binding for certain parameters (most notably, μ_0 and ρ).

To apply the selection procedure, we set $S = 20000$ and $R = 1000$. A larger value of S is used than was the case for the linear regression model, because the number of candidate statistics is much larger, and a larger sample is required to help ensure that $\hat{E}_S\left[\Theta | Z\left(\delta\right) = \tilde{z}^r\left(\delta\right)\right]$ in equation (3.5) is a reasonably accurate nonparametric fit. The loss function is the same mean absolute error criterion as was used in the previous example. We performed a battery of 100 runs of the selection procedure. The time to complete this entire computation, using parallel evaluation on 20 computational cores of a single powerful server, was approximately 250 minutes.

Table 2 lists the selected statistics from the best of the 100 runs, and describes each selected statistic. Of the 56 candidate statistics, 17 were selected. Perhaps the most important point to highlight is that the selected statistics are a subset of the candidate set which one would be unlikely to arrive at if unsystematic experimentation were done. It is encouraging to see that statistics indicative of jump activity (statistic 1), jump dynam-

Tab. 2: Selected statistics, jump-diffusion stochastic volatility model

Statistic	Description	Statistic	Description
1	jump detection	36	$\widehat{\sigma_\epsilon}$, aux. reg. 1
2	jump clustering	43	$\widehat{\alpha_2}$, aux. reg. 3
4	mean $r2$	44	$\widehat{\alpha_3}$, aux. reg. 3
6	mean BV	50	$\widehat{\alpha_3}$, aux. reg. 4
23	corr$(r, r2)$	52	$\widehat{\alpha_4}$, aux. reg. 4
25	corr(r,BV)	53	R^2, aux. reg. 4
30	corr(RV, BV)	55	corr$(\widehat{\epsilon_3},\widehat{\epsilon_4})$
31	corr(RV,IJ)	56	corr$(\widehat{\epsilon_2},\widehat{\epsilon_4})$
34	$\widehat{\alpha_2}$, aux. reg. 1		

ics (statistic 2) and leverage (statistics 25 and 55) are selected, because these features are present in the model. The means of $r2$ and BV are selected, but variances, skewness and kurtosis are never selected. Several correlations between variables are selected, and two of the cross equation residual correlations are selected. The four auxiliary regressions all contribute to the set of selected statistics.

Next, to investigate how well the selection procedure performs, we do a Monte Carlo exercise similar to that of CK15. The sample size is $n = 2000$, as in CK15. For each estimator explored, we perform only 100 Monte Carlo replications of the ABC estimator, as the goal is not to explore carefully the performance of the ABC estimator, but rather to focus on how well the selection procedure performs. The true parameter values are the estimated parameters for the 2008-2011 period, in Table 3 of CK15. Table 3 presents bias and root mean squared error for the prior, the ABC estimator using the selected statistics described in the previous paragraph (the "Baseline" results), the ABC estimator that uses all 56 statistics in the candidate set, and an ABC estimator that uses 17 randomly selected statistics, out of the 56 candidate statistics. We may note that the prior is essentially unbiased, except for μ_0 and ρ, for which the parameter bounds are sometimes binding. Comparing the "Prior" and "Baseline" columns, we see that the selection procedure yields an ABC estimator that reduces root mean squared error (RMSE), for all parameters. In a number of cases, the reduction is more than 50%. The "Baseline" results are more biased than the prior, for most parameters. This is not surprising, because the prior only has bias because of the enforcement of parameter bounds. However, the bias for α, ρ and λ_1 is perhaps larger than we would like to see. Comparing to the ABC estimator that uses all available statistics ("All" in the table), the selection procedure achieves a reduction, in some cases considerable, of RMSE, for most parameters. It is perhaps somewhat surprising how well the estimator that uses all statistics performs. This may be due to two factors: first, the set of candidate statistics may have been chosen well, in that most or all statistics are in fact are informative, and secondly, Algorithm 3, which is used for the ABC estimation, is successful in concentrating on an area of high posterior mass, even when the conditioning statistic has high dimension. Finally, the column labeled "Random" give results for an ABC estimator that uses 17 randomly selected statistics (this is

