arXiv:1712.07750v1 [stat.ME] 20 Dec 2017

Approximate Bayesian Forecasting*

David T. Frazier†, Worapree Maneesoonthorn‡, Gael M. Martin§ and Brendan P.M. McCabe¶

December 22, 2017

Abstract

Approximate Bayesian Computation (ABC) has become increasingly prominent as a method for conducting parameter inference in a range of challenging statistical problems, most notably those characterized by an intractable likelihood function. In this paper, we focus on the use of ABC not as a tool for parametric inference, but as a means of generating probabilistic forecasts; or for conducting what we refer to as 'approximate Bayesian forecasting'. The four key issues explored are: i) the link between the theoretical behavior of the ABC posterior and that of the ABC-based predictive; ii) the use of proper scoring rules to measure the (potential) loss of forecast accuracy when using an approximate rather than an exact predictive; iii) the performance of approximate Bayesian forecasting in state space models; and iv) the use of forecasting criteria to inform the selection of ABC summaries in empirical settings. The primary finding of the paper is that ABC can provide a computationally efficient means of generating probabilistic forecasts that are nearly identical to those produced by the exact predictive, and in a fraction of the time required to produce predictions via an exact method.

Keywords: Bayesian prediction**,** Likelihood-free methods, Predictive merging, Proper **s**coring rules, Particle filtering, Jump-diffusion models.

MSC2010 Subject Classification: 62E17, 62F15, 62F12

*This research has been supported by Australian Research Council Discovery Grants No. DP150101728 and DP170100729.

†Department of Econometrics and Business Statistics, Monash University, Australia. Corresponding author; email: david.frazier@monash.edu.

‡Melbourne Business School, University of Melbourne, Australia.

§Department of Econometrics and Business Statistics, Monash University, Australia.

¶Management School, University of Liverpool, U.K.

1 Introduction

Approximate Bayesian Computation (ABC) has become an increasingly prominent inferential tool in challenging problems, most notably those characterized by an intractable likelihood function. ABC requires only that one can simulate pseudo-data from the assumed model, for given draws of the parameters from the prior. Parameter draws that produce a 'match' between the pseudo and observed data - according to a given set of summary statistics, a chosen metric and a pre-specified tolerance - are retained and used to estimate the posterior distribution, with the resultant estimate of the exact (but inaccessible) posterior being conditioned on the summaries used in the matching. Various guiding principles have been established to select summary statistics in ABC (see, for instance, Joyce and Marjoram, 2008, or Fearnhead and Prangle, 2012) and we refer the reader to the reviews by Blum et al. (2013) and Prangle (2015) for discussions of these different approaches.

Along with the growth in applications of ABC (see Marin et al., 2012, Sisson and Fan, 2011, and Robert, 2016, for recent surveys), attention has recently been paid to the theoretical properties of the method, including the asymptotic behaviour of: ABC posterior distributions, point estimates derived from those distributions, and Bayes factors that condition on summaries. Notable contributions here are Marin et al. (2014), Creel et al. (2015), Jasra (2015), Li and Fearnhead (2015), Li and Fearnhead (2016), Frazier et al. (2016) and Martin et al. (2017), with Frazier et al. (2016) providing the full suite of asymptotic results pertaining to the ABC posterior - namely, Bayesian (or posterior) consistency, limiting posterior shape, and the asymptotic distribution of the posterior mean.

This current paper stands in contrast to existing ABC studies, with their focus on parametric inference and/or model choice. Our goal herein is to exploit ABC as a means of generating probabilistic *forecasts*; or for conducting what we refer to hereafter as 'approximate Bayesian forecasting' (ABF). Whilst ABF has particular relevance in scenarios in which the likelihood function and, hence, the exact predictive distribution, is inaccessible, we also give attention to cases where the exact predictive *is* able to be estimated (via a Monte Carlo Markov chain algorithm), but at a greater computational cost than that associated with ABF. That is, in part, we explore ABF as a computationally convenient means of constructing predictive distributions. We prove that, under certain regularity conditions, ABF produces forecasts that are asymptotically equivalent to those obtained from exact Bayesian methods, and illustrate numerically the close match that can occur between approximate and exact predictives, even when the corresponding approximate and exact posteriors for the parameters are very distinct. We also explore the application of ABF to state space models, in which the production of an approximate Bayesian predictive requires integration over both a small number of static parameters and a set of states with dimension equal to the sample

size.[1]

In summary, the four primary questions addressed in the paper are the following: i) What role does the asymptotic behavior of the ABC posterior - in particular Bayesian consistency - play in determining the accuracy of the approximate predictive as an estimate of the exact predictive? ii) Can we characterize the loss incurred by using the approximate rather than the exact predictive, using proper scoring rules? iii) How does ABF perform in state space models, and what role does (particle) filtering play therein? and iv) How can forecast accuracy be used to guide the choice of summary statistics in an empirical setting?

The remainder of the paper proceeds as follows. In Section 2 we first provide a brief overview of the method of ABC for producing estimates of an exact, but potentially inaccessible, posterior for the unknown parameters. The use of an ABC posterior to yield an approximate forecast distribution is then proposed. After a brief outline of existing asymptotic results pertaining to ABC in Section 3.1, the role played by Bayesian consistency in determining the accuracy of ABF is formally established in Section 3.2, with this building on earlier insights by Blackwell and Dubins (1962) and Diaconis and Freedman (1986) regarding the merging of predictive distributions. In Section 3.3, the concept of a proper scoring rule is adopted in order to formalize the loss incurred when adopting the approximate rather than the exact Bayesian predictive. The relative performance of ABF is then quantified in Section 3.4 using two simple examples: one in which an integer autoregressive model for count time series data is adopted as the data generating process (DGP), with a single set of summaries used to implement ABC; and a second in which a moving average (MA) model is the assumed DGP, and predictives based on alternative sets of summaries are investigated. In both examples there is little visual distinction between the approximate and exact predictives, despite enormous visual differences between the corresponding posteriors. Furthermore, the visual similarity between the exact and approximate predictives extends to forecast accuracy: using averages of various proper scores over a hold-out sample, we demonstrate that the predictive superiority of the exact predictive, over the approximate, is minimal in both examples. Moreover, we highlight the fact that all approximate predictives can be produced in a fraction of the time taken to produce the corresponding exact predictive.

In Section 4, we explore ABF in the context of a model in which latent variables feature. Using a simple stochastic volatility model for which the exact predictive is accessible via Markov chain Monte Carlo (MCMC), the critical importance (in terms of matching the exact predictive) of augmenting ABC inference on the static parameters with 'exact' inference on the states, via a particle filtering step, is made clear. An extensive empirical illustration is then undertaken in Section 5. Approximate predictives for both a financial return and its

[1]Throughout the paper, we use the terms 'forecast' and 'prediction', and their various adjectival forms and associated verb conjugations, synonymously, interchanging them for linguistic variety only.

volatility, in a dynamic jump diffusion model with α-stable volatility transitions, are produced, using a range of different summaries, including those extracted from simple auxiliary models with closed-form likelihood functions. Particular focus is given to using out-of-sample predictive performance to determine the 'best' set of summaries for driving ABC, in the case where prediction is the primary goal of the investigation. A discussion section concludes the paper in Section 6, and proofs are included in the Appendix.

2 Approximate Bayesian Computation (ABC): Inference and Forecasting

We observe a T-dimensional vector of data $\mathbf{y} = (y_1, y_2, ..., y_T)'$, assumed to be generated from some model with likelihood $p(\mathbf{y}|\theta)$, with θ a k_θ-dimension vector of unknown parameters, and where we possess prior beliefs on θ specified by $p(\theta)$. In this section, we propose a means of producing probabilistic forecasts for the random variables $Y_{T+k}, k = 1, ..., h$, in situations where $p(\mathbf{y}|\theta)$ is computationally intractable or numerically difficult to calculate. Before presenting this approach, we first give a brief overview of ABC-based inference for the unknown parameters θ.

2.1 ABC Inference: Overview

The aim of ABC is to produce draws from an approximation to the posterior distribution,

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta), \tag{1}$$

in the setting where both $p(\theta)$, and the assumed data generating process, $p(\mathbf{y}|\theta)$, can be simulated from, but where $p(\mathbf{y}|\theta)$ is typically intractable in some sense. These draws are, in turn, used to approximate posterior quantities of interest, and thereby form the basis for conducting inference about θ. The simplest (accept/reject) form of the algorithm proceeds as in Algorithm 1.

Algorithm 1 ABC accept/reject algorithm

1: Simulate θ^i, $i = 1, 2, ..., N$, from $p(\theta)$
2: Simulate $\mathbf{z}^i = (z_1^i, z_2^i, ..., z_T^i)'$, $i = 1, 2, ..., N$, from the likelihood, $p(.|\theta^i)$
3: Select θ^i such that:

$$d\{\eta(\mathbf{y}), \eta(\mathbf{z}^i)\} \leq \varepsilon, \tag{2}$$

where $\eta(.)$ is a (vector) statistic, $d\{.\}$ is a distance criterion, and, given N, the tolerance level ε is chosen to be small. (The Euclidean distance is used for all numerical illustrations in the paper.)

The algorithm thus samples θ and pseudo-data $\mathbf{z}$ from the joint posterior:

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

where $\mathbb{I}_\varepsilon[\mathbf{z}]:=\mathbb{I}[d\{\eta(\mathbf{y}), \eta(\mathbf{z})\} \leq \varepsilon]$ is one if $d\{\eta(\mathbf{y}), \eta(\mathbf{z})\} \leq \varepsilon$ and zero otherwise. When the vector of summary statistics, $\eta(\cdot)$, is sufficient for θ and ε is arbitrarily small,

$$p_\varepsilon(\theta|\eta(\mathbf{y})) = \int_{\mathbf{z}} p_\varepsilon(\theta, \mathbf{z}|\eta(\mathbf{y}))d\mathbf{z} \tag{3}$$

approximates the exact posterior, $p(\theta|\mathbf{y})$, and draws from $p_\varepsilon(\theta, \mathbf{z}|\eta(\mathbf{y}))$ can be used to estimate features of that exact posterior. In practice however, the complexity of the models to which ABC is applied implies that sufficiency is unattainable. Hence, as $\varepsilon \to 0$ the draws can be used to estimate features of $p(\theta|\eta(\mathbf{y}))$ only, with the 'proximity' of $p(\theta|\eta(\mathbf{y}))$ to $p(\theta|\mathbf{y})$ depending - in a sense that is not formally defined - on the 'proximity' to sufficiency of $\eta(\mathbf{y})$.

Unlike most existing studies on ABC, our end goal is *not* the quantification of uncertainty about θ, but the construction of probabilistic forecasts for future realizations of a random variable of interest, in which $p_\varepsilon(\theta|\eta(\mathbf{y}))$ expresses our uncertainty about θ. That is, in contrast to exact Bayesian forecasting, in which a (marginal) predictive distribution is produced by averaging the conditional predictive with respect to the exact posterior, $p(\theta|\mathbf{y})$, approximate Bayesian forecasting performs this integration step using the approximate posterior as the weighting function. This substitution (of $p(\theta|\mathbf{y})$ by $p_\varepsilon(\theta|\eta(\mathbf{y}))$) is most clearly motivated in cases where $p(\theta|\mathbf{y})$ is inaccessible, due to an intractable likelihood function. However, the use of $p_\varepsilon(\theta|\eta(\mathbf{y}))$ will also be motivated here by computational considerations alone.

2.2 Approximate Bayesian Forecasting (ABF)

Without loss of generality, we focus at this point on one-step-ahead forecasting in the context of a time series model.[2] Let Y_{T+1} denote a random variable that will be observed at time $T+1$, and which is generated from the (conditional) predictive density (or mass) function, $p(y_{T+1}|\theta, \mathbf{y})$, at some fixed value θ. The quantity of interest is thus

$$p(y_{T+1}|\mathbf{y}) = \int p(y_{T+1}|\theta, \mathbf{y})p(\theta|\mathbf{y})d\theta, \tag{4}$$

where $p(\theta|\mathbf{y})$ is the exact posterior defined in (1) and y_{T+1} denotes a value in the support of Y_{T+1}. The DGP, $p(\mathbf{y}|\theta)$, is required in closed form for numerical methods such as MCMC to be applicable to $p(\theta|\mathbf{y})$, in the typical case in which the latter itself cannot be expressed in

[2]Multi-step-ahead forecasting entails no additional conceptual challenges and, hence, is not treated herein.

a standard form.[3] Such methods yield draws from $p(\theta|\mathbf{y})$ that are then used to produce a simulation-based estimate of the predictive density as:

$$\widehat{p}(y_{T+1}|\mathbf{y}) = \frac{1}{M}\sum_{i=1}^{M} p(y_{T+1}|\theta^{(i)}, \mathbf{y}), \tag{5}$$

where the conditional predictive, $p(y_{T+1}|\theta^{(i)}, \mathbf{y})$, is also required to be known in closed-form for the 'Rao-Blackwellized' estimate in (5) to be feasible. Alternatively, draws of y_{T+1} from $p(y_{T+1}|\theta^{(i)}, \mathbf{y})$ can be used to produce a kernel density estimate of $p(y_{T+1}|\mathbf{y})$. Subject to convergence of the MCMC chain, either computation represents an estimate of the exact predictive that is accurate up to simulation error, and may be referred to as yielding the *exact* Bayesian forecast distribution as a consequence.

The motivation for the use of ABC in this setting is obvious: in cases where $p(\mathbf{y}|\theta)$ is not accessible, $p(\theta|\mathbf{y})$ itself is inaccessible (via an MCMC scheme of some sort, for example) and the integral in (4) that defines the exact predictive cannot be estimated via those MCMC draws in the manner described above. ABC enables approximate Bayesian *inference* about θ to proceed via a simulation-based estimate of $p(\theta|\eta(\mathbf{y}))$, for some chosen summary, $\eta(\mathbf{y})$. Hence, a natural way in which to approach the concept of approximate Bayesian *forecasting* is to define the quantity

$$g(y_{T+1}|\mathbf{y}) = \int_{\Theta} p(y_{T+1}|\theta, \mathbf{y}) p_{\varepsilon}(\theta|\eta(\mathbf{y}))d\theta, \tag{6}$$

with $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$ replacing $p(\theta|\mathbf{y})$ in (4). The conditional density function, $g(y_{T+1}|\mathbf{y})$, which is shown in the appendix to be a proper density function, represents an approximation of $p(y_{T+1}|\mathbf{y})$ that we refer to as the ABF density. This density can, in turn, be estimated via the sequential use of the ABC draws from $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$ followed by draws of y_{T+1} conditional on the draws of θ.

