A Likelihood-Free Reverse Sampler of the Posterior Distribution

Jean-Jacques Forneron[*] Serena Ng[†]

October 2015

Abstract

This paper considers properties of an optimization based sampler for targeting the posterior distribution when the likelihood is intractable. It uses auxiliary statistics to summarize information in the data and does not directly evaluate the likelihood associated with the specified parametric model. Our reverse sampler approximates the desired posterior distribution by first solving a sequence of simulated minimum distance problems. The solutions are then re-weighted by an importance ratio that depends on the prior and the volume of the Jacobian matrix. By a change of variable argument, the output are draws from the desired posterior distribution. Optimization always results in acceptable draws. Hence when the minimum distance problem is not too difficult to solve, combining importance sampling with optimization can be much faster than the method of Approximate Bayesian Computation that by-passes optimization.

JEL Classification: C22, C23.

Keywords: approximate Bayesian Computation, Indirect Inference, Importance Sampling.

[*] Department of Economics, Columbia University, 420 W. 118 St., New York, NY 10025. Email: jmf2209@columbia.edu

[†] Department of Economics, Columbia University, 420 W. 118 St. Room 1117, New York, NY 10025. Email Serena.Ng at Columbia.edu. Financial support is provided by the National Science Foundation, SES-0962431. We thank Christopher Drovandi, Neil Shephard, and two anonymous referees for many helpful comments. The second author would like to thank Aman Ullah for his support and guidance.

1 Introduction

Maximum likelihood estimation rests on the ability of a researcher to express the joint density of the data, or the likelihood, as a function of K unknown parameters $\boldsymbol{\theta}$. Inference can be conducted using classical distributional theory once the mode of the likelihood function is determined by numerical optimization. Bayesian estimation combines the likelihood with a prior to form the posterior distribution from which the mean and other quantities of interest can be computed. Though the posterior distribution may not always be tractable, it can be approximated by Monte Carlo methods provided that the likelihood is available. When the likelihood is intractable but there exists $L \geq K$ auxiliary statistics $\widehat{\boldsymbol{\psi}}$ with model analog $\boldsymbol{\psi}(\boldsymbol{\theta})$ that is analytically tractable, one can still estimate $\boldsymbol{\theta}$ by minimizing the difference between $\widehat{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}(\boldsymbol{\theta})$.

Increasingly, parametric models are so complex that neither the likelihood nor $\boldsymbol{\psi}(\boldsymbol{\theta})$ is tractable. But if the model is easy to simulate, the mapping $\boldsymbol{\psi}(\boldsymbol{\theta})$ can be approximated by simulations. Estimators that exploit this idea can broadly be classified into two types. One is simulated minimum distance estimator (SMD), a frequentist approach that is quite widely used in economic analysis. The other is the method of Approximate Bayesian Computation that is popular in other disciplines. This method, ABC for short, approximates the posterior distribution using auxiliary statistics $\widehat{\boldsymbol{\psi}}$ instead of the full dataset y. It takes draws of $\boldsymbol{\theta}$ from a prior distribution and keeps the draws that, when used to simulate the model, produces auxiliary statistics that are close to the sample estimates $\widehat{\boldsymbol{\psi}}$. Both the ABC and SMD can be regarded as likelihood free estimators in the sense that the likelihood that corresponds to the structural model of interest is not directly evaluated.

While both the SMD and ABC exploit auxiliary statistics to perform likelihood free estimation, there are important differences between them. The SMD solves for the $\boldsymbol{\theta}$ that makes $\widehat{\boldsymbol{\psi}}$ close to the average of $\boldsymbol{\psi}(\boldsymbol{\theta})$ over many simulated paths of the data. In contrast, the ABC evaluates $\boldsymbol{\psi}(\boldsymbol{\theta})$ for each draw from the prior and accepts the draw only if $\boldsymbol{\psi}(\boldsymbol{\theta})$ is close to $\widehat{\boldsymbol{\psi}}$. The ABC estimate is the average over the accepted draws, which is the posterior mean. In Forneron and Ng (2014), we focused on the case of exact identification and used a reverse sampler (RS) to better understand the difference between the two approaches. The RS approximates the posterior distribution by solving a sequence of SMD problems, each using only one simulated path of data. Using stochastic expansions as in Rilstone et al. (1996) and Bao and Ullah (2007), we reported that in the special case when $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ (i.e the auxiliary model is the assumed model), the SMD has an unambiguous bias advantage over the ABC. But in more general settings, the ABC can, by clever choice of prior, eliminate biases that are inherent in the SMD.

In this paper, we extend the analysis to over-identified models and provide a deeper understanding of the reverse sampler. The RS is shown to be an optimization-based importance sampler that transforms the density from draws of $\boldsymbol{\psi}$ to draws of $\boldsymbol{\theta}$ so that when multiplied by the prior and properly weighted, the draws follow the desired posterior distribution. Section 2 considers the exactly identified case and shows that the importance ratio is the determinant of the Jacobian matrix. Section 3 considers the over-identified case when the dimension of $\boldsymbol{\psi}(\boldsymbol{\theta})$ exceeds that of $\boldsymbol{\theta}$. Because of the need to transform densities of different dimensions, the determinant of the Jacobian matrix is replaced by its volume. Using analytically tractable models, we show that the RS exactly reproduces the desired posterior distribution.

The RS was initially developed as a framework to better understand the different approaches to likelihood free estimation. While not intended to compete with existing implementations of ABC, the use of optimization in RS turns out to have a property that is of independent interest. Creating a long sequence of ABC draws such that the simulated statistic $\widehat{\boldsymbol{\psi}}^b$ and the data $\widehat{\boldsymbol{\psi}}$ deviate by no more than δ can take infinite time if δ is set to exactly zero as theory suggests. This has generated interests within the ABC community to control for δ. The RS by-passes this problem because SMD estimation makes $\widehat{\boldsymbol{\psi}}^b$ as close to $\widehat{\boldsymbol{\psi}}^b$ as machine precision permits. We elaborate on this feature in Section 4. Of course, the RS is useful only when the SMD objective function is well behaved and easy to optimize, which may not always be the case. But allowing optimization to play a role in ABC can be useful, as independent work by Meeds and Welling (2015) also found.

1.1 Preliminaries

In what follows, we use a 'hat' to denote estimators that correspond to the mode (or extremum estimators) and a 'bar' for estimators that correspond to the posterior mean. We use (s, S) and (b, B) to denote the (specific, total number of) draws in frequentist and Bayesian type analyses respectively. A superscript s denotes a specific draw and a subscript S denotes the average over S draws. These parameters S and B have different roles. The SMD uses S simulations to approximate the mapping $\boldsymbol{\psi}(\boldsymbol{\theta})$, while the ABC uses B simulations to approximate the posterior distribution of the infeasible likelihood.

We assume that the data $\mathbf{y} = (y_1, \ldots, y_T)'$ have finite fourth moments and can be represented by a parametric model with probability measure $\mathcal{P}_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^K$, $\boldsymbol{\theta}_0$ is the true value. The likelihood $L(\boldsymbol{\theta}|\mathbf{y})$ is intractable. Estimation of $\boldsymbol{\theta}$ is based on $L \geq K$ auxiliary statistics $\widehat{\boldsymbol{\psi}}(\mathbf{y}(\boldsymbol{\theta}_0))$ which we simply denote by $\widehat{\boldsymbol{\psi}}$ when the context is clear. The model implies statistics $\boldsymbol{\psi}(\boldsymbol{\theta})$. The classical minimum distance estimator is

$$\widehat{\boldsymbol{\theta}}_{\text{CMD}} = \text{argmin}_{\boldsymbol{\theta}} J(\widehat{\boldsymbol{\psi}}, \boldsymbol{\psi}(\boldsymbol{\theta})) = \overline{g}(\boldsymbol{\theta})' \mathrm{W} \overline{g}(\boldsymbol{\theta}), \quad \overline{g}(\boldsymbol{\theta}) = \widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}(\boldsymbol{\theta}).$$

Assumption A :

i There exists a unique interior point $\boldsymbol{\theta}_0 \in \Theta$ (compact) that minimizes the population objective function $(\boldsymbol{\psi}(\boldsymbol{\theta}_0) - \boldsymbol{\psi}(\boldsymbol{\theta}))'\mathrm{W}(\boldsymbol{\psi}(\boldsymbol{\theta}_0) - \boldsymbol{\psi}(\boldsymbol{\theta}))$. The mapping $\boldsymbol{\theta} \to \boldsymbol{\psi}(\boldsymbol{\theta}) = \lim_{T\to\infty} \mathbb{E}[\widehat{\boldsymbol{\psi}}(\boldsymbol{\theta})]$ is continuously differentiable and injective. The $L \times K$ Jacobian matrix $\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ has full column rank, and the rank is constant in the neighborhood of $\boldsymbol{\theta}_0$.

ii There is an estimator $\widehat{\boldsymbol{\psi}}$ such that $\sqrt{T}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}(\boldsymbol{\theta}_0)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$.

iii W is a $L \times L$ positive definite matrix and $\mathrm{W}\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$ has rank K.

Assumption A ensures global identification and consistent estimation of $\boldsymbol{\theta}$, see Newey and McFadden (1994). In Gourieroux et al. (1993), the mapping $\boldsymbol{\psi} : \boldsymbol{\theta} \to \boldsymbol{\psi}(\boldsymbol{\theta})$ is referred to as the binding function while in Jiang and Turnbull (2004), $\boldsymbol{\psi}(\boldsymbol{\theta})$ is referred to as a bridge function. When $\boldsymbol{\psi}(\boldsymbol{\theta})$ is analytically intractable, the simulated minimum distance estimator (SMD) is

$$\widehat{\boldsymbol{\theta}}_{\mathrm{SMD}} \quad = \quad \mathrm{argmin}_{\boldsymbol{\theta}} J_S(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\psi}}_S(\boldsymbol{\theta})) = \mathrm{argmin}_{\boldsymbol{\theta}} \overline{g}_S(\boldsymbol{\theta})'\mathrm{W}\overline{g}_S(\boldsymbol{\theta}). \tag{1}$$

where $S \geq 1$ is the number of simulations,

$$\overline{g}_S(\boldsymbol{\theta}) = \widehat{\boldsymbol{\psi}} - \frac{1}{S}\sum_{s=1}^{S} \widehat{\boldsymbol{\psi}}^s(\mathbf{y}^s(\boldsymbol{\theta})).$$

Notably, the term $\mathbb{E}[\widehat{\boldsymbol{\psi}}(\boldsymbol{\theta})]$ in CMD estimation is approximated by $\frac{1}{S}\sum_{s=1}^{S} \widehat{\boldsymbol{\psi}}^s(\mathbf{y}^s(\boldsymbol{\theta}))$. The SMD was first used in Smith (1993). Different SMD estimators can be obtained by suitable choice of the moments $\overline{g}(\boldsymbol{\theta})$, including the indirect inference estimator of Gourieroux et al. (1993), the simulated method of moments of Duffie and Singleton (1993), and the efficient method of moments of Gallant and Tauchen (1996).

