MONASH University

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

On Consistency of Approximate Bayesian Computation

David T. Frazier, Gael M. Martin and Christian P. Robert

September 2015

On Consistency of Approximate Bayesian Computation[*]

David T. Frazier[†], Gael M. Martin[‡] and Christian P. Robert[§]

August 21, 2015

Abstract

Approximate Bayesian computation (ABC) methods have become increasingly prevalent of late, facilitating as they do the analysis of intractable, or challenging, statistical problems. With the initial focus being primarily on the practical import of ABC, exploration of its formal statistical properties has begun to attract more attention. The aim of this paper is to establish general conditions under which ABC methods are Bayesian consistent, in the sense of producing draws that yield a degenerate posterior distribution at the true parameter (vector) asymptotically (in the sample size). We derive conditions under which arbitrary summary statistics yield consistent inference in the Bayesian sense, with these conditions linked to identification of the true parameters. Using simple illustrative examples that have featured in the literature, we demonstrate that identification, and hence consistency, is unlikely to be achieved in many cases, and propose a simple diagnostic procedure that can indicate the presence of this problem. We also formally explore the link between consistency and the use of auxiliary models within ABC, and illustrate the subsequent results in the Lotka-Volterra predator-prey model.

Keywords: Bayesian consistency, likelihood-free methods, conditioning, auxiliary model-based ABC, ordinary differential equations, Lotka-Volterra model.

JEL Classification: C11, C15, C18

MSC2010 Subject Classification: 62F15, 62F12, 62C10

1   Introduction

The use of approximate Bayesian computation (ABC) methods in models with intractable likelihoods has gained increased momentum over recent years, extending beyond the original applications in the biological sciences. (See Marin *et al.,* 2011, and Sisson and Fan, 2011, for recent reviews.). Whilst ABC evolved initially as a practical tool, attention has begun to shift to the investigation of its formal statistical properties, in particular as they relate to the choice of summary statistics on which the technique typically relies; see for example, Fearnhead and Prangle (2012), Gleim and Pigorsch (2013), Marin *et al.* (2014), Martin *et al.* (2014) and Martin *et al.* (2014).

The aim of this paper is to establish general conditions under which summary statistic-based ABC methods are Bayesian consistent, in the sense of producing draws that yield a degenerate distribution at the true parameter (vector) in the (sample size) limit. This aim is much broader than that underlying Martin *et al.* (2014), in which standard quasi-likelihood conditions were invoked to establish the Bayesian consistency of auxiliary model-based versions of ABC. In particular, we derive the conditions under which *arbitrary* summary statistics yield consistent inference, with these conditions linked to the identification of the true parameters in any particular instance. Using simple illustrative examples that have featured in the literature, we demonstrate that consistency is *not* achieved in many cases. This finding calls into doubt routine applications of the ABC method that are driven primarily by the convenience with which simple summary statistics can be computed, without further thought being given to the information content of those summaries.

Consistency by its very nature is more of a "thought experiment" than a practical feature of an estimation procedure. Nonetheless, consistency is a useful metric with which to gauge the output of a given statistical procedure. Following Diaconis and Freedman (1986), we argue that regardless of Bayesian bearing, that is, whether one is a "Classical" Bayesian (who believes in a "true but unknown parameter which is to be estimated from the data") or a "Subjective" Bayesian (who does not believe in true models but, rather, thinks in terms of predictive distributions) consistency is important for verification and practical implementation of Bayesian procedures. That is, whilst consistency is a property that sits naturally within the Classical Bayesian paradigm, it can also be viewed as being important to Subjectivists. To wit, Blackwell and Dubins (1962) and Diaconis and Freedman (1986) argue that consistency can be viewed as a "merging of intersubjective opinions" and that consistency of the posterior implies that two separate subjective Bayesians with different prior beliefs will ultimately end up with similar predictive distributions.

In what follows, we only concern ourselves with the idea of consistency as it pertains to some true model that is known up to an unknown vector of parameters. In this setting Bayesian

consistency means that any Bayesian method should yield increasingly accurate posterior inference as the sample size increases. While the theory of Bayesian consistency for likelihood-based Bayesian methods is now well documented, at least in the finite-dimensional parameter case, a thorough study on Bayesian consistency of so-called likelihood free methods, such as ABC, has yet to be undertaken. This represents an important gap in the literature, and is one we look to fill.

Bayesian consistency for posteriors based on finite-dimensional parameters is often derived under boundedness conditions for the underlying density function of the true model; see, for example, Le Cam (1953), Ibragimov and Has'minskii (1981), and Ghosal et al. (1995). In the ABC setting however, conditions based on the underlying density function are not useful since by the nature of the very problems to which ABC is applied, the underlying density is typically unknown in closed form. To this end, we derive a set of conditions on the summary statistics chosen within the ABC procedure that, when satisfied, ensure consistency of the posterior obtained from ABC. These conditions are similar in spirit to those seen in the literature on indirect inference (Gouriéroux, *et al.*, 1993). Examples from the ABC literature are used to demonstrate how the aforementioned conditions can be verified in practice.

The paper proceeds as follows. In Section 2 we briefly outline the basic principles of ABC. In Section 3, we establish conditions under which ABC will be consistent for the unknown parameters, and simple examples that respectively do and do not satisfy these conditions are given. In Section 4 we then propose a practical technique for identifying, in any particular problem, when the conditions for consistency are (or are not) satisfied. The analysis in Sections 3 and 4 focuses on the typical application of ABC, whereby summary statistics are chosen that are deemed to contain some information about the parameters of the true model and, more often than not, are used to define a matching criterion that is a weighted function of sample moments. In Section 5 we couch the discussion in terms of a general criterion function, where the latter derives from an auxiliary model, and which may - but certainly does not need to - derive from the likelihood function of that approximating model. In Section 6 we pursue the matter of consistency when using ABC to conduct inference in systems of ordinary differential equations (ODEs), with the Lotka-Volterra system for predator and prey used for illustration, and demonstrate that a common method for obtaining ABC posterior estimates in this setting does not yield Bayesian consistent inference. Section 7 concludes. Proofs of two theorems and one corollary are provided in an appendix to the paper.

2   ABC: an Outline of the Basic Approach

Suppose we are interested in conducting Bayesian inference on a complex parametric model indexed by the unknown p-dimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{\Theta} \subset \mathbb{R}^p$ compact, and let $P_{\boldsymbol{\theta}}$ denote

the family of probability measures induced by the model. Assume $P_{\boldsymbol{\theta}}$ admits a corresponding conditional density $p(\cdot|\boldsymbol{\theta})$ and assume we have T observations on the stochastic process y_t, characterized by $p(\mathbf{y}|\boldsymbol{\theta})$, with $\mathbf{y} = (y_1, y_2, ..., y_T)'$ denoting the T-dimensional vector of observed data. The aim of ABC is to produce draws from an approximation to the posterior distribution of the unknown $\boldsymbol{\theta}$ given observed data $\mathbf{y}$,

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

in the case where both the prior, $p(\boldsymbol{\theta})$, and the likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$, can be easily simulated. These draws are used, in turn, to approximate posterior quantities of interest, including marginal posterior moments, marginal posterior distributions and predictive distributions. The simplest (accept/reject) form of the algorithm (Tavaré *et al.*, 1997, Pritchard *et al.*, 1999) is detailed in Algorithm 1.

Algorithm 1 ABC algorithm

1: Simulate $\boldsymbol{\theta}^i$, $i = 1, 2, ..., N$, from $p(\boldsymbol{\theta})$
2: Simulate $\mathbf{z}^i = (z_1^i, z_2^i, ..., z_T^i)'$, $i = 1, 2, ..., N$, from the likelihood, $p(.|\boldsymbol{\theta}^i)$
3: Select $\boldsymbol{\theta}^i$ such that:
$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\} \leq \varepsilon, \tag{1}$$
where $\boldsymbol{\eta}(\cdot)$ is a (vector) statistic, $d\{\cdot, \cdot\}$ is a distance function (or metric), and the tolerance level ε is chosen as small as the computing budget allows.

Algorithm 1 thus samples $\boldsymbol{\theta}$ and $\mathbf{z}$ from the joint posterior:

$$p_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\eta}(\mathbf{y})) = \frac{p(\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})\mathbb{I}_\varepsilon[\mathbf{z}(\boldsymbol{\theta})]}{\int_{\boldsymbol{\Theta}} \int_{\mathbf{z}} p(\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})\mathbb{I}_\varepsilon[\mathbf{z}(\boldsymbol{\theta})]d\mathbf{z}d\boldsymbol{\theta}},$$

where $\mathbb{I}_\varepsilon[\mathbf{z}(\boldsymbol{\theta})]:=\mathbb{I}[d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))\} \leq \varepsilon]$ is one if $d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))\} \leq \varepsilon$ and zero else.[1] Clearly, when $\boldsymbol{\eta}(\mathbf{y})$ is a sufficient statistic and ε is arbitrarily small,

$$p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y})) = \int_{\mathbf{z}} p_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\eta}(\mathbf{y}))d\mathbf{z} \tag{2}$$

approximates the exact posterior, $p(\boldsymbol{\theta}|\mathbf{y})$, and draws from $p_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\eta}(\mathbf{y}))$ can be used to estimate features of the true posterior. In practice however, the complexity of the models to which ABC is applied implies, almost by definition, that sufficiency is unattainable. Hence, in the limit, as $\varepsilon \to 0$, the draws can be used only to approximate features of $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$.

ABC-based estimates of $p(\boldsymbol{\theta}|\mathbf{y})$ thus suffer from three types of approximation error: one invoked by the use of summary statistics that are not sufficient for $\boldsymbol{\theta}$; another associated with the use in practice of a non-zero tolerance, ε, for selecting draws from $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$; and, thirdly, the error produced when using non-parametric density techniques to estimate $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ from a

---

[1] The notation $\mathbf{z}(\boldsymbol{\theta})$ is used to emphasize the dependence of the simulated $\mathbf{z}$ on $\boldsymbol{\theta}$.

given set of selected draws. For any level of overall computational burden (i.e., the total number of draws N), reducing ε comes at a cost of reducing the probability of a draw being accepted, thereby contributing to the third form of error. The problem is exacerbated the larger is the dimension of $\boldsymbol{\eta}(\mathbf{y})$; see Blum (2010), Blum *et al.* (2013) and Nott *et al.* (2014). In practice ε tends to be chosen such that, for a given value of N, a certain (small) proportion of draws of $\boldsymbol{\theta}^i$ are selected, with attempts then made to reduce the third form of error using a variety of post-sampling (kernel-based) corrections of the draws (Beaumont *et al.*, 2002, Blum, 2010, Blum and François, 2010). Other work gives emphasis to choosing $\boldsymbol{\eta}(\cdot)$ and/or the selection mechanism itself in such a way that $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ is a closer match to $p(\boldsymbol{\theta}|\mathbf{y})$, in some sense. This may involve the replacement of the basic accept/reject scheme with Markov chain Monte Carlo (MCMC) and/or sequential Monte Carlo (SMC) steps (Marjoram *et al.*, 2003, Sisson *et al.*, 2007, Beaumont *et al.*, 2009, Toni *et al.*, 2009 and Wegmann *et al.*, 2009); or the selection of a vector $\boldsymbol{\eta}(\cdot)$ that is more informative in some well-defined sense; see Joyce and Marjoram (2008), Wegmann *et al.* (2009), Blum (2010) and Fearnhead and Prangle (2012).

In this latter spirit - and mimicking the frequentist techniques of indirect inference (II) (Gouriéroux *et al.*, 1993, Heggland and Frigessi, 2004) and efficient method of moments (EMM) (Gallant and Tauchen, 1996), Drovandi *et al.* (2011), Gleim and Pigorsch (2013), Martin *et al.* (2014), Drovandi *et al.* (2015) and Creel and Kristensen (2015) exploit an approximating model to produce the summary statistic vector $\boldsymbol{\eta}(\cdot)$. Under certain conditions on the auxiliary model, asymptotic sufficiency (at least) is attainable via use of the maximum likelihood estimates of the auxiliary parameters as the matching statistics in the ABC algorithm. Martin *et al.* also prove (for $\varepsilon \to 0$) the (Bayesian) consistency of the ABC approach that uses the MLE of the parameters of the auxiliary model to define $\boldsymbol{\eta}(\cdot)$, under similar conditions to those used to prove the consistency of the II method. The authors demonstrate the equivalence (again, as the tolerance approaches zero) of inference based on the score of the auxiliary model to that based on the MLE. This equivalence holds for any sample size and, hence, ensures that consistency is maintained by the (computationally efficient) score-based approach on the satisfaction of the appropriate conditions.

In this paper we also address the issue of Bayesian consistency, but in the completely general setting in which $\boldsymbol{\eta}(\cdot)$ comprises an arbitrary vector summary statistic, with elements possibly including, but not limited to, sample moments of the data, and with $\boldsymbol{\eta}(\cdot)$ not necessarily having an explicit link to the parameters of an auxiliary model. In the particular situation where $\boldsymbol{\eta}(\cdot)$ forms a vector statistic composed of sample moments, ABC parallels the frequentist method of simulated moments (McFadden, 1989, Pakes and Pollard, 1989, Duffie and Singleton, 1993). In the following section we maintain full generality in terms of the definition of $\boldsymbol{\eta}(\cdot)$. In Section 5 we then consider the case where the matching criterion is explicitly defined with respect to an auxiliary model, highlighting the fact that the likelihood function of that model is by no means

the only possible criterion that can be adopted.

3 ABC and Consistency

3.1 Consistency and Summary Statistics

Herein, we will only concern ourselves with the Classical ideal of Bayesian consistency: namely, as more data accumulates the posterior should stabilize around some true value and eventually collapse to a point mass at the same true value. More formally, for some set $A \subset \Theta$, define the posterior probability of A as

$$\Pr(\boldsymbol{\theta} \in A | \mathbf{y}) = \int_A p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta},$$

we then have the following well-known definition:

Definition 1: For true value $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ is *Bayesian consistent* if for any $\delta > 0$ and $\aleph_\delta(\boldsymbol{\theta}^0)$ an open neighborhood of $\boldsymbol{\theta}^0$, $\Pr(\boldsymbol{\theta} \notin \aleph_\delta(\boldsymbol{\theta}^0)|\mathbf{y}) \xrightarrow{P} 0$ as $T \to \infty$.

Herein, the symbol $\xrightarrow{P}$ denotes convergence in probability, and the symbols $o_P(a_T), O_P(b_T), \text{plim}_{T\to\infty}$, to be used below, have the usual definition.

Unlike the notion of consistency defined above, Bayesian consistency of posterior densities obtained from ABC requires not only $T \to \infty$ but $\varepsilon \to 0$ and is particular to the choice of $\boldsymbol{\eta}(\cdot)$ (and, indeed $d\{\cdot, \cdot\}$). Given this fact, we require a separate definition of Bayesian consistency for ABC.

Definition 2: For true value $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ and (vector) summary statistics $\boldsymbol{\eta}(\cdot); \mathbb{R}^T \to \mathbf{B}$, where $\mathbf{B} \subset \mathbb{R}^d$ and $d \geq p$, the ABC-based posterior density $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ is *Bayesian consistent* if for any $\delta > 0$, $\Pr_\varepsilon(\boldsymbol{\theta} \notin \aleph_\delta(\boldsymbol{\theta}^0)|\boldsymbol{\eta}(\mathbf{y})) \xrightarrow{P} 0$ as $T \to \infty$ and $\varepsilon \to 0$.

In addition, and in common with the standard definition (Defn. 1), the prior density used in ABC must be positive at the true value $\boldsymbol{\theta}^0$ and so we will assume the following condition is satisfied.

Assumption [P]: The prior density $p(\cdot)$ is continuous and $p(\boldsymbol{\theta}^0) > 0$.