Certain natural questions become immediately relevant: First, what role, if any, do the properties of $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$ play in determining the accuracy of $g(y_{T+1}|\mathbf{y})$ as an estimate of $p(y_{T+1}|\mathbf{y})$? Second, can we formally characterize the anticipated loss associated with targeting $g(y_{T+1}|\mathbf{y})$ rather than $p(y_{T+1}|\mathbf{y})$? Third, in practical settings do conclusions drawn regarding Y_{T+1} from $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$ differ in any substantial way? These questions are tackled sequentially in the following Section 3.

[3]Pseudo-marginal MCMC methods may be feasible when certain components of the DGP are unavailable in closed form. For example, particle MCMC could be applied to state space models in which the state transitions are unavailable, but can be simulated from. However, the great majority of MCMC algorithms would appear to exploit full knowledge of the DGP in their construction.

3 Accuracy of ABF

It is well-known in the ABC literature that the posterior $p_\varepsilon(\theta|\eta(\mathbf{y}))$ is sometimes a poor approximation to $p(\theta|\mathbf{y})$ (Marin et al., 2012). What is unknown, however, is whether or not this same degree of inaccuracy will transfer to the ABC-based predictive. To this end, in Section 3.2, we begin by characterizing the difference between $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$ using the large sample behavior of $p_\varepsilon(\theta|\eta(\mathbf{y}))$ and $p(\theta|\mathbf{y})$. In so doing we demonstrate that if both $p_\varepsilon(\theta|\eta(\mathbf{y}))$ and $p(\theta|\mathbf{y})$ are Bayesian consistent for the true value θ_0, then the densities $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$ produce the same predictions asymptotically; that is, $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$ 'merge' asymptotically (Blackwell and Dubins, 1962; Diaconis and Freedman (1986)). Using the concept of a proper scoring rule, in Section 3.3 we quantify the loss in forecasting accuracy incurred by using $g(y_{T+1}|\mathbf{y})$ rather than $p(y_{T+1}|\mathbf{y})$.

To characterize the difference between $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$ in large samples, as is consistent with the standard approach to Bayesian asymptotic (van der Vaart, 1998 and Ghosh and Ramamoorthi, 2003), we view the conditioning values $\mathbf{y}$ as random and thus, by extension, $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$. However, for ease of notation, we continue to use the lower case notation $\mathbf{y}$ everywhere.

We first give a brief overview of certain existing results on the asymptotic properties of $p_\varepsilon(\theta|\eta(\mathbf{y}))$, which inform the theoretical results pertaining to prediction.

3.1 Asymptotic Properties of ABC posteriors

We briefly summarize recent results on this front as they pertain to our eventual goal of demonstrating the merging of $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$. To this end, we draw on the work of Frazier et al. (2016) but acknowledge here the important contributions by Li and Fearnhead (2015) and Li and Fearnhead (2016).

Establishing the asymptotic properties of $p_\varepsilon(\theta|\eta(\mathbf{y}))$ requires simultaneous asymptotics in the tolerance, ε, and the sample size, T. To this end, we denote a hypothetical $T-$dependent ABC tolerance by ε_T. Under relatively weak sufficient conditions on the prior $p(\theta)$ and the tail behavior of $\eta(\mathbf{y})$, plus an identification condition that is particular to the probability limit of $\eta(\mathbf{y})$, Frazier et al. (2016) prove the following results regarding the posterior produced from the ABC draws in Algorithm 1, as $T \to \infty$:

1. The posterior concentrates onto θ_0 (i.e. is Bayesian consistent) for any $\varepsilon_T = o(1)$;

2. The posterior is asymptotically normal for $\varepsilon_T = o(\nu_T^{-1})$, where ν_T is the rate at which the summaries $\eta(\mathbf{y})$ satisfy a central limit theorem.

In Section 3.2 we show that under Bayesian consistency, predictions generated from $g(y_{T+1}|\mathbf{y})$ will, to all intents and purposes, be identical to those generated from $p(y_{T+1}|\mathbf{y})$.

The asymptotic normality (i.e. a Bernstein-von Mises type of result) in 2. is applied in Section 3.3. Note that, without making this explicit, we assume that the tolerance underpinning an ABC posterior is specified in such a way that the theoretical properties invoked hold.

3.2 Merging of Approximate and Exact Predictives

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, with $\mathbb{P}$ a convex class of probability measures on $(\Omega, \mathcal{F})$. Define a filtration $\{\mathcal{F}_t : t \geq 0\}$ associated with the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let the sequence $\{y_t\}_{t\geq 1}$ be adapted to $\{\mathcal{F}_t\}$. Define, for $B \in \mathcal{F}$, the following predictive measures

$$P_{\mathbf{y}}(B) = \int_{\Omega}\int_{\Theta} p(y_{T+1}|\theta, \mathbf{y}) d\Pi[\theta|\mathbf{y}] d\delta_{y_{T+1}}(B)$$
$$G_{\mathbf{y}}(B) = \int_{\Omega}\int_{\Theta} p(y_{T+1}|\theta, \mathbf{y}) d\Pi[\theta|\eta(\mathbf{y})] d\delta_{y_{T+1}}(B),$$

where δ_x denotes the Dirac measure. $P_{\mathbf{y}}(\cdot)$ denotes the predictive distribution for the random variable Y_{T+1}, conditional on $\mathbf{y}$, and where parameter uncertainty - integrated out in the process of producing $P_{\mathbf{y}}(\cdot)$ - is described by the exact posterior distribution, $\Pi[\cdot|\mathbf{y}]$, with density $p(\theta|\mathbf{y})$ (with respect to the Lebesgue measure). $G_{\mathbf{y}}(\cdot)$ is the ABF predictive and differs from $P_{\mathbf{y}}(\cdot)$ in its quantification of parameter uncertainty, which is expressed via $\Pi[\cdot|\eta(\mathbf{y})]$ instead of $\Pi[\cdot|\mathbf{y}]$, where the former has density $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$.

The discrepancy between $G_{\mathbf{y}}$ and $P_{\mathbf{y}}$ is entirely due to the replacement of $\Pi[\theta|\mathbf{y}]$ by $\Pi[\theta|\eta(\mathbf{y})]$. In this way, noting that, for any $B \in \mathcal{F}$,

$$|G_{\mathbf{y}}(B) - P_{\mathbf{y}}(B)| \leq \int_{\Omega}\int_{\Theta} p(y_{T+1}|\theta, \mathbf{y}) d\delta_{y_{T+1}}(B) \left|p_{\varepsilon}(\theta|\eta(\mathbf{y})) - p(\theta|\mathbf{y})\right| d\theta,$$

it is clear that the difference between $G_{\mathbf{y}}$ and $P_{\mathbf{y}}$ is smaller, the smaller is the discrepancy between $p(\theta|\mathbf{y})$ and $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$.

Under regularity conditions (see, for example, Ghosal et al., 1995 or Ibragimov and Has'Minskii, 2013) the exact posterior $p(\theta|\mathbf{y})$ will concentrate onto θ_0 as $T \to \infty$. Hence, as long as the relevant conditions delineated in Frazier et al. (2016) for the Bayesian consistency of $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$ are satisfied, then $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$ will also concentrate onto θ_0 as $T \to \infty$. Consequently, the discrepancy between $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$ and $p(\theta|\mathbf{y})$ will disappear in large samples, and mitigate the discrepancy between $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$. The following theorem formalizes this intuition.

Theorem 1 *Under Assumption 1 in Appendix A, the predictive distributions $P_{\mathbf{y}}(\cdot)$ and $G_{\mathbf{y}}(\cdot)$ merge, in the sense that $\rho_{TV}\{P_{\mathbf{y}}, G_{\mathbf{y}}\} \to 0$ as $T \to \infty$, with $\mathbb{P}$-probability 1, and where $\rho_{TV}\{P_{\mathbf{y}}, G_{\mathbf{y}}\}$ denotes the total variation metric:* $\sup_{B\in\mathcal{F}} |P_{\mathbf{y}}(B) - G_{\mathbf{y}}(B)|$.

The merging of $P_{\mathbf{y}}$ and $G_{\mathbf{y}}$ is not without precedence and mimics early results on merging of predictive distributions due to Blackwell and Dubins (1962). A connection between merging of predictive distributions and Bayesian consistency was first discussed in Diaconis and Freedman (1986), with the authors viewing Bayesian consistency as implying a "merging of inter-subjective opinions". In their setting, Bayesian consistency implied that two separate Bayesians with different subjective prior beliefs would ultimately end up with the same predictive distribution. (See also Petrone et al., 2014, for related work).

Our situation is qualitatively different from that considered in Diaconis and Freedman (1986) in that we are not concerned with Bayesians who have different *prior* beliefs but Bayesians who are using completely different means of assessing the *posterior* uncertainty about the parameters θ. Given the nature of ABC, and the fact that under suitable conditions posterior concentration can be proven, we have the interesting result that, for a large enough sample, and under Bayesian consistency of both $p_\varepsilon(\theta|\eta(\mathbf{y}))$ and $p(\theta|\mathbf{y})$, conditioning inference about θ on $\eta(\mathbf{y})$ rather than $\mathbf{y}$ makes no difference to the probabilistic statements made about Y_{T+1}. In contexts where inference about θ is simply a building block for Bayesian predictions, and where sample sizes are sufficiently large, inference undertaken via (posterior consistent) ABC is sufficient to yield predictions that are virtually identical to those obtained by an exact (but potentially infeasible or, at the very least computationally challenging) method.

3.3 Proper Scoring Rules

The above merging result demonstrates that in large samples the difference between $p(y_{T+1}|\mathbf{y})$ and $g(y_{T+1}|\mathbf{y})$ is likely to be small. To formally quantify the loss in forecast accuracy incurred by using $g(y_{T+1}|\mathbf{y})$ rather than $p(y_{T+1}|\mathbf{y})$, we use the concept of a scoring rule. Heuristically, a scoring rule rewards a forecast for assigning a high density ordinate (or high probability mass) to the observed value (so-called 'calibration'), often subject to some shape or 'sharpness' criterion. (See Gneiting and Raftery, 2007 and Gneiting et al., 2007 for expositions). More specifically, we are interested in scoring rules $S : \mathbb{P} \times \Omega \mapsto \mathbb{R}$ whereby if the forecaster quotes the predictive distribution G and the value y eventuates, then the reward (or 'score') is $S(G, y)$. We then define the *expected* score under measure P of the probability forecast G, as

$$\mathbb{M}(G, P) = \int_{y \in \Omega} S(G, y) dP(y). \tag{7}$$

A scoring rule $S(\cdot, \cdot)$ is proper if for all $G, P \in \mathbb{P}$,

$$\mathbb{M}(P, P) \geq \mathbb{M}(G, P),$$

and is strictly proper, relative to P, if $\mathbb{M}(P, P) = \mathbb{M}(G, P)$ implies $G = P$. That is, a proper scoring rule is one whereby *if* the forecasters best judgment is indeed P there is no incentive to quote anything other than $G = P$.

Now define the true predictive distribution of the random variable Y_{T+1}, conditional on θ_0, as

$$F_{\mathbf{y}}(B) = \int_{\Omega}\int_{\Theta} p(y_{T+1}|\theta,\mathbf{y})d\delta_{\theta}(\theta_0)d\delta_{y_{T+1}}(B).$$

The following result builds on Theorem 1 and presents a theoretical relationship between the predictive density functions, $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$, in terms of the expectation of proper scoring rules with respect to $F_{\mathbf{y}}(\cdot)$.

Theorem 2 *Under Assumption 1 in Appendix A, if $S(\cdot,\cdot)$ is a strictly proper scoring rule,*

(i) $|\mathbb{M}(P_{\mathbf{y}},F_{\mathbf{y}}) - \mathbb{M}(G_{\mathbf{y}},F_{\mathbf{y}})| = o_{\mathbb{P}}(1)$*;*

(ii) $|\mathbb{E}\left[\mathbb{M}(P_{\mathbf{y}},F_{\mathbf{y}})\right] - \mathbb{E}\left[\mathbb{M}(G_{\mathbf{y}},F_{\mathbf{y}})\right]| = o(1)$*;*

(iii) The absolute differences in (i) and (ii) are identically zero if and only if $\eta(\mathbf{y})$ is sufficient for $\mathbf{y}$ and $\varepsilon = 0$.

The result in (i) establishes an asymptotic equivalence between the expected scores (under $F_{\mathbf{y}}$) of the exact and approximate predictives, where the expectation is with respect to Y_{T+1}, conditional on $\mathbf{y}$. Hence, the result establishes that (under regularity) as $T \to \infty$, there is no expected loss in accuracy from basing predictions on an approximation. The result in (ii) is marginal of $\mathbf{y}$ and follows from (i) and the monotonicity property of integrals. Part (iii) follows from the factorization theorem and the structure of $P_{\mathbf{y}}$ and $G_{\mathbf{y}}$. All results are, of course, consistent with the merging result demonstrated earlier, and with $P_{\mathbf{y}}$ and $G_{\mathbf{y}}$, by definition, equivalent for any T under sufficiency of $\eta(\mathbf{y})$.