The first ABC algorithm was implemented by Tavare et al. (1997) and Pritchard et al. (1996) to study population genetics. They draw $\boldsymbol{\theta}^b$ from the prior distribution $\pi(\boldsymbol{\theta})$, simulate the model under $\boldsymbol{\theta}^b$ to obtain data $\mathbf{y}^b$, and accept $\boldsymbol{\theta}^b$ if the vector of auxiliary statistics $\boldsymbol{\psi}(\boldsymbol{\theta}^b)$ deviates from $\widehat{\boldsymbol{\psi}}$ by no more than a tuning parameter δ. If $\widehat{\boldsymbol{\psi}}$ are sufficient statistics and $\delta = 0$, the procedure produces samples from the true posterior distribution if $B \to \infty$.

The Accept-Reject ABC: For $b = 1, \ldots, B$

i Draw $\boldsymbol{\vartheta}$ from $\pi(\boldsymbol{\theta})$ and $\boldsymbol{\varepsilon}^b$ from an assumed distribution $F_{\boldsymbol{\varepsilon}}$

ii Generate $\mathbf{y}^b(\boldsymbol{\varepsilon}^b, \boldsymbol{\vartheta})$ and $\widehat{\boldsymbol{\psi}}^b = \boldsymbol{\psi}(\mathbf{y}^b)$.

iii Accept $\boldsymbol{\theta}^b = \boldsymbol{\vartheta}$ if $J_1^b = \left(\widehat{\boldsymbol{\psi}}^b - \widehat{\boldsymbol{\psi}}\right)' \mathrm{W} \left(\widehat{\boldsymbol{\psi}}^b - \widehat{\boldsymbol{\psi}}\right) \leq \delta$.

The accept-reject method (hereafter, AR-ABC) simply keeps those draws from the prior distribution $\pi(\boldsymbol{\theta})$ that produce auxiliary statistics which are close to the observed $\widehat{\boldsymbol{\psi}}$. As it is not easy to choose δ a priori, it is common in AR-ABC to fix a desired quantile q, repeat the steps $[B/q]$ times. Setting δ to the q-th quantile of the sequence of J_1^b that will produce exactly B draws is analogous to the idea of keeping $k-$nearest neighbors considered in Gao and Hong (2014).

Since simulating from a non-informative prior distribution is inefficient, the accept-reject sampler can be replaced by one that targets at features of the posterior distribution. There are many ways to target the posterior distribution. We consider the MCMC implementation of ABC proposed in Marjoram et al. (2003) (hereafter, MCMC-ABC).

The MCMC-ABC: For $b = 1, \dots, B$ with $\boldsymbol{\theta}^0$ given and proposal density $q(\cdot|\boldsymbol{\theta}^b)$,

i Generate $\boldsymbol{\vartheta} \sim q(\boldsymbol{\vartheta}|\boldsymbol{\theta}^b)$

ii Draw errors $\boldsymbol{\varepsilon}^{b+1}$ from $F_{\boldsymbol{\varepsilon}}$ and simulate data $\mathbf{y}^{\mathbf{b+1}}(\boldsymbol{\varepsilon}^{b+1}, \boldsymbol{\vartheta})$. Compute $\widehat{\boldsymbol{\psi}}^{b+1} = \boldsymbol{\psi}(\mathbf{y}^{b+1})$.

iii Set $\boldsymbol{\theta}^{b+1}$ to $\boldsymbol{\vartheta}$ with probability $\rho_{ABC}(\boldsymbol{\theta}^b, \boldsymbol{\vartheta})$ and to $\boldsymbol{\theta}^{b+1}$ with probability $1 - \rho_{ABC}(\boldsymbol{\theta}^b, \boldsymbol{\vartheta})$ where

$$\rho_{ABC}(\boldsymbol{\theta}^b, \boldsymbol{\vartheta}) = \min\left(\mathbb{I}_{\|\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\psi}}^{b+1}\| \leq \delta} \frac{\pi(\boldsymbol{\vartheta}) q(\boldsymbol{\theta}^b|\boldsymbol{\vartheta})}{\pi(\boldsymbol{\theta}^b) q(\boldsymbol{\vartheta}|\boldsymbol{\theta}^b)}, 1\right) \tag{2}$$

The AR and MCMC both produce an approximation to the posterior distribution of $\boldsymbol{\theta}$. It is common to use the posterior mean of the draws $\overline{\boldsymbol{\theta}} = \frac{1}{B}\sum_{b=1}^B \boldsymbol{\theta}^b$ as the ABC estimate. The MCMC-ABC uses a proposal distribution to account for features of the data so that it is less likely to have proposed values with low posterior probability. The tuning parameter δ affects the bias of the estimates. Too small a δ may require making many draws which can be computationally costly.

The ABC samples from the joint distribution of $(\boldsymbol{\theta}^b, \boldsymbol{\psi}^b(\boldsymbol{\varepsilon}^b, \boldsymbol{\theta}^b))$ and then integrates out $\boldsymbol{\varepsilon}^b$. The posterior distribution is thus

$$p(\boldsymbol{\theta}^b|\widehat{\boldsymbol{\psi}}) \propto \int p(\boldsymbol{\theta}^b, \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)|\widehat{\boldsymbol{\psi}}) \mathbb{I}_{\|\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^b)\| < \delta} d\boldsymbol{\varepsilon}^b.$$

The indicator function (also the rectangular kernel) equals one if $\|\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^b\|$ does not exceed δ. The ABC draws are dependent due to the Markov nature of the MCMC-ABC sampler.

Both the SMD and ABC assume that simulations provide an accurate approximation of $\boldsymbol{\psi}(\boldsymbol{\theta})$ and that auxiliary statistics are chosen to permit identification of $\boldsymbol{\theta}$. Creel and Kristensen (2015) suggests a cross-validation method for selecting the auxiliary statistics. For the same choice of $\widehat{\boldsymbol{\psi}}$, the SMD finds the $\boldsymbol{\theta}$ that makes the average of the simulated auxiliary statistics close to $\widehat{\boldsymbol{\psi}}$. The ABC takes the average of $\boldsymbol{\theta}^b$, drawn from the prior, with the property that each $\boldsymbol{\psi}^b$ is close to $\widehat{\boldsymbol{\psi}}$. In an attempt to understand this difference, Forneron and Ng (2014), takes as starting point that each $\boldsymbol{\theta}^b$ in the above ABC algorithm can be reformulated as an SMD problem with $S = 1$. We consider an algorithm that solves the SMD problem many times to obtain a distribution for $\boldsymbol{\theta}^b$, each time using one simulated path. The sampler terminates with an evaluation of the prior probability, in contrast to the ABC which starts with a draw from the prior distribution. Hence we call our algorithm a reverse sampler (hereafter, RS). The RS produces a sequence of $\boldsymbol{\theta}^b$ that are independent optimizers and do not have a Markov structure.

In the next two sections, we explore additional features of the RS. As an overview, the distribution of draws that emerge from SMD estimation with $S = 1$ may not be from the desired posterior distribution. Hence the draws are re-weighted to target the posterior. In the exactly identified case, $\widehat{\boldsymbol{\psi}}^b$ can be made exactly equal to $\widehat{\boldsymbol{\psi}}$ by choosing the SMD estimate as $\boldsymbol{\theta}^b$. Thus the RS is simply an optimization based importance sampler using the determinant of Jacobian matrix as importance ratio. In the over-identified case, the volume of the (rectangular) Jacobian matrix is used in place of the determinant. Additional weighting is given to those $\widehat{\boldsymbol{\theta}}^b$ that yields $\widehat{\boldsymbol{\psi}}^b$ sufficiently close to $\widehat{\boldsymbol{\psi}}$.

2 The Reverse Sampler: Case $K = L$

The algorithm for the case of exact identification is as follows. For $b = 1, \dots, B$

i Generate $\boldsymbol{\varepsilon}^b$ from $F_{\boldsymbol{\varepsilon}}$.

ii Find $\boldsymbol{\theta}^b = \text{argmin}_{\boldsymbol{\theta}} J_1^b(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b), \widehat{\boldsymbol{\psi}})$ and let $\widehat{\boldsymbol{\psi}}^b = \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)$.

iii Set $w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) = \pi(\boldsymbol{\theta}^b)|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)|^{-1}$.

iv Re-weigh the $\boldsymbol{\theta}^b$ by $\frac{w(\boldsymbol{\theta}^b)}{\sum_{b=1}^B w(\boldsymbol{\theta}^b)}$.

Like the ABC, the draws $\boldsymbol{\theta}^b$ provides an estimate of the posterior distribution of $\boldsymbol{\theta}$ from which an estimate of the posterior mean:

$$\overline{\boldsymbol{\theta}}_{RS} = \sum_{b=1}^B \frac{w(\boldsymbol{\theta}^b)}{\sum_{b=1}^B w(\boldsymbol{\theta}^b)} \boldsymbol{\theta}^b$$

can be used as an estimate of $\boldsymbol{\theta}$. Each $\boldsymbol{\theta}^b$ is a function of the data $\widehat{\boldsymbol{\psi}}$ and the draws $\boldsymbol{\varepsilon}^b$ that minimizes $J_1^b(\boldsymbol{\psi}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b), \widehat{\boldsymbol{\psi}})$. The K first-order conditions are given by

$$\mathcal{F}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}) = \frac{\partial \overline{g}_1(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\theta}}' \mathrm{W} \overline{g}_1(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}) = 0 \tag{3}$$

where $\frac{\partial \overline{g}_1(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\theta}}$ is the $L \times K$ matrix of derivatives with respect to $\boldsymbol{\theta}$ evaluated at the arguments. It is assumed that, for all b, this derivative matrix has full column rank K. For SMD estimation, $\frac{\partial \overline{g}_1(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\theta}} = \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})$. This Jacobian matrix plays an important role in the RS.

The importance density denoted $h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}})$ is obtained by drawing $\boldsymbol{\varepsilon}^b$ from the assumed distribution $F_{\boldsymbol{\varepsilon}}$ and finding $\boldsymbol{\theta}^b$ such that $J(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b), \widehat{\boldsymbol{\psi}})$ is smaller than a pre-specified tolerance.

When $K = L$, this tolerance can be made arbitrarily small so that up to numerical precision, $\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) = \widehat{\boldsymbol{\psi}}$. This density $h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}})$ is related to $p_{\widehat{\boldsymbol{\psi}}^b,\boldsymbol{\varepsilon}^b}(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)) \equiv p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b)$ by a change of variable:

$$h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}}) = p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}}) \cdot |\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)|.$$

Now $p(\boldsymbol{\theta}^b, \widehat{\boldsymbol{\psi}}^b|\widehat{\boldsymbol{\psi}}) \propto p(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta}^b, \widehat{\boldsymbol{\psi}}^b)p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b|\boldsymbol{\theta}^b)\pi(\boldsymbol{\theta}^b)$ and $p(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta}^b, \widehat{\boldsymbol{\psi}}^b)$ is constant since $\widehat{\boldsymbol{\psi}}^b = \widehat{\boldsymbol{\psi}}$. Hence

$$\begin{aligned}
p(\boldsymbol{\theta}^b|\widehat{\boldsymbol{\psi}}) &\propto \int \pi(\boldsymbol{\theta}^b)p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}})\mathbb{I}_{\|\widehat{\boldsymbol{\psi}}-\widehat{\boldsymbol{\psi}}^b\|=0}d\boldsymbol{\varepsilon}^b \\
&= \int \pi(\boldsymbol{\theta}^b)|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|^{-1}h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}})\mathbb{I}_{\|\widehat{\boldsymbol{\psi}}-\widehat{\boldsymbol{\psi}}^b\|=0}d\boldsymbol{\varepsilon}^b \\
&= \int w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}})d\boldsymbol{\varepsilon}^b
\end{aligned}$$

where the weights are, assuming invertibility of the determinant:

$$w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) = \pi(\boldsymbol{\theta}^b)|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|^{-1}. \tag{4}$$

Note that in general, $\frac{w(\boldsymbol{\theta}^b)}{\sum_{b=1}^B w(\boldsymbol{\theta}^b)} \neq \frac{1}{B}$.

