International Journal of Forecasting 35 (2019) 521–539

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Approximate Bayesian forecasting☆

David T. Frazier [a,*], Worapree Maneesoonthorn [b], Gael M. Martin [a], Brendan P.M. McCabe [c]

[a] Department of Econometrics and Business Statistics, Monash University, Australia
[b] Melbourne Business School, University of Melbourne, Australia
[c] Management School, University of Liverpool, UK

ARTICLE INFO

Keywords:
Bayesian prediction
Likelihood-free methods
Predictive merging
Proper scoring rules
Particle filtering
Jump-diffusion models

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.

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, Drovandi, Pettitt, & Lee, 2015, Fearnhead & Prangle, 2012 and Joyce & Marjoram, 2008) and we refer the reader to reviews by Blum, Nunes, Prangle, and Sisson (2013) and Prangle (2015) for discussions of these different approaches.

Along with the growth in applications of ABC (see Marin, Pudlo, Robert, & Ryder, 2012, Robert, 2016, and Sisson & Fan, 2011, 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 Creel, Gao, Hong, and Kristensen (2015), Frazier, Martin, Robert, and Rousseau (2018), Jasra (2015), Li and Fearnhead (2018a, 2018b), Marin, Pillai, Robert, and

☆ We thank the Editor and two anonymous referees for very constructive and detailed comments on earlier drafts of the paper. This research has been supported by Australian Research Council Discovery Grants No. DP150101728 and DP170100729.
* Corresponding author.
E-mail address: david.frazier@monash.edu (D.T. Frazier).

https://doi.org/10.1016/j.ijforecast.2018.08.003
0169-2070/

Rousseau (2014) and Martin, McCabe, Frazier, Maneesoonthorn, and Robert (2018), with Frazier et al. (2018) 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 the vast majority of 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, are 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.[1]

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.

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? (iv) How can forecast accuracy be used to guide the choice of summary statistics in an empirical setting?

We note that, independent of this research, Canale and Ruggiero (2016) propose the use of ABC as a means of generating nonparametric forecasts of certain functional time series models with intractable likelihoods. In particular, Canale and Ruggiero use ABC sampling as a means of generating h-step-ahead point and interval forecasts for some underlying unknown curve of interest. The authors apply this methodology to the prediction of price dynamics in the Italian natural gas market. Whilst not pursuing the same lines of enquiry as in the current research, the Canale and Ruggiero paper highlights the usefulness of ABC as a forecasting tool in scenarios when exact Bayesian inference - and, hence, exact Bayesian prediction - is infeasible, and thereby provides further evidence of the practical importance of the results we provide herein.

[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.

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 volatility, in a dynamic jump-diffusion model with α-stable volatility transitions, are produced, using different sets of summary statistics, including those extracted from simple auxiliary models with closed-form likelihood functions. Particular focus is given to using out-of-sample predictive performance to choose 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. All Matlab code used in the production of the numerical results will be made available at http://users.monash.edu.au/~gmartin/.

2. Approximate Bayesian computation (ABC): Inference and forecasting

We observe a T-dimensional vector of data $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$, assumed to be generated from some model with likelihood $p(\mathbf{y}|\theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^{k_\theta}$ 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}, $k = 1, 2, \ldots, h$, 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 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:

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

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 small,

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

approximates $p(\theta|\mathbf{y})$ well, and draws from $p_\varepsilon(\theta|\eta(\mathbf{y}))$ 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 a low-dimensional vector of sufficient statistics does not exist. 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 ease and speed.

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

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

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 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

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

[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.

quantity

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

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 Sections 3.2, 3.3 and 3.4 respectively, after a brief review of existing asymptotic results pertaining to $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$ in Section 3.1.

However, before addressing the above questions, we acknowledge here that the ABC posterior $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$ is one of several posterior approximations that have been proposed in the literature. Other such approximations, for example, those produced by variational Bayes (Jaakkola & Jordan, 2000; Tran, Nott, & Kohn, 2017), Bayesian synthetic likelihood (Price, Drovandi, Lee, & Nott, 2018), Bayesian empirical likelihood (Mengersen, Pudlo, & Robert, 2013), or bootstrap methods (Zhu, Marin, & Leisen, 2016), could also be used to construct an approximate predictive. However, to formally characterize the accuracy of any such approximate predictive, relative to the exact predictive $p(y_{T+1}|\mathbf{y})$, we must know a good deal about the theoretical behavior of the posterior approximation itself. This requirement, and the ensuing regularity of the ABC posterior, partly motivates our focus on ABC as the inferential approach underpinning the production of an approximate predictive. In particular, the following section makes substantial use of the theoretical properties of the ABC posterior in characterizing the accuracy of ABF relative to exact Bayesian forecasting.