If, however, one is willing to make additional assumptions about the regularity of $p(\theta|\mathbf{y})$ and $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$, one can go further than the result in Theorem 2, to produce an actual *ranking* of $\mathbb{M}(P_{\mathbf{y}},F_{\mathbf{y}})$ and $\mathbb{M}(G_{\mathbf{y}},F_{\mathbf{y}})$, which should hold for large T with high probability. Heuristically, if both $p(\theta|\mathbf{y})$ and $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$ satisfy a Bernstein-von Mises result (invoking Result 2 in Section 3.1 in the latter case and standard regularity in the former): for $\phi_{\theta,V}$ a normal density function with mean θ and variance V,

$$p(y_{T+1}|\mathbf{y}) = \int_{\Theta} p(y_{T+1}|\theta,\mathbf{y})\phi_{\widehat{\theta},\mathcal{I}^{-1}}(\theta)d\theta + o_{\mathbb{P}}(T^{-1/2}) \tag{8}$$

$$g(y_{T+1}|\mathbf{y}) = \int_{\Theta} p(y_{T+1}|\theta,\mathbf{y})\phi_{\tilde{\theta},\mathcal{E}^{-1}}(\theta)d\theta + o_{\mathbb{P}}(T^{-1/2}), \tag{9}$$

where $\hat{\theta}$ is the maximum likelihood estimator (MLE), $\mathcal{I}$ is the Fisher information matrix (evaluated at θ_0), $\tilde{\theta}$ is the ABC posterior mean and $\mathcal{E}$ is the Fisher information conditional on the statistic $\eta(\mathbf{y})$ (evaluated at θ_0). We assume, for simplicity, that both $\mathcal{I}^{-1}$ and $\mathcal{E}^{-1}$ are

$O(T^{-1})$, where $\mathcal{I}^{-1} - \mathcal{E}^{-1}$ is negative semi-definite. Now, assuming validity of a second-order Taylor expansion for $p(y_{T+1}|\theta, \mathbf{y})$ in a neighborhood of θ_0, we can expand this function as

$$p(y_{T+1}|\theta, \mathbf{y}) = p(y_{T+1}|\hat{\theta}, \mathbf{y}) + \left.\frac{\partial p(y_{T+1}|\theta, \mathbf{y})}{\partial \theta'}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})' \left.\frac{\partial^2 p(y_{T+1}|\theta, \mathbf{y})}{\partial \theta \partial \theta'}\right|_{\theta=\theta^*} (\theta - \hat{\theta}), \tag{10}$$

for some consistent intermediate value θ^*. Substituting (10) into (8), and recognizing that $\int_\Theta (\theta - \hat{\theta}) \phi_{\widehat{\theta}, \mathcal{I}^{-1}}(\theta) d\theta = 0$, then yields

$$\begin{aligned} p(y_{T+1}|\mathbf{y}) &= \int_\Theta p(y_{T+1}|\hat{\theta}, \mathbf{y}) \phi_{\widehat{\theta}, \mathcal{I}^{-1}}(\theta) d\theta + \frac{1}{2} tr \left\{ \left.\frac{\partial^2 p(y_{T+1}|\theta, \mathbf{y})}{\partial \theta \partial \theta'}\right|_{\theta=\theta^*} \int_\Theta (\theta - \hat{\theta})(\theta - \hat{\theta})' \phi_{\widehat{\theta}, \mathcal{I}^{-1}}(\theta) d\theta \right\} \\ &\quad + o_{\mathbb{P}}(T^{-1/2}) \\ &= p(y_{T+1}|\hat{\theta}, \mathbf{y}) + O_{\mathbb{P}}(1) O(T^{-1}) + o_{\mathbb{P}}(T^{-1/2}) \\ &= p(y_{T+1}|\hat{\theta}, \mathbf{y}) + o_{\mathbb{P}}(1). \end{aligned}$$

Similarly, we have for $g(y_{T+1}|\mathbf{y})$ in (9):

$$g(y_{T+1}|\mathbf{y}) = p(y_{T+1}|\tilde{\theta}, \mathbf{y}) + o_{\mathbb{P}}(1).$$

Heuristically, for large T, under the approximate Gaussianity of $\hat{\theta}$ and $\tilde{\theta}$, we can view $p(y_{T+1}|\hat{\theta}, \mathbf{y}) - p(y_{T+1}|\theta_0, \mathbf{y})$ and $p(y_{T+1}|\tilde{\theta}, \mathbf{y}) - p(y_{T+1}|\theta_0, \mathbf{y})$ as approximately Gaussian with mean 0, but with the former having a smaller variance than the latter (even though these un-normalized quantities have variances that are both collapsing to zero as $T \to \infty$). Therefore, on average, the error $p(y_{T+1}|\hat{\theta}, \mathbf{y}) - p(y_{T+1}|\theta_0, \mathbf{y})$, should be smaller than the error $p(y_{T+1}|\tilde{\theta}, \mathbf{y}) - p(y_{T+1}|\theta_0, \mathbf{y})$, so that, for $S(\cdot, \cdot)$ a proper scoring rule, on average,

$$\begin{aligned} \int_\Omega S(p(y_{T+1}|\theta_0, \mathbf{y}), y_{T+1}) p(y_{T+1}|\theta_0, \mathbf{y}) dy_{T+1} &\geq \int_\Omega S(p(y_{T+1}|\hat{\theta}, \mathbf{y}), y_{T+1}) p(y_{T+1}|\theta_0, \mathbf{y}) dy_{T+1} \\ &\geq \int_\Omega S(p(y_{T+1}|\tilde{\theta}, \mathbf{y}), y_{T+1}) p(y_{T+1}|\theta_0, \mathbf{y}) dy_{T+1}. \end{aligned} \tag{11}$$

That is, using the notation defined in (7), one would expect that, for large enough T,

$$\mathbb{M}(P_{\mathbf{y}}, F_{\mathbf{y}}) \geq \mathbb{M}(G_{\mathbf{y}}, F_{\mathbf{y}}), \tag{12}$$

and - as accords with intuition - predictive accuracy to be greater when based on the exact predictive distribution.

In practice of course, in a situation in which exact inference is deemed to be infeasible, measurement of this loss is also infeasible, since $p(y_{T+1}|\mathbf{y})$ is inaccessible. However, it is of interest - in experimental settings, in which both $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$ *can* be computed

- to gauge the extent of this discrepancy, in particular for different choices of $\eta(\mathbf{y})$. This then gives us some insight into what might be expected in the more realistic scenario in which the exact predictive cannot be computed and the ABF density is the only option. Furthermore, even in situations in which $p(y_{T+1}|\mathbf{y})$ can be accessed, but only via a bespoke, finely-tuned MCMC algorithm, a finding that the approximate predictive produced via the simpler, more readily automated and less computationally burdensome ABC algorithm, is very similar to the exact, is consequential for practitioners. We pursue such matters in the following Section 3.4, with the specific matter of asymptotic merging - and the role played therein by Bayesian consistency - treated in Section 3.4.3.

3.4 Numerical Illustrations

3.4.1 Example: Integer Autoregressive Model

We begin by illustrating the approximate forecasting methodology for the case of a discrete random variable, in which case the object of interest is a predictive mass function. To do so, we adopt an integer autoregressive model of order one (INAR(1)) as the data generating process. The INAR(1) model is given as

$$y_t = \alpha \circ y_{t-1} + \varepsilon_t, \tag{13}$$

where $\circ$ is the binomial thinning operator defined as

$$\alpha \circ y_{t-1} = \sum_{j=0}^{y_{t-1}} B_j(\alpha), \tag{14}$$

and where $B_j(\alpha)$ are *i.i.d.* Bernoulli random variables with probability α. In the numerical illustration we take ε_t to be *i.i.d.* Poisson with intensity parameter λ.

The INAR(1) model sits within the broader class of integer-valued ARMA (INARMA) models, which has played a large role in the modeling and forecasting of count time series data. See Jung and Tremayne (2006) for a review, and Drost et al. (2009) and McCabe et al. (2011) for recent contributions. Of particular note is the work by Martin et al. (2014), in which the INARMA model is estimated 'indirectly' via efficient method of moments (Gallant and Tauchen, 1996), which is similar in spirit to ABC. No investigation of forecasting under this 'approximate' inferential paradigm is however undertaken.

Relevant also is the work of Neal and Rao (2007) in which an MCMC scheme for the INARMA class is devised, and from which an exact predictive could be estimated. However, given the very simple parameterization of (13), we evaluate the exact posterior for $\theta = (\alpha, \lambda)'$ numerically using deterministic integration, and estimate the exact predictive in (4) by taking a simple weighted average of the ordinates of the one-step-ahead conditional predictive

associated with the model. Given the structure of (13) this (conditional) predictive mass function is defined by the convolution of the two unobserved random variables, $\alpha \circ y_T$ and ε_T, as

$$P\left(Y_{T+1} = y_{T+1}|\mathbf{y},\theta\right) = \sum_{s=0}^{\min\{y_{T+1},y_T\}} P\left(B_{y_T}^{\alpha} = s\right) P(\varepsilon_{T+1} = y_{T+1} - s), \tag{15}$$

where $P\left[B_{y_T}^{\alpha} = s\right]$ denotes the probability that a binomial random variable associated with y_T replications (and a probability of 'success', α, on each replication) takes a value of s, and where $P(\varepsilon_{T+1} = y_{T+1} - s)$ denotes the probability that a Poisson random variable takes a value of $y_{T+1} - s$.

We generate a sample of size $T = 100$ from the model in (13) and (14), with $\theta_0 = (\alpha_0, \lambda_0)' = (0.4, 2)'$. Prior information on θ is specified as $U[0,1] \times U[0,10]$. ABC (via a nearest-neighbour version of Algorithm 1, with the smallest 1% of $N = 20,000$ draws retained) is based on a single vector of summary statistics comprising the sample mean of y, denoted as $\bar{y}$, and the first three sample autocovariances, $\gamma_l = cov(y_t, y_{t-l})$, $l = 1, 2, 3$: $\eta(\mathbf{y}) = (\bar{y}, \gamma_1, \gamma_2, \gamma_3)'$. Given the latent structure of (14) no reduction to sufficiency occurs; hence neither this, nor any other set of summaries will replicate the information in $\mathbf{y}$, and $p_\varepsilon(\theta|\eta(\mathbf{y}))$ will thus be distinct from $p(\theta|\mathbf{y})$. As is evident by the plots in Panels A and B of Figure 1, the exact and ABC posteriors for each element of θ are indeed quite different one from the other. In contrast, in Panel C the exact and approximate predictive mass functions (with the latter estimated by taking the average of the conditional predictives in (15) over the ABC draws of θ) are seen to be an extremely close match![4]

To illustrate the results of Theorem 2, we construct a series of 100 expanding window one-step-ahead predictive distributions (beginning with a sample size of $T = 100$), and report the average (over 100 one-step-ahead predictions) of the log score (LS) and quadratic score (QS) in Table 1, using the 'observed' value of y_{T+1} that is also simulated. (See Gneiting et al., 2007, for details of these particular scoring rules.) The assumptions under which Theorem 2 hold can be demonstrated analytically in this case, including the Bayesian consistency of $p_\varepsilon(\theta|\eta(\mathbf{y}))$. It is immediately obvious that, at least according to these two scoring rules, and to two decimal places, the predictive accuracy of $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$ is equivalent, even for this relatively small sample size!

In addition, it is important to note that the computational time required to produce the exact predictive, via rectangular integration over the prior grid, is just under four and a half minutes, which is approximately 18 times greater than the time required to construct the

[4]We point out that the predictives for y_{T+1} in Panel (C) of Figure 1 are mass functions, even though they are represented as smooth curves. The smooth curve representation has been used to help give a better sense of the closeness of the two predictives. We also note, with reference to the marginal posteriors of λ, that the ABC posterior places much more mass over the entire prior support for λ, given as $[0, 10]$, than does the exact posterior; hence the very marked difference in their shapes.

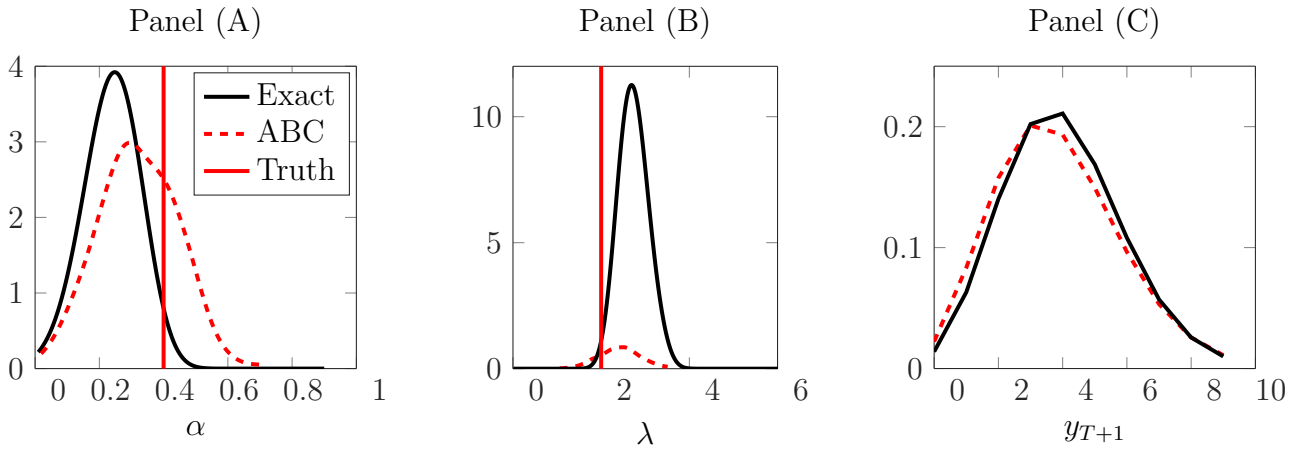

Figure 1: Panels (A) and (B) depict the marginal posteriors (exact and ABC) for α and λ, respectively. Panel (C) plots the one-step-ahead predictive density functions - both exact and approximate (ABC-based). The red vertical line (denoted by 'Truth' in the key) represents the true value of the relevant parameter in Panels (A) and (B).

approximate predictive via ABC. Therefore, in this simple example, we see that ABF offers a substantial speed improvement over the exact predictive, with no loss in predictive accuracy.[5]

Table 1: Log score (LS) and quadratic score (QS) associated with the approximate predictive $g(y_{T+1}|\mathbf{y})$, and the exact predictive, $p(y_{T+1}|\mathbf{y})$, each computed as an average over a series of (expanding window) 100 one-step-ahead predictions. The predictive with highest average score is in bold.