In the above, we have used the fact that $\mathbb{I}_{\|\widehat{\boldsymbol{\psi}}-\widehat{\boldsymbol{\psi}}^b\|=0}$ is 1 with probability one when $K = L$. The Jacobian of the transformation appears in the weights because the draws $\boldsymbol{\theta}^b$ are related to the likelihood via a change of variable. Hence a crucial aspect of the RS is that it re-weighs the draws of $\boldsymbol{\theta}^b$ from $h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon})$. Put differently, the unweighted draws will not, in general, follow the target posterior distribution.

Consider a weighted sample $(\boldsymbol{\theta}^b, w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}))$ with $w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)$ defined in (4). The following proposition shows that as $B \to \infty$, RS produces the posterior distribution associated with the infeasible likelihood, which is also the ABC posterior distribution with $\delta = 0$.

Proposition 1 *Suppose that* $\widehat{\boldsymbol{\psi}}^b : \boldsymbol{\theta} \to \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b)$ *is one-to-one and the determinant* $|\frac{\partial\psi^b(\boldsymbol{\theta},\boldsymbol{\varepsilon}^b,\widehat{\boldsymbol{\psi}})}{\partial\boldsymbol{\theta}}| = |\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|$ *is bounded away from zero around* $\boldsymbol{\theta}^b$*. For any measurable function* $\varphi(\boldsymbol{\theta})$ *such that* $\mathbb{E}_{p(\boldsymbol{\theta}|\widehat{\boldsymbol{\psi}})}(\varphi(\boldsymbol{\theta})) = \int \varphi(\boldsymbol{\theta})p(\boldsymbol{\theta}|\widehat{\boldsymbol{\psi}})d\boldsymbol{\theta}$ *exists, then*

$$\frac{\sum_b^B w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)\varphi(\boldsymbol{\theta}^b)}{\sum_b^B w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)} \xrightarrow{a.s.} \mathbb{E}_{p(\boldsymbol{\theta}|\widehat{\boldsymbol{\psi}})}(\varphi(\boldsymbol{\theta})).$$

Convergence to the target distribution follows from a strong law of large numbers. Fixing the event $\widehat{\boldsymbol{\psi}}^b = \widehat{\boldsymbol{\psi}}$ is crucial to this convergence result. To see why, consider first the numerator:

$$\begin{aligned}\frac{1}{B}\sum_b w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)\varphi(\boldsymbol{\theta}^b) &\xrightarrow{a.s.} \iint \varphi\left(\boldsymbol{\theta}\right) w\left(\boldsymbol{\theta}, \boldsymbol{\varepsilon}\right) p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b|\boldsymbol{\theta})|\widehat{\boldsymbol{\psi}}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}, \widehat{\boldsymbol{\psi}})| d\boldsymbol{\varepsilon}^b d\boldsymbol{\theta} \\ &= \iint \varphi\left(\boldsymbol{\theta}\right) \left|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}, \widehat{\boldsymbol{\psi}})\right|^{-1} \pi(\boldsymbol{\theta}) p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b|\boldsymbol{\theta}) \left|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}, \widehat{\boldsymbol{\psi}})\right| d\boldsymbol{\varepsilon}^b d\boldsymbol{\theta} \\ &= \iint \varphi\left(\boldsymbol{\theta}\right) \pi(\boldsymbol{\theta}) p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}|\boldsymbol{\theta}) d\boldsymbol{\varepsilon} d\boldsymbol{\theta} \\ &= \iint \varphi\left(\boldsymbol{\theta}\right) \pi(\boldsymbol{\theta}) p(\widehat{\boldsymbol{\psi}}, \boldsymbol{\varepsilon}|\boldsymbol{\theta}) d\boldsymbol{\varepsilon} d\boldsymbol{\theta} \\ &= \int \varphi\left(\boldsymbol{\theta}\right) \pi(\boldsymbol{\theta}) L(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta}) d\boldsymbol{\theta}.\end{aligned}$$

Furthermore, the denominator converges to the integrating constant since $\frac{1}{B}\sum_b w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}) \xrightarrow{a.s.} \int \pi(\boldsymbol{\theta}) L(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta}) d\boldsymbol{\theta}$. Proposition 1 implies that the weighted average of $\boldsymbol{\theta}^b$ converges to the posterior mean. Furthermore, the posterior quantiles produced by the reverse sampler tends to those of the infeasible posterior distribution $p(\boldsymbol{\theta}|\widehat{\boldsymbol{\psi}})$ as $B \to \infty$. As discussed in Forneron and Ng (2014), the ABC can be presented as an importance sampler. Hence the accept-reject algorithm in Tavare et al. (1997) and Pritchard et al. (1996), as well as the Sequential Monte-Carlo approach to ABC in Sisson et al. (2007); Toni et al. (2009) and Beaumont et al. (2009) are all important samplers. The RS differs in that it is optimization based. It is also developed independently in Meeds and Welling (2015).

We now use examples to illustrate how the RS works in the exactly identified case.

Example 1: Suppose we have one observation $y \sim \mathcal{N}\left(\theta, 1\right)$ or $y = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 1)$. The prior for θ is $\theta \sim \mathcal{N}\left(0, 1\right)$. By drawing, $\theta^b, \varepsilon^b \sim \mathcal{N}(0, 1)$, we obtain $y^b = \theta^b + \varepsilon^b \sim \mathcal{N}(0, 2)$. The ABC keeps $\theta^b|y^b = y$. Since (θ^b, y^b) are jointly normal with covariance of 1, we deduce that $\theta^b|y^b = y \sim \mathcal{N}(y/2, 1/2)$. The exact posterior distribution for θ is $\mathcal{N}(y/2, 1/2)$.

The RS draws $\varepsilon^b \sim \mathcal{N}\left(0, 1\right)$ and computes $\theta^b = y - \varepsilon^b$ which is $\mathcal{N}(y, 1)$ conditional on y. The Jacobian of the transformation is 1. Re-weighting according to the prior, we have:

$$\begin{aligned}p_{\text{RS}}(\theta|y) \propto \phi(\theta)\phi(\theta - y) \propto \exp\left(-\tfrac{1}{2}\left(\theta^2 + (\theta - y)^2\right)\right) \propto \exp\left(-\frac{1}{2}\left(2\theta^2 - 2\theta y\right)\right) \\ \propto \exp\left(-\frac{2}{2}\left(\theta - y/2\right)^2\right).\end{aligned}$$

This is the exact posterior distribution as derived above.

Example 2 Suppose $y = Q(u, \theta), \varepsilon \sim \mathcal{U}_{[0,1]}$ and Q is a quantile function that is invertible and differentiable in both arguments.[1] For a single draw, y is a sufficient statistic. The likelihood-

[1] We thank Neil Shephard for suggesting the example.

based posterior is:

$$p(\theta|y) \propto \pi(\theta) f(y|\theta).$$

The RS simulates $y^b(\theta) = Q(\varepsilon^b|\theta)$ and sets $Q(\varepsilon^b|\theta^b) = y$. Or, in terms of the CDF:

$$\varepsilon^b = F(y|\theta^b)$$

Consider a small perturbation to y holding u^b fixed:

$$0 \;=\; dy\frac{dF(y|\theta^b)}{dy} + d\theta^b\frac{dF(y|\theta^b)}{d\theta^b} = dyF'_y(y|\theta^b) + d\theta^b F'_{\theta^b}(y|\theta^b).$$

In the above, $f \equiv F'_y(\cdot)$ is the density of y given θ. The Jacobian is:

$$\left|\frac{d\theta^b}{dy}\right| = \left|\frac{F'_y(y|\theta^b)}{F'_{\theta^b}(y|\theta^b)}\right| = \left|\frac{f(y|\theta^b)}{F'_{\theta^b}(y|\theta^b)}\right|.$$

To find the distribution of θ^b conditional on y, assume $F(y,.)$ is increasing in θ:

$$\begin{aligned}\mathbb{P}\left(\theta^b \le t|y\right) &= \mathbb{P}\Big(F(y|\theta^b) \le F(y|t)|y\Big)\\ &= \mathbb{P}\left(\varepsilon^b \le F(y|t)|y\right)\\ &= F(y|t).\end{aligned}$$

By construction, $f(\theta|y) = F'_\theta(y|\theta)$.[2] Putting things together,[3]

$$p_{\text{RS}}(\theta|y) \propto \pi(\theta)|F'_\theta(y|\theta)|\left|\frac{f(y|\theta)}{F'_\theta(y|\theta)}\right| = \pi(\theta)f(y|\theta) \propto p(\theta|y).$$

Example 3: Normal Mean and Variance We now consider an example in which the estimators can be derived analytically, and given in Forneron and Ng (2014). We assume $y_t = \boldsymbol{\varepsilon}_t \sim N(m, \sigma^2)$. The parameters of the model are $\boldsymbol{\theta} = (m, \sigma^2)'$. We consider the auxiliary statistics: $\widehat{\boldsymbol{\psi}}(\mathbf{y})' = \begin{pmatrix}\overline{y} & \widehat{\sigma}^2\end{pmatrix}$. The parameters are exactly identified.

The MLE of $\boldsymbol{\theta}$ is

$$\widehat{m} = \frac{1}{T}\sum_{t=1}^{T} y_t, \qquad \widehat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \overline{y})^2.$$

We consider the prior $\pi(m,\sigma^2) = (\sigma^2)^{-\alpha}\mathbb{I}_{\sigma^2>0}$, $\alpha > 0$ so that the log posterior distribution is

$$\log p(\boldsymbol{\theta}|\widehat{m},\widehat{\sigma}^2) \propto \frac{-T}{2}\log(2\pi)\sigma^2 - \alpha\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - m)^2.$$

[2] If $F(y,\cdot)$ is decreasing in θ, we have $\mathbb{P}(\theta^b \le t|y) = 1 - F(y,t)$.

[3] An alternative derivation is to note that $t = \mathbb{P}\left(u \le t|y\right) = \mathbb{P}\left(u = F(y,\theta^b) \le t|y\right) = \mathbb{P}\left(\theta^b \le F^{-1}(y,t) = t'|y\right)$. Hence $f(\theta^b|y) = \frac{dt}{dt'} = \frac{1}{(F^{-1})'_\theta(y,t)} = F'_2(y,t)$ as above.