As a heuristic for what Bayesian consistency in the ABC setting entails, consider the following simple example. Assume $\Theta := [0,1]$ and that $\theta \in \Theta$ has uniform prior probability. Consider some $\delta > 0$, $\varepsilon > 0$, and assume we have data $\mathbf{y}$ that is generated according to true (scalar) value θ^0. For $N = 5$ simulations the output is contained in Figure 1. For the particular $\delta > 0$ and $\varepsilon > 0$ chosen, two points lie within $\aleph_\delta(\theta^0)$ and three points lie outside. Clearly, the ABC-based posterior density $p_\varepsilon(\theta|\boldsymbol{\eta}(\mathbf{y}))$ only places mass on θ_1 and θ_4, as these points lead to a distance less
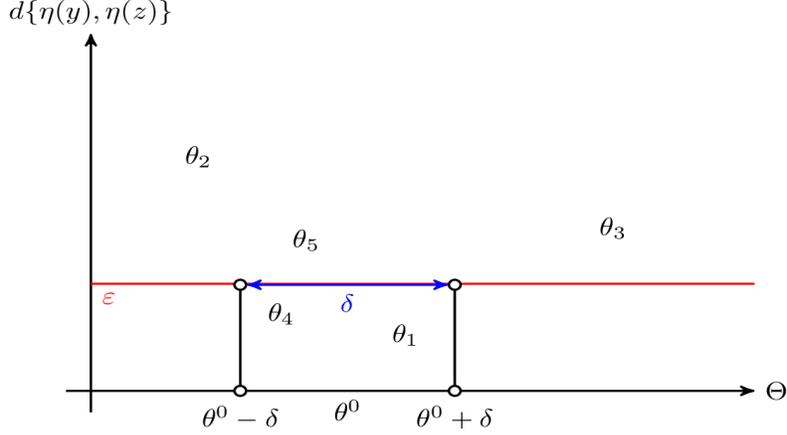
Figure 1: An illustration of ABC output for $N = 5$ simulations and given values of $\theta^0, \delta, \varepsilon,$ and $\boldsymbol{\eta}(\mathbf{y})$.

than ε, and zero mass is placed on the remaining three points. The ABC-based posterior density $p_\varepsilon(\theta|\boldsymbol{\eta}(\mathbf{y}))$ will be Bayesian consistent if for any arbitrary $\delta > 0$ and some $\varepsilon > 0$ similar behavior to that observed in Figure 1 holds as $T \to \infty$. This requires the following to be satisfied: one, for any given $\delta > 0$, we must be able to simulate draws within $\aleph_\delta(\theta^0)$ (guaranteed by Assumption [P]); two, for any T, including large T, there must exist a value of ε such that the only draws satisfying $d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z})\} \leq \varepsilon$ are those in $\aleph_\delta(\theta^0)$; three, for the value of ε in two, there must exist a corresponding number of simulation draws $N(\varepsilon)$ such that at least one simulated $\theta^i \in \{\theta^i\}_{i=1}^{N(\varepsilon)}$ satisfying $d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\theta^i))\} \leq \varepsilon$ occurs, else $p_\varepsilon(\theta|\boldsymbol{\eta}(\mathbf{y}))$ will not exist.

The formalization of these statements, along with the precise set of assumptions that a vector of summary statistics, $\boldsymbol{\eta}(\mathbf{y})$, should satisfy in order for ABC to yield consistent inference, is the content of Theorem 1 and its proof. Subsequent to the presentation of the theorem, we provide a simple example in which the conditions are satisfied, followed by a second example in which they are not. The way in which an increase in the dimension of $\boldsymbol{\eta}(\mathbf{y})$ can be used to retrieve consistency in the latter case, is then illustrated.

Define the limiting value of the summary statistic based on observed data (respectively, simulated data) as $\mathbf{b}(\boldsymbol{\theta}^0)$ (respectively, $\mathbf{b}(\boldsymbol{\theta}^i)$), and let $\| \cdot \|$ denote the Euclidean norm.

Theorem 1 *Let $d\{\cdot, \cdot\}$ be an induced metric on the normed space $(\mathbf{B}, \|\cdot\|_*)$. Given summary statistics $\boldsymbol{\eta}(\mathbf{y})$, assume that the following conditions are satisfied:*

[S0] *The DGP for $\mathbf{y}$ is uniquely defined at $\boldsymbol{\theta}^0$.*

[S1] $\|\boldsymbol{\eta}(\mathbf{y}) - \mathbf{b}(\boldsymbol{\theta}^0)\| = o_P(1).$

[S2] *The map $\boldsymbol{\theta}^i \mapsto \mathbf{b}(\boldsymbol{\theta}^i)$ is deterministic, continuous, and satisfies*

7

[S2(1)] $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta})) - \mathbf{b}(\boldsymbol{\theta})\| = o_P(1)$.

[S2(2)] $\mathbf{b}(\cdot)$ *is one-to-one in* $\boldsymbol{\theta}^i$.

If [S0]-[S2] above and [P] are satisfied, then, for any $\delta > 0$, $Pr_\varepsilon(\boldsymbol{\theta} \notin \aleph_\delta(\boldsymbol{\theta}^0) | \boldsymbol{\eta}(\mathbf{y})) \xrightarrow{P} 0$ *as* $T \to \infty$ *and* $\varepsilon \to 0$.

Remark 1 Theorem 1 states that, for ABC based on $\boldsymbol{\eta}(\cdot)$ to be consistent the limit map $\boldsymbol{\theta}^i \mapsto \mathbf{b}(\boldsymbol{\theta}^i)$ must act in the same manner as the "binding function" in indirect inference; see Gouriéroux *et al.* (1993) and Gouriéroux and Monfort (1996) for a general discussion of binding functions.

Remark 2 Since we are only concerned with Classical Bayesian consistency, Assumption [S0] is implicit and therefore not explicitly required. However, Assumption [S0] is a deep identification condition that may not be satisfied in all circumstances and is therefore maintained to illustrate the scope of the models to which this result will apply. Assumption [S1] is often satisfied under general conditions restricting the dependence in the observed data. Assumption [S2(1)] requires that for all $\xi > 0$,

$$\lim_{T \to \infty} \Pr\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta})) - \mathbf{b}(\boldsymbol{\theta})\| > \xi\right) = 0,$$

and is generally referred to as uniform convergence. This stronger notion of convergence is required to ensure that the simulated paths $\mathbf{z}(\boldsymbol{\theta})$, and the subsequent $\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))$, are well-behaved over $\boldsymbol{\Theta}$. General conditions determining satisfaction of [S2(1)] are now well-known and a great many results can be obtained from the empirical process literature; see, for instance, Pollard (1990). In particular, [S2(1)] is likely to be satisfied for many different types of summary statistics so long as the prior density $p(\cdot)$ admits values of $\boldsymbol{\theta}$ that do not allow the simulated data to display too much persistence.[2]

Remark 3 Theorem 1 requires that the (vector of) summary statistics based on observed data converges, with respect to $d\{\cdot, \cdot\}$, to a fixed quantity and the corresponding vector of statistics based on simulated data $\mathbf{z}^i = \mathbf{z}(\boldsymbol{\theta}^i)$ converges (uniformly), with respect to $d\{\cdot, \cdot\}$, to a deterministic function of $\boldsymbol{\theta}^i$. Consistency thus depends not only on the choice of $\boldsymbol{\eta}(\mathbf{y})$ but also on the precise choice of $d\{\cdot, \cdot\}$, with convergence in one metric not necessarily implying convergence in another. However, restricting $d : \mathbf{B} \times \mathbf{B} \to \mathbb{R}_+$ to be an induced metric on the normed space $(\mathbf{B}, \|\cdot\|_*)$ - i.e., for $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathbf{B}$, requiring that $d\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2\} = \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\|_*$ for some norm $\|\cdot\|_*$ - relieves the convergence issue since all norms on $\mathbf{B}$ are equivalent to the Euclidean norm $\|\cdot\|$. The requirement that $d\{\cdot, \cdot\}$ be an induced metric is not restrictive as the most common choices of $d\{\cdot, \cdot\}$ satisfy this condition.

---

[2] Technically, conditions [S0] and [S2(1)] imply condition [S1]. However, the authors believe it is helpful to specify separate conditions on the statistics associated with observed and simulated data.

Remark 4 Bayesian consistency says that for any $\delta > 0$, $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ will attribute zero probability, as $T \to \infty$, to points outside $\aleph_\delta(\boldsymbol{\theta}^0)$; it does not say *anything* about how well $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ approximates the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ or even the partial posterior density $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$. Specifically, the demonstration of Bayesian consistency is distinct from existing theoretical work on ABC that shows $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ is consistent for $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$, as $N \to \infty$ and $\varepsilon \to 0$, for any $\boldsymbol{\theta} \in \Theta$ and for any fixed T. To prove the latter form of result, researchers have borrowed from the literature on nonparametric density estimation and relied on the idea of mean squared error (MSE) consistency, which requires the bias and variance of $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y})) - p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ to approach zero as $N \to \infty$ and $\varepsilon \to 0$; see, for example, Blum (2010) and Biau *et al.* (2015). In particular, MSE consistency requires a specific rate condition between N and ε to ensure that the variance of $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y})) - p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ shrinks to zero fast enough. As noted in the proof of Theorem 1, Bayesian consistency still requires N to increase as $\varepsilon \to 0$, but only to ensure that $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ exists for small ε, and any T. In this way, the particular relationship between N and ε is independent of the sample size T. This lack of any T-dependent condition for ε contrasts with the need for such a condition when deriving results for the asymptotic distribution of ABC point estimators; see, for example, Li and Fearnhead (2015). We elaborate further on this distinction in an on-line supplementary appendix to the paper.[3]

3.2   Success and Failure of Summary Statistic-based ABC

Consistency of ABC based on $\boldsymbol{\eta}(\cdot)$ hinges on the particular form of $\mathbf{b}(\boldsymbol{\theta}^i)$. If $\mathbf{b}(\cdot)$ is one-to-one, i.e., the map $\boldsymbol{\theta}^i \mapsto \mathbf{b}(\boldsymbol{\theta}^i)$ satisfies [S2(2)], and the remaining assumptions in Theorem 1 hold, ABC based on $\boldsymbol{\eta}(\cdot)$ will be consistent. There is generally no guarantee that $\mathbf{b}(\cdot)$ will be one-to-one and satisfaction of this condition depends on both the true structural model and the particular choice of summary statistics. Examples 1 and 2 illustrate a case where [S2(2)] is and is not satisfied, respectively. Example 3 illustrates the impact on identification and, hence, the attainment of consistency, of adding summary statistics to an initial set.

Example 1 (Satisfaction of S2(2)) *Consider the following autoregressive (AR) model of order one:*

$$y_t = \theta y_{t-1} + \nu_t,$$

where $\nu_t \sim i.i.d. N(0,1)$ and $|\theta| < 1, \theta \neq 0$. Whilst the likelihood for this model is known in closed form and, hence, exact Bayesian inference is perfectly feasible, for the sake of illustration, consider Algorithm 2 based on the summary statistic $\eta(\mathbf{y}) = \frac{1}{T}\sum_{t=2}^{T} y_t y_{t-1}$.

Assume that some true value θ^0 has generated the observed sample $\mathbf{y}$. For $p_\varepsilon(\theta|\eta(\mathbf{y}))$ to be degenerate at θ^0 it must be that $d\{b(\theta^0), b(\theta^i)\} = 0$ has a unique solution $\theta^i = \theta^0$. By the

---

[3]This document is available at: http://users.monash.edu.au/~gmartin/FMR_Supplementary_Appendix.pdf).

Algorithm 2 ABC algorithm: AR(1) example

1: Simulate the AR(1) coefficient θ^i from, for instance, a uniform prior over $(-1, 1)$;
2: Generate an i.i.d. sequence $\{\nu_t\}_{t=1}^T$;
3: Produce a simulated series $\{z_t^i(\theta^i)\}_{t=1}^T$;
4: Accept the value simulated in Step (1) if $d\{\eta(\mathbf{y}), \eta(\mathbf{z}^i)\} \leq \varepsilon$ for $\varepsilon > 0$ and small.

---

weak law of large numbers $\eta(\mathbf{y}) \xrightarrow{P} E[y_t y_{t-1}] = b(\theta^0) = \theta^0/(1 - (\theta^0)^2)$ *and* $\eta(\mathbf{z}^i) \xrightarrow{P} E[z_t^i z_{t-1}^i] = b(\theta^i) = \theta^i/(1 - (\theta^i)^2)$, *so* $d\{b(\theta^0), b(\theta^i)\} = 0$ *requires that*

$$0 = b(\theta^0) - b(\theta^i) = (\theta^i)^2 \theta^0 + \theta^i(1 - (\theta^0)^2) - \theta^0$$

has unique solution $\theta^i = \theta^0$. *This quadratic equation in* θ^i *has two solutions:* $\theta^i = \theta^0$ *and* $\theta^i = -1/\theta^0$. *However, given that* $|\theta^0| < 1$, $\theta^0 \neq 0$, *the second solution is not in the feasible region for* θ^i *and so ABC based on* $\eta(\mathbf{y}) = \frac{1}{T} \sum_{t=2}^T y_t y_{t-1}$ *satisfies the conditions of Theorem 1.*

Example 2 (Failure of S2(2)) *Consider now the moving average (MA) model of order two:*

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, \tag{3}$$

where $e_t \sim i.i.d. N(0, 1)$ *and* θ_1, θ_2 *satisfy the following invertibility conditions*

$$-2 < \theta_1 < 2, \ \theta_1 + \theta_2 > -1, \theta_1 - \theta_2 < 1. \tag{4}$$

Following Marin et al. (2011), we choose as summary statistics the sample autocovariances $\eta_j(\mathbf{y}) = \frac{1}{T} \sum_{t=1+j}^T y_t y_{t-j}$, *for* $j = 0, 1, 2..., K$. *Consider, initially, Algorithm 3, based on* $\boldsymbol{\eta}(\mathbf{y}) = (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}))'$.

Algorithm 3 ABC algorithm: MA(2) example

1: Simulate the MA(2) coefficients $\boldsymbol{\theta}^i$ from $p(\boldsymbol{\theta})$ satisfying (4), where $\boldsymbol{\theta} = (\theta_1, \theta_2)'$;
2: Generate an i.i.d. sequence $\{e_t\}_{t=1}^T$;
3: Produce a simulated series $\{z_t^i(\boldsymbol{\theta}^i)\}_{t=1}^T$;
4: Accept the value generated in Step (1) if $d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\} \leq \varepsilon$ for $\varepsilon > 0$ and small.

Assume that true value $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0)'$ *has generated the observed data* $\mathbf{y}$. *By the weak law of large numbers* $\eta_0(\mathbf{y}) \xrightarrow{P} E[y_t^2] = 1 + (\theta_1^0)^2 + (\theta_2^0)^2$ *and* $\eta_1(\mathbf{y}) \xrightarrow{P} E[y_t y_{t-1}] = \theta_1^0(1 + \theta_2^0)$. *In addition, conditional on* $\boldsymbol{\theta}^i = (\theta_1^i, \theta_2^i)'$ *satisfying equation* (4), $\eta_0(\mathbf{z}^i) \xrightarrow{P} 1 + (\theta_1^i)^2 + (\theta_2^i)^2$ *and* $\eta_1(\mathbf{z}^i) \xrightarrow{P} \theta_1^i(1 + \theta_2^i)$. *For* $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ *obtained from the above algorithm to be degenerate at* $\boldsymbol{\theta}^0$ *it must be that for all* $\boldsymbol{\theta}^i, \boldsymbol{\theta}^0 \in \boldsymbol{\Theta}$, $0 = \mathbf{b}(\boldsymbol{\theta}^0) - \mathbf{b}(\boldsymbol{\theta}^i)$ *has unique solution* $\boldsymbol{\theta}^i = \boldsymbol{\theta}^0$. *Clearly,*

$$0 = \mathbf{b}(\boldsymbol{\theta}^0) - \mathbf{b}(\boldsymbol{\theta}^i) = \begin{pmatrix} 1 + (\theta_1^0)^2 + (\theta_2^0)^2 \\ \theta_1^0(1 + \theta_2^0) \end{pmatrix} - \begin{pmatrix} 1 + (\theta_1^i)^2 + (\theta_2^i)^2 \\ \theta_1^i(1 + \theta_2^i) \end{pmatrix}.$$

As in Marin et al. (2011), take $\theta_1^0 = .6, \theta_2^0 = .2$. Then the question becomes, does there exist $\theta_1^i \neq .6, \theta_2^i \neq .2$ such that

$$0 = \mathbf{b}(\theta^0) - \mathbf{b}\left(\theta^i\right) = \begin{pmatrix} 1 + (.6)^2 + (.2)^2 \\ .6(1 + .2) \end{pmatrix} - \begin{pmatrix} 1 + (\theta_1^i)^2 + (\theta_2^i)^2 \\ \theta_1^i(1 + \theta_2^i) \end{pmatrix}? \tag{5}$$

Simple numerical calculations reveal that (5) has two solutions: $\theta_1^i = .6, \theta_2^i = .2$ and $\theta_1^i \approx .5453, \theta_2^i \approx .3204$, where the latter solution remains in the feasible region for $\theta^i = (\theta_1^i, \theta_2^i)'$. Therefore, $\mathbf{b}(\cdot)$ is not one-to-one and the ABC-based posterior will not converge to $\theta^0 = (.6, .2)'$.