Tab. 3: Monte Carlo results, Stochastic volatility model

		Prior		Baseline		All		Random	
Param.	True	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
μ_0	-0.021	0.012	0.052	-0.007	0.026	-0.015	0.023	-0.022	0.029
μ_1	-0.013	0.000	0.036	-0.017	0.027	-0.021	0.039	-0.013	0.031
α	0.785	-0.003	0.357	0.103	0.176	0.136	0.232	0.227	0.303
κ	0.036	0.003	0.021	0.007	0.008	0.007	0.013	0.009	0.011
σ	0.204	0.000	0.048	-0.010	0.025	-0.007	0.029	0.014	0.028
ρ	-0.713	0.014	0.147	0.027	0.089	-0.046	0.117	0.023	0.132
λ_0	0.016	0.000	0.008	0.006	0.006	0.004	0.005	0.004	0.005
λ_1	1.152	0.000	0.331	0.074	0.126	0.054	0.130	0.088	0.142
μ_J	-0.002	0.000	0.012	-0.001	0.003	-0.001	0.003	0.000	0.003
σ_J	1.472	0.000	0.072	0.014	0.023	0.002	0.018	-0.005	0.025

the same number of statistics as was selected for the "Baseline" results). We note that this estimator has considerably larger RMSE than does the estimator that uses the "Baseline" set, which is indicative that the selection procedure does in fact succeed in choosing informative statistics out of a candidate set. The "Random" results do give a reduction in RMSE, compared to the prior. This is not surprising, because the candidate set was chosen to include statistics that one would hope are informative. Also, one should note that the results for the "Random" set are only indicative, because they depend on the specific set of 17 statistics that were drawn in a single case. The main point is that the selection procedure performs considerably better than the random selection.

The ABC estimator using the "Baseline" set of statistics exhibits some biases that we might like to try to address. It is possible to modify the criterion that the selection procedure minimizes to focus on the mean absolute error (MAE) of one or more parameters, rather than on all of the parameters. The default criterion minimizes the scaled MAE, averaged over all of the parameters: the loss function L in equation (3.5) is set to $L(\theta) = \frac{1}{\dim\theta}\sum_{j=1}^{\dim\theta}\frac{1}{\sigma_j}\left|\theta_j\right|$. If we are interested in selecting statistics to fit well a certain parameter, for example, the fourth parameter, without regard to the other parameters, we can set the loss function to $L(\theta) = \frac{1}{\sigma_4}\left|\theta_4\right|$. This targets only the selected parameter. It may also be useful to introduce the penalty term $p(\delta)$ in equation (3.5), by setting a in equation to a number greater than zero, in order to find a small set of statistics that are most informative for the chosen parameter. We did this for the parameters α, ρ and λ_1, each in turn, setting $a = 0.05$. For α, the selected statistics were 5 and 48. For ρ, the selected statistics were 30, 34 and 55. For λ_1, only statistic 25 was selected. Statistics 25, 30, 34 and 55 were already selected in the baseline set, so for ρ and λ_1, the targeted procedure did not find new statistics to add to the baseline set. For α, however, new statistics were selected when using targeting. Table 4 presents ABC Monte Carlo results when statistics 5 and 48 are added to the baseline set, in order to target α. In this case, the bias of the ABC estimator drops from 0.103 ("Baseline" results in Table 3) to 0.018, in Table 4. RMSE also drops slightly. For the other parameters, bias and RMSE are not affected in any sys-

Tab. 4: Monte Carlo results, Stochastic volatility model, adding statistics targeted to α

Param.	True	Bias	RMSE
μ_0	-0.021	-0.010	0.030
μ_1	-0.013	-0.016	0.023
α	0.785	0.018	0.174
κ	0.036	0.009	0.011
σ	0.204	-0.000	0.030
ρ	-0.713	0.015	0.098
λ_0	0.016	0.006	0.006
λ_1	1.152	0.079	0.124
μ_J	-0.002	-0.000	0.003
σ_J	1.472	0.013	0.023

tematic way by the inclusion of the two additional statistics. Some variation is expected, due to the fact that only 100 Monte Carlo replications were used. This exercise illustrates the fact that the selection procedure can be performed for all parameters jointly, which is the normal usage of most general interest, and it can be targeted to certain parameters, if these parameters are of special interest, or if one has reason to believe that the overall selection procedure may not have found statistics that are sufficiently informative for these parameters. The final set of statistics used for ABC estimation can be determined using both sources of information.