Since $\widehat{\boldsymbol{\psi}}(\mathbf{y})$ are sufficient statistics, the RS coincides with the likelihood-based Bayesian estimator, denoted B below. This is also the infeasible ABC estimator. We focus discussion on estimators for σ^2 which have more interesting properties. Under a uniform prior, we obtain

$$\begin{aligned}
\overline{\sigma}^2_B &= \widehat{\sigma}^2 \frac{T}{T-5} \\
\widehat{\sigma}^2_{\text{SMD}} &= \frac{\widehat{\sigma}^2}{\frac{1}{ST}\sum_{s=1}^S \sum_{t=1}^T (\boldsymbol{\varepsilon}_t^s - \overline{\boldsymbol{\varepsilon}}^s)^2} \\
\widehat{\sigma}^2_{RS} &= \sum_{b=1}^B \frac{\frac{\widehat{\sigma}^2}{[\sum_{t=1}^T (\boldsymbol{\varepsilon}_t^b - \overline{\boldsymbol{\varepsilon}}^b)^2/T]^2}}{\sum_{k=1}^B \frac{1}{\sum_{t=1}^T (\boldsymbol{\varepsilon}_t^k - \overline{\boldsymbol{\varepsilon}}^k)^2/T}}
\end{aligned}$$

In this example, the RS is also the ABC estimator with $\delta = 0$. It is straightforward to show that the bias reducing prior is $\alpha = 1$ and coincides with the SMD. Table 2 shows that the estimators are asymptotically equivalent but can differ for fixed T.

Table 1: Properties of the Estimators

Estimator	Prior	$\mathbf{E}[\widehat{\boldsymbol{\theta}}]$	Bias	Variance	MSE
$\widehat{\boldsymbol{\theta}}_{ML}$	-	$\sigma^2 \frac{T-1}{T}$	$-\frac{\sigma^2}{T}$	$2\sigma^4 \frac{T-1}{T^2}$	$2\sigma^4 \frac{2T-1}{2T^2}$
$\overline{\boldsymbol{\theta}}_B$	1	$\sigma^2 \frac{T-1}{T-5}$	$\frac{2\sigma^2}{T-5}$	$2\sigma^4 \frac{T-1}{(T-5)^2}$	$2\sigma^4 \frac{T+1}{(T-5)^2}$
$\overline{\boldsymbol{\theta}}_{RS}$	1	$\sigma^2 \frac{T-1}{T-5}$	$\frac{2\sigma^2}{T-5}$	$2\sigma^4 \frac{T-1}{(T-5)^2}$	$2\sigma^4 \frac{T+1}{(T-5)^2}$
$\widehat{\boldsymbol{\theta}}_{\text{SMD}}$	-	$\sigma^2 \frac{S(T-1)}{S(T-1)-2}$	$\frac{2\sigma^2}{S(T-1)-2}$	$2\sigma^4 \kappa_1 \frac{1}{T-1}$	$2\sigma^4 \frac{\kappa_1}{T-1} + \frac{4\sigma^4}{(S(T-1)-2))^2}$

where $\kappa_1(S,T) = \frac{(S(T-1))^2(T-1+S(T-1)-2)}{(S(T-1)-2)^2(S(T-1)-4)} > 1$, κ_1 tends to one as S tend to infinity.

To highlight the role of the Jacobian matrix in the RS, the top panel of Figure 2 plots the exact posterior distribution and the one obtained from the reverse sampler. They are indistinguishable. The bottom panel shows an incorrectly constructed reverse sampler that does not apply the Jacobian transformation. Notably, the two distributions are not the same. Re-weighting by the Jacobian matrix is crucial to targeting the desired posterior distribution.

Figure 1 presents the likelihood based posterior distribution, along with the likelihood free ones produced by ABC and the RS-JI (just identified) for one draw of the data. The ABC results are based on the accept-reject algorithm. The numerical results corroborate with the analytical ones: all the posterior distributions are very similar. The RS-JI posterior distribution is very close to the exact posterior distribution. Figure 1 also presents results for the over-identified case (denoted RS-OI) using two additional auxiliary statistics: $\widehat{\boldsymbol{\psi}} = (\overline{y}, \widehat{\sigma}^2_y, \widehat{\mu}_3/\widehat{\sigma}^2_y, \widehat{\mu}_4/\widehat{\sigma}^4_y)$ where $\mu_k = \mathbb{E}(y^k)$. The weight matrix is $\text{diag}(1,1,1/2,1/2)$. The posterior distribution is very close to RS-JI obtained for exact identification. We now explain how the posterior distribution for the over-identified case is obtained.

3 The RS: Case $L \geq K$:

The idea behind the RS is the same when we go from the case of exact to overidentification. The precise implementation is as follows. Let $\mathbb{K}_\delta(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\psi}}^b)$ be a kernel function and δ be a tolerance level such that $\mathbb{K}_0(\widehat{\boldsymbol{\psi}}, \boldsymbol{\psi}^b) = \mathbb{I}_{\|\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b\| = 0}$.

For $b = 1, \ldots, B$

i Generate $\boldsymbol{\varepsilon}^b$ from $F_{\boldsymbol{\varepsilon}}$.

ii Find $\boldsymbol{\theta}^b = \text{argmin}_{\boldsymbol{\theta}} J_1^b(\widehat{\boldsymbol{\psi}}^b, \widehat{\boldsymbol{\psi}})$ where $\widehat{\boldsymbol{\psi}}^b = \widehat{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b)$;

iii Set $w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) = \pi(\boldsymbol{\theta}^b)\text{vol}\left(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\right)^{-1} \mathbb{K}_\delta(J_1^b(\widehat{\boldsymbol{\psi}}^b, \widehat{\boldsymbol{\psi}}))$ where: $\text{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}) = \sqrt{\left|\widehat{\boldsymbol{\psi}}^{b\prime}_{\boldsymbol{\theta}} \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}\right|}$.

iv Re-weigh $\boldsymbol{\theta}^b$ by $\frac{w(\boldsymbol{\theta}^b)}{\sum_{b=1}^B w(\boldsymbol{\theta}^b)}$.

We now proceed to explain the two changes:- the use of volume in place of determinant in the importance ratio, and the need for $L - K$ dimensional kernel smoothing.

The usual change of variable formula evaluates the absolute value of the determinant of the Jacobian matrix when the matrix is square. The determinant then gives the infinitesimal dilatation of the volume element in passing from one set of variables to another. The main issue in the case of overidentification is that the determinant of a rectangular Jacobian matrix is not well defined. However, as shown in Ben-Israel (1999), the determinant can be replaced by the volume when transforming from sets of a higher dimension to a lower one.[4] For a $L \times K$ matrix A, its volume, denoted vol(A), is the product of the (non-zero) singular values of A:

$$\text{vol}(A) = \begin{cases} \sqrt{|A'A|} & L \geq K, \ \text{rank}(A) = K \\ \sqrt{|AA'|} & L \leq K, \ \text{rank}(A) = L. \end{cases}$$

Furthermore, if $A = BC$, $\text{vol}(A) = \text{vol}(B)\text{vol}(C)$.

To verify that our target distribution is unaffected by whether we calculate the volume or the determinant of the Jacobian matrix when $K = L$, observe that

$$\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b(\widehat{\boldsymbol{\psi}}), \boldsymbol{\varepsilon}^b) = \frac{\partial \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})}{\partial \widehat{\boldsymbol{\psi}}} \frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \boldsymbol{\theta}^b}. \tag{5}$$

The K first order conditions defined by (3) become:

$$\mathcal{F}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}) = \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})' \mathrm{W}\left(\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)\right) = 0. \tag{6}$$

[4] From Ben-Israel (2001), $\int_V f(v)dv = \int_U f(\phi(u))\text{vol}\Big(\phi_u(u)\Big)du$ for a real valued function f integrable on V. See also `http://www.encyclopediaofmath.org/index.php/Jacobian`.

Since $L = K$, W can be set to an identity matrix I_K. Furthermore, $\boldsymbol{\psi}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}) = \widehat{\boldsymbol{\psi}}$ since $J_1^b(\boldsymbol{\theta}^b) = 0$ under exact identification. As $\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}$ is a square matrix when $K = L$, we can directly use the fact that $\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})d\boldsymbol{\theta} + \mathcal{F}_{\psi}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})d\widehat{\boldsymbol{\psi}} = 0$ to obtain the required determinant:

$$|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|^{-1} = \mathrm{I}_K \cdot |\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}| = | - \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})^{-1} \mathcal{F}_{\widehat{\boldsymbol{\psi}}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|. \tag{7}$$

Now to use the volume result, put $\mathrm{A} = \mathrm{I}_K$, $\mathrm{B} = \frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}$ and $\mathrm{C} = \frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \boldsymbol{\theta}}$. But A is just a K-dimensional identity matrix. Hence $\mathrm{vol}(\mathrm{I}_K) = \mathrm{vol}\bigg(\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}\bigg)\mathrm{vol}\bigg(\frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \boldsymbol{\theta}}\bigg)$ which evaluates to

$$\mathrm{vol}\bigg(\frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \boldsymbol{\theta}}\bigg)^{-1} = \mathrm{vol}\bigg(\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}\bigg), \qquad \text{or} \qquad \bigg|\frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \boldsymbol{\theta}}\bigg|^{-1} = \bigg|\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}\bigg|$$

which is precisely $|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon})|^{-1}$ as given in (7)[5]. Hence in the exactly identified case, there is no difference whether one evaluates the determinant or the volume of the Jacobian matrix.

Next, we turn to the role of the kernel function $\mathbb{K}_\delta(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\psi}}^b)$. The joint density $h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)$ is related to $p_{\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b}(\widehat{\boldsymbol{\psi}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)) = p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b)$ through a change a variable now expressed in terms of volume:

$$h(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}}) = p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}}) \cdot \mathrm{vol}\left(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\right)$$

When $L \geq K$, the objective function $\|\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b\|_{\mathrm{W}} = J_1^b \geq 0$ measures the extent to which $\widehat{\boldsymbol{\psi}}$ deviates from $\widehat{\boldsymbol{\psi}}^b$ when the objective function at its minimum. Consider the thought experiment that $J_1^b = 0$ with probability 1, such as enabled by a particular draw of $\boldsymbol{\varepsilon}^b$. Then the arguments above for $K = L$ would have applied. We would still have $p(\boldsymbol{\theta}^b|\widehat{\boldsymbol{\psi}}) = \int \pi(\boldsymbol{\theta}^b) p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}}) \mathbb{I}_{\|\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b\| = 0} d\boldsymbol{\varepsilon}^b = \int w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b|\widehat{\boldsymbol{\psi}}) d\boldsymbol{\varepsilon}^b$, except that the weights are now defined in terms of volume. Proposition 1 would then extend to the case with $L \geq K$.