Example 3 (Effect of Additional Statistics) *Consider the same MA(2) model as in Example 2, but now consider the use of the three-dimensional vector of summary statistics:*

$$\boldsymbol{\eta}\left(\mathbf{y}\right) = (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}), \eta_2(\mathbf{y}))'.$$

In the language of the generalized method of moments (GMM) literature, the summary statistics of which $\boldsymbol{\eta}(\mathbf{y})$ is comprised "over-identify" θ^0. In this case, [S2(2)] will be satisfied if the following equation has a unique solution for all $\theta^i, \theta^0 \in \Theta$:

$$0 = \mathbf{b}(\theta^0) - \mathbf{b}(\theta^i) = \begin{pmatrix} 1 + (\theta_1^0)^2 + (\theta_2^0)^2 \\ \theta_1^0(1 + \theta_2^0) \\ \theta_2^0 \end{pmatrix} - \begin{pmatrix} 1 + (\theta_1^i)^2 + (\theta_2^i)^2 \\ \theta_1^i(1 + \theta_2^i) \\ \theta_2^i \end{pmatrix}.$$

The additional (linear) restriction, $0 = \theta_2^0 - \theta_2^i$, ensures that the only value that satisfies $0 = \mathbf{b}(\theta^0) - \mathbf{b}(\theta^i)$ is now $\theta^0 = (\theta_1^0, \theta_2^0)'$, and consistency will be achieved as a consequence.

Simply adding summary statistics to the ABC procedure is, however, not guaranteed to yield consistent inference: the chosen summary statistics must be informative about the underlying parameters θ governing the statistical properties of the structural model. To illustrate this point, consider again the above example, but with the three-dimensional vector summary statistic:

$$\boldsymbol{\eta}\left(\mathbf{y}\right) = (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}), \eta_3(\mathbf{y}))',$$

where $\eta_3(\mathbf{y}) = \frac{1}{T} \sum_{t=4}^T y_t y_{t-3}$. Given the nature of the structural model, $\eta_3(\mathbf{y}) \xrightarrow{P} E[y_t y_{t-3}] = 0$ and by construction $\eta_3(\mathbf{z}^i) \xrightarrow{P} E[z_t^i z_{t-3}^i] = 0$ for all θ^i. Hence, the summary statistic $\eta_3(\mathbf{y})$ yields no new information about θ^0 and does not therefore produce a mapping $\theta^i \longmapsto \mathbf{b}(\theta^i)$ that is one-to-one.

We illustrate the theoretical results in Examples 2 and 3 graphically in Figure 2, denoting the three relevant vectors of summary statistics as:

$$\begin{aligned} \boldsymbol{\eta}^1\left(\mathbf{y}\right) &= (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}))', \\ \boldsymbol{\eta}^2\left(\mathbf{y}\right) &= (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}), \eta_2(\mathbf{y}))', \\ \boldsymbol{\eta}^3\left(\mathbf{y}\right) &= (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}), \eta_3(\mathbf{y}))'. \end{aligned}$$
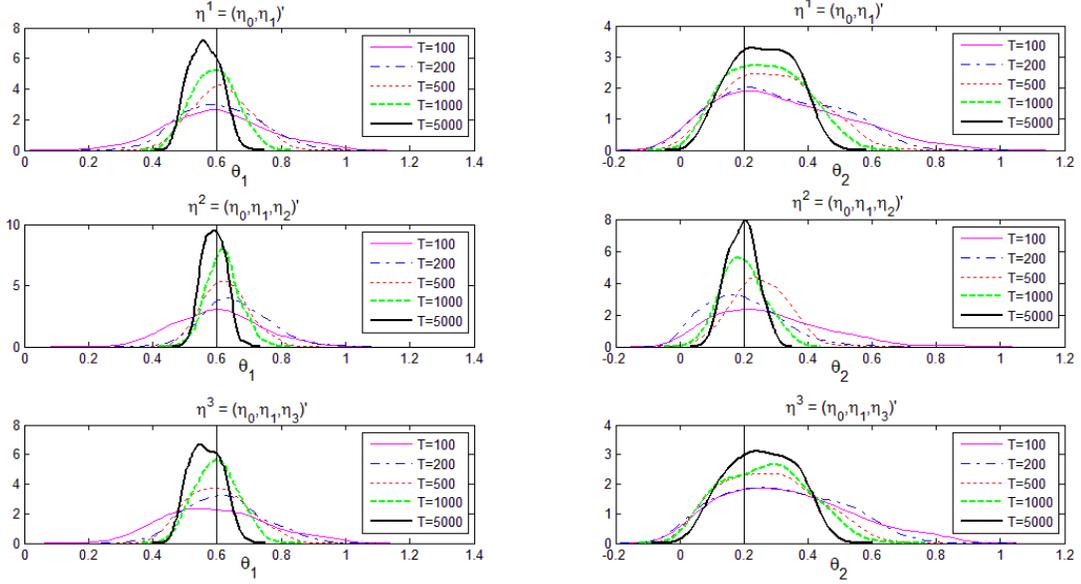
11

ABC marginal posteriors: MA(2) model

Figure 2: ABC-based estimates of the marginal posterior densities for the parameters of the MA(2) model, θ_1 and θ_2, with varying sample sizes. Top two panels: summary statistic vector of $\boldsymbol{\eta}^1(\mathbf{y}) = (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}))'$; Middle two panels: summary statistic vector of $\boldsymbol{\eta}^2(\mathbf{y}) = (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}), \eta_2(\mathbf{y}))'$; Bottom two panels: summary statistic vector of $\boldsymbol{\eta}^3(\mathbf{y}) = (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}), \eta_3(\mathbf{y}))'$. The true parameter values are $\theta_1^0 = 0.6$ and $\theta_2^0 = 0.2$.

Using the true parameter vector $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0)' = (0.6, 0.2)'$, a vector of 'observed' data, $\mathbf{y} = (y_1, y_2, ..., y_T)'$ is generated, for $T = 100, 200, 500, 1,000$ and $5,000$. For each given sample of size T, $p(\boldsymbol{\theta}|\mathbf{y})$ is then estimated via the ABC method, using $N = 50,000$ simulated draws from uniform priors satisfying (4), and with the tolerance ε_j, $j = 1, 2, 3$, chosen so that only one-percent of the simulated draws are accepted. The top two panels of Figure 2 plot $p_{\varepsilon_1}(\theta_1|\boldsymbol{\eta}^1(\mathbf{y}))$ and $p_{\varepsilon_1}(\theta_2|\boldsymbol{\eta}^1(\mathbf{y}))$ respectively, where the notation here indicates the kernel density estimate of the relevant marginal density, conditional on $\boldsymbol{\eta}^1(\mathbf{y})$, and as defined for the given ε_1. As the sample size increases both estimated marginals become more concentrated, but not around the true values of 0.6 and 0.2. In contrast, the plots in the two middle panels demonstrate the consistency that obtains when conditioning on $\boldsymbol{\eta}^2(\mathbf{y})$, a result that is not replicated in the two bottom panels, in which the three-dimensional conditioning vector is $\boldsymbol{\eta}^3(\mathbf{y})$.[4]

Remark 5 The above example illustrates that adding additional summary statistics to an ABC procedure may or may not aid researchers in obtaining consistent inference. In particular, adding summary statistics will only be helpful if the additional statistics contain infor-

---

[4]Whilst we have not pursued this in any formal way, the indications are that in the two cases in which identification (of the true parameters) does not obtain, the marginal posteriors are some form of mixture distribution, each with a mode (or modes) that reflects (reflect) the location of the two pairs of parameter values that satisfy (5).

mation about the parameters that is not accounted for by the summary statistics already used in the analysis. Therefore, *arbitrarily* adding summary statistics will not necessarily yield valid inference. Moreover, and as was noted in Section 2, given that adding summary statistics hampers our ability to accurately estimate the associated conditional density, adding summary statistics to any initial ABC procedure should be embarked upon with care.

Remark 6 It is also important to note that no link is to be expected between the particular model at hand and the likelihood of Assumption [S2(2)] being satisfied. As the above examples illustrate, it is the *combination* of the model structure and the choice of summary statistics that determines Bayesian consistency via ABC.

4 Detecting Consistency

4.1 Preliminaries

Beyond understanding the theoretical conditions that must hold in order for a particular set of summary statistics to yield valid inference, and noting that in complex settings verifying the conditions of Theorem 1 will typically not be possible via analytical means, it is useful to have some way of ascertaining numerically whether those conditions actually hold in any given case. To this end, we present a diagnostic tool that can be used to determine if the estimated posterior obtained using a specific set of summary statistics, say $\boldsymbol{\eta}(\mathbf{y})$, is Bayesian consistent.

The key insight to understanding the diagnostic procedure is that if the true value $\boldsymbol{\theta}^0$ were known, we would only require a local version of the identification condition (Assumption [S2(2)]); i.e., we would only need to check that there existed no $\boldsymbol{\theta}^*$, with $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^0$, for which $\mathbf{b}(\boldsymbol{\theta}^*) = \mathbf{b}(\boldsymbol{\theta}^0)$. However, because $\boldsymbol{\theta}^0$ is unknown, a sufficient condition to ensure that the above holds is that the map $\boldsymbol{\theta} \mapsto \mathbf{b}(\boldsymbol{\theta})$ is one-to-one; i.e., that $\mathbf{b}(\boldsymbol{\theta}^0) - \mathbf{b}(\boldsymbol{\theta}^*) = \mathbf{0}$ yields the unique solution $\boldsymbol{\theta}^* = \boldsymbol{\theta}^0$ for each and every possible value of $\boldsymbol{\theta}^0$. In this way, detecting Bayesian consistency in ABC reduces to detecting satisfaction of the one-to-one mapping assumption. The diagnostic procedure we propose seeks to verify this condition, and hence the consistency of ABC posterior estimates, in two stages: firstly, as it is applied to the observed data $\mathbf{y}$ (Section 4.2), and secondly, in terms of its repeated application to data sets artificially generated from the assumed true data generating process and across the feasible parameter space (Section 4.3).

The verification procedure exploits the following two facts: 1) under the conditions of Theorem 1, the possible set of solutions for which $d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))\} = o_P(1)$ always includes $\boldsymbol{\theta} = \boldsymbol{\theta}^0$; 2) if $d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))\} = o_P(1)$ uniquely at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, then an ABC procedure based on an augmented vector of summary statistics, $\boldsymbol{\gamma}(\mathbf{y}) = (\boldsymbol{\eta}(\mathbf{y})', \mathbf{g}(\mathbf{y})')'$, will yield a posterior that is Bayesian consistent, so long as $\mathbf{g}(\cdot)$ satisfies conditions [S1] and [S2(1)] of Theorem 1. To state these results more

formally, assume $\boldsymbol{\eta}(\mathbf{y})$ (respectively, $\mathbf{g}(\mathbf{y})$) has a well-defined limit $\mathbf{b}_1(\boldsymbol{\theta}^0)$ (respectively, $\mathbf{b}_2(\boldsymbol{\theta}^0)$) and denote the limit quantity of $\boldsymbol{\eta}(\mathbf{z}^i)$ (respectively, $\mathbf{g}(\mathbf{z}^i)$) as $\mathbf{b}_1(\boldsymbol{\theta}^i)$ (respectively, $\mathbf{b}_2(\boldsymbol{\theta}^i)$).

Corollary 1 *Given summary statistics* $\boldsymbol{\gamma}(\mathbf{y}) = (\boldsymbol{\eta}(\mathbf{y})', \mathbf{g}(\mathbf{y})')'$, *assume that the following conditions are satisfied:*

[C0] *The DGP for* $\mathbf{y}$ *is uniquely defined at* $\boldsymbol{\theta}^0$.

[C1] *For* $\mathbf{b}(\boldsymbol{\theta}^0) = (\mathbf{b}_1(\boldsymbol{\theta}^0)', \mathbf{b}_2(\boldsymbol{\theta}^0)')'$, *we have* $\|\boldsymbol{\gamma}(\mathbf{y}) - \mathbf{b}(\boldsymbol{\theta}^0)\| = o_P(1)$.

[C2] *The map* $\boldsymbol{\theta}^i \mapsto \mathbf{b}(\boldsymbol{\theta}^i)$ *is deterministic, continuous, exists for all* $\boldsymbol{\theta}^i \in \boldsymbol{\Theta}$, *and satisfies*

> **[C2(1)]** $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{\gamma}(\mathbf{z}(\boldsymbol{\theta})) - \mathbf{b}(\boldsymbol{\theta})\| = o_P(1)$,
>
> **[C2(2)]** $\mathbf{b}_1(\cdot)$ *is one-to-one in* $\boldsymbol{\theta}^i$.

If [C0]-[C2] and [P] are satisfied, then, for all $\delta > 0$, $Pr_\varepsilon(\boldsymbol{\theta} \notin \aleph_\delta(\boldsymbol{\theta}^0)|\boldsymbol{\gamma}(\mathbf{y})) \xrightarrow{P} 0$ *as* $T \to \infty$ *and* $\varepsilon \to 0$.

4.2 Use of the Observed Data

To understand the implications of Corollary 1, consider the case where we have already obtained $p_{\varepsilon_1}(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$, for some tolerance ε_1, with $\boldsymbol{\eta}(\mathbf{y})$ based on a value of T that is assumed to be large enough for large sample behavior to be in evidence. Now, if we were to run ABC again using the joint summary statistic $\boldsymbol{\gamma}(\mathbf{y})$, Corollary 1 implies that one of two things will happen: either the posterior $p_{\varepsilon_2}(\boldsymbol{\theta}|\boldsymbol{\gamma}(\mathbf{y}))$, computed for some tolerance ε_2, will be located in a very similar position to $p_{\varepsilon_1}(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$, only potentially flatter or with a slightly different shape, a consequence of the increased dimensionality,[5] or the high mass region of $p_{\varepsilon_2}(\boldsymbol{\theta}|\boldsymbol{\gamma}(\mathbf{y}))$ will be located in a distinctly different part of the support from that of $p_{\varepsilon_1}(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$. We refer to this latter event as one of $p_{\varepsilon_2}(\boldsymbol{\theta}|\boldsymbol{\gamma}(\mathbf{y}))$ "jumping away" from $p_{\varepsilon_1}(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ and, according to Corollary 1, see the occurrence of this event as evidence that the initial summary statistics did *not* yield a posterior that is Bayesian consistent. If, on the other hand, the addition of $\mathbf{g}(\mathbf{y})$ does not cause the mass of $p_{\varepsilon_2}(\boldsymbol{\theta}|\boldsymbol{\gamma}(\mathbf{y}))$ to jump in relation to $p_{\varepsilon_1}(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$, then this suggests that the initial choice of summary statistics, $\boldsymbol{\eta}(\mathbf{y})$, may have yielded valid inference. The use of the word 'may' reflects the fact that there is no guarantee possible, via use of the observed data alone, that consistency has been achieved, since there is no guarantee *a priori* that $d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))\} = o_P(1)$ has a *unique* solution $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. It is this point that is addressed in next subsection.

Meanwhile, we illustrate this preliminary diagnostic exercise via the MA(2) model, in which case (from Example 3) we have an analytical result that establishes that consistency for the true

---

[5]Simulation evidence suggests that the increased flatness (or otherwise) of the subsequent posterior estimates depends on the nature of the information about $\boldsymbol{\theta}^0$ contained in the additional summary statistics.