	ABF	MCMC
LS	**-1.89**	**-1.89**
QS	**0.17**	**0.17**

We do emphasize at this point that refinements of Algorithm 1 based on either post-sampling corrections (Beaumont et al., 2002, Blum, 2010), or the insertion of MCMC or sequential Monte Carlo steps (Marjoram et al., 2003, Sisson et al., 2007, Beaumont et al., 2009) may well improve the accuracy with which the exact posteriors are approximated. However, the key message - both here and in what follows - is that a poor match between exact and approximate posteriors does not translate into a corresponding poor match at the predictive level. The degree to which the ABC algorithm is optimized, or not, is not germane to that message; hence, we choose to use the simplest form of the algorithm in all illustrations.

[5]Given the independent nature of ABC sampling, we are able to exploit parallel computing. This is done using the standard 'parfor' function in MATLAB. All computations are conducted on an Intel Xeon E5-2630 2.30GHz dual processors (each with 6 cores) with 16GB RAM. Note that all computation times quoted in the paper are 'time elapsed' or 'wall-clock' time.

3.4.2 Example: Moving Average Model

We now explore an example from the canonical class of time series models for a *continuous* random variable, namely the Gaussian autoregressive moving average (ARMA) class. We simulate $T = 500$ observations from an invertible moving average model of order 2 (MA(2)),

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}, \tag{16}$$

where $\varepsilon_t \sim N(0, \sigma)$, and the true values of the unknown parameters are given by $\theta_{10} = 0.8$, $\theta_{20} = 0.6$ and $\sigma_0 = 1.0$. Inference on $\theta = (\theta_1, \theta_2, \sigma)'$ is conducted via ABC using the sample autocovariances as summary statistics, with $\eta^{(l)}(\mathbf{y}) = (\gamma_0, \gamma_1, ..., \gamma_l)'$, and $\gamma_l = cov(y_t, y_{t-l})$. Four alternative sets of $\eta^{(l)}(\mathbf{y})$ are considered in this case, with $l = 1, 2, 3, 4$, with the one-step-ahead predictive distribution $g^{(l)}(y_{T+1}|\mathbf{y})$ estimated for each set by using the selected draws, θ^i, $i = 1, 2, ..., N$, (again, via a nearest-neighbour version of Algorithm 1) from $p_\varepsilon(\theta|\eta^{(l)}(\mathbf{y}))$ to define $p(y_{T+1}|\theta^i, \mathbf{y})$, from which draws y^i_{T+1}, $i = 1, 2, ..., N$, are taken and used to produce a kernel density estimate of $g^{(l)}(y_{T+1}|\mathbf{y})$. We note that the moving average dependence in (16) means that reduction to a sufficient set of statistics of dimension smaller than T is not feasible. Hence, none of the sets of statistics considered here are sufficient for θ and $p_\varepsilon(\theta|\eta^{(l)}(\mathbf{y}))$ is, once again, distinct from $p(\theta|\mathbf{y})$ for all l.

Panels (A)-(C) in Figure 2 depict the marginal posteriors for each of the three parameters: the four ABC posteriors are given by the dotted and dashed curves of various types, with the relevant summary statistic (vector) indicated in the key appearing in Panel A. The exact marginals (the full curves) for all parameters are computed using the sparse matrix representation of the MA(2) process in an MCMC algorithm comprised of standard Gibbs-Metropolis-Hastings (MH) steps (see, in particular, Chan, 2013). All five densities are computed using 500 draws of the relevant parameter. For the ABC densities this is achieved by retaining (approximately) the smallest 0.5% of the distances in Algorithm 1, based on $N = 111,803$ total draws.[6] For the exact posterior this is achieved by running the chain for $N = 20,000$ iterates (after a burn-in of 5000) and selecting every 40^{th} draw.

Panel (D) of Figure 2 plots the one-step-ahead predictive densities - both approximate and exact. As is consistent with the previous example, the contrast between the two sets of graphs in Figure 1 is stark. The ABC posteriors in Panels (A)-(C) are all very inaccurate representations of the corresponding exact marginals, in addition to being, in some cases, very different one from the other. In contrast, in Panel D three of the four ABF predictives (associated with $\eta^{(1)}(\mathbf{y})$, $\eta^{(2)}(\mathbf{y})$ and $\eta^{(3)}(\mathbf{y})$) are all very similar, one to the other, and *extremely accurate* as representations of the exact predictive; indeed, the approximate predictive generated by $\eta^{(4)}(\mathbf{y})$ is also relatively close to all other densities.

[6] An explanation of this particular choice for the selected proportion (and, hence, N) is provided in the next section.

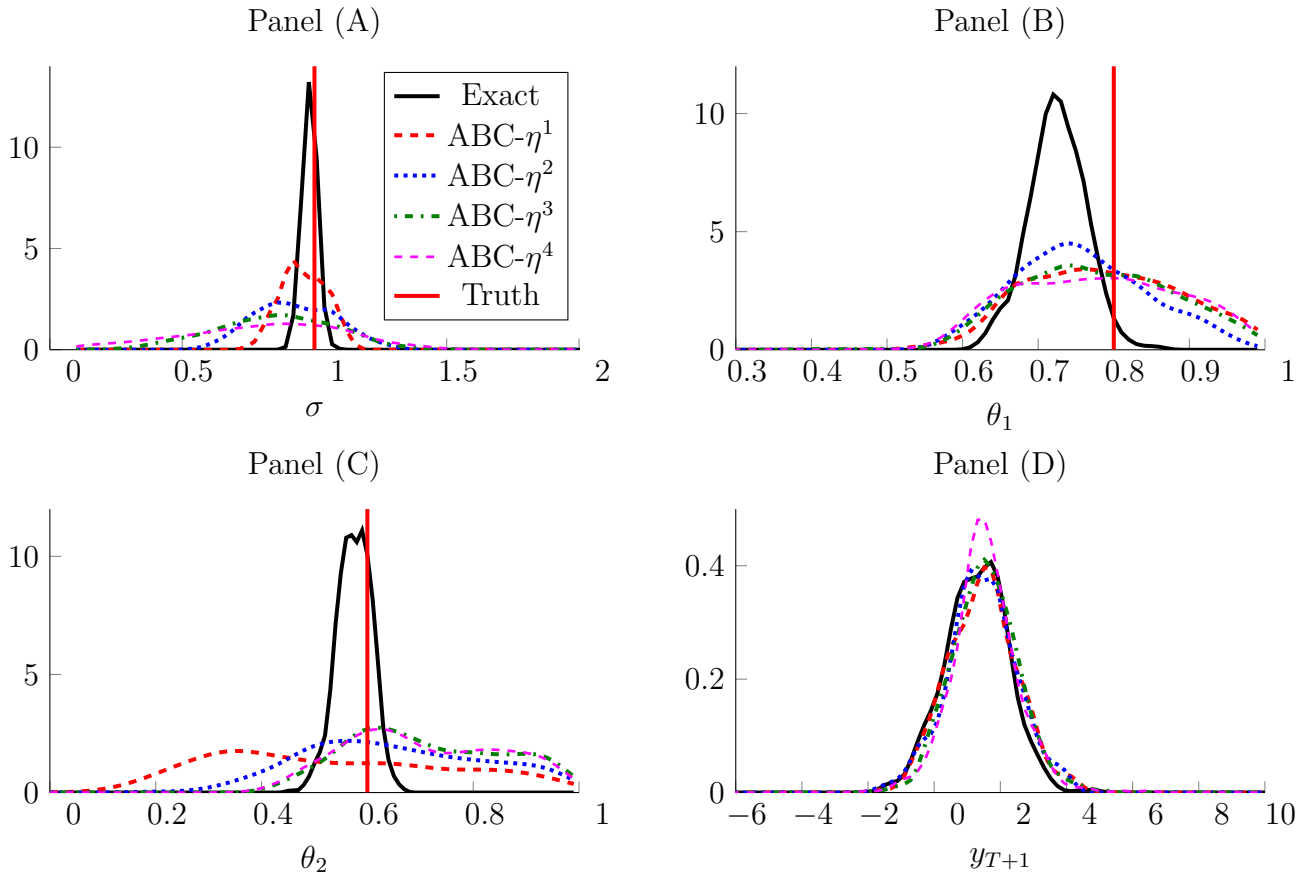

Figure 2: Panels (A), (B) and (C) depict the marginal posteriors (exact and ABC) for σ, θ_1 and θ_2 respectively. The four approximate posteriors are based on the sets of summaries indicated in the key included in Panel A. Panel (D) plots the one-step-ahead predictive densities - both exact and approximate (ABC-based). The red vertical line (denoted by 'Truth' in the key) represents the true value of the relevant parameter in Panels (A), (B) and (C).

We now numerically illustrate the content of Section 3.3, by performing a similar exercise to that undertaken in the previous section: we construct a series of 500 expanding window one-step-ahead predictive distributions (beginning with a sample size of $T = 500$) and record the average LS, QS and cumulative rank probability score (CRPS) for each case in Table 2. It is clear that the MCMC-based predictive, which serves as a simulation-based estimate of $p(y_{T+1}|\mathbf{y})$, generates the highest average score, as is consistent with (12). Nevertheless, the ABF predictives yield average scores that are *nearly identical* to those based on MCMC, indeed in one case (for $l = 2$) equivalent to two decimal places. That is, the extent of the loss associated with the use of insufficient summaries is absolutely minimal. Moreover, we note that the computational time required to produce the MCMC-based estimate of the exact predictive for the case of $T = 500$ is just over 6 minutes, which is approximately 115

times greater than that required to produce any of the approximate predictives! In any real-time exercise in which repeated production of such predictions were required, the *vast* speed improvement yielded by ABF in this example, and with such minimal loss of accuracy, could be of enormous practical benefit.

Table 2: Log score (LS), quadratic score (QS) and cumulative rank probability score (CRPS) associated with the approximate predictive density $g^{(l)}(y_{T+1}|\mathbf{y})$, $l = 1, 2, 3, 4$, and the exact MCMC-based predictive, $p(y_{T+1}|\mathbf{y})$, each computed as an average over a series of 500 (expanding window) one-step-ahead predictions. The predictive with highest average score is in bold.

	$l=1$	$l=2$	$l=3$	$l=4$	MCMC
LS	-1.43	-1.42	-1.43	-1.43	**-1.40**
QS	0.28	0.28	0.28	0.28	**0.29**
CRPS	-0.57	**-0.56**	-0.57	-0.57	**-0.56**

3.4.3 Numerical evidence of merging

In this final sub-section we illustrate the matter of predictive merging and posterior consistency. To this end, we now consider data $\mathbf{y}$ simulated from (16), using increasing sample sizes: $T = 500$, $T = 2000$, $T = 4000$ and $T = 5000$. We also now make explicit that, of the four sets of summaries that we continue to use in the illustration, the three sets, $\eta^{(2)}(\mathbf{y})$, $\eta^{(3)}(\mathbf{y})$ and $\eta^{(4)}(\mathbf{y})$ are such that $p_\varepsilon(\theta|\eta^{(l)}(\mathbf{y}))$ is Bayesian consistent, whilst $\eta^{(1)}(\mathbf{y})$ can be readily shown to *not* yield Bayesian consistency.

We document the merging across four separate measures; with all results represented as averages over 100 synthetic samples. We compute the RMSE based on the distance between the CDF for the approximate and exact predictives, as a numerical approximation of

$$\int \left(dP_{\mathbf{y}} - dG_{\mathbf{y}}\right)^2 d\mu, \tag{17}$$

for μ the Lebesgue measure. Similarly, we compute (numerical approximations of) the total variation metric,

$$\rho_{TV}\{P_{\mathbf{y}}, G_{\mathbf{y}}\} = \sup_{B\in\mathcal{F}} |P_{\mathbf{y}}(B) - G_{\mathbf{y}}(B)|, \tag{18}$$

the Hellinger distance,

$$\rho_H\{P_{\mathbf{y}}, G_{\mathbf{y}}\} = \left\{\frac{1}{2}\int \left[\sqrt{dP_{\mathbf{y}}} - \sqrt{dG_{\mathbf{y}}}\right]^2 d\mu\right\}^{1/2}, \tag{19}$$

and the overlapping measure (OVL) (see Blomstedt and Corander, 2015) defined as,

$$\left[\int \min\{p(y_{T+1}|\mathbf{y}), g(y_{T+1}|\mathbf{y})\} dy_{T+1}\right]^2. \tag{20}$$