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, 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, in Section 3.2 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 & Dubins, 1962; Diaconis & 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 $p(y_{T+1}|\mathbf{y})$. In Section 3.4 we then provide numerical illustrations of $g(y_{T+1}|\mathbf{y})$ and $p(y_{T+1}|\mathbf{y})$ for particular models, and for particular choices of summary statistics in the production of $g(y_{T+1}|\mathbf{y})$.

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 approximate forecasting.

3.1. Asymptotic properties of ABC posteriors

We briefly summarize recent theoretical results for ABC 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. (2018) but acknowledge here the important contributions by Li and Fearnhead (2018a, 2018b). As is consistent with the standard approach to Bayesian asymptotics (Ghosh & Ramamoorthi, 2003; van der Vaart, 1998), 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.

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. (2018) 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 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 $\{\mathcal{F}_t : t \geq 1\}$ 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). The distribution $G_{\mathbf{y}}(\cdot)$ 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 $B \in \mathcal{F}$,

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

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, Ghosh, & Samanta, 1995 or Ibragimov & Has'Minskii, 2013) the exact posterior $p(\theta|\mathbf{y})$ will concentrate onto θ_0 as $T \to \infty$. As long as the relevant conditions delineated in Frazier et al. (2018) 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 the Appendix, 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 $T \to \infty$ and $\varepsilon_T \to 0$, with $\mathbb{P}$-probability 1, 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, Rousseau, & Scricciolo, 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 forecaster 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, Balabdaoui, & Raftery, 2007 and Gneiting & Raftery, 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 the 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} 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 the Appendix, 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 $\varepsilon_T = 0$.*

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 $\eta(\mathbf{y})$.

However, if additional assumptions regarding the regularity of $p(\theta|\mathbf{y})$ and $p_\varepsilon(\theta|\eta(\mathbf{y}))$ are satisfied, a ranking between $\mathbb{M}(P_{\mathbf{y}}, F_{\mathbf{y}})$ and $\mathbb{M}(G_{\mathbf{y}}, F_{\mathbf{y}})$, which will hold for large T and with high probability, can be deduced. In particular, a

ranking can be obtained if both $p(\theta|\mathbf{y})$ and $p_\varepsilon(\theta|\eta(\mathbf{y}))$ satisfy a Bernstein-von Mises result (which requires invoking Result 2 in Section 3.1 in the latter case and standard regularity in the former). For $\phi_{\theta,V}$ denoting a normal density function with mean θ and variance–covariance matrix V, a Bernstein-von Mises result for $p(\theta|\mathbf{y})$ and $p_\varepsilon(\theta|\eta(\mathbf{y}))$ implies that

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

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

where $\hat\theta$ is the maximum likelihood estimator, $\mathcal{I}$ 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). For regular models with weakly dependent data, both $\mathcal{I}^{-1}$ and $\mathcal{E}^{-1}$ are $O(T^{-1})$, and $\mathcal{I}^{-1} - \mathcal{E}^{-1}$ 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 $\hat\theta$, we can expand this function as

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

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

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

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,

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

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

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

and - as accords with intuition - predictive accuracy to be greater when based on the exact predictive distribution.[4]

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

$$y_t = \rho \circ y_{t-1} + \varepsilon_t, \quad (13)$$

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

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

and where $B_1(\rho), B_2(\rho), \ldots, B_{y_{t-1}}(\rho)$ are *i.i.d.* Bernoulli random variables each with

$$\Pr(B_j(\rho)=1) = 1 - \Pr(B_j(\rho)=0) = \rho.$$

In the numerical illustration we take ε_t to be *i.i.d.* Poisson with intensity parameter λ. Note that it is with a slight abuse of notation that, in this and later models, we use the notation ε_t to denote a random error term. It will always be clear from the context when the symbol is being used in this way, rather than referring to the ABC tolerance.