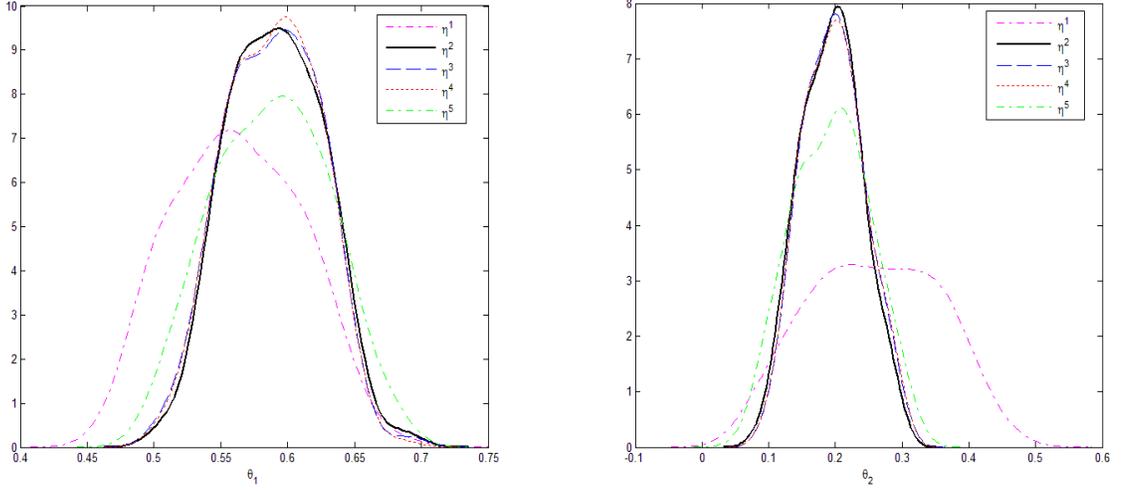**ABC marginal posteriors: MA(2) model, adding summary statistics**

Figure 3: ABC-based estimates of the marginal posterior densities for the parameters of the MA(2) model, θ_1 and θ_2, with $T = 5000$. The key for both panels indicates the summary statistic vector used: $\boldsymbol{\eta}^1(\mathbf{y}) = (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}))'$; $\boldsymbol{\eta}^2(\mathbf{y}) = (\boldsymbol{\eta}^1(\mathbf{y})', \eta_2(\mathbf{y}))'$; $\boldsymbol{\eta}^3(\mathbf{y}) = (\boldsymbol{\eta}^2(\mathbf{y})', \eta_3(\mathbf{y}))'$; $\boldsymbol{\eta}^4(\mathbf{y}) = (\boldsymbol{\eta}^3(\mathbf{y})', \frac{1}{T}\sum_{t=1}^{T} y_t)'$; $\boldsymbol{\eta}^5(\mathbf{y}) = (\boldsymbol{\eta}^4(\mathbf{y})', \frac{1}{T}\sum_{t=1}^{T} y_t^3)'$. The statistics $\boldsymbol{\eta}^2(\mathbf{y})$ to $\boldsymbol{\eta}^5(\mathbf{y})$ yield Bayesian consistency. The true parameter values are $\theta_1^0 = 0.6$ and $\theta_2^0 = 0.2$.

$\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0)' = (.6, .2)'$ is achieved via a particular choice of summary statistics. We adopt five different choices of summary statistics for use in the illustration:

$$
\begin{aligned}
\boldsymbol{\eta}^1(\mathbf{y}) &= (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}))' \\
\boldsymbol{\eta}^2(\mathbf{y}) &= (\boldsymbol{\eta}^1(\mathbf{y})', \eta_2(\mathbf{y}))' \\
\boldsymbol{\eta}^3(\mathbf{y}) &= (\boldsymbol{\eta}^2(\mathbf{y})', \eta_3(\mathbf{y}))' \\
\boldsymbol{\eta}^4(\mathbf{y}) &= (\boldsymbol{\eta}^3(\mathbf{y})', \frac{1}{T}\sum_{t=1}^{T} y_t)' \\
\boldsymbol{\eta}^5(\mathbf{y}) &= (\boldsymbol{\eta}^4(\mathbf{y})', \frac{1}{T}\sum_{t=1}^{T} y_t^3)',
\end{aligned}
$$

where $\eta_j(\mathbf{y}) = \frac{1}{T}\sum_{t=1+j}^{T} y_t y_{t-j}$, for $j = 0, 1, 2, 3$. We set the sample size to $T = 5,000$, consider $N = 50,000$ simulations and set the tolerance ε_j, $j = 1, ..., 5$, so that we retain one-percent of the simulated draws for each choice of summary statistics. From our previous theoretical analysis we know that $\boldsymbol{\eta}^1(\mathbf{y})$ will not yield an estimated posterior, $p_{\varepsilon_1}(\boldsymbol{\theta}|\boldsymbol{\eta}^1(\mathbf{y}))$, that is Bayesian consistent, while the remaining sets will yield posteriors that *are* Bayesian consistent, due to the inclusion of $\eta_2(\mathbf{y})$. Therefore, after adding $\eta_2(\mathbf{y})$ to our initial choice of summary statistics, $\boldsymbol{\eta}^1(\mathbf{y})$, the estimated posterior $p_{\varepsilon_2}(\boldsymbol{\theta}|\boldsymbol{\eta}^2(\mathbf{y}))$ should be centered around the true values, or thereabouts (given the still finite value of T); that is, the main mass of the posterior computed using $\boldsymbol{\eta}^2(\mathbf{y})$ should

"jump" away from the main mass of the posterior computed using $\boldsymbol{\eta}^1(\mathbf{y})$. Subsequently, the posteriors based on summary statistics $\boldsymbol{\eta}^3(\mathbf{y}), \boldsymbol{\eta}^4(\mathbf{y})$ and $\boldsymbol{\eta}^5(\mathbf{y})$ should not move much, if at all, in relation to $p_{\varepsilon_2}(\boldsymbol{\theta}|\boldsymbol{\eta}^2(\mathbf{y}))$ but may possibly become flatter, and possibly change shape, with each additional summary statistic. Figure 3 illustrates these points exactly. The estimated posterior $p_{\varepsilon_2}(\boldsymbol{\theta}|\boldsymbol{\eta}^2(\mathbf{y}))$ is seen to shift substantially in relation to the estimated posterior $p_{\varepsilon_1}(\boldsymbol{\theta}|\boldsymbol{\eta}^1(\mathbf{y}))$. In turn, adding $\eta_3(\mathbf{y})$ and $\frac{1}{T}\sum_{t=1}^{T} y_t$ to $\boldsymbol{\eta}^2(\mathbf{y})$ causes minimal change, and certainly no discernible change in location. The location of the high mass point is preserved by the subsequent addition of $\frac{1}{T}\sum_{t=1}^{T} y_t^3$; however at this point the dimension of the full statistic $\boldsymbol{\eta}^5(\mathbf{y})$ appears to cut in, with the accuracy of the kernel density estimation adversely affected.[6]
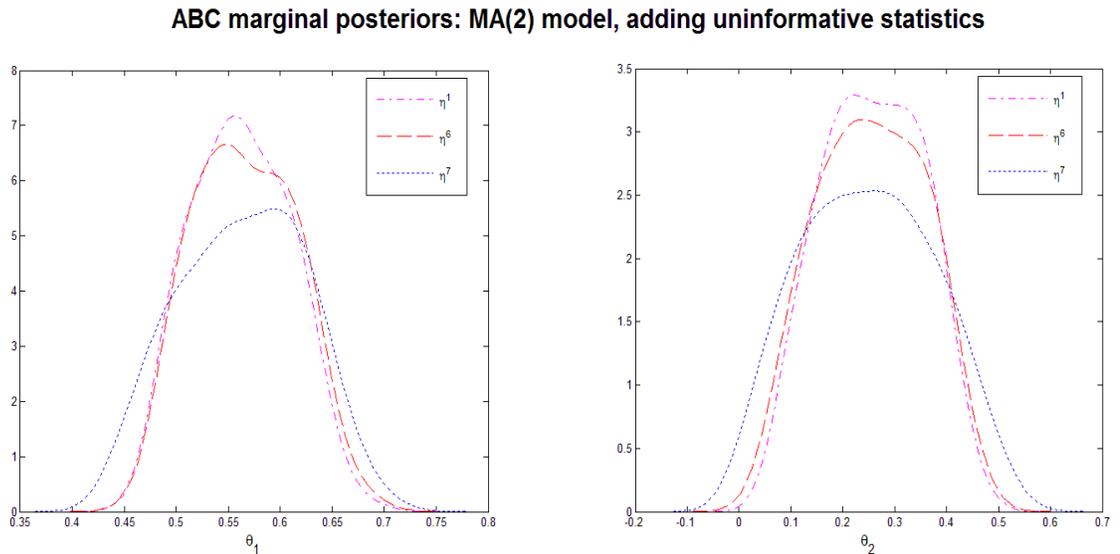
ABC marginal posteriors: MA(2) model, adding uninformative statistics

Figure 4: ABC-based estimates of the marginal posterior densities for the parameters of the MA(2) model, θ_1 and θ_2, with $T = 5000$. The key for both panels indicates the summary statistic vector used: $\boldsymbol{\eta}^1(\mathbf{y}) = (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}))'$; $\boldsymbol{\eta}^6(\mathbf{y}) = (\boldsymbol{\eta}^1(\mathbf{y}), \eta_3(\mathbf{y}))'$; $\boldsymbol{\eta}^7(\mathbf{y}) = (\boldsymbol{\eta}^6(\mathbf{y}), \frac{1}{T}\sum_{t=1}^{T} y_t^3)'$. All three summary statistics do not yield Bayesian consistency. The true parameter values are $\theta_1^0 = 0.6$ and $\theta_2^0 = 0.2$.

Let us now consider a similar exercise with summary statistics $\boldsymbol{\eta}^1(\mathbf{y}) = (\eta_0(\mathbf{y}), \eta_1(\mathbf{y}))'$, $\boldsymbol{\eta}^6(\mathbf{y}) = (\boldsymbol{\eta}^1(\mathbf{y}), \eta_3(\mathbf{y}))'$, $\boldsymbol{\eta}^7(\mathbf{y}) = (\boldsymbol{\eta}^6(\mathbf{y}), \frac{1}{T}\sum_{t=1}^{T} y_t^3)'$ and $\boldsymbol{\eta}^8(\mathbf{y}) = (\boldsymbol{\eta}^7(\mathbf{y}), \eta_2(\mathbf{y}))'$, where we deliberately use different notation to distinguish these statistics from those used in the illustration above. In this particular setup the only set of summary statistics that will yield consistent inference is $\boldsymbol{\eta}^8(\mathbf{y})$, and the aim of the exercise is to illustrate the *differential* impact of adding non-informative and informative summary statistics to an initial set that does not yield identification. First, consider Figure 4, which plots posteriors based only on $\boldsymbol{\eta}^1(\mathbf{y}), \boldsymbol{\eta}^6(\mathbf{y})$ and $\boldsymbol{\eta}^7(\mathbf{y})$. Adding $\eta_3(\mathbf{y})$ to

---

[6]Results for $T = 1,000$ and $T = 10,000$ were also considered. The resulting plots paint a similar picture and hence have not been included for brevity.

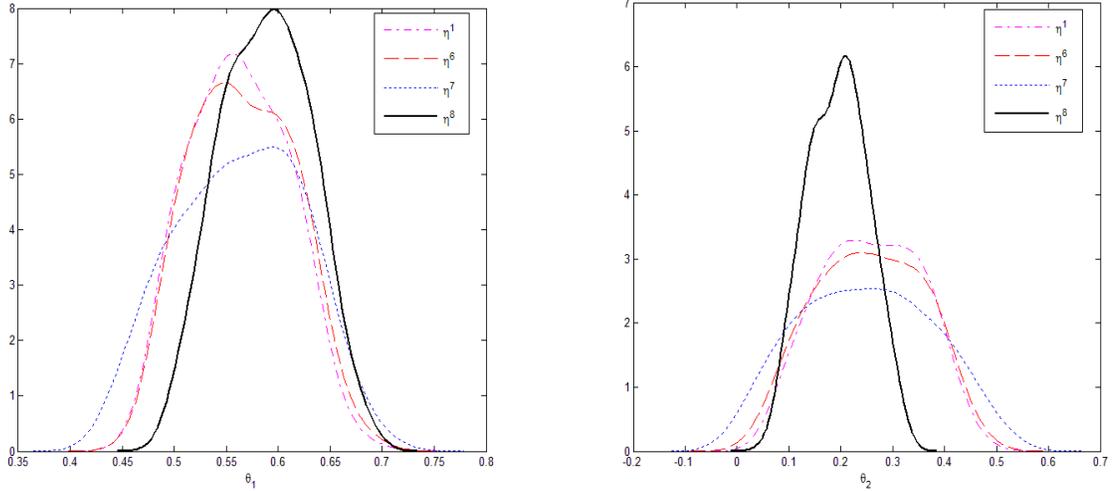**ABC marginal posteriors: MA(2) model, adding informative statistics**

Figure 5: ABC-based estimates of the marginal posterior densities for the parameters of the MA(2) model, θ_1 and θ_2, with $T = 5000$. The key for both panels indicates the summary statistic vector used: $\boldsymbol{\eta}^1(\mathbf{y}) = (\ \eta_0(\mathbf{y}), \quad \eta_1(\mathbf{y})\ )'$; $\boldsymbol{\eta}^6(\mathbf{y}) = (\boldsymbol{\eta}^1(\mathbf{y}), \eta_3(\mathbf{y}))'$; $\boldsymbol{\eta}^7(\mathbf{y}) = (\boldsymbol{\eta}^6(\mathbf{y}), \frac{1}{T}\sum_{t=1}^{T} y_t^3)'$; $\boldsymbol{\eta}^8(\mathbf{y}) = (\boldsymbol{\eta}^7(\mathbf{y}), \eta_2(\mathbf{y}))'$. The first three summary statistics do not yield Bayesian consistency. The fourth is associated with consistency and the marginal posterior estimates are shown to differ markedly from the first three as a consequence. The true parameter values are $\theta_1^0 = 0.6$ and $\theta_2^0 = 0.2$.

$\boldsymbol{\eta}^1(\mathbf{y})$ to produce $\boldsymbol{\eta}^6(\mathbf{y})$ (which we know does not yield identification) causes the estimated posterior $p_{\varepsilon_6}(\boldsymbol{\theta}|\boldsymbol{\eta}^6(\mathbf{y}))$ to flatten out compared to $p_{\varepsilon_1}(\boldsymbol{\theta}|\boldsymbol{\eta}^1(\mathbf{y}))$, and to shift slightly. Now, adding the statistic $\frac{1}{T}\sum_{t=1}^{T} y_t^3$ to $\boldsymbol{\eta}^6(\mathbf{y})$, knowing as we do that this statistic will also not aid in identification[7], the posterior $p_{\varepsilon_7}(\boldsymbol{\theta}|\boldsymbol{\eta}^7(\mathbf{y}))$ becomes even flatter (reflecting the increased dimension) and continues to shift away from the posterior mode of both previously estimated posteriors, a clear indication that we did not yet have a valid set of summary statistics on the previous rounds.

In Figure 5 we then superimpose on these three plots the estimated marginal posterior based on $\boldsymbol{\eta}^8(\mathbf{y})$, where we know that the combination of $\boldsymbol{\eta}^7(\mathbf{y})$ and $\eta_2(\mathbf{y})$ (which defines $\boldsymbol{\eta}^8(\mathbf{y})$) contains sufficient information for the parameters to now be identified. The change in the estimated posterior $p_{\varepsilon_8}(\boldsymbol{\theta}|\boldsymbol{\eta}^8(\mathbf{y}))$, relative to the existing three, is marked, with a clear peak observed around the true values, $\theta_1^0 = 0.6$ and $\theta_2^0 = 0.2$. Subsequent additions of statistics to this set will, along the lines illustrated in Figure 3, produce posteriors that now remain reasonably fixed at the same modal value and that vary only in terms of dispersion, if at all.[8]

---

[7] For the MA(2) model in (3), with $e_t \sim i.i.d.N(0,1)$, $E(y_t^3) = E(e_t + \theta_1 e_{t-1} + \theta_2 e_{t-1})^3$ is composed of four different types of moments: $E(e_{t-k}^3), E(e_{t-k}^2 e_{t-j}), E(e_{t-k}e_{t-j})$ and $E(e_{t-k}e_{t-j}e_{t-l})$, for $l \neq k \neq j$, which are all zero for any value of $\boldsymbol{\theta} = (\theta_1, \theta_2)'$.