But in general $J_1^b \neq 0$ almost surely. Nonetheless, we can use only those draws that yield $J_1^b(\boldsymbol{\theta}^b)$ that are sufficiently close to zero. The more draws we make, the tighter this criterion can be. Suppose there is a symmetric kernel $\mathbb{K}_\delta(\cdot)$ satisfying conditions in Pagan and Ullah (1999, p.96) for consistent estimation of conditional moments non-parametrically. Analogous to Proposition 1, the volume $\mathrm{vol}\big(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\big)$ is assumed to be bounded away from zero. Then as the number of draws $B \to \infty$, the bandwidth $\delta(B) \to 0$ and $B\delta(B) \to \infty$ with

$$w_{\delta(B)}(\boldsymbol{\theta}^b, \widehat{\boldsymbol{\varepsilon}}^b) \quad = \quad \pi(\boldsymbol{\theta}^b)\mathrm{vol}\bigg(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\bigg)^{-1} \mathbb{K}_{\delta(B)}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\psi}}^b), \tag{8}$$

[5] Using the implicit function theorem to compute the gradient gives the same result. Since $\widehat{\boldsymbol{\psi}}^b = \widehat{\boldsymbol{\psi}}$ we have: $\mathcal{F}_{\boldsymbol{\theta}} = -\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})'\mathrm{W}\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}) + \sum_j \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}, \boldsymbol{\theta}_j}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) W \left(\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\right) = -\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})' W \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})$. Then $\mathrm{vol}(\mathcal{F}_{\boldsymbol{\theta}}^{-1} \mathcal{F}_{\widehat{\psi}}) = \mathrm{vol}(\mathcal{F}_{\boldsymbol{\theta}}^{-1}) \mathrm{vol}(\mathcal{F}_{\widehat{\psi}}) = \mathrm{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}))^{-1} |W|^{-1} \mathrm{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}))^{-1} \mathrm{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}))^{-1} |W| = \mathrm{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}))^{-1}$. Hence the weights are the same when we only consider the draws where $J_1^b = 0$ which are the draws we are interested in.

a result analogous to Proposition 1 can be obtained:

$$\begin{aligned}
&\frac{1}{B}\sum_b w_{\delta(B)}(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)\varphi(\boldsymbol{\theta}^b)\xrightarrow{p}\iint \varphi(\boldsymbol{\theta})w_0(\boldsymbol{\theta},\boldsymbol{\varepsilon})\text{vol}\Big(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\varepsilon}^b;\widehat{\boldsymbol{\psi}})\Big)p(\widehat{\boldsymbol{\psi}},\boldsymbol{\varepsilon}^b|\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\varepsilon}^b\\
&=\iint \varphi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathbb{1}_{\|\widehat{\boldsymbol{\psi}}-\widehat{\boldsymbol{\psi}}^b\|=0}\text{vol}\Big(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\varepsilon}^b;\widehat{\boldsymbol{\psi}})\Big)^{-1}p(\widehat{\boldsymbol{\psi}}^b,\boldsymbol{\varepsilon}^b|\boldsymbol{\theta})\text{vol}\Big(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\varepsilon}^b;\widehat{\boldsymbol{\psi}})\Big)d\boldsymbol{\theta}d\boldsymbol{\varepsilon}^b\\
&=\iint \varphi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})1_{\|\widehat{\boldsymbol{\psi}}-\widehat{\boldsymbol{\psi}}^b\|=0}p(\widehat{\boldsymbol{\psi}},\boldsymbol{\varepsilon}^b|\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\varepsilon}^b\\
&=\int \varphi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})L(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta})d\boldsymbol{\theta}.
\end{aligned}$$

Similarly, the integrating constant is consistent as $\frac{1}{B}\sum_b w_{\delta(B)}(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)\xrightarrow{p}\int\pi(\boldsymbol{\theta})L(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta})d\boldsymbol{\theta}$. Hence, the RS sampler still recovers the posterior distribution with the infeasible likelihood. Note that the kernel function was introduced for developing a result analogous to Proposition 1, but no kernel smoothing is required in practical implementation. What is needed for the RS in the over-identified case is B draws with sufficiently small $J_1(\boldsymbol{\theta}^b)$. Hence, we can borrow the idea used in the AR-ABC. Specifically, we fix a quantile q, repeat $[B/q]$ times until the desired number of draws is obtained. Discarding some draws seems necessary in many ABC implementations.

In summary, there are two changes in implementation of the RS in the over-identified case: the volume and the kernel function. Kernel smoothing has no role in the RS when $K = L$. It is interesting to note that while the ABC and RS both rely on the kernel $\mathbb{K}_\delta$ to keep draws close to $\widehat{\boldsymbol{\psi}}^b$ in the over-identified case, the non-parametric rate at which the sum converges to the integral are different. The RS uses the first order conditions $\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)'\mathrm{W}\left(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}\right)=0$ to indicate which K combinations of $\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}$ are set to zero, rendering the dimension of the smoothing problem $L-K$. To see this, note first that each draw $\boldsymbol{\theta}^b$ from the RS is consistent for $\boldsymbol{\theta}_0$ and asymptotically normal as shown in Forneron and Ng (2014). In consequence, the first order condition (FOC) can be re-written as: $\left(\frac{d\boldsymbol{\psi}(\boldsymbol{\theta})}{d\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}+O_p(\frac{1}{\sqrt{T}})\right)'\mathrm{W}\left(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}\right)=0$, or

$$\frac{d\boldsymbol{\psi}(\boldsymbol{\theta})}{d\boldsymbol{\theta}}|'_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\mathrm{W}\left(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}\right)=o_p(\frac{1}{\sqrt{T}}).$$

Since $\frac{d\boldsymbol{\psi}(\boldsymbol{\theta})}{d\boldsymbol{\theta}}|'_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\mathrm{W}$ is full rank, there exists a subspace of dimension K such that $\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}$ is zero asymptotically. Hence the kernel smoothing problem is effectively $L-K$ dimensional. The ABC does not use the FOC. Even in the exactly identified case, the kernel smoothing is a $L = K$ dimensional problem. In general, the convergence rate of the ABC is $L \geq K$, the dimension of $\widehat{\boldsymbol{\psi}}$.

The following two examples illustrate the properties of the ABC and RS posterior distributions. The first example uses sufficient statistics and the second example does not. Both the ABC and RS achieve the desired number of draws by setting the quantile, as discussed in Section 2.

Example 4: Exponential Distribution Let $y_1, \dots, y_T \sim \mathcal{E}(\theta), T = 5, \theta_0 = 1/2$. Now $\widehat{\psi} = \overline{y}$ is a sufficient statistic for $y_1, \dots, y_T$. For a flat prior $\pi(\theta) \propto 1_{\theta \geq 0}$ we have:

$$p(\theta|\overline{y}) \propto p(\theta|y_1, \dots, y_T) = \theta^T \exp(-\theta^T \overline{y}) \sim \Gamma(T+1, T\overline{y})$$

In the just identified case, we let $u_t^b \sim \mathcal{U}_{[0,1]}$ and $y_t^b = -\log(1-u_t^b)/\theta^b$. This gives:

$$\widehat{\psi}^b = \frac{1}{T}\sum_{t=1}^T y_t^b = -\frac{1}{T}\sum_{t=1}^T \frac{\log(1-u_t^b)}{\theta^b}.$$

Since $\overline{y}^b = \overline{y}$, the Jacobian matrix is:

$$\widehat{\psi}_b(\theta^b) = \frac{d\widehat{\psi}^b(\theta)}{d\theta}\bigg|_{\theta^b} = \frac{1}{T}\sum_{t=1}^T \frac{\log(1-u_t^b)}{[\theta^b]^2} = -\frac{\overline{y}}{\theta^b}.$$

Hence for a given T, the weights are: $w(\theta^b, u^b) \propto \mathbb{I}_{\boldsymbol{\theta}^b \geq 0} \frac{\theta^b}{\overline{y}^b} = \frac{\theta^b}{\overline{y}}$. We verified that the numerical results agree with this analytical result.

In the over identified case, we consider two moments:

$$\widehat{\boldsymbol{\psi}}^b = \begin{pmatrix} \overline{y}^b \\ \widehat{\sigma}_y^{b,2} \end{pmatrix} = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T y_t^b \\ \frac{1}{T}\sum_{t=1}^T (y_t^b)^2 - (\frac{1}{T}\sum_{t=1}^T y_t^b)^2 \end{pmatrix}.$$

Since $\frac{dy_t^b}{d\theta} = \frac{\log(1-u_t^b)}{(\theta^b)^2} = -\frac{y_t^b}{\theta^b}$. If $\delta = 0$, the Jacobian matrix is

$$\widehat{\boldsymbol{\psi}}_\theta^b = -\begin{pmatrix} \frac{1}{T}\sum_{t=1}^T \frac{y_t^b}{\theta} \\ \frac{2}{\theta^b}\frac{1}{T}\sum_{t=1}^T (y_t^b)^2 - \frac{2}{\theta^b}\left[\frac{1}{T}\sum_{t=1}^T y_t^b\right]^2 \end{pmatrix} = -\begin{pmatrix} \frac{\overline{y}}{\theta^b} \\ \frac{2(\widehat{\sigma}_y)^2}{\theta^b} \end{pmatrix}.$$

The volume to be computed is $\text{vol}(\widehat{\boldsymbol{\psi}}_\theta^b) = \sqrt{|\widehat{\boldsymbol{\psi}}_\theta^{b\prime}\widehat{\boldsymbol{\psi}}_\theta^b|}$, as stated in the algorithm. Even if W = I, the volume is the determinant of $\widehat{\boldsymbol{\psi}}_\theta^b$ in the exactly identified case, plus a term relating to the variance of y^b. We computed $\widehat{\boldsymbol{\psi}}_\theta^b$ for draws with $J_1^b \approx 0$ using numerical differentiation[6] and verified that the values are very close to the ones computed analytically for this example.

Figure 3 depicts a particular draw of the ABC posterior distribution (which coincides with the likelihood-based posterior since the statistics are sufficient), along with two generated by the RS sampler. The first one uses the sample mean as auxiliary statistic and hence is exactly identified. The second uses two auxiliary statistics: the sample mean and the sample variance. For the AR-ABC, we draw from the prior ten million times and keep the ten thousand nearest draws. This corresponds to a value of $\delta = 0.0135$. For the RS, we draw one million times[7] and

[6] In practice, since the mapping $\theta \to \widehat{\boldsymbol{\psi}}^b(\theta)$ is not known analytically, the derivatives are approximated using finite differences: $\partial_{\theta_j}\widehat{\boldsymbol{\psi}}^b(\theta) \simeq \frac{\widehat{\boldsymbol{\psi}}^b(\theta+e_j\varepsilon)-\widehat{\boldsymbol{\psi}}^b(\theta-e_j\varepsilon)}{2\varepsilon}$ for $\varepsilon \simeq 0$.

[7] This means that we solve the optimization problem one million times. Given that the optimization problem is one dimensional, the one dimensional R optimization routine *optimize* is used. It performs a combination of the golden section with parabolic interpolations. The optimum is found, up to a given tolerance level (the default is 10^{-4}), over the interval $[0, 10]$.

keep the ten thousand nearest draws which corresponds to a $\delta = 0.0001$. As for the weight matrix W, if we put $W_{11} > 0$ and zero elsewhere, we will recover the exactly identified distribution. Here, we intentionally put a positive weight on the variance (which is not a sufficient statistic) to check the effect on the posterior mean. With $\mathrm{W}_{11} = 1/5$ and $\mathrm{W}_{22} = 4/5$, the RS posterior means are 0.7452 and 0.7456 for the just and overidentifed cases. The corresponding values are are 0.7456 and .7474 for the exact posterior and the ABC-AR. They are very similar.

Example 5: ARMA(1,1): For $t = 1, \dots, T = 200$ and $\boldsymbol{\theta}_0 = (\alpha_0, \theta_0, \sigma_0) = (0.5, 0.5, 1.0)$, the data are generated as

$$y_t = \alpha y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \sigma^2).$$

Least squares estimation of the auxiliary model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + u_t$$

yields $L = 5 > K = 3$ auxiliary parameters

$$\widehat{\boldsymbol{\psi}} = (\widehat{\phi}_1, \widehat{\phi}_2, \widehat{\phi}_3, \widehat{\phi}_4, \widehat{\sigma}_u^2).$$

We let $\pi(\alpha, \theta, \sigma) = \mathbb{I}_{\alpha,\theta \in [-1,1], \sigma \geq 0}$ and $\mathrm{W} = \mathrm{I}_5$ which is inefficient. In this example, $\widehat{\boldsymbol{\psi}}$ are not sufficient statistics since y_t has an infinite order autoregressive representation.