6 Conclusion

This paper presents a method of selection of statistics for ABC (and simulation-based GMM estimators) that is conceptually simple and easy to implement. It is designed to select statistics for accurate point inference, throughout the support of the prior (or on a parameter space). While one might hope that this would lead to an accurate posterior, no claims are made that this is the case. The method relies on the widely accepted criterion of minimizing mean absolute error (or a similar measure, such as root mean squared error), possibly with a penalty to encourage parsimonious selection. The criterion to be minimized is directly related to a point estimator associated with the ABC or GMM estimator of interest - there is no projection or dimension reduction step that intervenes. The criterion must be minimized over a discrete search space that has many elements. Simulated annealing, which is a widely used and well known tool, was used to address this problem. In principle, any global minimization algorithm that operates on a discrete choice set could be used in place of simulated annealing. Simulated annealing was chosen simply because it is well known in the economics literature and related fields, and because it works well for this application. It is a simple matter to run multiple simulated annealing minimizations in parallel, to speed up the search for a set of statistics that minimize the criterion.

It is to be emphasized that other researchers who wish to use the method for their own

ABC research can very easily make use of the provided software. The only requirement is that the user provides a set of parameters, drawn from the prior, and the corresponding candidate statistics, $\{(\theta^s, W^s)\}$. This is easy to do, as it is an essential part of any ABC method that is based on statistics. The software provided will then select Z, the subset of W, according to the default criterion, or the criterion that the user specifies. This takes very little effort on the part of the user.

The method has been tested using a simple example, for which theory tells us which statistics should be selected, and it is found to be able to select these statistics with good accuracy. In spite of the simplicity of the example, it is important to note that the model contains 6 parameters, and a candidate set of statistics with 35 elements. The stochastic volatility example presents a model with 10 parameters and 56 candidate statistics. In this case, theory does not tell us what are the optimal statistics. However, the Monte Carlo results show that the selection method results in precise estimation, compared to results using different sets of statistics. This second example also shows that the proposed methods can be applied when the model is complex, with numerous parameters and a large set of candidate statistics.

The criterion function that is minimized focuses on minimizing the integrated Bayes expected loss of simulated posterior mean. As explained in Section 4, it is a simple matter to change the criterion function in order to focus on other features of the posterior distribution, including the full distribution, and still use the simulated annealing algorithm for minimization. In order to do so, one would simply edit the function "select_obj" which is near the top of the file "SelectionAlgorithm.jl" and replace the least-squares criterion by the KLIC criterion developed in Section 4. The performance of this alternative selection method is left for future research.

`[insert online link to SelectionAlgorithm.jl accompanying text file about here]`

7 Acknowledgements

Kristensen acknowledges research support by the Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation, the ESRC through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001 and the European Research Council (grant no. ERC-2012-StG 312474). These sponsoring agencies had no role in the specific research conducted.

References

[1] Andersen, T. G., T. Bollerslev and F. X. Diebold, 2007, "Roughing it up: including jump components in the measurement, modeling and forecasting of return volatility", *Review of Economics and Statistics* 89, 701–720.

[2] Andrews, D.W.K. and B. Lu, 2001, "Consistent model and moment selection

procedures for GMM estimation with application to dynamic panel data models, *Journal of Econometrics*, 101, 123-164.

[3] Beaumont, M., W. Zhang and D. Balding, 2002, "Approximate Bayesian computation in population genetics", *Genetics*, 162, 2025-2035.

[4] Beaumont, M.A., J.-M. Cornuet, J.-M. Marin, and C.P. Robert, 2009, "Adaptive approximate Bayesian computation", *Biometrika,* 96, 983-990.

[5] Blum, M.G.B., M. A. Nunes, D. Prangle and S. A. Sisson, 2013, "A comparative review of dimension reduction methods in approximate Bayesian computation", *Statistical Science*, 28, 189–208.

[6] Barnes, C.P, S. Filippi, M.P.H. Stumpf and T. Thorne, 2012, "Considerate approaches for constructing summary statistics for ABC model selection", *Statistics and Computing*, 22, 1181-1197.

[7] Černý, V. (1985), "Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm", *Journal of Optimization Theory and Applications,* 45, 41-51.

[8] Cheng, Xu and Z. Liao, 2012, "Select the valid and relevant moments: one-step procedure for GMM with many moments", PIER Working paper 12-045.

[9] Chernozhukov, V. and H. Hong, 2003, "An MCMC approach to classical estimation", *Journal of Econometrics* 115, 293-346.