[8] The experiments in Section 4.2 are conducted using raw distances (no component scaling). However, the results were also conducted using distances scaled by the sample covariance matrix of the summary statistics, and

The results of this section are summed up in the following remarks:

Remark 7 Corollary 1 says that if ABC based on summary statistics $\boldsymbol{\eta}(\mathbf{y})$ yields consistent inference, adding more information, in the form of additional summary statistics, will never invalidate the inference. Typically, adding further statistics can 'dull' the inference, in terms of producing a more dispersed posterior, or a posterior with slightly different shape; but it will not shift the mode. Hence, repeated augmentation of an initial choice of statistics, whereby the mode of the estimated posterior eventually 'settles' at a particular location, should instill some confidence in the mind of the investigator that consistency may have been achieved.

Remark 8 Whilst the impact of adding non-informative statistics to an initially non-informative set is likely to be problem-specific, we speculate that small and continual changes in both location and dispersion are indicative that a sufficiently informative set of summary statistics has not yet been located. A more substantial shift at some point, followed by a lack of change in the location at least, with the subsequent additions of statistics, is indicative that identification and, hence, consistency, may have been achieved. As flagged above, however, important caveats pertaining to this statement are pursued in the following section.

Remark 9 The above procedure has a similar flavor to the stepwise search algorithm proposed in Joyce and Marjoram (2008). Despite this apparent similarity however, the two procedures differ in terms of their details, as well as having very different objectives. To wit, whilst the approach outlined above is concerned with obtaining a vector of summary statistics that yield consistent inference, that of Joyce and Marjoram is concerned with obtaining a vector of summary statistics that is as informative as possible (or 'approximately sufficient' to use their terminology) for any given T.

4.3   Use of Repeated Simulation

We have demonstrated how the numerical procedure proposed above can determine with some certainty whether or not $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ - for some choice of $\boldsymbol{\eta}(\mathbf{y})$ - is concentrating at $\boldsymbol{\theta}^0$, in the artifical scenario in which $\boldsymbol{\theta}^0$ is known. In practice of course, the true value $\boldsymbol{\theta}^0$ is unknown, and the proposed method is not capable of distinguishing between $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ concentrating at $\boldsymbol{\theta}^0$ and $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ concentrating at some other value $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^0$, satisfying $\mathbf{b}(\boldsymbol{\theta}^*) = \mathbf{b}(\boldsymbol{\theta}^0)$. However, if the binding function is one-to-one, perverse situations such as the above can be ruled out.

For a fixed (vector) summary statistic, $\boldsymbol{\eta}(\mathbf{y})$, it turns out that verifying whether or not the binding function is one-to-one is, in principle, possible. To understand how we can verify this

---

with individual elements scaled by their simulated variance. Results based on these alternative scaling measures are not qualitatively different from those presented herein. The results are available from the authors upon request.

condition, first recall that $\mathbf{b}(\boldsymbol{\theta})$ is simply the limit, as $T \to \infty$, of the simulated summary statistics $\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))$, and note that because $\mathbf{z}(\boldsymbol{\theta})$ is simulated from the structural model, $\mathbf{z}(\boldsymbol{\theta})$ is no longer restricted to be of the same length as the observed sample $\mathbf{y}$. From these facts, we see that our ability to obtain $\mathbf{b}(\boldsymbol{\theta})$ is limited only by computational power and time; i.e., we are limited only by our ability to simulate (very) long trajectories for $\mathbf{z}(\boldsymbol{\theta})$. In addition, the entire map $\boldsymbol{\theta} \mapsto \mathbf{b}(\boldsymbol{\theta})$ can be obtained simply by simulating long trajectories of $\mathbf{z}(\boldsymbol{\theta})$, forming $\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))$, and repeating the exercise at every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Therefore, with enough computing power (and time), it is theoretically possible to verify whether or not $\boldsymbol{\theta} \mapsto \mathbf{b}(\boldsymbol{\theta})$ is one-to-one.

While the above logic demonstrates that it is theoretically possible to verify the one-to-one condition, it is not practically possible as this approach (technically) requires simulating an infinite number of infinite series. However, when the data is stationary and the parameter space relatively small, we can approximately check this condition through the following steps:

Algorithm 4 One-to-one verification

1: Use the empirical procedure in Section 4.2 to identify a (vector) summary statistic of interest, hereafter denoted $\boldsymbol{\eta}(\cdot)$.
2: Choose K^* distinct parameter values with which to simulate data from the structural model, call them $\boldsymbol{\theta}^{0,1}, ..., \boldsymbol{\theta}^{0,K^*}$. Choose a large integer $T^* >> T >> 0$.
3: Simulate $\widetilde{\mathbf{z}}^k = (z_1(\boldsymbol{\theta}^{0,k}), ..., z_{T^*}(\boldsymbol{\theta}^{0,k}))'$ and form the series $\left\{\boldsymbol{\eta}(\widetilde{\mathbf{z}}^k)\right\}_{k=1}^{K^*}$; $\left\{\boldsymbol{\eta}(\widetilde{\mathbf{z}}^k)\right\}_{k=1}^{K^*}$ constitutes a discrete approximation to $\boldsymbol{\theta} \mapsto \mathbf{b}(\boldsymbol{\theta})$
4: Determine whether or not $\left\{\boldsymbol{\eta}(\widetilde{\mathbf{z}}^k)\right\}_{k=1}^{K^*}$ contains K^* unique elements.

For K^* and T^* large enough, if $\boldsymbol{\eta}(\cdot)$ satisfies Step (4), *and* if the estimated $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$, as based on the observed data, $\mathbf{y}$, is also collapsing toward some point, one should conclude in favor of consistency. Of course, this process leaves much left unspecified, with the most critical issues being how to span the parameter space, how to selected the set of pre-specified statistics, and the order in which they are to be explored, plus the manner in which degeneracy of the estimated posterior is tested for as T is allowed to increase. However, providing guidelines for and proving the theoretical properties of any such search procedure would require several layers of formalization and the introduction of new terms and concepts that would detract from the current message of the paper. Hence, at this stage we simply emphasize that a completely satisfactory assessment of consistency would appear to require both the use of the observed data and repeated application of data from the assumed process; and suggest that the sort of exercise we are proposing here, albeit informal, is a sensible one to pursue.

5 Criterion Functions based on an Auxiliary Model

5.1 Consistency of Auxiliary Model-based ABC

At a minimum, implementation of ABC requires some means of generating summary statistics $\boldsymbol{\eta}(\mathbf{y})$ that are "informative" about the unknown parameters of the underlying structural model; whereby "informative" it is generally meant that the summary statistics are a useful way of characterizing the information contained in the observed data. However, the previous sections demonstrate that care must be taken to ensure that the chosen summary statistics yield consistent inference.

An alternative way of obtaining informative summary statistics is through the use of an auxiliary model that depends on parameters $\boldsymbol{\beta} \in \mathbf{B} \subset \mathbb{R}^{d_\beta}$, where $d_\beta \geq p = \dim(\boldsymbol{\Theta})$, and for which the likelihood function of the auxiliary model, denoted by $L(\mathbf{y}; \boldsymbol{\beta})$, is known in closed form. Given a simple auxiliary likelihood $L(\mathbf{y}; \boldsymbol{\beta})$, a growing literature suggests using summary statistics generated from $L(\mathbf{y}; \boldsymbol{\beta})$; for example, one can choose $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y})$, where $\widehat{\boldsymbol{\beta}}(\mathbf{y}) = \arg\max_{\beta \in \mathbf{B}} L(\mathbf{y}; \boldsymbol{\beta})$, or $\boldsymbol{\eta}(\mathbf{y})$ equivalent to the vector score of $L(\mathbf{y}; \boldsymbol{\beta})$ evaluated at $\widehat{\boldsymbol{\beta}}(\mathbf{y})$. However, by its very nature the auxiliary model, and by proxy the summary statistics derived from $L(\mathbf{y}; \boldsymbol{\beta})$, is (are) likely to describe only *certain* salient features of the underlying structural model. In particular, there is generally no reason to believe that the auxiliary model should "nest" the true structural model in some well-defined sense. Indeed, if it does so then this suggests either that the structural model itself is tractable - hence excluding the need for ABC - or that the nesting model is highly parameterized, thereby inducing a $\boldsymbol{\eta}(\cdot)$ of high dimension and the associated problems for accuracy.

Given then that a typical auxiliary model is capable of representing only certain salient features of the DGP, there is nothing particularly special about choosing the auxiliary *likelihood function* to generate summary statistics for use within ABC. Moreover, in many cases a realistic auxiliary model may yield a likelihood function that is itself too complicated for ABC, from a purely computational standpoint, whilst an alternative criterion function, based on the same auxiliary model, may yield computationally simpler summary statistics. For example, alternative criterion functions - other than an auxiliary likelihood - that could be used inside an ABC algorithm include: sums of squared errors, least absolute deviations, and even quadratic functions of sample moments (conditional and unconditional) from an auxiliary model, with the latter used to define an MSM-type of approach, but with moments of the auxiliary rather than the true model defining the selection mechanism.

However, as in the previous section, conditions need to be placed on the relevant criterion function to ensure the resultant ABC procedure yields consistent inference. This is the content of Theorem 2. Begin by defining a sample criterion function based on observed data $\mathbf{y}$ (respectively, simulated data $\mathbf{z}^i = \mathbf{z}(\boldsymbol{\theta}^i)$) $Q(\mathbf{y}; \boldsymbol{\beta})$ (respectively, $Q(\mathbf{z}^i; \boldsymbol{\beta})$) and define $\widehat{\boldsymbol{\beta}}(\mathbf{y})$ (respectively, $\widehat{\boldsymbol{\beta}}(\mathbf{z}^i)$)

as the minimizer of $Q(\mathbf{y}; \boldsymbol{\beta})$ (respectively, $Q(\mathbf{z}^i; \boldsymbol{\beta})$). For a particular choice of $Q(\cdot; \boldsymbol{\beta})$ an ABC algorithm could be based on the summary statistics $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y})$, $\boldsymbol{\eta}(\mathbf{z}^i) = \widehat{\boldsymbol{\beta}}(\mathbf{z}^i)$.

The above intuition yields ABC Algorithm 5 based on generic criterion $Q(.; \boldsymbol{\beta})$:

Algorithm 5 ABC algorithm: auxiliary criterion function

1: Obtain $\widehat{\boldsymbol{\beta}}(\mathbf{y}) = \arg\min_{\beta \in \mathbf{B}} Q(\mathbf{y}; \boldsymbol{\beta})$,
2: Simulate $\boldsymbol{\theta}^i$, $i = 1, 2, ..., N$, from $p(\boldsymbol{\theta})$
3: Simulate $\mathbf{z}^i = (z_1^i, z_2^i, ..., z_T^i)'$, $i = 1, 2, ..., N$, from the likelihood, $p(.|\boldsymbol{\theta}^i)$
4: Select $\boldsymbol{\theta}^i$ such that:
$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\} = d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \widehat{\boldsymbol{\beta}}(\mathbf{z}^i)\} \leq \varepsilon, \tag{6}$$
where $d\{\cdot, \cdot\}$ is a distance function (or metric), and the tolerance level ε is chosen as small as the computing budget allows.

Denote the posterior obtained from the above algorithm as $p_\varepsilon^Q(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$. The following result gives conditions under which $\Pr_\varepsilon^Q(\boldsymbol{\theta} \notin \aleph_\delta(\boldsymbol{\theta}^0)|\boldsymbol{\eta}(\mathbf{y})) \xrightarrow{P} 0$ as $T \to \infty$ and $\varepsilon \to 0$.

Theorem 2 *For an auxiliary model with parameters $\boldsymbol{\beta} \in \mathbf{B}$, $\mathbf{B} \subset \mathbb{R}^{d_\beta}$ compact, assume that the following are satisfied:*

[G1] *There exists a deterministic limit criterion function $Q_\infty(\boldsymbol{\theta}^i; \boldsymbol{\beta})$ such that*

> **[G1(1)]** $Q_\infty(\boldsymbol{\theta}^i; \boldsymbol{\beta})$ *is continuous as a function of $\boldsymbol{\beta}$, uniformly in $\boldsymbol{\theta}^i$.*
>
> **[G1(2)]** $\sup_{\beta \in \mathbf{B}} |Q(\mathbf{y}; \boldsymbol{\beta}) - Q_\infty(\boldsymbol{\theta}^0; \boldsymbol{\beta})| = o_P(1)$ *and* $\sup_{\theta \in \Theta, \beta \in \mathbf{B}} |Q(\mathbf{z}(\boldsymbol{\theta}); \boldsymbol{\beta}) - Q_\infty(\boldsymbol{\theta}; \boldsymbol{\beta})| = o_P(1)$

[G2] $Q_\infty(\boldsymbol{\theta}^i; \boldsymbol{\beta})$ *has a unique minimum $\mathbf{b}(\boldsymbol{\theta}^i)$ for all $\boldsymbol{\theta}^i \in \Theta$; i.e., for all $\boldsymbol{\theta}^i \in \Theta$, $\mathbf{b}(\boldsymbol{\theta}^i) := \arg\min_{\beta \in \mathbf{B}} Q_\infty(\boldsymbol{\theta}^i; \boldsymbol{\beta})$ and $\boldsymbol{\beta}^0 := \mathbf{b}(\boldsymbol{\theta}^0)$.*

[G3] $\mathbf{b}(\cdot)$ *is one-to-one in $\boldsymbol{\theta}^i$; i.e., $\boldsymbol{\beta} = \mathbf{b}(\boldsymbol{\theta}^i)$ has a unique solution for all $\boldsymbol{\theta}^i \in \Theta$.*

If [G1]-[G3] and [P] are satisfied $\Pr_\varepsilon^Q(\boldsymbol{\theta} \notin \aleph_\delta(\boldsymbol{\theta}^0)|\boldsymbol{\eta}(\mathbf{y})) \to_p 0$ as $T \to \infty$ and $\varepsilon \to 0$.

Remark 10 The above result states that, so long as $Q(.; \boldsymbol{\beta})$ satisfies standard properties ([G1], [G2]), and if the so-called binding function $\mathbf{b}(\cdot)$ is one-to-one, an ABC algorithm that uses as summary statistics the minimizers of $Q(.; \boldsymbol{\beta})$ will yield a posterior that is degenerate at $\boldsymbol{\theta}^0$. For a specific objective function, conditions [G1] and [G2] are generally satisfied under more primitive conditions; see, for example, Jennrich (1969) in the setting where $Q(.; \boldsymbol{\beta})$ is the nonlinear least squares criterion, and Newey and McFadden (1994) in the case where $Q(.; \boldsymbol{\beta})$ is a minimum distance criterion. While the result of Theorem 2 is intuitive it is nonetheless important as it illustrates that we are not confined to using simple summary statistics of the data or the log-likelihood function $L(\mathbf{y}; \boldsymbol{\beta})$ of the auxiliary model within ABC. Instead, any criterion function satisfying [G1]-[G3] can be used to generate valid summary statistics for use in ABC.