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

[4] We reiterate that the derivation of the result in (12) is based on asymptotic approximations of the *unscaled* quantities, $p(y_{T+1}|\mathbf{y})$ and $g(y_{T+1}|\mathbf{y})$, which (in common with all asymptotic results pertaining to unnormalized quantities) is valid for large but finite T.

series data. See Jung and Tremayne (2006) for a review, and Drost, van den Akker, and Werker (2009) and McCabe, Martin, and Harris (2011) for contributions. Of particular note is the work by Martin, Tremayne, and Jung (2014), in which the INARMA model is estimated 'indirectly' via efficient method of moments (Gallant & 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 $\theta = (\rho, \lambda)'$ 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, $\rho \circ y_T$ and ε_T, as

$$\Pr\left(Y_{T+1} = y_{T+1}|\mathbf{y}, \theta\right) = \sum_{s=0}^{\min\{y_{T+1}, y_T\}} \Pr\left(B^{\rho}_{y_T} = s\right) \Pr(\varepsilon_{T+1} = y_{T+1} - s), \qquad (15)$$

where $\Pr\left(B^{\rho}_{y_T} = s\right)$ denotes the probability that a binomial random variable associated with y_T replications (and a probability of 'success', ρ, on each replication) takes a value of s, and where $\Pr(\varepsilon_{T+1} = y_{T+1} - s)$ 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 $\theta_0 = (\rho_0, \lambda_0)' = (0.4, 2)'$. Prior information on θ is specified as $U[0, 1] \times U[0, 10]$.[5] We implement ABC using a nearest-neighbour version of Algorithm 1. This version of ABC replaces Step-3 in Algorithm 1 with the following selection step:

3. Select all θ^i associated with the $\alpha = \delta/N$ smallest distances $d\{\eta(\mathbf{z}^i), \eta(\mathbf{y})\}$ for some δ.

For this experiment, the nearest-neighbour version of ABC is implemented by retaining the simulated draws that lead to the smallest $\alpha = 0.01$ of the $N = 20{,}000$ simulated draws based on a single vector of summary statistics comprising the sample mean of $\mathbf{y}$, 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 Fig. 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.[6]

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 the highest average score is in bold.

	ABF	Exact
LS	**−1.89**	**−1.89**
QS	**0.17**	**0.17**

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 the quadratic score (QS) in Table 1, using the 'observed' value of y_{T+1} that is also simulated.[7] (See Gneiting et al., 2007, for details of these particular scoring rules.) The assumptions under which Theorem 2 holds can be demonstrated analytically in this case, including the Bayesian consistency of $p_{\varepsilon}(\theta|\eta(\mathbf{y}))$; see Appendix A.4.1. 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.[8]

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 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.[9]

We do emphasize at this point that refinements of Algorithm 1 based on either post-sampling corrections (Beaumont, Zhang, & Balding, 2002; Blum, 2010), or the insertion of MCMC or sequential Monte Carlo steps (Beaumont, Cornuet, Marin, & Robert, 2009; Marjoram, Molitor, Plagnol, & Tavaré, 2003; Sisson, Fan, & Tanaka, 2007) may well improve the accuracy with which the exact posteriors are

[5] In this and the following sections we use the simplest possible priors, including truncated uniform priors on location parameters. We acknowledge that these prior choices will have some influence on the posterior densities produced, both exact and approximate. However, given that the sample sizes are reasonable (and large in some cases) we do not expect that influence to be substantial, nor for the conclusions regarding predictive performance to qualitatively alter. In particular we emphasize that the *same* priors are used to generate both the exact and approximate posteriors in all cases.

[6] We 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.

[7] In an expanding window one-step-ahead prediction scheme, the initial sample, say from period 1 to T, is used to produce a one-step-ahead prediction for period $T + 1$. At the next iteration, we use observations up to, and including, time $T + 1$ to produce a prediction for period $T + 2$. This expanding window procedure then iterates until some pre-specified period, say $T + K$.

[8] Additional simulation results, not reported for brevity, demonstrate that the qualitative nature of this result is not sensitive to the choice of θ_0.

[9] 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 2.30 GHz dual processor (each processor with 6 cores) with 16 GB RAM. Note that all computation times quoted in the paper are 'time elapsed' or 'wall-clock' time.