We draw σ from a uniform distribution on $[0, 3]$ since $\mathcal{U}_{[0,\infty]}$ is not a proper density. The weights of the RS are obtained by numerical differentiation. The likelihood based posterior is computed by MCMC using the Kalman Filter with initial condition $\varepsilon_0 = 0$. As mentioned above, the desired number of draws is obtained by setting the quantile instead of setting the tolerance δ. For the RS, we keep the 1/10=10% closest draws corresponding to a $\delta = 0.0007$. The Sequential Monte-Carlo implementation of ABC (SMC-ABC) is more efficient at targeting the posterior than the ABC-AR. Hence we also compare the RS with SMC-ABC as implemented in the Easy-ABC package of Lenormand et al. (2013).[8] The requirement for 10,000 posterior draws are as follows:

	AR-ABC	SMC-ABC	RS	Likelihood
Computation Time (hours)	63	25	5	0.1
Effective number of draws	100,000,000	36,805,000	10,153,108	
δ	0.0132	0.0283	0.0007	

[8] We implemented the SMC-ABC in two ways. First, we use the procedure inVo et al. (2015) using code generously provided by Christopher Drovandi. We also use the Easy-ABC package in R of Lenormand et al. (2013). We thank an anonymous referee for this suggestion.

The difference, both in terms of computation time and number of model simulations, is notable. As shown in figure 4 the quality of the approximation is also different, especially for α and σ. The difference can be traced to δ. The δ used for the SMC-ABC is effectively much larger than for the RS. A better approximation requires a smaller δ which implies longer computational time. Alternatively stated, the acceptance rate at a low value of δ is very low. The caveat is that the speed gain is possible only if the optimization problem can be solved in a few iterations and reasonably fast. In practice, there will be a trade-off between the number of draws and the number of iterations in the optimization step as we further explore below.

4 Acceptance Rate

The RS was initially developed in Forneron and Ng (2014) as a framework to help understand frequentist (SMD) and the Bayesian (ABC) way of likelihood-free estimation. But it turns out that the RS has one computation advantage that is worth highlighting. The issue pertains to the low acceptance rate of the ABC.

As noted above, the ABC exactly recovers the posterior distribution associated with the infeasible likelihood if $\widehat{\boldsymbol{\psi}}$ are sufficient statistics and $\delta = 0$ as noted in Blum (2010). Of course, $\delta = 0$ is an event of measure zero, and the ABC has an approximation bias that depends on δ. In theory, a small δ is desired. The ABC needs a large number of draws to accurately approximate the posterior and can be computationally costly.

To illustrate this point, consider estimating the mean m in Example 3 with $\sigma^2 = 1$ assumed to be known, and $\pi(m) \propto 1$. All computations are based on the software package R. From a previous draw m^b, a random walk step gives $m^\star = m^b + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}(0,1)$. For small δ, we can assume $m^\star|\widehat{m} \sim \mathcal{N}(\widehat{m}, 1/T)$. From a simulated sample of T observations, we get an estimated mean $\widehat{m}^\star \sim \mathcal{N}(m^\star, 1/T)$. As is typical of MCMC chains, these draws are serially correlated. To see that the algorithm can be stuck for a long time if m^* is far from $\widehat{m}$, observe that the event $\widehat{m}^\star \in [\widehat{m} - \delta,\, \widehat{m} + \delta]$ occurs with probability

$$\mathbb{P}(\widehat{m}^\star \in [\widehat{m}-\delta,\, \widehat{m}+\delta]) = \Phi\left(\sqrt{T}(\widehat{m} + \delta - m^\star)\right) - \Phi\left(\sqrt{T}(\widehat{m} - \delta - m^\star)\right) \approx 2\sqrt{T}\delta\phi\left(\sqrt{T}(\widehat{m} - m^\star)\right).$$

The acceptance probability $\int_{m^*} \mathbb{P}(\widehat{m}^\star \in [\widehat{m} - \delta,\, \widehat{m} + \delta])dm^*$ is thus approximately linear in δ. To keep the number of accepted draws constant, we need to increase the number of draws as we decrease δ.

This result that the acceptance rate is linear in δ also applies in the general case. Assume that $\widehat{\boldsymbol{\psi}}^\star(\boldsymbol{\theta}^\star) \sim \mathcal{N}(\boldsymbol{\psi}(\boldsymbol{\theta}^\star), \Sigma/T)$. We keep the draw if $\|\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^\star(\boldsymbol{\theta}^\star)\| \leq \delta$. The probability of this event can be bounded above by $\sum_{j=1}^K \mathbb{P}\left(|\widehat{\boldsymbol{\psi}}_j - \widehat{\boldsymbol{\psi}}_j^\star(\boldsymbol{\theta}^\star)| \leq \delta\right)$ i.e.:

$$\sum_{j=1}^K \Phi\left(\frac{\sqrt{T}}{\sigma_j}\left(\widehat{\boldsymbol{\psi}}_j + \delta - \boldsymbol{\psi}_j(\boldsymbol{\theta}^\star)\right)\right) - \Phi\left(\frac{\sqrt{T}}{\sigma_j}\left(\widehat{\boldsymbol{\psi}}_j - \delta - \boldsymbol{\psi}_j(\boldsymbol{\theta}^\star)\right)\right) \approx 2\sqrt{T}\delta\sum_{j=1}^K \frac{\phi}{\sigma_j}\left(\frac{\sqrt{T}}{\sigma_j}\left(\widehat{\boldsymbol{\psi}}_j - \boldsymbol{\psi}_j(\boldsymbol{\theta}^\star)\right)\right).$$

The acceptance probability is still at best linear in δ. In general we need to increase the number of draws at least as much as δ declines to keep the number of accepted draws fixed.

Table 2: Acceptance Probability as a function of δ

δ	10	1	0.1	0.01	0.001
$\mathbb{P}(\lVert\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b\rVert_W \leq \delta)$	0.72171	0.16876	0.00182	0.00002	<0.00001

Table 2 shows the acceptance rate for Example 3 for $\boldsymbol{\theta}_0 = (m_0, \sigma_0^2) = (0, 2)$, $T = 20$, and weighting matrix $W = \text{diag}(\widehat{\sigma}^2, 2\widehat{\sigma}^4)/T, \pi(m, \sigma^2) \propto \mathbb{I}_{\sigma^2 \geq 0}$. The results confirm that for small values of δ, the acceptance rate is approximately linear in δ. Even though in theory, the targeted ABC posterior should be closer to the true posterior when δ is small, this may not be true in practice because of the poor properties of the MCMC chain. At least for this example, the MCMC chain with moderate value of δ provides a better approximation to the true posterior density.

To overcome the low acceptance rate issue, Beaumont et al. (2002) suggests to use local regression techniques to approximate $\delta = 0$ without setting it equal to zero. The convergence rate is then non-parametric. Gao and Hong (2014) analyzes the estimator of Creel and Kristensen (2013) and finds that to compensate for the large variance associated with the kernel smoothing, the number of simulations need to be larger than $T^{K/2}$ to achieve $\sqrt{T}$ convergence, where K is the number of regressors. Other methods that aim to increase the acceptance rate include the ABC-SMC algorithm of Sisson et al. (2007); Sisson and Fan (2011), as well as the adaptive weighting variant due to Bonassi and West (2015), referred to below as SMC-AW. These methods build a sequence of proposals to more efficiently target the posterior. The acceptance rate still declines rapidly with δ, however.

The RS circumvents this problem because each $\boldsymbol{\theta}^b$ is accepted by virtue of being the solution of an optimization problem, and hence $\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b)$ is the smallest possible. In fact, in the exactly identified case, $\delta = J_1^b = 0$. Furthermore, the sequence of optimizers are independent, and the sampler cannot be stuck. We use two more examples to highlight this feature.

Example 6: Mixture Distribution Consider the example in Sisson et al. (2007), also considered in Bonassi and West (2015). Let $\pi(\theta) \propto 1_{\theta \in [-10,10]}$ and

$$x|\theta \sim 1/2\mathcal{N}(\theta, 1) + 1/2\mathcal{N}(\theta, 1/100)$$

Suppose we observe one draw $x = 0$. Then the true posterior is $\theta|x \sim 1/2\mathcal{N}(0, 1) + 1/2\mathcal{N}(0, 1/100)$ truncated to $[-10, 10]$. As in Sisson et al. (2007) and Bonassi and West (2015), we choose three tolerance levels: $(2, 0.5, 0.025)$ for AR-ABC. Figure 5 shows that the ABC posterior distributions computed using accept-reject sampling with $\delta = 0.025$ are similar to the ones using SMC with and without adaptive weighting. The RS posterior distribution is close to both ABC-SMC and ABC-SMC-AW, and all similar to Figure 3 reported in Bonassi and West (2015). However, they are quite different from the AR-ABC with $\delta = 2$ and 0.5 are 2, showing that the choice of δ is important in ABC.

While the SMC, RS, and ABC-AR sampling schemes can produce similar posterior distributions, Table 3 shows that their computational time differ dramatically. The two SMC algorithms need to sample from a multinomial distribution which are evidently more time consuming. When $\delta = 0.25$, the AR-ABC posterior distribution is close to the ones produced by the SMC samplers and the RS, but the computational cost is still high. The AR-ABC is computationally efficient when δ is large, but as seen from Figure 5, the posterior distribution is quite poorly approximated. The RS takes 0.0017 seconds to solve, which is amazingly fast because for this example, the solution is available analytically. No optimization is involved, and there is no need to evaluate the Jacobian because the model is linear. Of course, in cases when the SMD problem is numerically challenging, numerical optimization can be time consuming as well. Our results nonetheless suggest a role for optimization in Bayesian computation; they need not be mutually exclusive. Combining the ideas is an interesting topic for future research.

Table 3: Computation Time (in seconds)

RS	ABC-AR			ABC-SMC	
	δ=2	δ=.5	δ=.025	Sisson et-al	Bonassi-West
.0017	0.4973	1.6353	33.8136	190.1510	199.1510

Example 7: Precautionary Savings The foregoing examples are simple and are serve illustrative purposes. We now consider an example that indeed has an infeasible likelihood. In Deaton (1991), agents maximize expected utility $\mathbb{E}_0\left(\sum_{t=0}^{\infty}\beta^t u(c_t)\right)$ subject to the constraint that assets $a_{t+1} = (1+r)(a_t + y_t - c_t)$ are bounded below by zero, where r is interest rate, y is income and c consumption. The desire for precautionary saving interacts with borrowing constraints to generate a policy function that is not everywhere concave, but is a piecewise linear when cash-on-hand is below an endogenous threshold. The policy function can only be solved numerically at assumed parameter values. SMD estimation thus consists of solving the model and simulating S auxiliary statistics at each guess $\boldsymbol{\theta}$. Michaelides and Ng (2000) evaluate the finite sample properties of several SMD estimators using a model with similar features. Since the likelihood for this model is not available analytically. Hence Bayesian estimation of this model has not been implemented. Here, we use the RS to approximate the posterior distribution.