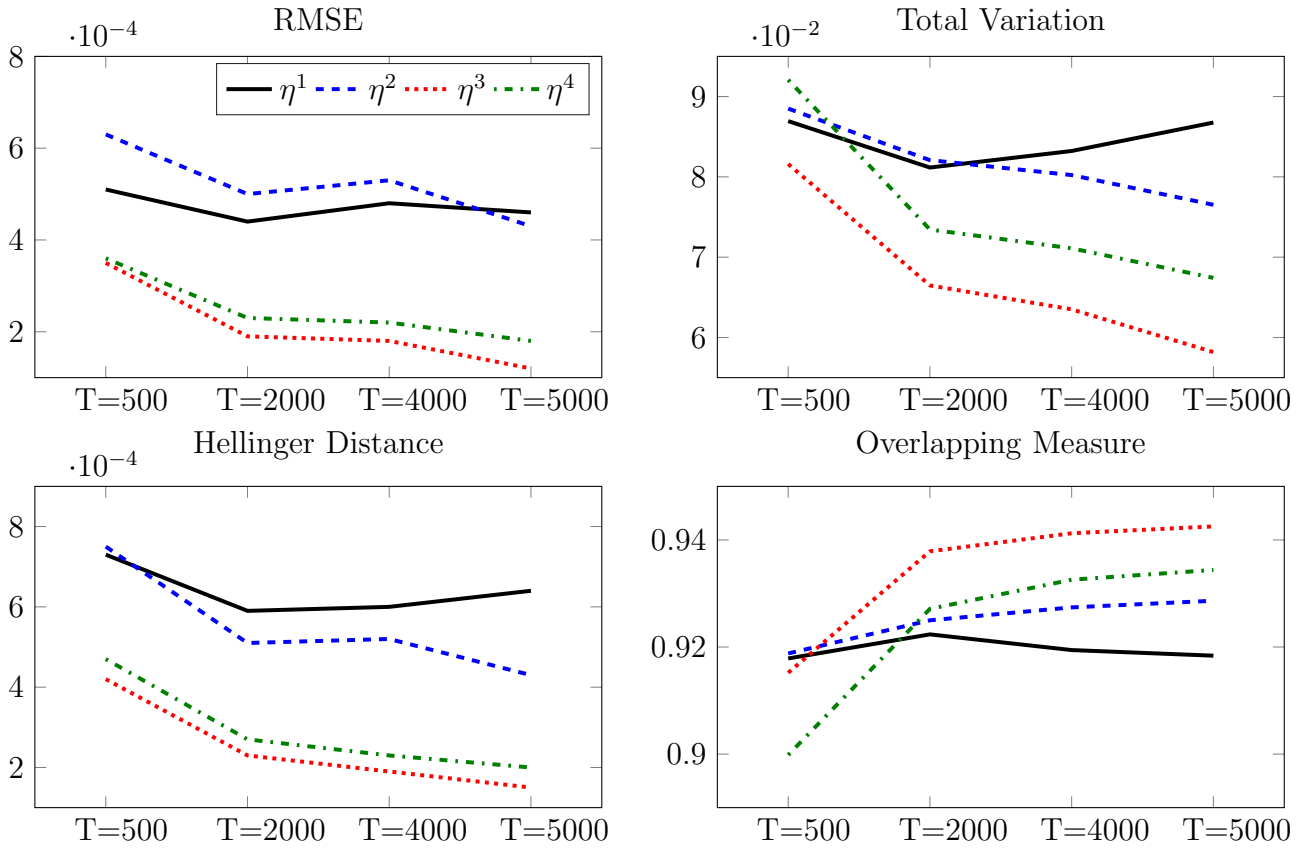

Figure 3: The four panels depict numerical approximations to the measures in (17)-(20). The key in the upper left-hand panel indicates the set of summaries that underpins the ABC-based predictive used in each sequence of computations over T.

Small RMSE, supremum and Hellinger distances indicate closeness of the approximate and exact predictive distributions, while large values of OVL indicate a large degree of overlap between the two distributions. These four measures are presented graphically in Figure 3.

All four panels in Figure 3 illustrate precisely the role played by Bayesian consistency in producing a merging of predictive distributions, in accordance with Theorem 1. Specifically, the RMSE, total variation and Hellinger distances uniformly decrease, while the OVL measure uniformly increases, as T increases, for the cases of ABF conducted with $\eta^{(l)}(\mathbf{y})$ for $l = 2, 3, 4$ (all of which are associated with Bayesian consistent inference). Only in the case of ABF based on $\eta^{(1)}(\mathbf{y})$ (for which $p_\varepsilon(\theta|\eta^{(1)}(\mathbf{y}))$ is not Bayesian consistent) is a uniform decline for RMSE and the total variation and Hellinger distances, not in evidence, and a uniform increase in OVL not observed.

We comment here that in order to satisfy the theoretical results discussed in Section 3.1, we require that the number of draws taken for the ABC algorithm increases with T. This is a consequence of replacing the acceptance step in Algorithm 1 by a nearest-neighbour selection

step, with draws of θ being retained only if they fall below a certain left-hand-tail quantile of the simulated distances. The theoretical results in Frazier et al. (2016) remain valid under this more common implementation of ABC, but they must be cast in terms of the limiting behaviour of the acceptance probability $\alpha_T = Pr\left[d\{\eta(\mathbf{y}), \eta(\mathbf{z})\} \leq \varepsilon_T\right]$. Under this nearest-neighbour interpretation, to ensure consistency, and for N_T denoting the number of Monte Carlo draws used in ABC, we choose $N_T = 500/\alpha_T$ and $\alpha_T = 50T^{-3/2}$. In contrast, the number of MCMC draws used to produce the exact predictives for each sample size remains fixed at 20,000 draws, with a burn-in of 5,000 iterations. However, despite the vast increase in the total number of Monte Carlo draws used in ABC, as T increases, the computation gains in using the ABC algorithm to produce predictive distributions remains marked. In accordance with the result reported in Section 3.4.2, for $T = 500$ the ABF computation is approximately 115 times faster than the exact computation. The relative computational gain factors for $T = 2000$ and $T = 4000$ are 21 and 9, respectively, while a gain of a factor of almost 5 is still achieved at $T = 5000$.[7]

4 ABF in State Space Models

Thus far the focus has been on the case in which the vector of unknowns, θ, is a k_θ-dimensional set of parameters for which informative summary statistics are sought for the purpose of generating probabilistic predictions. By implication, and certainly in the case of both the INAR(1) and MA(2) examples, the elements of θ are static in nature, with k_θ small enough for a set of summaries of manageable dimension to be defined with relative ease.

State space models, in which the set of unknowns is augmented by a vector of random parameters that are of dimension equal to or greater than the sample size, present additional challenges for ABC (Creel and Kristensen, 2015; Martin et al., 2017), in terms of producing an ABC posterior for the static parameters, θ, that is a good match for the exact. However, the results in the previous section highlight that accuracy at the posterior level is not necessary for agreement between the approximate and exact predictives. This suggests that we may be able to choose a crude, but computationally convenient, method of generating summaries for θ in a state space model, and still yield predictions that are close to those given by exact methods. The results below confirm this intuition, as well as making it clear that exact

[7] The requirement that N_T diverge, at a particular rate, is intimately related to the inefficient nature of the basic accept/reject ABC approach. In large samples, it is often useful to use more refined sampling techniques within ABC, as these approaches can often lead to faster estimates of the ABC posterior than those obtained via the accept/reject approach. Thus, at least in large samples, utilizing more efficient ABC approaches will lead to a decrease in ABF computing times, which will lead to an even higher computational gain over MCMC-based approaches. See Li and Fearnhead (2015) for alternative sampling schemes that only require $N_T \to \infty$ very slowly.

posterior inference on the full vector of states (and the extra computational complexities that such a procedure entails) is not required for this accuracy to be achieved.

We illustrate these points in the context of a very simple state space model, namely a stochastic volatility model for a financial return, y_t, in which the logarithm of the random variance, V_t, follows a simple autoregressive model of order 1 (AR(1)):

$$y_t = \sqrt{V_t}\varepsilon_t; \qquad \varepsilon_t \sim i.i.d.N(0,1) \tag{21}$$

$$\ln V_t = \theta_1 \ln V_{t-1} + \eta_t; \qquad \eta_t \sim i.i.d.N(0,\theta_2) \tag{22}$$

with $\theta = (\theta_1, \theta_2)'$. To generate summary statistics for the purpose of defining $p_\varepsilon(\theta|\eta(\mathbf{y}))$, we begin by adopting the following auxiliary generalized autoregressive conditional heteroscedastic model with Gaussian errors (GARCH-N):

$$y_t = \sqrt{V_t}\varepsilon_t; \qquad \varepsilon_t \sim i.i.d.N(0,1) \tag{23}$$

$$V_t = \beta_1 + \beta_2 V_{t-1} + \beta_3 y_{t-1}^2. \tag{24}$$

As a computationally efficient summary statistic vector for use in ABF we use the score of the GARCH-N likelihood function, computed using the simulated and observed data, with both evaluated at the (quasi-) maximum likelihood estimator of $\beta = (\beta_1, \beta_2, \beta_3)'$ (see, for example, Drovandi et al., 2015, and Martin et al., 2017).

The exact predictive, $p(y_{T+1}|\mathbf{y})$, requires integration with respect to both the static and latent parameters, including the value of the latent variance at time $T+1$. Defining $p(V_{T+1}, \mathbf{V}, \theta|\mathbf{y})$ as the joint posterior for this full set of unknowns, and recognizing the Markovian structure in the (log) variance process, we can represent this predictive as

$$\begin{aligned} &p(y_{T+1}|\mathbf{y}) \\ &= \int_{V_{T+1}} \int_{\mathbf{V}} \int_{\theta} p(y_{T+1}|V_{T+1}) p(V_{T+1}, \mathbf{V}, \theta|\mathbf{y}) d\theta d\mathbf{V} dV_{T+1} \\ &= \int_{V_{T+1}} \int_{\mathbf{V}} \int_{\theta} p(y_{T+1}|V_{T+1}) p(V_{T+1}|V_T, \theta, \mathbf{y}) p(\mathbf{V}|\theta, \mathbf{y}) p(\theta|\mathbf{y}) d\theta d\mathbf{V} dV_{T+1}. \end{aligned} \tag{25}$$

A hybrid Gibbs-MH MCMC algorithm is applied to yield posterior draws of θ and $\mathbf{V}$. We apply the sparse matrix sampling algorithm of Chan and Jeliazkov (2009) to sample $\mathbf{V}$, and a standard Gibbs algorithm to sample from the conditional posterior of θ given the states. Conditional on the draws of θ and V_T (in particular), draws of V_{T+1} and y_{T+1} are produced directly from $p(V_{T+1}|V_T, \theta, \mathbf{y})$ and $p(y_{T+1}|V_{T+1})$ respectively, and the draws of y_{T+1} used to produce an estimate of $p(y_{T+1}|\mathbf{y})$.

Replacing $p(\theta|\mathbf{y})$ in (25) by $p_\varepsilon(\theta|\eta(\mathbf{y}))$, the approximate predictive is then defined as

$$\begin{aligned} &g(y_{T+1}|\mathbf{y}) \\ &= \int_{V_{T+1}} \int_{\mathbf{V}} \int_{\theta} p(y_{T+1}|V_{T+1}) p(V_{T+1}|V_T, \theta, \mathbf{y}) p(\mathbf{V}|\theta, \mathbf{y}) p_\varepsilon(\theta|\eta(\mathbf{y})) d\theta d\mathbf{V} dV_{T+1}. \end{aligned} \tag{26}$$

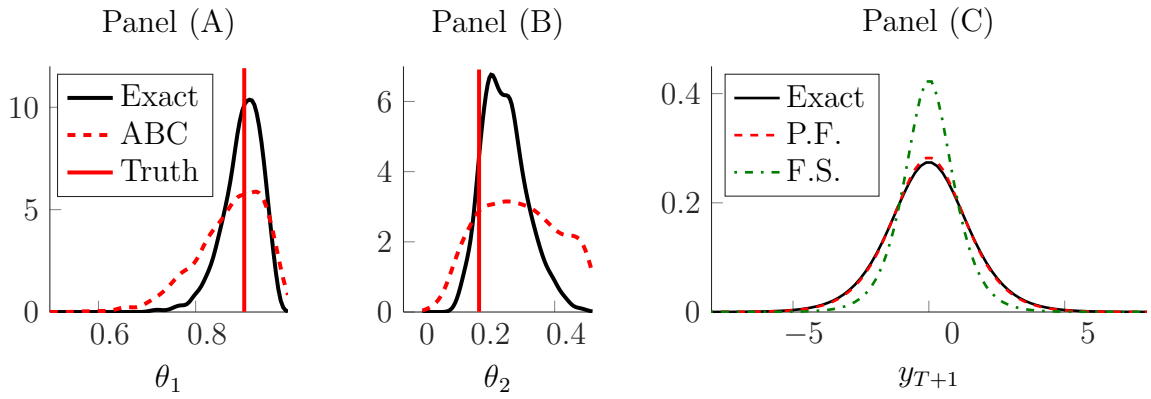

Figure 4: Panels (A) and (B) depict the marginal posteriors (exact and ABC) for θ_1 and θ_2, respectively. Panel (C) plots the one-step-ahead predictive density functions - both exact and approximate (ABC-based). P.F. indicates the approximate predictive computed using the particle filtering step; F.S. indicates the approximate predictive computed using a forward simulation step (for the latent variance) only. The red vertical line (denoted by 'Truth' in the key) represents the true value of the relevant parameter in Panels (A) and (B).

In this case, however, draws are produced from $p_\varepsilon(\theta|\eta(\mathbf{y}))$ via Algorithm 1, separately from the treatment of $\mathbf{V}$. That is, posterior draws of $\mathbf{V}$, including V_T, are not an automatic output of a simulation algorithm applied to the joint set of unknowns θ and $\mathbf{V}$, as was the case in the estimation of (25). However, the estimation of $g(y_{T+1}|\mathbf{y})$ requires only that posterior draws of V_T and θ are produced; that is, posterior inference on the *full vector* $\mathbf{V}$, as would require a backward smoothing step of some sort to be embedded within the simulation algorithm, is not necessary. All that is required is that $\mathbf{V}_{1:T-1}$ are integrated out, and this can occur simply via a forward filtering step. The implication of this is that, conditional on a simple *i.i.d.* version of Algorithm 1 being adopted (i.e., that no MCMC modifications of ABC are employed), a simulation-based estimate of the approximate predictive can still be produced using *i.i.d.* draws only. As such, the great gains in computational speed afforded by the use of ABC - including access to parallelization - continue to obtain even when latent variables characterize the true DGP.

Panels (A) and (B) of Figure 4 depict the marginal ABC posteriors of θ_1 and θ_2 alongside the MCMC-based comparators. The dashed curve in Panel (C) of Figure 4 then represents the estimate of (26), in which the particle filter is used to integrate out the latent variances, and the full curve represents the MCMC-based estimate of (25). As is consistent with the numerical results recorded earlier for the INAR(1) and MA(2) examples, the difference between the approximate and exact posteriors is marked, whilst - at the same time - the approximate predictive is almost equivalent to the exact, and having been produced using a much simpler algorithm, and in a fraction of the time!