Remark 11 An alternative to Algorithm 5 is to replace the summary statistics $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y})$, $\boldsymbol{\eta}(\mathbf{z}^i) = \widehat{\boldsymbol{\beta}}(\mathbf{z}^i)$ in Step (3) with a distance based on $(\partial/\partial\beta)Q(\mathbf{z}^i; \widehat{\boldsymbol{\beta}}(\mathbf{y}))$; e.g.,

$$d\left\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\right\} = \sqrt{\left\{(\partial/\partial\boldsymbol{\beta})Q(\mathbf{z}^i; \widehat{\boldsymbol{\beta}}(\mathbf{y}))\right\}' \widehat{\Omega}(\mathbf{y}) \left\{(\partial/\partial\boldsymbol{\beta})Q(\mathbf{z}^i; \widehat{\boldsymbol{\beta}}(\mathbf{y}))\right\}}, \qquad (7)$$

for some positive definite weighting matrix $\widehat{\Omega}(\mathbf{y})$. Such an algorithm would be quite useful in situations where $(\partial/\partial\beta)Q(\mathbf{z}^i; \widehat{\boldsymbol{\beta}}(\mathbf{y}))$ is known in closed form and would (in all cases) lead to an ABC algorithm that is several orders of magnitude faster than one based on computing $\widehat{\boldsymbol{\beta}}(\mathbf{z}^i)$ at every value $\boldsymbol{\theta}^i$. Under conditions similar to those in Theorem 2, a consistency result will hold for the posterior obtained from an ABC algorithm that uses the distance measure in (7). We omit this proof for brevity.

5.2 The Role of the Auxiliary Model

Intimately tied to the idea of choosing a suitable criterion function is the choice of the auxiliary model from which the criterion function is computed. If the chosen auxiliary model is a poor representation of the observed data it is likely that no criterion function, likelihood or otherwise, will produce adequate summary statistics upon which to base our ABC algorithm. In this way using summary statistics from an auxiliary model inside of ABC is not a panacea.

ABC algorithms based on an auxiliary model and with summary statistics derived from a criterion function $Q(.;\boldsymbol{\beta})$ can fail for precisely the same reason ABC based on arbitrary summary statistics can fail, namely, failure of [G3] (respectively [S2(2)]). Satisfaction of [G3] is affected by both the choice of the auxiliary model and the subsequent criterion function used to obtain $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y})$. Since the choice of auxiliary model and criterion $Q(.;\boldsymbol{\beta})$ are user and example specific, attempting to give hard and fast guidelines for how one should choose either is a research topic in its own right. Rather, we simply advocate that validation of [G1]-[G3] should at least be attempted for any specified combination (of model and criterion function) before implementing an ABC algorithm. In the following example we provide support for this statement by illustrating a case in which consistency is not yielded via what seems to be a sensible ABC specification: namely the use of an AR(2) auxiliary model along with an OLS criterion function to produce inference about the true parameters of a MA(2) model.

Example 4 *Consider again the MA(2) model from Example 2. Instead of a summary statistic based ABC approach, consider implementing ABC using summary statistics generated via the OLS criterion function for the AR(2) auxiliary model: $y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \nu_t$, with $\nu_t \sim (0, 1)$. Using*

$$Q(\mathbf{y}; \boldsymbol{\beta}) = \frac{1}{T}\sum_{t=3}^{T}(y_t - \beta_1 y_{t-1} - \beta_2 y_{t-2})^2, \qquad (8)$$

22

and OLS estimator $\widehat{\boldsymbol{\beta}}(\mathbf{y}) = (\widehat{\beta}_1(\mathbf{y}), \widehat{\beta}_2(\mathbf{y}))'$, the summary statistic $\boldsymbol{\eta}(\mathbf{y}) = (\widehat{\beta}_1(\mathbf{y}), \widehat{\beta}_2(\mathbf{y}))'$, which has a simple closed form, can be used to build a computationally simple ABC algorithm.

Given the particular structure of $Q(\cdot; \boldsymbol{\beta})$ in (8) and under conditions (4) for $\boldsymbol{\theta}$, [G1] and [G2] are satisfied. Therefore, all that remains is to verify [G3]. Differentiating the limit criterion $Q_\infty(\boldsymbol{\theta}^i; \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ yields the following equations

$$
\begin{aligned}
E(y_{t-1}(z_t^i - \beta_1 z_{t-1}^i - \beta_2 z_{t-2}^i)) = 0, \\
E(y_{t-2}(z_t^i - \beta_1 z_{t-1}^i - \beta_2 z_{t-2}^i)) = 0.
\end{aligned} \tag{9}
$$

Defining the autocovariances based on $\boldsymbol{\theta}^i$ as $\gamma_0^i = (1 + (\theta_1^i)^2 + (\theta_2^i)^2), \gamma_1^i = (\theta_1^i + \theta_1^i \theta_2^i)$ and $\gamma_2^i = \theta_2^i$, we can re-write (9) as

$$
\begin{aligned}
\gamma_1^i - \beta_1 \gamma_0^i - \beta_2 \gamma_1^i = 0, \\
\gamma_2^i - \beta_1 \gamma_1^i - \beta_2 \gamma_0^i = 0.
\end{aligned} \tag{10}
$$

Solving for β_1, β_2 in (10) yields the following:

$$
\beta_1(\boldsymbol{\theta}^i) := \left[\gamma_1^i - \left(\frac{\gamma_1^i \gamma_2^i}{\gamma_0^i} \right) \right] \Big/ \left[\gamma_0^i - \left(\frac{(\gamma_1^i)^2}{\gamma_0^i} \right) \right], \beta_2(\boldsymbol{\theta}^i) := \frac{\gamma_2^i}{\gamma_0^i} - \frac{\gamma_1^i}{\gamma_0^i} \beta_1(\boldsymbol{\theta}^i).
$$

Interestingly, and as an illustration of the point made in Section 4.1, the binding function $\mathbf{b}(\boldsymbol{\theta}^i) = (\beta_1(\boldsymbol{\theta}^i), \beta_2(\boldsymbol{\theta}^i))'$ does not admit a unique solution to $0 = \mathbf{b}(\boldsymbol{\theta}^0) - \mathbf{b}(\boldsymbol{\theta}^i)$, for all $\boldsymbol{\theta}^0, \boldsymbol{\theta}^i \in \Theta$. For instance, simple numerical calculations reveal that if $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0)' = (.6, .2)'$ the equation $0 = \mathbf{b}(\boldsymbol{\theta}^0) - \mathbf{b}(\boldsymbol{\theta}^i)$ has a unique solution satisfying the conditions of (4), namely $\boldsymbol{\theta}^i = \boldsymbol{\theta}^0$ (a second solution, $\boldsymbol{\theta}^i = (3, 5)'$, exists but does not satisfy the parameter restrictions (4)). However, if $\boldsymbol{\theta}^0 = (.5, .5)'$, the equation $0 = \mathbf{b}(\boldsymbol{\theta}^0) - \mathbf{b}(\boldsymbol{\theta}^i)$ has two solutions satisfying the conditions of (4), $\boldsymbol{\theta}^i = (.5, .5)'$ and $\boldsymbol{\theta}^i = (1, 2)$! Therefore, $\mathbf{b}(\boldsymbol{\theta}^i) = (\beta_1(\boldsymbol{\theta}^i), \beta_2(\boldsymbol{\theta}^i))'$ is not a one-to-one function and hence will not yield consistent inference in general.

6 Consistency of ABC in Ordinary Differential Equations Models

In this section we investigate the ability of ABC to yield Bayesian consistent inference for parameters governing a system of ordinary differential equations (ODEs). As will be demonstrated, this particular type of application, which has been given some attention in the ABC literature (see, for example, Toni *et al.*, 2009, Sun *et al.*, 2014, Prangle, 2015), highlights certain important issues related to Bayesian consistency of ABC-based posterior estimates. In particular, by checking the conditions of Theorem 2 in a simple deterministic system, we demonstrate that ABC can yield inconsistent inference in such settings, highlighting the importance of these conditions for verifying the validity of ABC-based inference. While we specifically focus on a simple deterministic system, these findings can easily be generalized to other ODEs.

Specifically, we give our attention to the Lotka-Volterra (LV) model, which describes the interaction between a species x_1, referred to as the prey species, and a species x_2, referred to as

the predator species. For $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ unknown, we consider the deterministic LV model defined through the system of ODEs:

$$\frac{dx_1}{dt} = \theta_1 x_1 - x_1 x_2, \tag{11}$$
$$\frac{dx_2}{dt} = \theta_2 x_1 x_2 - x_2.$$

For any point t_i in the interval $[0, T]$, the vector $\mathbf{x}(t_i) = (x_1(t_i), x_2(t_i))'$ is the solution to the above ODEs, with initial value t_0. Typically, it is assumed that we do not observe $\mathbf{x}(t_i)$; rather, we observe a quantity corresponding to $\mathbf{x}(t_i)$ that is measured with error that is both additive and independent over observational points, see, for example, Beck and Arnold (1977). Following this usual practice then, we specify a measurement equation of the form

$$\mathbf{y}(t_i) = \mathbf{x}(t_i) + \boldsymbol{\nu}(t_i), \tag{12}$$

where $\boldsymbol{\nu}(t_i) \sim i.i.d.(0, \Sigma_v)$ and Σ_v is diagonal.[9]

Assume we have an observed sample of size R_T from (12), with corresponding design points $t_1, ..., t_{R_T}$, fixed or random. Our goal is to estimate the posterior density of $\boldsymbol{\theta}$ using the observed sample $\{\mathbf{y}(t_i)\}_{i=1}^{R_T}$, and prior density $p(\boldsymbol{\theta})$. Toni *et al.* (2009) propose to estimate these posteriors via ABC using the squared distance between the observed and simulated samples. Specifically, for $\{\mathbf{y}(t_i)\}_{i=1}^{R_T}$ the observed sample and $\{\mathbf{z}(t_i; \boldsymbol{\theta})\}_{i=1}^{R_T}$ the simulated sample, obtained by solving equation (11) at $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, ABC is based on the distance

$$\rho\{\mathbf{y}, \mathbf{z}(\boldsymbol{\theta})\} = \frac{1}{R_T} \sum_{j=1}^{2} \sum_{i=1}^{R_T} (y_j(t_i) - z_j(t_i; \boldsymbol{\theta})))^2. \tag{13}$$

That is, draws of $\boldsymbol{\theta}$ are retained according to the proximity of the stochastic quantity $y_j(t_i)$ to the deterministic quantity $z_j(t_i; \boldsymbol{\theta})$.

It is critical to note, however, that choosing values of $\boldsymbol{\theta}^i$ such that $\rho\{\mathbf{y}, \mathbf{z}(\boldsymbol{\theta}^i)\} \leq \varepsilon$ will not yield an ABC-based posterior that is Bayesian consistent. This can be seen by noting that as $R_T \to \infty$, even if we select $\boldsymbol{\theta}^i = \boldsymbol{\theta}^0$, and so $x_j(t_i) = z_j(t_i; \boldsymbol{\theta}^0)$ for all t_i and $j = 1, 2$, it will be the case that $\lim_{R_T \to \infty} \rho\{\mathbf{y}, \mathbf{z}(\boldsymbol{\theta}^0)\} \xrightarrow{P} E(\nu_1^2(t_i) + \nu_2^2(t_i)) > \varepsilon$, for ε arbitrarily small. Therefore, there exists no value of $\boldsymbol{\theta}^i \in \boldsymbol{\Theta}$ for which $\rho\{\mathbf{y}, \mathbf{z}(\boldsymbol{\theta}^i)\} \leq \varepsilon$ as $R_T \to \infty$ and $\varepsilon \to 0$, and so the ABC-based posterior defined by the distance $\rho\{\mathbf{y}, \mathbf{z}(\boldsymbol{\theta})\}$ can not be Bayesian consistent.

However, an alternative to the "distance" $\rho\{\mathbf{y}, \mathbf{z}(\boldsymbol{\theta})\}$ in (13) is a metric based on statistics obtained from minimizing an objective function representing the data in equation (12). A common means of obtaining (frequentist) point estimates for parameters defined by ODEs is

---

[9]It could be assumed that the evolution of x_1 and x_2 is stochastic rather than deterministic; however the adoption (or not) of this assumption is not germane to our discussion and we thus retain the ODE structure for the states.

nonlinear least squares (NLS), whereby the squared distance between the observed and simulated solutions is minimized (see Beck and Anrnold, 1977, for a discussion). This then motivates us to consider the consistency properties of an ABC method that mimics the spirit of NLS. As such, we consider as summary statistics for use in ABC, the parameters that minimize the ordinary least squares (OLS) criterion

$$Q(\mathbf{y}; \boldsymbol{\beta}) = \frac{1}{R_T} \sum_{j=1}^{2} \frac{1}{\beta_{2,j}} \sum_{i=1}^{R_T} (y_j(t_i) - \beta_{1,j})^2,$$

with respect to $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, $\boldsymbol{\beta}_1 = (\beta_{1,1}, \beta_{1,2})'$ and $\boldsymbol{\beta}_2 = (\beta_{2,1}, \beta_{2,2})'$, which defines $\widehat{\beta}_{1,j} = \frac{1}{R_T} \sum_{i=1}^{R_T} y_j(t_i)$ as the sample mean and $\widehat{\beta}_{2,j} = \frac{1}{R_T} \sum_{i=1}^{R_T} (y_j(t_i) - \widehat{\beta}_{1,j})^2$ as the sample variance. ABC can then be conducted using $\widehat{\boldsymbol{\beta}}(\mathbf{y}) = (\widehat{\boldsymbol{\beta}}_1'(\mathbf{y}), \widehat{\boldsymbol{\beta}}_2'(\mathbf{y}))'$ and its simulated counterpart $\widehat{\boldsymbol{\beta}}(\mathbf{z}^i) = (\widehat{\boldsymbol{\beta}}_1'(\mathbf{z}^i), \widehat{\boldsymbol{\beta}}_2'(\mathbf{z}^i))'$, with a distance of the form specified in (6) adopted. Further alternatives can be defined by basing ABC on matching $\widehat{\boldsymbol{\beta}}_1(\mathbf{y})$ alone (respectively, $\widehat{\boldsymbol{\beta}}_2(\mathbf{y})$) with its simulated counterpart $\widehat{\boldsymbol{\beta}}_1(\mathbf{z}^i)$ (respectively, $\widehat{\boldsymbol{\beta}}_2(\mathbf{z}^i)$), with the use of $\widehat{\boldsymbol{\beta}}_1(\cdot)$ alone as the matching statistic being closest in spirit to NLS.

Sufficient conditions guaranteeing that ABC will yield consistent inference are given in Theorem 2 and must be verified, for each version of $\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)$ obtained from $Q(\cdot; \boldsymbol{\beta})$. Whilst formal verification of the identification condition in this case is complicated by the fact that $\mathbf{x}(t_i)$ has no closed form, some analytical insights are attainable, by noting the following. Define $\bar{x}_j = \lim_{R_T \to \infty} \frac{1}{R_T} \sum_{i=1}^{R_T} x_j(t_i)$, $\bar{x}_j^2 = \lim_{R_T \to \infty} \frac{1}{R_T} \sum_{i=1}^{R_T} x_j^2(t_i)$, and corresponding simulated counterparts $\bar{z}_j(\boldsymbol{\theta}) = \lim_{R_T \to \infty} \frac{1}{R_T} \sum_{i=1}^{R_T} z_j(t_i; \boldsymbol{\theta})$, $\bar{z}_j^2(\boldsymbol{\theta}) = \lim_{R_T \to \infty} \frac{1}{R_T} \sum_{i=1}^{R_T} z_j^2(t_i; \boldsymbol{\theta})$, for $j = 1, 2$. Assuming these quantities exist, it can be shown that

$$\lim_{R_T \to \infty} \widehat{\beta}_{2,j}(\mathbf{y}) \xrightarrow{P} \bar{x}_j^2 + \sigma_j^2 - (\bar{x}_j)^2,$$

where σ_j^2 is the (j, j) element of Σ_v. However, it is also the case that

$$\lim_{R_T \to \infty} \widehat{\beta}_{2,j}(\mathbf{z}(\boldsymbol{\theta})) \xrightarrow{P} \bar{z}_j^2(\boldsymbol{\theta}) - (\bar{z}_j(\boldsymbol{\theta}))^2.$$

Hence, as $R_T \to \infty$

$$\left\| \widehat{\boldsymbol{\beta}}_2(\mathbf{y}) - \widehat{\boldsymbol{\beta}}_2(\mathbf{z}(\boldsymbol{\theta}^0)) \right\| \neq o_P(1)$$

and so there is no hope that ABC based on $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}_2(\mathbf{y})$ will yield consistent inference. That is, and reverting to the general notation of the previous section, $Q(\mathbf{y}; \boldsymbol{\beta})$ and $Q(\mathbf{z}(\boldsymbol{\theta}^0); \boldsymbol{\beta})$ do not have corresponding limit $Q_\infty(\boldsymbol{\theta}^0; \boldsymbol{\beta})$, which violates Assumption [G1] of Theorem 2. This same point pertains to the case in which the augmented statistic $\widehat{\boldsymbol{\beta}}(\mathbf{y})$ is used.