We generate $T = 400$ observations assuming that $U(c) = \frac{c^{1-\gamma}-1}{1-\gamma}$, $y_t \sim$ iid $\mathcal{N}(\mu, \sigma^2)$ with $r = 0.05$, $\beta = 10/11$, $\mu = 100, \sigma = 10, \gamma = 2$ as true values. We estimate $\boldsymbol{\theta} = (\gamma, \mu, \sigma)$ and assume (β, r) are known. We use 10 auxiliary statistics:

$$\widehat{\boldsymbol{\psi}} = \Big(\overline{y} \quad \widehat{\Gamma}_{yy}(0) \quad \widehat{\Gamma}_{aa}(0) \quad \widehat{\Gamma}_{cc}(0) \quad \widehat{\Gamma}_{cc}(1) \quad \widehat{\Gamma}_{aa}(1) \quad \widehat{\Gamma}_{cc}(2) \quad \widehat{\Gamma}_{aa}(2) \quad \widehat{\Gamma}_{cy}(0) \quad \widehat{\Gamma}_{ay}(0)\Big)'$$

where $\widehat{\Gamma}_{ab}(j) = \frac{1}{T}\sum_{t=1}^{T}(a_t - \overline{a})(b_{t-j} - \overline{b})$. We generate $B = 13,423$ draws and keep the $3,356$ (25%) nearest draws to $\widehat{\boldsymbol{\psi}}$. After weighting using the volume of the Jacobian matrix we have an effective sample size of $1,421$ draws.[9] We use an identity weighting matrix so $J_{RS}(\boldsymbol{\theta}) = \overline{g}(\boldsymbol{\theta})'\overline{g}(\boldsymbol{\theta})$. The Jacobian is computed using finite differences for the RS. As benchmark, we also compute an SMD with $S = 100$, $J_S = \overline{g}_S(\boldsymbol{\theta})'\overline{g}_S(\boldsymbol{\theta})$. In this exercise, the SMD only needs to solve for the policy function once at each step of the optimization. Hence the binding function can be approximated using simulated data at a low cost. For this example, the programs are coded in PYTHON. The Nelder-Mead method is used for optimization.

Table 4: Deaton Model: RS, SMD with W = I

	Posterior Mean/Estimate			Posterior SD/SE		
	γ	μ	σ	γ	μ	σ
RS	1.86	99.92	10.48	0.19	0.84	0.37
SMD	1.76	99.38	10.31	0.12	0.60	0.34

Figure 6 shows the posterior distribution of the RS (blue) along with the SMD distribution (purple) as approximated by $\mathcal{N}(\widehat{\boldsymbol{\theta}}_{\text{SMD}}, \widehat{\text{V}}_{\text{SMD}}/T)$ according to asymptotic theory. Table 4 shows that the two sets of point estimates are similar. As explained in Forneron and Ng (2014), the SMD uses simulations to approximate the binding function while the RS (and by implication the ABC) uses simulations to approximate the infeasible posterior distribution. In this example, the difference in bias is quite small. We should note that the RS took well over a day to solve while the SMD took less than three hours to compute. Whether we use our own code for the ABC-MCMC or from packages available, the acceptance rate is too low for the exercise to be feasible.

5 Conclusion

This paper studies properties of the reverse sampler considered in Forneron and Ng (2014) for likelihood-free estimation. The sampler produce draws from the infeasible posterior distribution by solving a sequence of frequentist SMD problems. We showed that the reverse sampler uses the Jacobian matrix as importance ratio. In the over-identified case, the importance ratio can be computed using the volume of the Jacobian matrix. The reverse sampler does not suffer from the problem of low acceptance rate that makes the ABC computationally demanding.

[9] The effective sample size is computed as $1/\sum_{b=1}^{B} w_b^2$ where the weights satisfy $\sum_{b=1}^{B} w^b = 1$.

References

Bao, Y. and Ullah, A. 2007, The Second-Order Bias and Mean-Squared Error of Estimators in Time Series Models, *Journal of Econometrics* **140**(2), 650–669.

Beaumont, M., Corneut, J., Marin, J. and Robert, C. 2009, Adaptive Approximate Bayesian Computation, *Biometrika* **96**(4), 983–990.

Beaumont, M., Zhang, W. and Balding, D. 2002, Approximate Bayesian Computation in Population Genetics, *Genetics* **162**, 2025–2035.

Ben-Israel, A. 1999, The Change of Variables Forumla Using Matrix Volume, *SIAM Journal of Matrix Analysis* **21**, 300–312.

Ben-Israel, A. 2001, An Application of the Matrix Volume in Probability, *Linear Algebra and Applications* **321**, 9–25.

Blum, M. 2010, Approximate Bayesian Computation: A Nonparametric Perspective, *Journal of the American Statistical Association* **105**(491), 1178–1187.

Bonassi, F. and West, M. 2015, Sequential Monte Carlo With Adaptive Weights for Approximate Bayesian Computation, *Bayesian Analysis* **10**(1), 171–187.

Creel, M. and Kristensen, D. 2013, Indirect Likelihood Inference, *mimeo, UCL.*

Creel, M. and Kristensen, D. 2015, On Selection of Statistics for Approximate Bayesian Computing or the Method of Simulated Moments, *Computational Statistics and Data Analysis.*

Deaton, A. S. 1991, Saving and Liqudity Constraints, *Econometrica* **59**, 1221–1248.

Duffie, D. and Singleton, K. 1993, Simulated Moments Estimation of Markov Models of Asset Prices, *Econometrica* **61**, 929–952.

Forneron, J. J. and Ng, S. 2014, The ABC of Simulation Estimation with Auxiliary Statistics, arXiv:1501.01265.

Gallant, R. and Tauchen, G. 1996, Which Moments to Match, *Econometric Theory* **12**, 657–681.

Gao, J. and Hong, H. 2014, A Computational Implementation of GMM, SSRN Working Paper 2503199.

Gourieroux, C., Monfort, A. and Renault, E. 1993, Indirect Inference, *Journal of Applied Econometrics* **85**, 85–118.

Jiang, W. and Turnbull, B. 2004, The Indirect Method: Inference Based on Intermediate Statistics- A Synthesis and Examples, *Statistical Science* **19**(2), 239–263.

Lenormand, M., Jabot, F. and Deffuant, G. 2013, Adaptive Approximate Bayesian Computation Computation for Complex Models, *Journal of Computational and Graphical Statistics* **28**(6), 2777–2796.

Marjoram, P., Molitor, J., Plagnol, V. and Tavare, S. 2003, Markov Chain Monte Carlo Without Likelihoods, *Procedings of the National Academy of Science* **100**(26), 15324–15328.

Meeds, E. and Welling, M. 2015, Optimization Monte Carlo: Efficient and Embarrassingly Parallel Likelihood-Free Inference, arXiv:1506:03693v1.

Michaelides, A. and Ng, S. 2000, Estimating the Rational Expectations Model of Speculative Storage: A Monte Carlo Comparison of Three Simulation Estimators, *Journal of Econometrics* **96:2**, 231–266.

Newey, W. and McFadden, D. 1994, Large Sample Estimation and Hypothesis Testing, *Handbook of Econometrics*, Vol. 36:4, North Holland, pp. 2111–2234.

Pagan, A. and Ullah, A. 1999, *Nonparametric Econometrics*, Vol. Themes in Modern Econometrics, Cambridge University Press.

Pritchard, J., Seielstad, M., Perez-Lezman, A. and Feldman, M. 1996, Population Growth of Human Y chromosomes: A Study of Y Chromosome MicroSatellites, *Molecular Biology and Evolution* **16**(12), 1791–1798.

Rilstone, P., Srivastara, K. and Ullah, A. 1996, The Second-Order Bias and Mean Squared Error of Nonlinear Estimators, *Journal of Econometrics* **75**, 369–385.

Sisson, S. and Fan, Y. 2011, Likelihood Free Markov Chain Monte Carlo, *in* S. Brooks, A. Celman, G. Jones and X.-L. Meng (eds), *Handbook of Markov Chain Monte Carlo*, Vol. Chapter 12, pp. 313–335. arXiv:10001.2058v1.

Sisson, S., Fan, Y. and Tanaka, M. 2007, Sequential Monte Carlo Without Likelihood, *Procedings of the National Academy of Science* **104**(6), 1760–1765.

Smith, A. 1993, Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions, *Journal of Applied Econometrics* **8**, S63–S84.

Tavare, S., Balding, J., Griffiths, C. and Donnelly, P. 1997, Inferring Coalescence Times From DNA Sequence Data, *Genetics* **145**, 505–518.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. 2009, Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamic Models, *Journal of the Royal Society Inference* **6**, 187–222.

Vo, R., Drovandi, C., Pettitt, A. and Pettet, G. 2015, Melanoma Cell Colony Expansion Parameters Revealed by Approximate Bayesian Computation, eprints.qut.edu/au/83824.