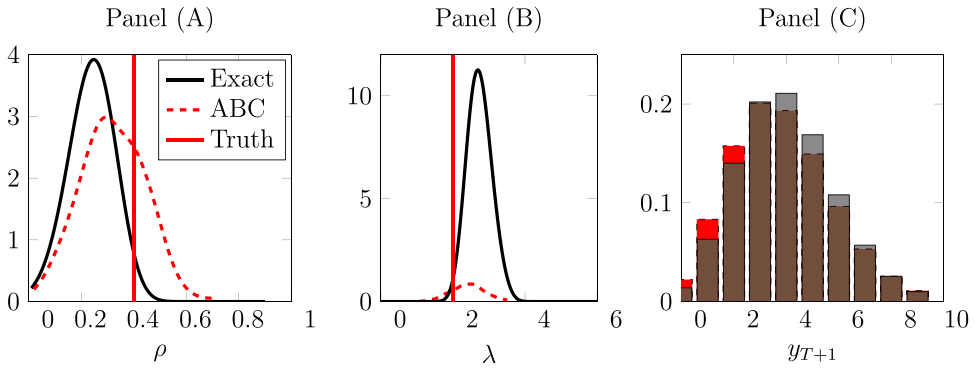

Fig. 1. Panels (A) and (B) depict the marginal posteriors (exact and ABC) for ρ and λ, respectively. The red vertical line (denoted by 'Truth' in the key) represents the true value of the relevant parameter in both panels. Panel (C) plots the one-step-ahead predictive mass functions - both exact and approximate (ABC-based). The brown shading corresponds to an overlap of exact and approximate predictive probabilities. The red shading indicates when the approximate probabilities exceed the exact, with the grey shading indicating the reverse situation.

approximated. However, the key message - both here and in what follows - is that a poor match between exact and approximate posteriors does not necessarily translate into a corresponding poor match at the predictive level; hence, we choose to use the simplest form of the algorithm in all illustrations.

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 $\varepsilon_t \sim i.i.d.N(0, \sigma^2)$, and the true values of the unknown parameters are given by $\theta_{10} = 0.8$, $\theta_{20} = 0.6$ and $\sigma_0 = 1.0$.[10] We specify the following priors: $\theta_1 \sim U(0, 0.99)$, $\theta_2 \sim U(0, 0.99)$ and $\sigma \sim U(0.1, 3)$. Inference on $\theta = (\theta_1, \theta_2, \sigma)'$ is conducted via ABC using the sample autocovariances as summary statistics, with $\eta^{(l)}(\mathbf{y}) = (\gamma_0, \gamma_1, \ldots, \gamma_l)'$, and $\gamma_l = cov(y_t, y_{t-l})$. Four alternative sets of $\eta^{(l)}(\mathbf{y})$ are considered in this case, with $l = 1, 2, 3, 4$. The one-step-ahead approximate predictive densities are estimated for each set by using the selected draws, θ^i, $i = 1, 2, \ldots, N$, (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}, $i = 1, 2, \ldots, N$, 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 Fig. 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.[11] 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 40th draw.

Panel (D) of Fig. 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 Fig. 2 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.

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 continuous 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

[10] Similar to the INAR example, additional simulation results in this MA(2) example, not reported for brevity, demonstrate that the qualitative results are not sensitive to the choice of θ_0.