The critical insight from the above illustration, as it pertains to ABC, is that a mismatch between the assumed processes for the observed and simulated data, with the latter failing to replicate the stochastic nature of the former, can create a fundamental disconnect between

the matching statistics formed from the two sets of data, so that they will never coincide, no matter the proximity of the drawn parameter vector to the truth. We now contrast this with an alternative approach in which we deliberately draw simulated data according to the measurement equation

$$\mathbf{z}(t_i; \boldsymbol{\theta}) = \mathbf{w}(t_i; \boldsymbol{\theta}) + \tilde{\boldsymbol{\nu}}(t_i), \tag{14}$$

where $\mathbf{w}(t_i; \boldsymbol{\theta})$ is the numerical solution of the ODE at parameter value $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\nu}}(t_i)$ is a random error drawn from the same distribution as $\boldsymbol{\nu}(t_i)$. In this case, it is easy to verify that the simulated statistics $\widehat{\boldsymbol{\beta}}_2(\mathbf{z}^i)$ and $\widehat{\boldsymbol{\beta}}(\mathbf{z}^i)$ depend on the measurement error variance Σ_v in the same manner as the observed data, with Assumption [G1] of Theorem 2 no longer violated as a consequence, and so, as $R_T \to \infty$

$$\left\| \widehat{\boldsymbol{\beta}}_2(\mathbf{y}) - \widehat{\boldsymbol{\beta}}_2(\mathbf{z}(\boldsymbol{\theta}^0)) \right\| = o_P(1), \text{ and } \left\| \widehat{\boldsymbol{\beta}}(\mathbf{y}) - \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta}^0)) \right\| = o_P(1).$$

Once again, since no closed form solution exists for the state process, establishing the identification condition analytically, as in the previous examples, is not feasible. However, numerical exploration indicates the existence of consistency for matching statistics $\widehat{\boldsymbol{\beta}}_1(y)$, $\widehat{\boldsymbol{\beta}}_2(y)$ and $\widehat{\boldsymbol{\beta}}(y)$ when data is simulated according to (14).[10]

It has generally been recognized that consistent inference for point estimates of parameters in ODEs is due to the additive nature of the measurement error, combined with the fact that the measurement error has mean zero, known variance, and is independent of the data, as well as the satisfaction of identification conditions guaranteeing the existence of a unique minimum at $\boldsymbol{\theta}^0$; see, for example, Beck and Arnold (1977). However, when conducting inference for ODEs via ABC, we see that in addition to these conditions (or variants thereof), care must be taken to ensure that data is simulated in such a way that it matches the observed data. It is the price we pay for conducting complete inference using a simulation-based procedure.

7   Discussion

Consistency is one of the most fundamental properties with which to gauge the output of a statistical inference procedure. With our focus on Bayesian consistency, we demonstrate that in the limit (as both $T \to \infty$ and $\varepsilon \to 0$) the ABC posterior estimate will be degenerate at the true parameter (vector) if (and only if!) the summary statistics upon which ABC is based are appropriately chosen. Conditions guaranteeing Bayesian consistency of ABC posterior estimates for a wide range of summary statistics, with and without respect to an auxiliary model, with the former defined with respect to an arbitrary criterion function, are developed and several examples featured in the literature are used to illustrate these conditions. The results are less heartening than expected and demonstrate that consistent inference in ABC is in no way guaranteed. In

---

[10]Numerical results illustrating consistency are available from the authors upon request.

general, we find that ABC will only yield consistent inference when a judicious choice of summary statistics has been employed, subsequently calling into question a large collection of ABC results based on arbitrary summary statistics, as well as those generated from well-specified auxiliary models. In addition, our results highlight the need both to specify a proper distance measure and to ensure an exact match between the process assumed to have generated the observed data and that used to produce simulated samples, in order to have any hope of yielding consistent inference.

To determine if ABC will be Bayesian consistent in practice, we develop a useful and computationally simple diagnostic procedure that can be applied to any given data set and any choice of summary statistics. This procedure constitutes an important first step in determining, in any practical situation, whether ABC will yield consistent inference. Formalization of this diagnostic procedure, as well as work detailing its theoretical properties, is a topic of ongoing research by the authors.

Before closing, we re-emphasize the fact that the results presented herein, while cast within the framework of the ABC accept/reject algorithm, apply to the more sophisticated variants of the ABC method. In particular, the results are applicable to ABC algorithms that generate summary statistics through various simulation-based approximations, as well as algorithms that utilize more efficient methods of post-sampling density estimation. Given this fact, the results discussed herein can be used to form the basic foundation for determining Bayesian consistency for all summary statistic-based ABC algorithms. Further, the key issue that we have emphasized throughout, namely the need to verify the relevant conditions for Bayesian consistency, including the required one-to-one property of the (implied) binding function, is just as pertinent, of course, to related frequentist simulation-based inference methods. In particular, the development of a formal and rigorous method for confirming the one-to-one property of the binding function is as critical to the establishment of the (asymptotic) validity of all other such methods as it is to ABC.

8 Appendix: Proofs

Proof of Theorem 1.   The proof is broken into three parts: first, we show that the only value $\boldsymbol{\theta}^i$ that will be selected for all $\varepsilon \geq 0$ as $T \to \infty$ is $\boldsymbol{\theta}^i = \boldsymbol{\theta}^0$; second, we demonstrate that for any $\varepsilon > 0$ there exists some $N(\varepsilon)$ such that if $N > N(\varepsilon)$ the posterior density $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ has a well-defined probability limit; lastly, we use these two pieces to demonstrate that for any $\aleph_\delta(\boldsymbol{\theta}^0)$ and $\Lambda_\delta := \Theta/\aleph_\delta(\boldsymbol{\theta}^0)$, the posterior probability $\Pr_\varepsilon(\boldsymbol{\theta} \in \Lambda_\delta|\boldsymbol{\eta}(\mathbf{y})) \xrightarrow{P} 0$.

Part 1:

By the triangle inequality

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\} \leq d\{\boldsymbol{\eta}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^i)\} + d\{\mathbf{b}(\boldsymbol{\theta}^i), \boldsymbol{\eta}(\mathbf{z}^i)\}. \tag{15}$$

Applying the triangle inequality again to the first term on the right-hand-side of (15) yields

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\} \leq d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} + d\{\boldsymbol{\eta}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^0)\} + d\{\mathbf{b}(\boldsymbol{\theta}^i), \boldsymbol{\eta}(\mathbf{z}^i)\}.$$

By [S1], $\boldsymbol{\eta}(\mathbf{y}) \overset{P}{\to} \mathbf{b}(\boldsymbol{\theta}^0)$ and so $d\{\boldsymbol{\eta}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^0)\} = o_P(1)$. In addition,

$$d\{\boldsymbol{\eta}(\mathbf{z}^i), \mathbf{b}(\boldsymbol{\theta}^i)\} \leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} d\{\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta})), \mathbf{b}(\boldsymbol{\theta})\},$$

and by [S2(1)] $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} d\{\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta})), \mathbf{b}(\boldsymbol{\theta})\} = o_P(1)$. Combining these facts yields

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\} \leq d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} + o_P(1). \tag{16}$$

For fixed $\varepsilon > 0$, as $T \to \infty$ a value $\boldsymbol{\theta}^i$ will be selected if

$$d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} + o_P(1) \leq \varepsilon.$$

By [S2(2)] the only value of $\boldsymbol{\theta}^i \in \boldsymbol{\Theta}$ for which $\mathbf{b}(\boldsymbol{\theta}^i) = \mathbf{b}(\boldsymbol{\theta}^0)$ is $\boldsymbol{\theta}^i = \boldsymbol{\theta}^0$. Therefore, as $T \to \infty$, the only value of $\boldsymbol{\theta}^i$ satisfying $d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} \leq \varepsilon$ for any $\varepsilon \geq 0$ is $\boldsymbol{\theta}^i = \boldsymbol{\theta}^0$.

Part 2:

Part 1 suggests that for small enough ε the posterior density $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ will be zero for values $\boldsymbol{\theta} \in \Lambda_\delta$ as $T \to \infty$. However, because $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ is built from N random draws, for any $\varepsilon > 0$ we must ensure that N can be chosen large enough so that $\text{plim}_{T \to \infty} p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ exists for any $\varepsilon > 0$.

By compactness of $\boldsymbol{\Theta}$ and Assumption [P], for any $r > 0$ there exists a finite integer $N(r)$ and points $\{\boldsymbol{\theta}^i\}_{i=1}^{N(r)}$, each drawn according to $\boldsymbol{\theta}^i \sim p(\boldsymbol{\theta})$, such that

$$\boldsymbol{\Theta} = \bigcup_{i=1}^{N(r)} \aleph_r(\boldsymbol{\theta}^i) \text{ w.p.1.}$$

By continuity of $\mathbf{b}(\cdot)$, there exists an $r(\varepsilon) > 0$ such that, $\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| < r(\varepsilon)$ implies $d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta})\} < \varepsilon$ for any $\varepsilon > 0$. Combining the two ideas we see that for any $\varepsilon > 0$, we can cover $\boldsymbol{\Theta}$ with $N(r(\varepsilon))$ balls w.p.1. Now, note that by Assumption [P] and the above argument, for any $\varepsilon > 0$, we can find a radius $r(\varepsilon)$ such that $\boldsymbol{\theta}^0 \in \aleph_{r(\varepsilon)}(\boldsymbol{\theta}^i)$ for some $\boldsymbol{\theta}^i$, and $d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} < \varepsilon$ w.p.1 as a consequence. In addition, by Assumption [S1], [S2(1)] and Part 1 of the proof,

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}^i))\} = d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} + o_P(1)$$

for any $\boldsymbol{\theta}^i \in \boldsymbol{\Theta}$ and so

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}^i))\} < \varepsilon + o_P(1)$$

for $\boldsymbol{\theta}^i$ such that $\boldsymbol{\theta}^0 \in \aleph_{r(\varepsilon)}(\boldsymbol{\theta}^i)$. Using this fact, we have that

$$
\begin{aligned}
\mathbb{I}_\varepsilon[\mathbf{z}(\boldsymbol{\theta}^i)] = \mathbb{I}[d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}^i))\} \leq \varepsilon] &= \mathbb{I}[d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} \leq \varepsilon] + o_P(1), \\
&= 1 + o_P(1).
\end{aligned}
$$

From here we see that for T arbitrarily large and any $\varepsilon > 0$, there exists some $N(\varepsilon) := N(r(\varepsilon))$ such that for $N > N(\varepsilon)$

$$p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y})) = \int_{\mathbf{z}} \frac{p(\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})\mathbb{I}_\varepsilon[\mathbf{z}(\boldsymbol{\theta})]}{\left\{ \int_{\boldsymbol{\Theta}} \int_{\mathbf{z}} p(\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})\mathbb{I}_{A_\varepsilon}[\mathbf{z}(\boldsymbol{\theta})]d\mathbf{z}d\boldsymbol{\theta} \right\}} d\mathbf{z} \tag{17}$$

exists.

Part 3:

We now use Parts 1 and 2 to show that $\Pr_\varepsilon(\boldsymbol{\theta} \in \Lambda_\delta|\boldsymbol{\eta}(\mathbf{y})) \xrightarrow{P} 0$, where $\Lambda_\delta := \boldsymbol{\Theta}/\aleph_\delta(\boldsymbol{\theta}^0)$. By Markov's inequality

$$\lim_{T \to \infty} \Pr\left\{ \Pr_\varepsilon(\boldsymbol{\theta} \in \Lambda_\delta|\boldsymbol{\eta}(\mathbf{y})) > \xi \right\} \leq \lim_{T \to \infty} \mathrm{E}\left\{ \Pr_\varepsilon(\boldsymbol{\theta} \in \Lambda_\delta|\boldsymbol{\eta}(\mathbf{y})) \right\}/\xi, \tag{18}$$

for all $\xi > 0$, and the result follows if the left-hand side of (18) is zero. By the definition of $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$, $\Pr_\varepsilon(\boldsymbol{\theta} \in A|\boldsymbol{\eta}(\mathbf{y})) < 1$ for any $A \subset \boldsymbol{\Theta}$, $\varepsilon > 0$, and $T \geq 1$. By the bounded convergence theorem

$$\lim_{T \to \infty} \mathrm{E}\left[\Pr_\varepsilon(\boldsymbol{\theta} \in \Lambda_\delta|\boldsymbol{\eta}(\mathbf{y}))\right] = \mathrm{E}\left[\mathrm{plim}_{T \to \infty}\Pr_\varepsilon(\boldsymbol{\theta} \in \Lambda_\delta|\boldsymbol{\eta}(\mathbf{y}))\right].$$

By the definition of $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ in (17), $\Pr_\varepsilon(\boldsymbol{\theta} \in \Lambda_\delta|\boldsymbol{\eta}(\mathbf{y})) = o_P(1)$ only if, for some, $\varepsilon > 0$

$$\sup_{\boldsymbol{\theta} \in \Lambda_\delta} \mathbb{I}\left[d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))\} \leq \varepsilon\right] = o_P(1).$$

For any $\delta > 0$, if $\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| \geq \delta$, by injectivity of $\mathbf{b}(\cdot)$, it follows that $d\{\mathbf{b}(\boldsymbol{\theta}), \mathbf{b}(\boldsymbol{\theta}^0)\} \geq \varepsilon_*$ for some $\varepsilon_* > 0$. By compactness of Λ_δ and continuity of $\mathbf{b}(\cdot)$, there exists some $\boldsymbol{\theta}_*$ (not necessarily unique) such that

$$\boldsymbol{\theta}_* = \arg\inf_{\boldsymbol{\theta} \in \Lambda_\delta} d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta})\}, \tag{19}$$

and, by injectivity of $\mathbf{b}(\cdot)$, for some $\varepsilon_* > 0$

$$d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}_*)\} \geq \varepsilon_* > 0. \tag{20}$$

Moreover, by Assumptions [S1], [S2(1)], and equation (16) in Part 1,

$$\sup_{\boldsymbol{\theta} \in \Lambda_\delta} |\mathbb{I}_\varepsilon[\mathbf{z}(\boldsymbol{\theta})] - \mathbb{I}_\varepsilon[\mathbf{b}(\boldsymbol{\theta})]| = \sup_{\boldsymbol{\theta} \in \Lambda_\delta} |\mathbb{I}[d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))\} \leq \varepsilon] - \mathbb{I}[d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta})\} \leq \varepsilon]| = o_P(1). \tag{21}$$

From equation (21) it follows that

$$\mathbb{I}[d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}_*))\} \le \varepsilon] = \mathbb{I}[d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}_*)\} \le \varepsilon] + o_P(1).$$

By the definition of $\boldsymbol{\theta}_*$ in (19) and equation (20), for any $\varepsilon^* < \varepsilon_*$,

$$\mathbb{I}[d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}_*)\} \le \varepsilon^*] = 0.$$

We can then conclude that

$$\mathbb{I}[d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}_*))\} \le \varepsilon^*] = o_P(1).$$

Moreover, from equation (19)

$$\inf_{\theta \in \Lambda_\delta} d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta})\} = d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}_*)\} \ge \varepsilon_* > \varepsilon^*, \tag{22}$$

and so it follows from (21) and equation (22)

$$\sup_{\theta \in \Lambda_\delta} \mathbb{I}[d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\theta}))\} \le \varepsilon^*] = o_P(1).$$

Therefore, for $\varepsilon \le \varepsilon^*$ and a corresponding $N > N(\varepsilon)$ number of simulated draws, which exists by Part 2, $\mathrm{Pr}_\varepsilon(\boldsymbol{\theta} \in \Lambda_\delta | \boldsymbol{\eta}(\mathbf{y})) = o_P(1)$ and the result follows. ■

Proof of Corollary 1. We have two cases to consider: one, the vector $\mathbf{b}(\boldsymbol{\theta}^i)$ is one-to-one in $\boldsymbol{\theta}^i$ and two, only the sub-vector $\mathbf{b}_1(\boldsymbol{\theta}^i)$ is one-to-one in $\boldsymbol{\theta}^i$. Clearly, if the first case obtains then the result follows from Theorem 1 and so we can focus on the latter case.