[10] Creel, M. and D. Kristensen, 2011 (unpublished), "Indirect likelihood inference", Dynare Working Papers (CEPREMAP), Working Paper 8.

[11] Creel, M. and D. Kristensen, 2013 (unpublished), "Indirect likelihood inference (revised)," Working Paper 931, UFAE and IAE Working Papers.

[12] Creel, M. and D. Kristensen, 2015, "ABC of SV: limited information likelihood inference in stochastic volatility jump-diffusion models", *Journal of Empirical Finance*, 31, 85-108. doi:10.1016/j.jempfin.2015.01.002.

[13] 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.

[14] Drovandi, C.C. and A.N. Pettitt, 2011, "Estimation of parameters for macroparasite population evolution using approximate Bayesian Computation", *Biometrics*, 67, 225-233.

[15] 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, 419–474.

[16] Forneron, J.-J., and S. Ng (2015), "The ABC of Simulation Estimation with Auxiliary Statistics," working paper, Columbia University.

[17] Gao, J. and H. Hong (2015), "A Computational Implementation of GMM," working paper, Stanford University.

[18] Gallant, A. R. and G. Tauchen, 1996, "Which moments to match?" *Econometric Theory* 12, 657-681.

[19] Goffe, W.L., G. D. Ferrier and J. Rogers (1994), "Global optimization of statistical functions with simulated annealing", *Journal of Econometrics,* **60**, 65-99.

[20] Gouriéroux, C., A. Monfort, and E. Renault, 1993, "Indirect inference," *Journal of Applied Econometrics*, 8, S85-S118.

[21] Hall, A. R., A. Inoue, J. M. Nason, and B. Rossi, 2012, "Information criteria for impulse response function matching estimation of DSGE Models," *Journal of Econometrics*, 170, 499-518.

[22] Hall, P., J. Racine and Q. Li , 2004, "Cross-Validation and the Estimation of Conditional Probability Densities", *Journal of the American Statistical Association*, 99, 1015-1026.

[23] Hansen, L. P., 1982, "Large sample properties of generalized method of moments estimators," *Econometrica,* 50, 1029-1054.

[24] Härdle, W., P. Hall and J.S. Marron, 1988, "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" *Journal of the American Statistical Association*, 83, 86- 95.

[25] Joyce, P. and P. Marjoram, 2008, "Approximately sufficient statistics and Bayesian computation", *Statistical Applications in Genetics and Molecular Biology*, 7, Art. 26.

[26] McFadden, D., 1989, "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, 57, 995–1026.

[27] Marin, J.-M., P. Rudlo, C. Robert and R. Ryder, 2012, "Approximate Bayesian computational methods, *Statistics and Computing*, 22, 1167-1180.

[28] Marjoram, P., J. Molitor, V. Plagonal, V. and S. Tavaré, 2003, "Markov chain Monte Carlo without likelihoods", *Proceedings of the National Academy of Science*, 100, 15324- 15328.

[29] Marron, J.S. and D. Nolan, 1989,"Canonical kernels for density estimation", *Statistics and Probability Letters*, 7, 195-199.

[30] 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, Art. 34.

[31] Örkcü, H., 2013, "Subset selection in multiple linear regression models: A hybrid of genetic and simulated annealing algorithms", *Applied Mathematics and Computation*, 219, 11018-11028.

[32] Sisson, S., Y. Fan, and M. Tanaka, 2007, "Sequential Monte Carlo without likelihoods", *Proceedings of the National Academy of Science*, 104, 1760-1765.

[33] Smith, A., 1993, "Estimating nonlinear time series models using simulated vector autoregressions," *Journal of Applied Econometrics*, 8, S63-S84.

[34] Stock, J., J. Wright and M. Yogo, 2002, "A survey of weak instruments and weak identification in generalized method of moments", *Journal of Business & Economic Statistics*, 20, 518-529.

[35] Stone, M., 1974, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society, Series B*, 36, 111-147.

[36] Tauchen, G., 1986, "Statistical properties of generalized method-of-moments estimators of structural parameters obtained from financial market data", *Journal of Business & Economic Statistics*, 4, 397-416.

[37] Wu, L. and Y. Wang, 1998, "An introduction to simulated annealing algorithms for the computation of economic equilibrium", *Computational Economics*, 12, 151-169.