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

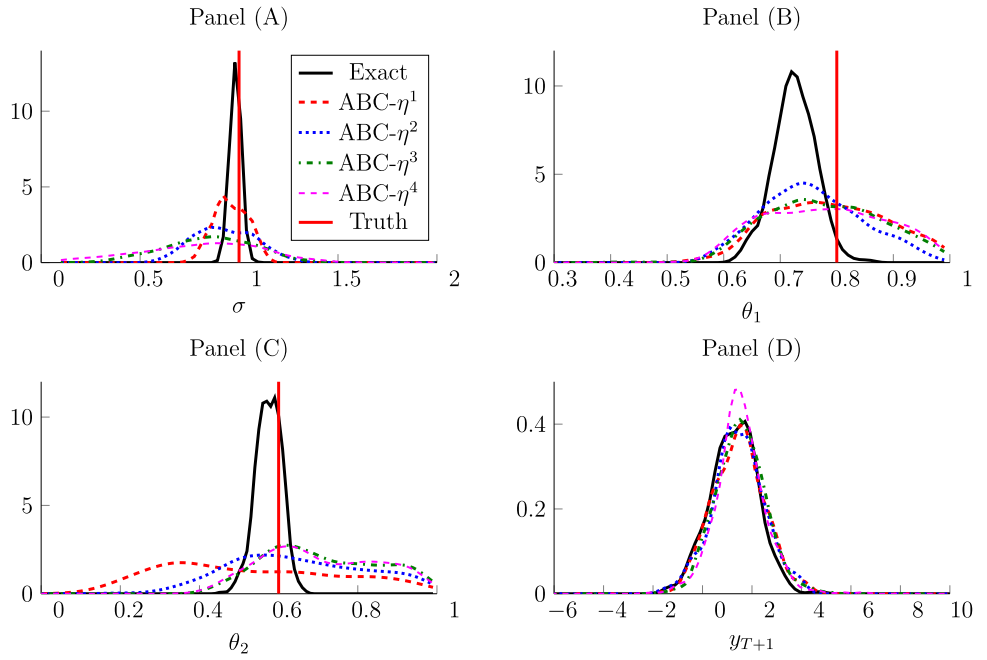

Fig. 2. Panels (A), (B) and (C) depict the marginal posteriors (exact and ABC) for σ, θ_1 and θ_2 respectively. The four approximate posteriors are based on the sets of summaries indicated in the key included in Panel A. Panel (D) plots the one-step-ahead predictive densities - both exact and approximate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

Table 2
Log score (LS), quadratic score (QS) and continuous 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 the 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**

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 min, 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.

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 (see Appendix A.4.2 for this demonstration), whilst $\eta^{(1)}(\mathbf{y})$ can be readily shown to *not* satisfy the sufficient conditions that guarantee 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 & Corander, 2015) defined as,

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

Small RMSE, total variation 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 Fig. 3.

All four panels in Fig. 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

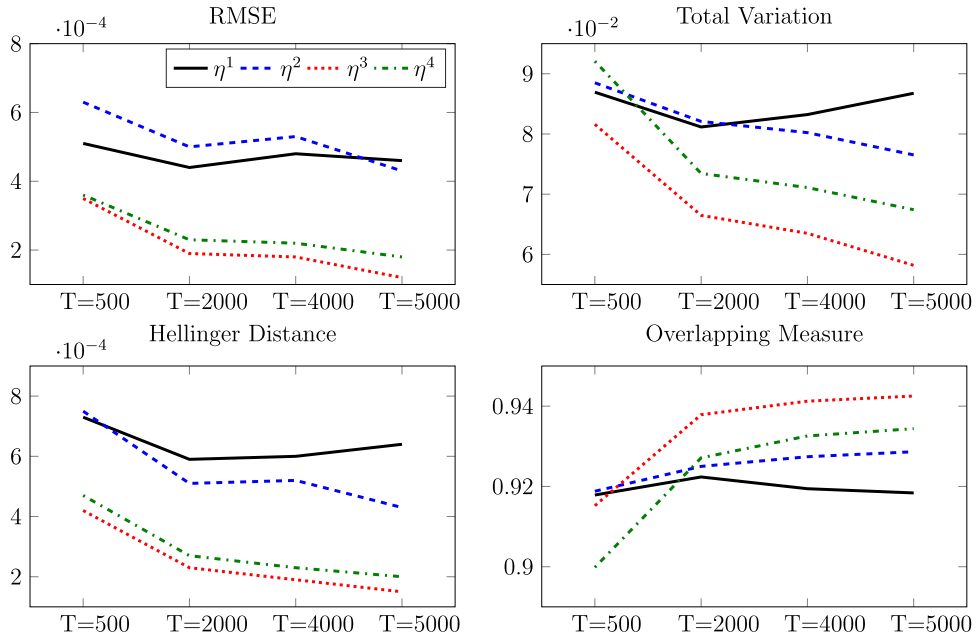

Fig. 3. The four panels depict numerical approximations to the measures in (17)–(20). The key in the upper-left-hand panel indicates the set of summaries that underpins the ABC-based predictive used in each sequence of computations over T.

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. (2018) remain valid under this more common implementation of ABC, but they must be cast in terms of the limiting behaviour of the acceptance probability $\alpha_T = \Pr[d\{\eta(\mathbf{y}), \eta(\mathbf{z})\} \leq \varepsilon_T]$. Under this nearest-neighbour interpretation, Corollary 1 in Frazier et al. (2018) demonstrates that consistency requires $\alpha_T \to 0$ as $T \to \infty$, and, in particular, we require that $\alpha_T \asymp T^{-k_\theta/2}$, where $\asymp$ can be understood as "equal" in an order sense. Moreover, for N_T denoting the number of Monte Carlo draws used in ABC, it must also be the case that $N_T \to \infty$ as $\alpha_T \to 0$. To jointly satisfy these conditions 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 5000 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 $T = 5000$.[12]

Before concluding, we note that even though the predictive based on the ABC posterior $p_\varepsilon(\theta|\eta^{(1)}(\mathbf{y}))$ does not exhibit evidence of merging, as is clear from Panel D in Fig. 2, for $T = 500$ this approximate predictive is still very accurate as an estimate of the exact predictive. Therefore, we conjecture that, in relatively small samples Bayesian consistency may not be a necessary condition for ABC to yield predictives that are close to the exact. However, the numerical merging results demonstrate that this accuracy would degrade as the sample size increased if the ABC posterior were not consistent.