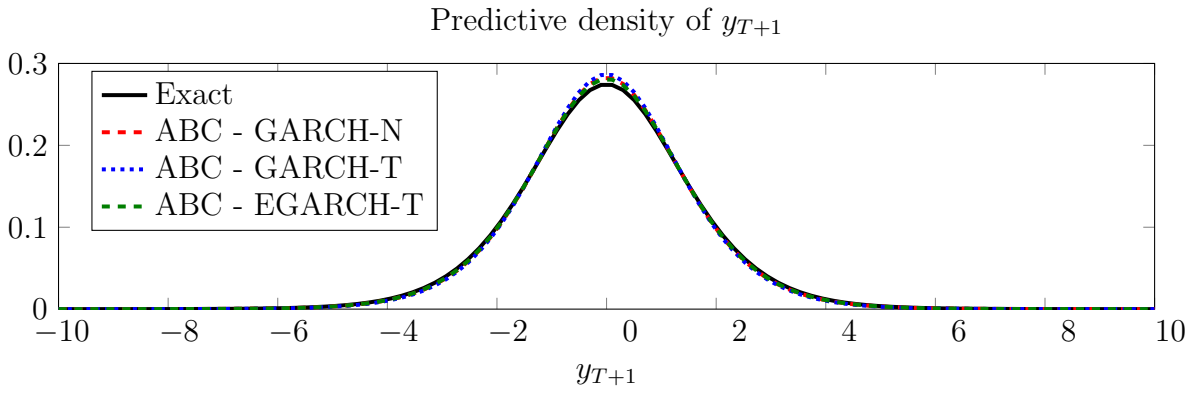

Figure 5: Exact and approximate (ABC-based) predictives. The three approximate posteriors are based on the auxiliary models indicated in the key. All approximate predictives use the particle filtering step.

The importance of the particle filtering step in obtaining this (near) equivalence is highlighted by the inclusion of a third predictive (the finely dashed curve) in Panel (C) of Figure 4, which is constructed by replacing the particle filtering step by a simple forward simulation of the latent variance model in (22) - conditional on the ABC draws of θ - such that inference on V_T is itself conditioned on $\eta(\mathbf{y})$, rather than $\mathbf{y}$. Without full posterior inference on V_T, the gains obtained by ABC inference on θ (in terms of computational speed and ease) are achieved only at the cost of producing an inaccurate estimate of the exact predictive. We reiterate, however, that full posterior inference on V_T (as reflected in the very accurate dashed curve in Panel (C) of Figure 4) requires *only* a particle filtering step.

To further highlight the apparent second-order importance of *static* parameter inference on the predictive, along with the exact and approximate predictives reproduced in Panel (C) of Figure 4 (namely the full and dashed curves), Figure 5 plots two alternative approximate predictives, that use different auxiliary models to define the summary statistics. The GARCH-T auxiliary model employs the structure as defined in (23) and (24), but with a Student-t error term, $\varepsilon_t \sim i.i.d.\ t(\nu)$, used to accommodate extra leptokurtosis in the return. The EGARCH-T auxiliary model also employs Student-t errors, but with skewness in the return also modeled via an asymmetric specification for the conditional variance:

$$\ln V_t = \beta_0 + \beta_1 \ln V_{t-1} + \beta_2 \left(|\varepsilon_{t-1}| - \mathbb{E}(|\varepsilon_{t-1}|)\right) + \beta_3 \varepsilon_{t-1}.$$

As is clear, given the inclusion of the particle filtering step, the choice of auxiliary model has little impact on the nature of the resultant predictive, with all three auxiliary models generating approximate predictives that are extremely close.

This robustness of prediction to the choice of summary statistics augers well for the automated use of ABC as a method for generating Bayesian predictions in models where

finely-tuned, specialized algorithms have been viewed as an essential ingredient up to now. It also suggests that Bayesian predictions that are close to exact can be produced in models in which exact prediction is infeasible, that is, in models where the DGP - and hence, the exact predictive itself - is unavailable. It is precisely such a case that we explore in the following empirical section, with performance now gauged not in terms of the accuracy with which any particular $g(y_{T+1}|\mathbf{y})$ matches $p(y_{T+1}|\mathbf{y})$, but in terms of out-of-sample predictive accuracy.

5 Empirical Illustration: Forecasting Financial Returns and Volatility

5.1 Background, model and computational details

The effective management of financial risk entails the ability to plan for unexpected, and potentially large, movements in asset prices. Central to this is the ability to accurately quantify the probability distribution of the future return on the asset, including its degree of variation, or volatility. The stylized features of time-varying and autocorrelated volatility, allied with non-Gaussian return distributions, are now extensively documented in the literature (Bollerslev et al., 1992); with more recent work focusing also on random 'jump' processes, both in the asset price itself and its volatility (Broadie et al., 2007, Bandi and Renò, 2016, and Maneesoonthorn et al., 2017). Empirical regularities documented in the option pricing literature (see Garcia et al., 2011 for a recent review), most notably implied volatility 'smiles', are also viewed as evidence that asset prices do not adhere to the geometric Brownian motion assumption underlying the ubiquitous Black-Scholes option price, and that the processes driving asset returns are much more complex in practice.

Motivated by these now well-established empirical findings, we explore here a state space specification for financial returns on the S&P500 index, in which both stochastic volatility *and* random jumps are accommodated. To do so, we supplement a measurement equation for the daily return, in which a dynamic jump process features, with a second measurement equation based on bipower variation, constructed using five-minute intraday returns over the trading day (Barndorff-Nielsen and Shephard, 2004). Such a model is representative of models used recently to capture returns data in which *clustering* of jumps features, in addition to the stylized autocorrelation in the diffusive variance (Fulop et al., 2014; Aït-Sahalia et al., 2015; Bandi and Renò, 2016; Maneesoonthorn et al., 2017). It also reflects the recent trend of exploiting high frequency data to construct - and use as additional measures in state space settings - nonparametric measures of return variation, including jumps therein (Koopman and Scharth, 2012; Maneesoonthorn et al., 2012; Maneesoonthorn et al.,

2017). To capture the possibility of extreme movements in volatility, and in the spirit of Lombardi and Calzolari (2009) and Martin et al. (2017), we adopt an α-stable process for the volatility innovations. Despite the lack of a closed-form transition density, the α-stable process presents no challenges for ABC-based inference and forecasting, given that such a process can still be simulated via the algorithm of Chambers et al. (1976).

In summary, the assumed data generating process comprises two measurement equations: one based on daily logarithmic returns, r_t

$$r_t = \exp\left(\frac{h_t}{2}\right)\varepsilon_t + \Delta N_t Z_t, \tag{27}$$

where $\varepsilon_t \sim i.i.d.N(0,1)$, h_t denotes the latent logarithmic variance process, ΔN_t the latent jump occurrence and Z_t the latent jump size; and a second using logarithmic bipower variation,

$$\ln BV_t = \psi_0 + \psi_1 h_t + \sigma_{BV}\zeta_t, \tag{28}$$

where $BV_t = \frac{\pi}{2}\left(\frac{M}{M-1}\right)\sum_{i=2}^{M}|r_{t_i}|\left|r_{t_{i-1}}\right|$, with r_{t_i} denoting the i^{th}, of M equally-spaced returns observed during day t, and $\zeta_t \sim i.i.d.N(0,1)$. As is now well-known (Barndorff-Nielsen and Shephard, 2004), under certain conditions BV_t is a consistent, but potentially biased (for finite M), estimate of integrated volatility over day t, with h_t here being a discretized representation of the (logarithm of the) latter. The latent states in equations (27) and (28), h_t, Z_t and ΔN_t, evolve, respectively, according to

$$h_t = \omega + \rho h_{t-1} + \sigma_h \eta_t \tag{29}$$

$$Z_t \sim N(\mu, \sigma_z^2) \tag{30}$$

$$Pr(\Delta N_t = 1|\mathcal{F}_{t-1}) = \delta_t = \delta + \beta\delta_{t-1} + \gamma\Delta N_{t-1}, \tag{31}$$

where $\eta_t \sim i.i.d.\mathcal{S}(\alpha, -1, 0, dt = 1)$. We note that the model for the jump intensity, δ_t, is the conditionally deterministic Hawkes structure adopted by Fulop et al. (2014), Aït-Sahalia et al. (2015) and Maneesoonthorn et al. (2017). We estimate δ (in (31)) indirectly via the unconditional intensity implied by this particular structure, namely, $\delta^0 = \delta/(1-\beta-\gamma)$.

Exact inference on the full set of static parameters,

$$\theta = (\psi_0, \psi_1, \sigma_{BV}, \omega, \rho, \sigma_h, \alpha, \delta^0, \beta, \gamma, \mu, \sigma_z)', \tag{32}$$

is challenging, not only due to the overall complexity of the model, but in particular as a consequence of the presence of α-stable (log) volatility transitions. Hence, ABC is a natural choice for inference on θ. Moreover, given the previously presented evidence regarding the accuracy with which ABC-based predictives match the predictive that *would* be yielded by an exact method, one proceeds with some confidence to build Bayesian predictives via ABC posteriors.

To measure the predictive performance of our ABF approach, we consider an out-of-sample predictive exercise, whereby we assess the accuracy of the approximate predictives as based on different choices of summaries, $\eta(\mathbf{y})$. We make two comments here. First, and as highlighted in the previous section, a forward particle filtering step (conditional on draws from the ABC posterior) is required to produce the full posterior inference on the latent state, h_T, that is required to construct $g(y_{T+1}|\mathbf{y})$ under any choice for $\eta(\mathbf{y})$. We adopt the bootstrap particle filter of Gordon et al. (1993) for this purpose.[8] Second, when the data generating process is correctly specified, and if the conditions for Bayesian consistency and asymptotic normality of both the exact and ABC posteriors are satisfied, then the out-of-sample accuracy of $g(y_{T+1}|\mathbf{y})$ is bounded above by that of $p(y_{T+1}|\mathbf{y})$, as measured by some proper scoring rule.[9] Hence, in maximizing predictive accuracy via the choice of $\eta(\mathbf{y})$, we are - in spirit - choosing an approximate predictive that is as close as possible to the inaccessible exact predictive.

We consider observed data from 26 February 2010 to 7 February 2017, comprising 1750 daily observations on both r_t and BV_t. We reserve the most recent 250 observations (approximately one trading year) for one-step-ahead predictive assessments, using an expanding window approach. In the spirit of the preceding section, we implement ABC using the scores of alternative auxiliary GARCH models fitted to daily returns. In this case, however, we must also conduct inference on the parameters of the additional measurement equation, (28), and the jump processes in (30) and (31); hence we supplement the auxiliary model scores with additional summary statistics based on both BV_t as well as the realized jump variation measure, $JV_t = \max(RV_t - BV_t, 0)$, where $RV_t = \sum_{i=1}^{M} r_{t_i}$ defines so-called realized variance for day t.

We consider four auxiliary models: GARCH with normal and Student-t errors (GARCH-N and GARCH-T, respectively), threshold GARCH with Student-t errors (TARCH-T), and the realized GARCH (RGARCH) model of Hansen et al. (2012). Table 3 details these four models, plus the additional summary statistics that we employ in each case. In particular, we note that the RGARCH model itself incorporates a component in which $\ln BV_t$ is modeled; hence, in this case we do not adopt additional summary statistics based on this measure. We adopt independent uniform priors for all static parameters in the structural model, subject to relevant model-based restrictions, with the lower and upper bounds for each given in Table 4.

[8]Note that the conditionally deterministic structure in (31) means that no additional filtering step is required in order to model the jump intensity at time T.

[9]See Section 3.3 for a heuristic derivation of this result, and in particular equations (11) and (12).

5.2 Empirical forecasting results

In Table 5 we report the marginal ABC posterior means (MPM) and the 95% highest posterior density (HPD) intervals for the elements of θ, based on the four choices of summaries. The posterior results obtained via the first three sets (based, in turn, on the GARCH-N, GARCH-T and TARCH-T auxiliary models) are broadly similar, except for the TARCH-T auxiliary model producing noticeable narrower 95% HPD intervals for ω, μ and σ_z than the other auxiliary models. In contrast to the relative conformity of these three sets of results, the RGARCH auxiliary model (augmented by the additional summaries) produces ABC posteriors that differ quite substantially. Most notably, and with reference to the latent process for h_t in (29), ABC based on this fourth set of summaries produces a larger MPM for ω, a lower MPM of ρ, and a smaller MPM for σ_h than do the other instances of ABC. In addition this version produces a larger point estimate for the mean jump size, μ, plus a smaller point estimate of the jump size variation, σ_z. These differences imply somewhat different conclusions regarding the process generating returns than those implied by the other three sets of ABC posterior results. As a consequence there would be differing degrees of concordance between the four sets of ABC posteriors and the corresponding exact, unattainable, posteriors. The question of interest here is the extent to which such differences translate into substantial differences at the predictive level, where a judgment is made solely in terms of out-of-sample predictive accuracy, given our lack of access to the exact predictive.

To summarize predictive performance over the out-of-sample period, average LS, QS and CRPS values for each of the four approximate predictives are reported in Table 6, with the largest figure in each case indicated in bold. The results indicate that the predictive distribution for r_t generated via the TARCH-T auxiliary model (and additional summaries) performs best according to all three score criteria. The GARCH-N auxiliary model (and additional summaries) generates the best-performing predictive distribution for $\ln BV_t$ according to LS and QS, but with CRPS still suggesting that the TARCH-T-based predictive performs the best. It is interesting to note that the set of statistics that generates the worst overall predictive performance (with the lowest predictive scores in most cases) is that which includes the RGARCH auxiliary model - i.e. the set that resulted in ABC marginal posteriors that were distinctly different from those obtained via the other three statistic sets.

In summary, these predictive outcomes - in which the T-GARCH-based set of summaries performs best - suggest that this choice of summaries be the one settled upon. Repeating the point made above, for any finite sample the predictive performance of any approximate predictive will (under appropriate regularity conditions) be bounded above by that of the exact predictive; however, this difference is likely to be minor under correct specification of the DGP.

Table 3: Auxiliary model specifications for ABC posterior inference for the model in (27)-(31). The final column gives the set of supplementary summary statistics used alongside the scores from each auxiliary model. The error terms, ε_t and ζ_t, are specified as *i.i.d.* The notation $\hat{\sigma}_t$ in the third column refers to fitted volatility from the corresponding volatility equation in the auxiliary model.