Figure 1: Normally Distributed data

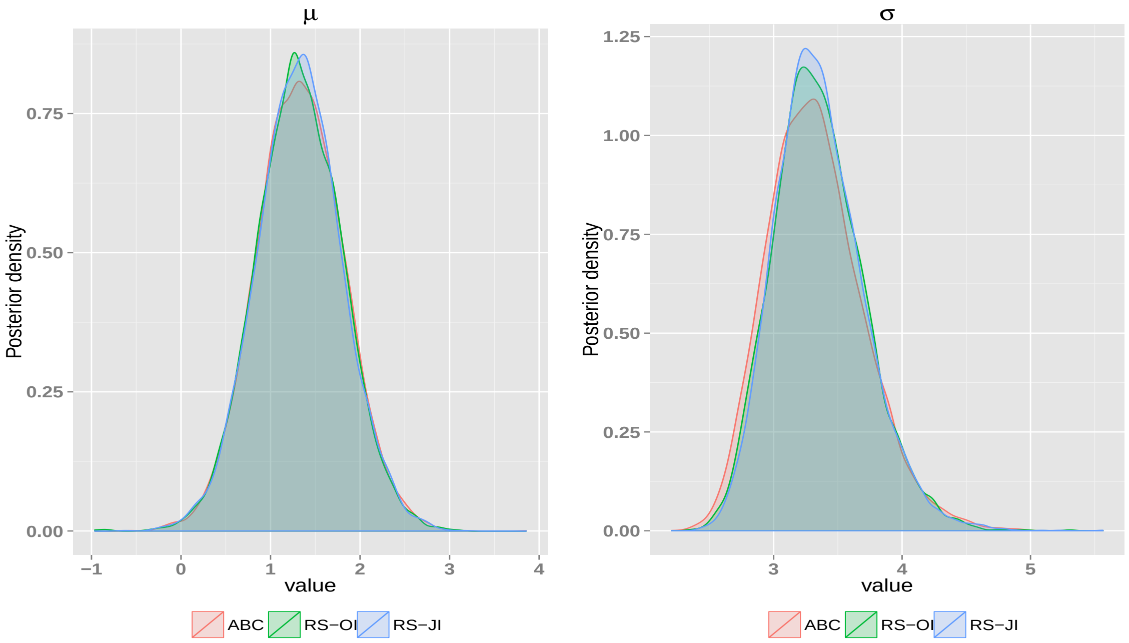


Figure 2: The Importance Weights in RS

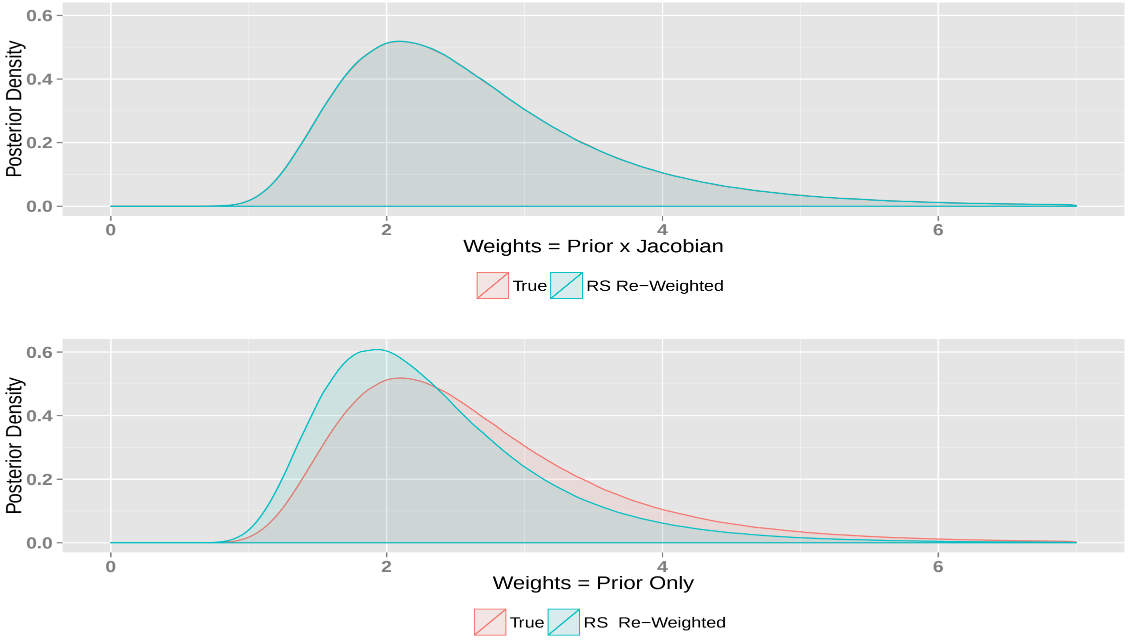


Figure 3: Exponential Distribution

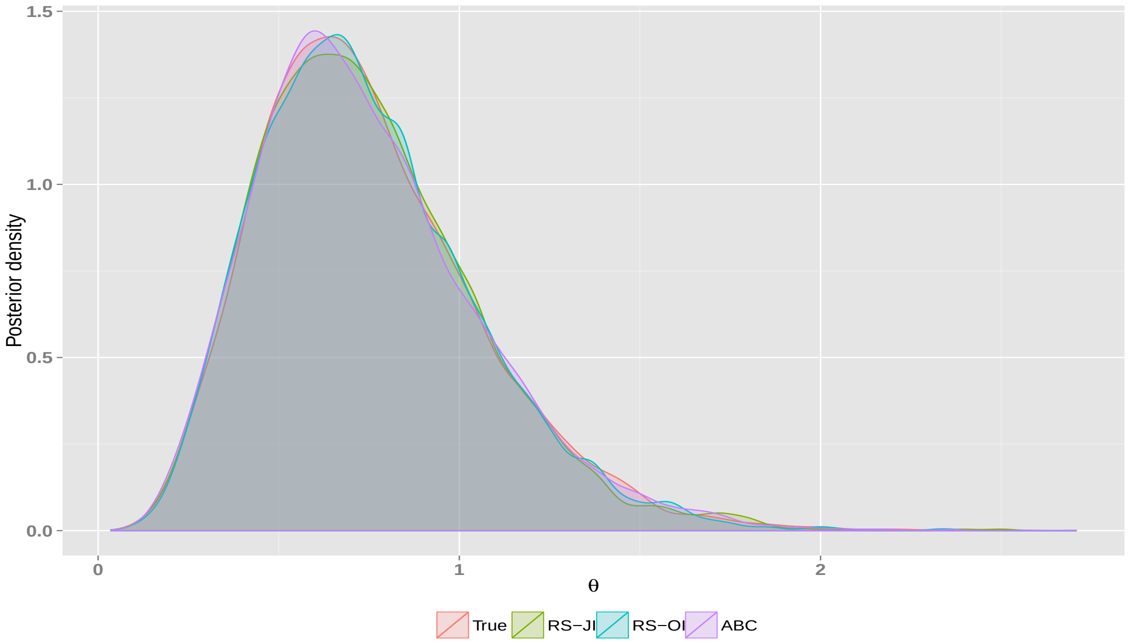


Figure 4: ARMA Model

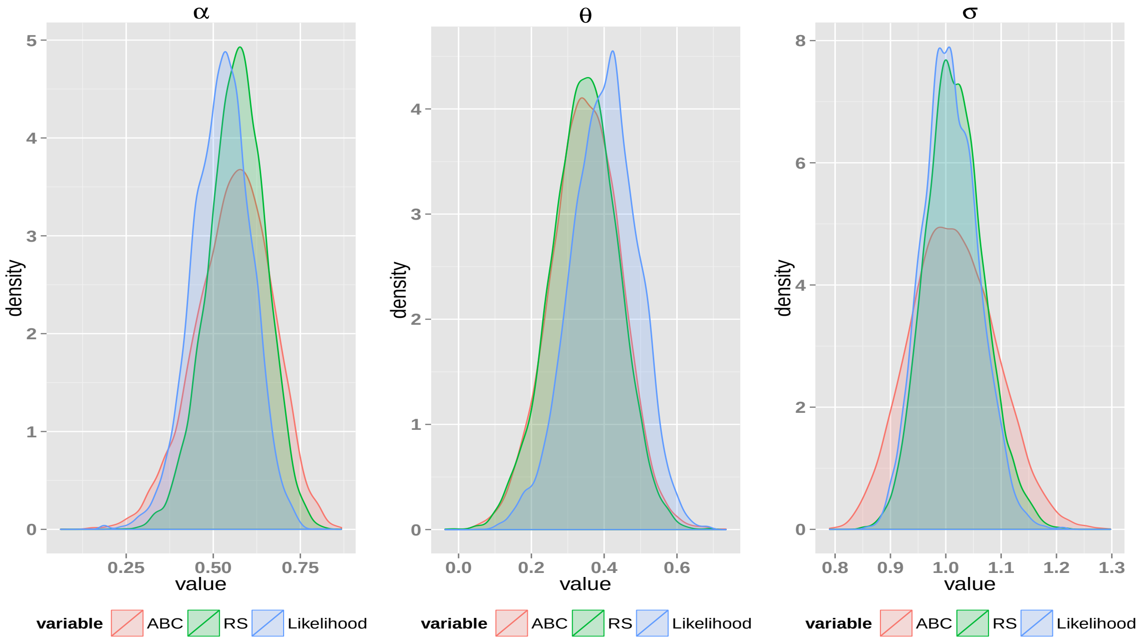


Figure 5: Mixture Distribution

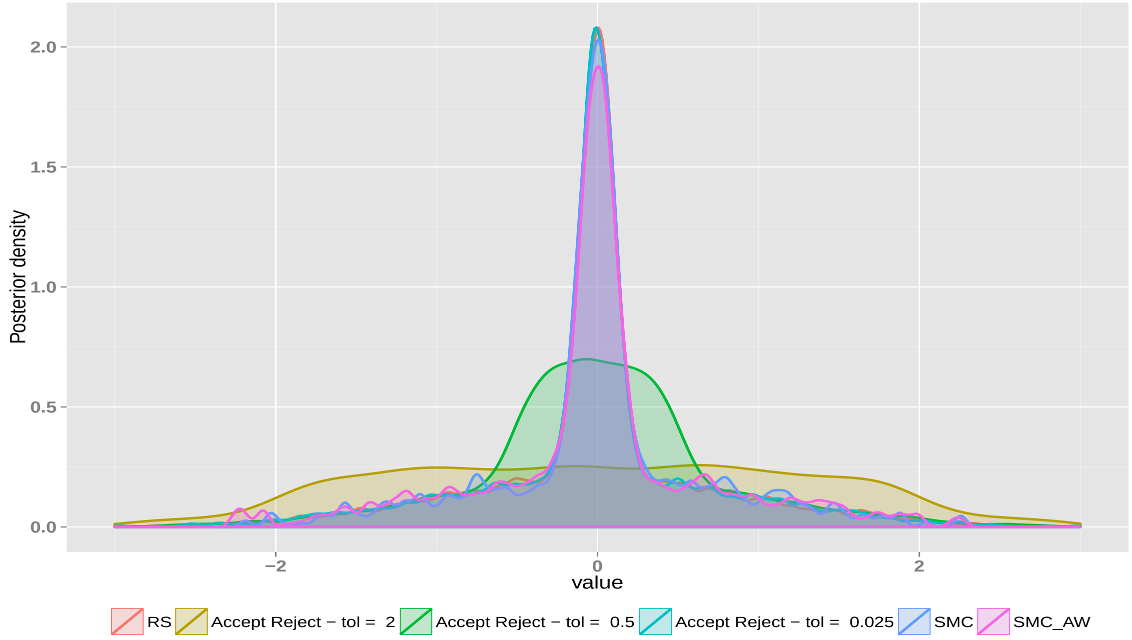
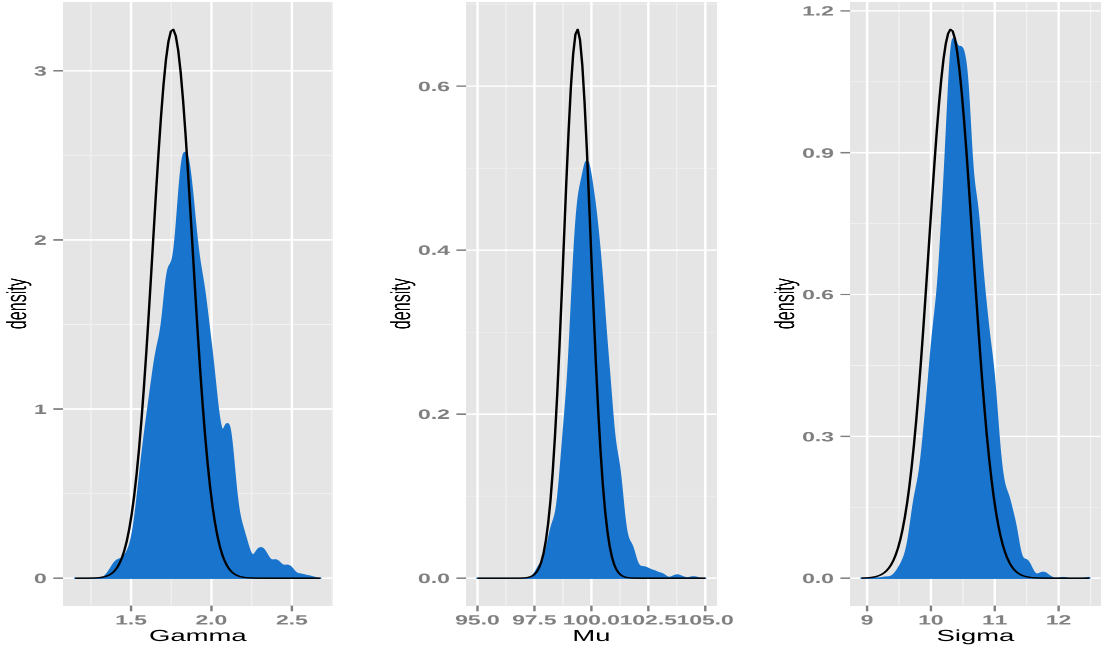


Figure 6: Deaton Model: RS and SMD



Note: Blue density: RS posterior, Black line: large sample approximation for the SMD estimator (identity weighting matrix).