For the second case then, by the triangle inequality

$$d\{\boldsymbol{\gamma}(\mathbf{y}), \boldsymbol{\gamma}(\mathbf{z}^i)\} \le d\{\boldsymbol{\gamma}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^i)\} + d\{\mathbf{b}(\boldsymbol{\theta}^i), \boldsymbol{\gamma}(\mathbf{z}^i)\}. \tag{23}$$

Using the same arguments as in Theorem 1, equation (23) can be restated as

$$d\{\boldsymbol{\gamma}(\mathbf{y}), \boldsymbol{\gamma}(\mathbf{z}^i)\} \le d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} + o_P(1).$$

By assumption $d\{\cdot, \cdot\}$ is an induced metric, and so for vectors x and z (of the same dimension) $d\{x, z\} = 0$ if and only if $x = z$. Using this fact we see that

$$d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta})\} = 0 \iff \begin{pmatrix} \mathbf{b}_1(\boldsymbol{\theta}^0) \\ \mathbf{b}_2(\boldsymbol{\theta}^0) \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1(\boldsymbol{\theta}) \\ \mathbf{b}_2(\boldsymbol{\theta}) \end{pmatrix}.$$

The key observation is that the set $S_{\boldsymbol{\theta}} := \{\theta \in \Theta : \|\mathbf{b}_2(\boldsymbol{\theta}^0) - \mathbf{b}_2(\boldsymbol{\theta})\| = 0\}$ always includes the point $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, but can include other points since $\mathbf{b}_2(\cdot)$ need not be one-to-one. However, by [C2(2)], we know that $\mathbf{b}_1(\cdot)$ is one-to-one and so the only value of $\boldsymbol{\theta}$ for which

$$\begin{pmatrix} \mathbf{b}_1(\boldsymbol{\theta}^0) \\ \mathbf{b}_2(\boldsymbol{\theta}^0) \end{pmatrix} - \begin{pmatrix} \mathbf{b}_1(\boldsymbol{\theta}) \\ \mathbf{b}_2(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

is $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. The result then follows by the same arguments as in Theorem 1. ∎

Proof of Theorem 2.   By the triangle inequality

$$d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \widehat{\boldsymbol{\beta}}(\mathbf{z}^i)\} \leq d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^i)\} + d\{\mathbf{b}(\boldsymbol{\theta}^i), \widehat{\boldsymbol{\beta}}(\mathbf{z}^i)\}. \tag{24}$$

Before proceeding further we must show that, under the maintained assumptions,[11]

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})) - \mathbf{b}(\boldsymbol{\theta}) \right\| = o_P(1). \tag{25}$$

Define the following terms:

$$\begin{aligned} \widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\beta}) &= Q(\mathbf{z}(\boldsymbol{\theta}), \boldsymbol{\beta}) - Q_\infty(\boldsymbol{\theta}, \mathbf{b}(\boldsymbol{\theta})), \\ \widetilde{Q}_\infty(\boldsymbol{\theta}, \boldsymbol{\beta}) &= Q_\infty(\boldsymbol{\theta}, \boldsymbol{\beta}) - Q_\infty(\boldsymbol{\theta}, \mathbf{b}(\boldsymbol{\theta})). \end{aligned}$$

Note that, by [G1(1)], for all $\delta > 0$, if $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})) - \mathbf{b}(\boldsymbol{\theta})\| > \delta$, there exists $\epsilon(\delta) > 0$, such that

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|Q_\infty(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta}))) - Q_\infty(\boldsymbol{\theta}, \mathbf{b}(\boldsymbol{\theta}))\| = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\widetilde{Q}_\infty(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| > \epsilon(\delta).$$

From here, note that

$$\Pr\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})) - \mathbf{b}(\boldsymbol{\theta})\| > \delta \right) \leq \Pr\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\widetilde{Q}_\infty(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| > \epsilon(\delta) \right).$$

The result in (25) then follows if $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\widetilde{Q}_\infty(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| = o_P(1)$.

Uniformly in $\boldsymbol{\theta}$,

$$\begin{aligned} \|\widetilde{Q}_\infty(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| &\leq \|\widetilde{Q}_\infty(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta}))) - \widetilde{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| + \|\widetilde{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| \\ &= \|Q_\infty(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta}))) - Q(\mathbf{z}(\boldsymbol{\theta}), \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| + \|\widetilde{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| \\ &\leq \sup_{\boldsymbol{\beta} \in \mathbf{B}} \|Q_\infty(\boldsymbol{\theta}, \boldsymbol{\beta}) - Q(\mathbf{z}(\boldsymbol{\theta}), \boldsymbol{\beta})\| + \|\widetilde{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| \\ &\leq o_P(1) + \|\widetilde{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\|. \tag{26} \end{aligned}$$

The first inequality follows from the triangle inequality, the second from the definition of $\widetilde{Q}_\infty(\boldsymbol{\theta}, \boldsymbol{\beta})$ and $\widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\beta})$, the third from the definition of sup, and the last from Assumption [G1(2)].

From (26), the result follows if

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\widetilde{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| = o_P(1).$$

---

[11]Recall that $d\{\cdot, \cdot\}$ is an induced metric and hence convergence in $\|\cdot\|$ will imply convergence in $d\{\cdot, \cdot\}$.

By the definition of $\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta}))$, uniformly in $\boldsymbol{\theta}$,

$$
\begin{aligned}
\|\widetilde{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})))\| &\leq \inf_{\boldsymbol{\beta} \in \mathbf{B}} \|\widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\beta})\| + o_P(1) \\
&\leq \inf_{\boldsymbol{\beta} \in \mathbf{B}} \|\widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\beta}) - \widetilde{Q}_\infty(\boldsymbol{\theta}, \boldsymbol{\beta})\| + \inf_{\boldsymbol{\beta} \in \mathbf{B}} \|\widetilde{Q}_\infty(\boldsymbol{\theta}, \boldsymbol{\beta})\| + o_P(1) \\
&\leq \sup_{\boldsymbol{\beta} \in \mathbf{B}} \|Q(\mathbf{z}(\boldsymbol{\theta}), \boldsymbol{\beta}) - Q_\infty(\boldsymbol{\theta}, \boldsymbol{\beta})\| + 0 + o_P(1) \\
&\leq o_P(1),
\end{aligned}
\tag{27}
$$

with the last inequality following from [G1(2)]. Combining equations (26) and (27) yields $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\|\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})) - \mathbf{b}(\boldsymbol{\theta})\right\| = o_P(1)$, and we can conclude

$$
d\{\widehat{\boldsymbol{\beta}}(\mathbf{z}^i), \mathbf{b}(\boldsymbol{\theta}^i)\} \leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} d\{\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\theta})), \mathbf{b}(\boldsymbol{\theta})\} = o_P(1).
\tag{28}
$$

Applying equation (28) to equation (24) we have

$$
d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \widehat{\boldsymbol{\beta}}(\mathbf{z}^i)\} \leq d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^i)\} + o_P(1).
\tag{29}
$$

Applying the triangle inequality to $d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^i)\}$ yields

$$
d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^i)\} \leq d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^0)\} + d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} + o_P(1),
$$

By [G1] and [G2], $\left\|\widehat{\boldsymbol{\beta}}(\mathbf{y}) - \mathbf{b}(\boldsymbol{\theta}^0)\right\| = o_P(1)$, and so

$$
d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \mathbf{b}(\boldsymbol{\theta}^i)\} \leq d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} + o_P(1).
\tag{30}
$$

From (29) and (30) we thus have

$$
d\{\widehat{\boldsymbol{\beta}}(\mathbf{y}), \widehat{\boldsymbol{\beta}}(\mathbf{z}^i)\} \leq d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} + o_P(1).
$$

For fixed $\varepsilon \geq 0$, as $T \to \infty$ a value $\boldsymbol{\theta}^i$ will be selected if and only if

$$
d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} + o_P(1) \leq \varepsilon.
$$

By [G3] the only value of $\boldsymbol{\theta}^i \in \boldsymbol{\Theta}$ for which $\mathbf{b}\left(\boldsymbol{\theta}^i\right) = \mathbf{b}\left(\boldsymbol{\theta}^0\right)$ is $\boldsymbol{\theta}^i = \boldsymbol{\theta}^0$. Therefore, the only value of $\boldsymbol{\theta}^i$ satisfying $d\{\mathbf{b}(\boldsymbol{\theta}^0), \mathbf{b}(\boldsymbol{\theta}^i)\} \leq \varepsilon$ as $\varepsilon \to 0$ is $\boldsymbol{\theta}^i = \boldsymbol{\theta}^0$. The result follows using similar arguments to those of Theorem 1. ∎

References

[1] Beaumont, M.A., Cornuet, J-M., Marin, J-M. and Robert, C.P. 2009. Adaptive Approximate Bayesian Computation, *Biometrika*, 96, 983–990.

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

[3] Beck, J.V. and Arnold, K.J. 1977. Parameter Estimation in Engineering and Science, *John Wiley & Sons.*

[4] Biau, G., Cérou, F. and Guyader, A. 2015. New insights into Approximate Bayesian Computation. *Ann. Inst. H. Poincaré Probab. Statist.*, 51, 376–403.

[5] Blackwell, D. and Dubins, L. 1962. Merging of Opinions with Increasing Information, *Ann. Statist.*, 33, 882–886.

[6] Blum, M.G.B. 2010. Approximate Bayesian Computation: a Nonparametric Perspective, *Journal of the American Statistical Association*, 105, 1178-1187.

[7] Blum, M.G.B. and François, O. 2010. Non-linear Regression Models for Approximate Bayesian Computation, *Statistics and Computing*, 20, 63–73.

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

[9] Creel, M. and Kristensen, D. 2015. ABC of SV: Limited Information Likelihood Inference in Stochastic Volatility Jump-Diffusion Models, *Journal of Empirical Finance,* 31, 85-108.

[10] Diaconis, P. and Freedman, D. 1986. On the Consistency of Bayes Estimates, *Ann. Statist.*, Vol 14, no. 1, 1–26.

[11] Drovandi, C.C., Pettitt, A.N. and Faddy, M.J. 2011. Approximate Bayesian Computation Using Indirect Inference, *J. Royal Statistical Soc. Series C*, 60 1 –21.

[12] Drovandi, C. C., Pettitt, A. N. and Lee, A. 2015. Bayesian Indirect Inference using a Parametric Auxiliary Model. *Statistical Science*, 30, 72–95.

[13] Duffie, D. and Singleton, K.J. 1993. Simulated Moments Estimation of Markov Models of Asset Prices, *Econometrica*, 64, 929–952.

[14] Fearnhead, P. and Prangle, D. 2012. Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic Approximate Bayesian Computation. *J. Royal Statistical Soc. Series B*, 74, 419–474.

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

[16] Ghosal, S., Ghosh, J.K. and Samanta, T. 1995. On Convergence of Posterior Distributions. *Ann. Statist.*, 23, 2145–2152.

[17] Gleim, A. and Pigorsch, C. 2013. Approximate Bayesian Computation with Indirect Summary Statistics. *Draft paper: http://ect-pigorsch.mee.uni-bonn.de/data/research/papers/.*

[18] Gouriéroux, C. and Monfort, A. 1996. *Simulation-based Econometric Methods*, OUP.

[19] Gouriéroux, C., Monfort, A. and Renault, E. 1993. Indirect Inference, *Journal of Applied Econometrics*, 85, S85–S118.

[20] Heggland, K. and Frigessi, A. 2004. Estimating Functions in Indirect Inference, *J. Royal Statistical Soc. Series B*, 66, 447–462.

[21] Ibragimov, I. A. and Has'minskii, R. Z. 1981. *Statistical Estimation: Asymptotic Theory.* Springer, New York.

[22] Jennrich, R. I. 1969. Asymptotic Properties of Non-Linear Least Squares Estimators. *Ann. Math. Statist.*, 40, 633–643.

[23] Joyce, P. and Marjoram, P. 2008. Approximately Sufficient Statistics and Bayesian Computation. *Statistical applications in genetics and molecular biology*, 7, 1–16.

[24] Le Cam. 1953. On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates. *University of California Publications in Statistics*, 1, 277–330.

[25] Li, W. and Fearnhead, P. 2015. Behaviour of ABC for Big Data, *http://arxiv.org/abs/1506.03481.*

[26] Marin, J-M., Pudlo, P., Robert, C.P. and Ryder, R. 2011. Approximate Bayesian Computation Methods. *Statistics and Computing*, 21, 289–291.

[27] Marin, J-M., Pillai, N., Robert, C.P. and Rousseau, J. 2014. Relevant statistics for Bayesian model choice. *J. Royal Statistical Soc. Series B*, 76, 833–859.

[28] Martin, G.M., McCabe, B.P.M., Maneesoonthorn, O. and Robert, C.P. 2014. Approximate Bayesian Inference in State Space Models, *http://arxiv.org/abs/1409.8363.*

[29] Martin, J.S., Jasra, A., Singh, S.S., Whiteley, N., Del Morale, P. and McCoy, E. 2014. Approximate Bayesian Computation for Smoothing. *Stochastic Analysis and Applications,* 32, 397-420.

[30] Marjoram, P., Molitor, J., Plagonal, V. and Tavaré, S. 2003. Markov Chain Monte Carlo Without Likelihoods, *Proceedings of the National Academie of Science USA*, 100, 15324–15328.

[31] McFadden, D. 1989. A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration, *Econometrica*, 57, 995–1026.

[32] Milstein, G. 1978. A Method of Second Order Accuracy Integration of Stochastic Differential Equations, *Theory of Probability and Its Applications*, 23, 396–401.

[33] Newey, W.K. and McFadden, D. 1994. Large Sample Estimation and Hypothesis Testing, In *Handbook of Econometrics* (Eds. Engle and McFadden), Amsterdam: Elsevier Science.

[34] Nott D., Fan, Y., Marshall, L. and Sisson, S. 2014. Approximate Bayesian Computation and Bayes Linear Analysis: Towards High-dimensional ABC, *Journal of Computational and Graphical Statistics*, 23, 65–86.

[35] Pakes, A. and Pollard, D. 1989. Simulation and the Asymptotics of Optimization Estimators, *Econometrica*, 57, 1027–1057.

[36] Pollard , D. 1990. Empirical Processes: Theory and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics, 2, 1–86.

[37] Prangle, D. 2015. Adapting the ABC distance function, *http://arxiv.org/pdf/1507.00874.*

[38] Pritchard, J.K., Seilstad, M.T., Perez-Lezaun, A. and Feldman, M.W. 1999. Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites, *Molecular Biology and Evolution*, 16, 1791–1798.

[39] Sisson S. and Fan, Y. 2011. Likelihood-free Markov Chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo* (Eds. Brooks, Gelman, Jones, Meng). Chapman and Hall/CRC Press.

[40] Sisson, S., Fan, Y. and Tanaka, M. 2007. Sequential Monte Carlo without Likelihoods, *Proceedings of the National Academie of Science USA*, 104, 1760–1765.

[41] Sun, L., Lee, C. and Hoeting, J.A. 2014. Parameter Inference and Model Selection in Deterministic and Stochastic Dynamical Models via Approximate Bayesian Computation: Modeling a Wildlife Epidemic, http://arxiv.org/pdf/1409.7715.pdf.

[42] Tavaré, S., Balding, D.J., Griffiths, R.C. and Donnelly, P. 1997. Inferring Coalescence Times from DNA Sequence Data, *Genetics*, 145, 505–518.

[43] Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M.P.H. 2009. Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems, *JRSS (Interface)*, 6, 187–202.

[44] Wegmann, D., Leuenberger, C. and Excoffier, L. 2009. Efficient Approximate Bayesian Computation Coupled with Markov chain Monte Carlo with Likelihood, *Genetics*, 182, 1207–1218.