Auxiliary Model	Auxiliary Model Specification	Supplementary Statistics
GARCH-N	$r_t = \sigma_t \varepsilon_t,\ \varepsilon_t \sim N(0,1)$ $\sigma_t^2 = \gamma_0 + \gamma_1 r_{t-1}^2 + \gamma_2 \sigma_{t-1}^2$	$Mean(sign(r_t)\sqrt{JV_t})$, $Var(JV_t)$ $Corr(JV_t, JV_{t-1})$ Skewness($\ln BV_t$), Kurtosis($\ln BV_t$) Regression coefficients from $\ln BV_t = \kappa_0 + \kappa_1 \ln \hat{\sigma}_t^2 + \kappa_3 \zeta_t$
GARCH-T	$r_t = \sigma_t \varepsilon_t,\ \varepsilon_t \sim t(\nu)$ $\sigma_t^2 = \gamma_0 + \gamma_1 r_{t-1}^2 + \gamma_2 \sigma_{t-1}^2$	$Mean(sign(r_t)\sqrt{JV_t})$, $Var(JV_t)$ $Corr(JV_t, JV_{t-1})$ Skewness($\ln BV_t$), Kurtosis($\ln BV_t$) Regression coefficients from $\ln BV_t = \kappa_0 + \kappa_1 \ln \hat{\sigma}_t^2 + \kappa_3 \zeta_t$
TARCH-T	$r_t = \sigma_t \varepsilon_t,\ \varepsilon_t \sim t(\nu)$ $\sigma_t^2 = \gamma_0 + \gamma_1 r_{t-1}^2 + \gamma_2 I_{(r_{t-1}<0)} r_{t-1}^2$ $+\gamma_3 \sigma_{t-1}^2$	$Mean(sign(r_t)\sqrt{JV_t})$, $Var(JV_t)$ $Corr(JV_t, JV_{t-1})$ Skewness($\ln BV_t$), Kurtosis($\ln BV_t$) OLS regression coefficients from $\ln BV_t = \kappa_0 + \kappa_1 \ln \hat{\sigma}_t^2 + \kappa_3 \zeta_t$
RGARCH	$r_t = \sigma_t \varepsilon_t,\ \varepsilon_t \sim N(0,1)$ $\ln \sigma_t^2 = \gamma_0 + \gamma_1 \ln BV_{t-1} + \gamma_2 \ln \sigma_{t-1}^2$ $\ln BV_t = \gamma_3 + \gamma_4 \ln \sigma_{t-1}^2 + \gamma_5 \varepsilon_t$ $+\gamma_6 (\varepsilon_t^2 - 1) + \gamma_7 u_t,\ u_t \sim N(0,1)$	$Mean(sign(y_t)\sqrt{JV_t})$, $Var(JV_t)$ $Corr(JV_t, JV_{t-1})$, Kurtosis($\ln BV_t$)

Table 4: Lower and upper bounds on the uniform prior specifications used for each element of θ, as defined in (32).

Parameter	ψ_0	ψ_1	σ_{bv}	ω	ρ	σ_h	α	δ	β	γ	μ	σ_z
Lower	-0.50	0.50	0.001	-1	0.50	0.001	1.50	0.001	0.50	0.001	-1	.50
Upper	0.50	1.50	1	1	0.99	0.30	2	0.30	0.99	0.20	1	3

Table 5: Marginal posterior means (MPM) and 95% highest posterior density (HPD) intervals for each of the elements of θ, as defined in (32), obtained from ABC posterior inference using the four auxiliary models and supplementary statistics defined in Table 3.

	GARCH-N		GARCH-T		TARCH-T		RGARCH	
	MPM	95% HPD	MPM	95% HPD	MPM	95% HPD	MPM	95% HPD
ψ_0	-0.019	(-0.466,0.465)	-0.004	(-0.485,0.461)	-0.011	(-0.471,0.475)	-0.012	(-0.479,0.471)
ψ_1	1.261	(0.774,1.492)	1.250	(0.829,1.488)	1.195	(0.725,1.485)	0.963	(0.513,1.448)
σ_{BV}	0.447	(0.021,0.955)	0.471	(0.027,0.953)	0.478	(0.021,0.952)	0.553	(0.038,0.987)
ω	-0.038	(-0.680,0.384)	-0.101	(-0.338,0.203)	-0.171	(-0.483,-0.007)	0.192	(-0.945,0.967)
ρ	0.935	(0.811,0.985)	0.930	(0.825,0.986)	0.919	(0.811,0.983)	0.788	(0.519,0.985)
σ_h	0.196	(0.078,0.293)	0.208	(0.078,0.297)	0.202	(0.061,0.296)	0.132	(0.006,0.291)
α	1.757	(1.521,1.984)	1.756	(1.520,1.985)	1.771	(1.517,1.989)	1.797	(1.521,1.992)
δ^0	0.113	(0.007,0.283)	0.111	(0.010,0.273)	0.103	(0.007,0.281)	0.102	(0.007,0.268)
β	0.692	(0.510,0.901)	0.689	(0.507,0.909)	0.691	(0.509,0.898)	0.688	(0.516,0.897)
γ	0.124	(0.019,0.197)	0.122	(0.018,0.197)	0.121	(0.015,0.195)	0.132	(0.029,0.196)
μ	0.071	(-0.874,0.940)	0.048	(-0.856,0.900)	0.115	(-0.809,0.881)	0.234	(-0.691,0.941)
σ_z	1.208	(0.523,2.566)	1.229	(0.529,2.721)	1.144	(0.526,2.486)	1.011	(0.525,2.153)

Table 6: Average predictive log score (LS), quadratic score (QS) and cumulative rank probability score (CRPS) for the one-step-ahead predictive distributions of r_t and $\ln BV_t$, evaluated between 11 February 2016 and 7 February 2017. The figures in bold indicate the largest average score amongst the four sets of summaries.

		GARCH-N	GARCH-T	TARCH-T	RGARCH
r_t	LS	-1.571	-1.280	**-1.202**	-1.945
	QS	0.377	0.474	**0.515**	0.274
	CRPS	-1.515	-1.052	**-0.989**	-2.103
$\ln BV_t$	LS	**-2.732**	-2.757	-2.928	-2.827
	QS	**0.095**	0.049	0.016	0.094
	CRPS	-2.038	-1.416	**-1.377**	-2.570

6 Discussion

As far as we are aware, this is the first paper to explore the use of approximate Bayesian computation (ABC) in generating probabilistic forecasts and to propose the concept of approximate Bayesian forecasting (ABF). Theoretical and numerical evidence has been presented which indicates that if the assumed data generating process (DGP) is correctly specified, very little is lost - in terms of forecast accuracy - by conducting approximate inference (only) on the unknowns that characterize the DGP. A caveat here applies to latent variable models, in that exact inference on the conditioning latent state(s) would appear to be important. However, even that requires only independent particle draws, to supplement the computationally fast and simple independent draws of the static parameters via ABC; detracting little from the overall conclusion that ABC represents a powerful base on which to produce accurate Bayesian forecasts in rapid time. Whilst the asymptotic results based on merging formally exploit the property of Bayesian consistency, numerical evidence suggests that lack of consistency, or evidence thereof, for the ABC posteriors does not preclude the possibility of close matches to the exact predictive being yielded in any given finite sample. The theoretical results presented regarding expected scores is also borne out in the numerical illustrations, with minor - if any - forecasting loss incurred by moving from exact to approximate prediction, for the sample sizes considered.

Importantly, in an empirical setting where the exact predictive is unattainable, the idea of choosing ABC summaries to produce the best performing approximate predictive is a sensible approach to adopt when predictive accuracy is the primary goal, and when the true DGP is of course unknown. What remains the subject of on-going investigation by the authors, is the interplay between new results on the impact on ABC inference of model misspecification (Frazier et al., 2017) and the performance of ABC in a forecasting setting in which misspecification of the DGP is explicitly acknowledged. The outcomes of this exploration are reserved for future research output.

References

Aït-Sahalia, Y., Cacho-Diaz, J., and Laeven, R. J. (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 117(3):585–606.

Bandi, F. M. and Renò, R. (2016). Price and volatility co-jumps. *Journal of Financial Economics*, 119(1):107–146.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1):1–37.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990.

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

Blackwell, D. and Dubins, L. (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33(3):882–886.

Blomstedt, P. and Corander, J. (2015). Posterior predictive comparisons for the two-sample problem. *Communications in Statistics-Theory and Methods*, 44(2):376–389.

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

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. *Statist. Sci.*, 28(2):189–208.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). Arch modeling in finance: A review of the theory and empirical evidence. *Journal of econometrics*, 52(1-2):5–59.

Broadie, M., Chernov, M., and Johannes, M. (2007). Model specification and risk premia: Evidence from futures options. *The Journal of Finance*, 62(3):1453–1490.

Chambers, J. M., Mallows, C. L., and Stuck, B. (1976). A method for simulating stable random variables. *Journal of the American Statistical Association*, 71(354):340–344.

Chan, J. C. (2013). Moving average stochastic volatility models with application to inflation forecast. *Journal of Econometrics*, 176(2):162–172.

Chan, J. C.-C. and Jeliazkov, I. (2009). MCMC estimation of restricted covariance matrices. *Journal of Computational and Graphical Statistics*, 18(2):457–480.

Creel, M., Gao, J., Hong, H., and Kristensen, D. (2015). Bayesian indirect inference and the ABC of GMM. *arXiv preprint arXiv:1512.07385*.

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.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, pages 1–26.

Drost, F. C., Akker, R. v. d., and Werker, B. J. (2009). Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models. *Journal of the Royal Statistical Society: Series B*, 71(2):467–485.

Drovandi, C. C., Pettitt, A. N., and Lee, A. (2015). Bayesian indirect inference using a parametric auxiliary model. *Statistical Science*, 30(1):72–95.

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

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2016). Asymptotic properties of approximate Bayesian computation. *arXiv preprint arXiv:1607.06903*.

Frazier, D. T., Robert, C. P., and Rousseau, J. (2017). Model misspecification in ABC: consequences and diagnostics. *arXiv preprint arXiv:1708.01974*.

Fulop, A., Li, J., and Yu, J. (2014). Self-exciting jumps, learning, and asset pricing implications. *The Review of Financial Studies*, 28(3):876–912.

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

Garcia, R., Lewis, M.-A., Pastorello, S., and Renault, É. (2011). Estimation of objective and risk-neutral distributions based on moments of integrated volatility. *Journal of Econometrics*, 160(1):22–32.

Ghosal, S., Ghosh, J. K., and Samanta, T. (1995). On convergence of posterior distributions. *The Annals of Statistics*, pages 2145–2152.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer-Verlag New York, Inc.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET.

Hansen, P. R., Huang, Z., and Shek, H. H. (2012). Realized garch: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, 27(6):877–906.

Ibragimov, I. A. and Has'Minskii, R. Z. (2013). *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media.

Jasra, A. (2015). Approximate Bayesian computation for a class of time series models. *International Statistical Review*, 83(3):405–435.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical applications in Genetics and Molecular Biology*, 7(1):1–16.

Jung, R. C. and Tremayne, A. (2006). Binomial thinning models for integer time series. *Statistical Modelling*, 6(2):81–96.

Koopman, S. J. and Scharth, M. (2012). The analysis of stochastic volatility in the presence of daily realized measures. *Journal of Financial Econometrics*, 11(1):76–115.

Li, W. and Fearnhead, P. (2015). On the asymptotic efficiency of ABC estimators. *arXiv preprint arXiv:1506.03481*.

Li, W. and Fearnhead, P. (2016). Improved convergence of regression adjusted approximate Bayesian computation. *arXiv preprint arXiv:1609.07135*.

Lombardi, M. J. and Calzolari, G. (2009). Indirect estimation of α-stable stochastic volatility models. *Computational Statistics & Data Analysis*, 53(6):2298–2308.

Maneesoonthorn, W., Forbes, C. S., and Martin, G. M. (2017). Inference on self-exciting jumps in prices and volatility using high-frequency measures. *Journal of Applied Econometrics*, 32(3):504–532.

Maneesoonthorn, W., Martin, G. M., Forbes, C. S., and Grose, S. D. (2012). Probabilistic forecasts of volatility and its risk premia. *Journal of Econometrics*, 171(2):217–236.

Marin, J.-M., Pillai, N. S., Robert, C. P., and Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B*, 76(5):833–859.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Martin, G., M., McCabe, B. P., Frazier, D. T., Maneesoonthorn, W., and Robert, C. P. (2017). Auxiliary likelihood-based approximate Bayesian computation in state space models. *arXiv preprint arXiv:1604.07949*.

Martin, V. L., Tremayne, A. R., and Jung, R. C. (2014). Efficient method of moments estimators for integer time series models. *Journal of Time Series Analysis*, 35(6):491–516.

McCabe, B. P. M., Martin, G. M., and Harris, D. (2011). Efficient probabilistic forecasts for counts. *Journal of the Royal Statistical Society: Series B*, 73(2):253–272.

Neal, P. and Rao, T. S. (2007). MCMC for integer-valued ARMA processes. *Journal of Time Series Analysis*, 28(1):92–110.

Petrone, S., Rousseau, J., and Scricciolo, C. (2014). Bayes and empirical Bayes: do they merge? *Biometrika*, 101(2):285–302.

Prangle, D. (2015). Summary statistics in approximate Bayesian computation. *arXiv preprint arXiv:1512.05633*.

Robert, C. P. (2016). *Approximate Bayesian Computation: A Survey on Recent Results*, pages 185–205. Springer International Publishing, Cham.

Sisson, S. A. and Fan, Y. (2011). *Likelihood-free MCMC*. Chapman & Hall/CRC, New York.[839].

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.

A Proofs

Let $\{\mathcal{F}_t : t \geq 0\}$ be a filtration associated with the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sequence $\{y_t\}_{t\geq 1}$ is adapted to the filtration $\{\mathcal{F}_t\}$. Let $P(\cdot|\theta)$ denote the generative model for $\mathbf{y}$. Define

$$F_{\mathbf{y}} = P(\cdot|\theta_0, \mathbf{y})$$

to be the true conditional predictive distribution.

Throughout the remainder, let y_{T+1} denote a point of support for the random variable Y_{T+1}. Recall the definitions

$$P_{\mathbf{y}} = \int_{\Theta} P(\cdot|\theta, \mathbf{y}) d\Pi[\theta|\mathbf{y}], \quad G_{\mathbf{y}} = \int_{\Theta} P(\cdot|\theta, \mathbf{y}) d\Pi[\theta|\eta(\mathbf{y})].$$

The results of this section hold under the following high-level assumptions. Lower level sufficient conditions for these assumptions can easily be given, however, such a goal is not germane to the discussion at hand.

Assumption 1 *The following are satisfied: (1) $p(\mathbf{y}|\theta)$ is $\mathcal{F}_T$ measurable for all $\theta \in \Theta$ and for all $T \geq 1$; (2) For all $\theta \in \Theta$ and all $T \geq 1$, $0 < p(\mathbf{y}|\theta) < \infty$; (3) There exists a unique $\theta_0 \in \Theta$, such that $\mathbf{y} \sim P(\cdot|\theta_0) \in \mathbb{P}$; (4) For any $\epsilon > 0$, and $A_\epsilon := \{\theta \in \Theta : \|\theta - \theta_0\| > \epsilon\}$, $\Pi[A_\epsilon|\mathbf{y}] \to_{\mathbb{P}} 0$ and $\Pi[A_\epsilon|\eta(\mathbf{y})] \to_{\mathbb{P}} 0$, i.e. Bayesian consistency of $\Pi[A_\epsilon|\mathbf{y}]$ and $\Pi[A_\epsilon|\eta(\mathbf{y})]$ holds.*

A.1 Lemma

We begin the proof by first showing the following Lemma.

Lemma *$G_{\mathbf{y}}$ is a conditional measure and $G_{\mathbf{y}}(\Omega) = 1$.*

Proof. The result follows by verifying the required conditions for a probability measure.

(1) For any $B \in \mathcal{F}$, $\mathbb{1}[Y \in B]g(Y|\mathbf{y}) \geq 0$ and hence

$$G_{\mathbf{y}}(B) = \int_\Omega \mathbb{1}[Y \in B]g(Y|\mathbf{y})dY \geq 0.$$

(2) By definition, $G_{\mathbf{y}}(\{\emptyset\}) = 0$.

(3) Let $E_k = [a_k, b_k)$, $k \geq 1$, be a collection of disjoint sets (in $\mathcal{F}$). By construction, for all $\omega \in \Omega$, $\mathbb{1}[Y \in E_k]g(Y|\mathbf{y}(\omega)) \geq 0$ and hence

$$\begin{aligned} G_{\mathbf{y}}\left(\bigcup_{k=1}^{\infty} E_k\right) &= \int \mathbb{1}\left[Y \in \bigcup_{k=1}^{\infty} E_k\right] g(Y|\mathbf{y})dY = \int \sum_{k=1}^{\infty} \mathbb{1}[Y \in E_k]g(Y|\mathbf{y})dY \\ &= \sum_{k=1}^{\infty} \int \mathbb{1}[Y \in E_k]g(Y|\mathbf{y})dY, \end{aligned}$$

where the last line follows by Fubini's theorem.

(4) All that remains to be shown is that $G_{\mathbf{y}}(\Omega) = 1$. By definition

$$G_{\mathbf{y}}(\Omega) = \int_\Omega g(Y|\mathbf{y})dY = \int_\Omega \int_\Theta p(Y|\theta, \mathbf{y})d\Pi[\theta|\eta(\mathbf{y})]dY.$$

By Fubini's Theorem,

$$G_{\mathbf{y}}(\Omega) = \int_\Theta \left(\int_\Omega \frac{p(Y, \mathbf{y}, \theta)}{p(\mathbf{y}, \theta)} dY\right) d\Pi[\theta|\eta(\mathbf{y})] = \int_\Theta \frac{p(\mathbf{y}, \theta)}{p(\mathbf{y}, \theta)} d\Pi[\theta|\eta(\mathbf{y})] = \Pi[\Theta|\eta(\mathbf{y})] = 1$$

■

A.2 Theorem 1

Proof. Define ρ_H to be the Hellinger metric, that is, for absolutely continuous probability measures P and G,

$$\rho_H\{P,G\} = \left\{\frac{1}{2}\int\left[\sqrt{dP}-\sqrt{dG}\right]^2 d\mu\right\}^{1/2}, \quad 0 \leq \rho_H\{P,G\} \leq 1$$

for μ the Lebesgue measure, and define ρ_{TV} to be the total variation metric,

$$\rho_{TV}\{P,G\} = \sup_{B\in\mathcal{F}} |P(B)-G(B)|, \quad 0 \leq \rho_{TV}\{P,G\} \leq 2$$

Recall that, according to the definition of merging in Blackwell and Dubins (1962), two predictive measures $P_{\mathbf{y}}$ and $G_{\mathbf{y}}$ are said to merge if

$$\rho_{TV}\{P_{\mathbf{y}}, G_{\mathbf{y}}\} = o_{\mathbb{P}}(1).$$

Fix $\epsilon > 0$ and define the set $V_\epsilon := \{\theta \in \Theta : \rho_H\{F_{\mathbf{y}}, P(\cdot|\mathbf{y},\theta)\} > \epsilon/2\}$. By convexity of $\rho_H\{F_{\mathbf{y}},\cdot\}$ and Jensen's inequality,

$$\begin{aligned}
\rho_H\{F_{\mathbf{y}}, P_{\mathbf{y}}\} &\leq \int_\Theta \rho_H\{F_{\mathbf{y}}, P(\cdot|\mathbf{y},\theta)\}d\Pi[\theta|\mathbf{y}] \\
&\leq \int_{V_\epsilon} \rho_H\{F_{\mathbf{y}}, P(\cdot|\mathbf{y},\theta)\}d\Pi[\theta|\mathbf{y}] + \int_{V_\epsilon^c} \rho_H\{F_{\mathbf{y}}, P(\cdot|\mathbf{y},\theta)\}d\Pi[\theta|\mathbf{y}] \\
&\leq \Pi[V_\epsilon|\mathbf{y}] + \frac{\epsilon}{2}\Pi[V_\epsilon^c|\mathbf{y}].
\end{aligned}$$

By definition, $\theta^0 \notin V_\epsilon$ and therefore, by Assumption 1 Part (4), $\Pi[V_\epsilon|\mathbf{y}] = o_{\mathbb{P}}(1)$. Hence, we can conclude:

$$\rho_H\{F_{\mathbf{y}}, P_{\mathbf{y}}\} \leq o_{\mathbb{P}}(1) + \frac{\epsilon}{2}. \tag{33}$$

Now, apply the triangle inequality to obtain $\rho_H\{P_{\mathbf{y}}, G_{\mathbf{y}}\} \leq \rho_H\{F_{\mathbf{y}}, P_{\mathbf{y}}\} + \rho_H\{F_{\mathbf{y}}, G_{\mathbf{y}}\}$. Using (33), convexity of ρ_H, and Jensen's inequality,

$$\begin{aligned}
\rho_H\{P_{\mathbf{y}}, G_{\mathbf{y}}\} &\leq o_{\mathbb{P}}(1) + \frac{\epsilon}{2} + \int_\Theta \rho_H\{F_{\mathbf{y}}, P(\cdot|\mathbf{y},\theta)\}d\Pi[\theta|\eta(\mathbf{y})] \\
&\leq o_{\mathbb{P}}(1) + \frac{\epsilon}{2} + \int_{V_\epsilon^c} \rho_H\{F_{\mathbf{y}}, P(\cdot|\mathbf{y},\theta)\}d\Pi[\theta|\eta(\mathbf{y})] + \int_{V_\epsilon} \rho_H\{F_{\mathbf{y}}, P(\cdot|\mathbf{y},\theta)\}d\Pi[\theta|\eta(\mathbf{y})] \\
&\leq o_{\mathbb{P}}(1) + \frac{\epsilon}{2} + \frac{\epsilon}{2}\Pi[V_\epsilon^c|\eta(\mathbf{y})] + \Pi[V_\epsilon|\eta(\mathbf{y})].
\end{aligned}$$

From the Bayesian consistency of $\Pi[\cdot|\eta(\mathbf{y})]$, for any $\epsilon' \leq \epsilon$, $A_{\epsilon'}^c \not\subset \limsup_{T\to\infty} V_\epsilon$, where we recall that the set A_ϵ was defined previously as $A_\epsilon := \{\theta \in \Theta : \|\theta-\theta_0\| > \epsilon\}$. Applying again Assumption 1 Part (4), $\Pi[V_\epsilon|\eta(\mathbf{y})] = o_{\mathbb{P}}(1)$, and we can conclude

$$\rho_H\{P_{\mathbf{y}}, G_{\mathbf{y}}\} \leq o_{\mathbb{P}}(1) + \epsilon.$$

For probability distributions P, G, recall that

$$0 \leq \rho_{TV}\{P, G\} \leq \sqrt{2} \cdot \rho_H\{P, G\}.$$

Applying the relationship between ρ_H and ρ_{TV}, yields the stated result. ■

A.3 Theorem 2

Proof.

Part (i): Under correct model specification, for any $B \in \mathcal{F}_{T+1}$

$$\begin{aligned} P_{\mathbf{y}}(B) &= \int_{\Omega}\int_{\Theta} p(Y|\mathbf{y},\theta)d\Pi[\theta|\mathbf{y}]d\delta_Y(B) \\ &= \int_{\Omega}\int_{\Theta} p(Y|\mathbf{y},\theta)d\delta_\theta(\theta_0)d\delta_Y(B) + \int_{\Omega}\int_{\Theta} p(Y|\mathbf{y},\theta)\{d\Pi[\theta|\mathbf{y}] - d\delta_\theta(\theta_0)\}d\delta_Y(B) \\ &= F_{\mathbf{y}}(B) + \int_{\Omega}\int_{\Theta} p(Y|\mathbf{y},\theta)d\delta_Y(B)\{d\Pi[\theta|\mathbf{y}] - d\delta_\theta(\theta_0)\}. \end{aligned} \tag{34}$$

Analysing the second term in $P_{\mathbf{y}}$, $\int_{\Theta}\int_{\Omega} p(Y|\mathbf{y},\theta)d\delta_Y(B)\{d\Pi[\theta|\mathbf{y}] - d\delta_\theta(\theta_0)\}$, we see that the inner integral is a bounded and continuous function of θ for each $\mathbf{y}$. From Assumption 1 part (4), it follows that $P_{\mathbf{y}} = F_{\mathbf{y}} + o_{\mathbb{P}}(1)$, from which we can conclude

$$\mathbb{M}(P_{\mathbf{y}}, F_{\mathbf{y}}) = \int_{\Omega} S(P_{\mathbf{y}}, Y)dF_{\mathbf{y}}(Y) = \int_{\Omega} S(F_{\mathbf{y}}, Y)dF_{\mathbf{y}}(Y) + o_{\mathbb{P}}(1).$$

Similarly, from Assumption 1 part (4), $G_{\mathbf{y}} = F_{\mathbf{y}} + o_{\mathbb{P}}(1)$, and

$$\mathbb{M}(G_{\mathbf{y}}, F_{\mathbf{y}}) = \int_{\Omega} S(G_{\mathbf{y}}, Y)dF_{\mathbf{y}}(Y) = \int_{\Omega} S(F_{\mathbf{y}}, Y)dF_{\mathbf{y}}(Y) + o_{\mathbb{P}}(1).$$

Therefore,

$$\mathbb{M}(P_{\mathbf{y}}, F_{\mathbf{y}}) - \mathbb{M}(G_{\mathbf{y}}, F_{\mathbf{y}}) = o_{\mathbb{P}}(1).$$

Part (ii): Define the random variables, $\hat{Y} = S(P_{\mathbf{y}}, Y_{T+1})$ and $\hat{X} = S(G_{\mathbf{y}}, Y_{T+1})$. The result of **Part (i)** can then be stated as, up to an $o_{\mathbb{P}}(1)$ term, $\mathbb{E}\left[\hat{Y}|\mathbf{y}\right] = \mathbb{E}\left[\hat{X}|\mathbf{y}\right]$. Therefore, up to an $o(1)$ term,

$$\mathbb{E}\left[\hat{Y}\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{Y}|\mathbf{y}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{X}|\mathbf{y}\right]\right] = \mathbb{E}\left[\hat{X}\right].$$

Part (iii): For $\eta_0 = \eta(\mathbf{y})$, rewrite $g(y_{T+1}|\mathbf{y})$ as

$$\begin{aligned} g(y_{T+1}|\mathbf{y}) &= \int_{\Theta} \frac{p(y_{T+1},\theta,\mathbf{y})}{p(\theta,\mathbf{y})} \frac{p(\eta_0|\theta)p(\theta)}{\int_\theta p(\eta_0|\theta)p(\theta)d\theta}d\theta = \int_{\Theta} \frac{p(y_{T+1},\theta,\mathbf{y})}{p(\mathbf{y}|\theta)p(\theta)} \frac{p(\eta_0|\theta)p(\theta)}{\int_\theta p(\eta_0|\theta)p(\theta)d\theta}d\theta \\ &= \int_{\Theta} \frac{p(y_{T+1},\theta,\mathbf{y})}{\int_\theta p(\eta_0|\theta)p(\theta)d\theta} \frac{p(\eta_0|\theta)}{p(\mathbf{y}|\theta)}d\theta \end{aligned}$$

Likewise, $p(y_{T+1}|\mathbf{y})$ can be rewritten as $p(y_{T+1}|\mathbf{y}) = \int_\Theta p(y_{T+1},\theta,\mathbf{y})d\theta / \int_\theta p(\mathbf{y}|\theta)p(\theta)d\theta$. The result follows if and only if $p(\mathbf{y}|\theta) = p(\mathbf{y})p(\eta_0|\theta)$. ■