4. ABF in state space models

So 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.

[12] 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 (2018a) for alternative sampling schemes that only require $N_T \to \infty$ very slowly.

State space models, in which the set of unknowns is augmented by a vector of random parameters that is of dimension equal to or greater than the sample size, present additional challenges for ABC (Creel & Kristensen, 2015; Martin et al., 2018), 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 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 $\theta = (\theta_1, \theta_2)'$. Prior specifications $\theta_1 \sim U(0.5, 0.99)$ and $\theta_2 \sim U(0.05, 0.5)$ are employed. 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., 2018).

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 $T+1$, V_{T+1}. Defining $p(V_{T+1}, \mathbf{V}, \theta|\mathbf{y})$ as the joint posterior for this full set of unknowns (with $\mathbf{V} = (V_1, V_2, \ldots, V_T)'$), and recognizing the Markovian structure in the (log) variance process, we can represent this predictive as

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

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

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

In this case, however, draws are produced from $p_\varepsilon(\theta|\eta(\mathbf{y}))$ via the nearest-neighbour version of Algorithm 1 (with $\alpha = 0.01$ and $N = 50{,}000$), 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 sampling step to be embedded within the simulation algorithm, is not necessary. All that is required is that $\mathbf{V}_{1:T-1}$ is integrated out, and this can occur via a forward filtering step alone. 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 Fig. 4 depict the marginal ABC posteriors of θ_1 and θ_2 alongside the MCMC-based comparators. The dashed curve in Panel (C) of Fig. 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.

The importance of the particle filtering step in obtaining this (near) equivalence is highlighted by the inclusion of a third predictive (the dot-dashed curve) in Panel (C) of Fig. 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 Fig. 4) requires *only* a particle filtering step.

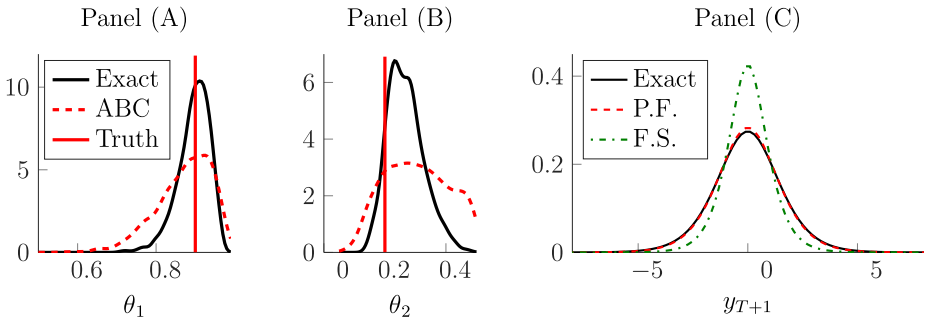

Fig. 4. Panels (A) and (B) depict the marginal posteriors (exact and ABC) for θ_1 and θ_2, respectively. Panel (C) plots the one-step-ahead predictive density functions - both exact and approximate (ABC-based). P.F. indicates the approximate predictive computed using the particle filtering step; F.S.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

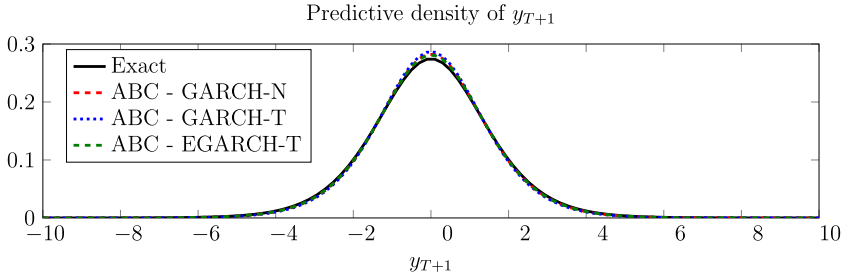

Fig. 5. Exact and approximate (ABC-based) predictives. The three ABC-based predictives are based on the auxiliary models indicated in the key. All approximate predictives use the particle filtering step.

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 Fig. 4 (namely the full and dashed curves), Fig. 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 modeled via an asymmetric specification for the (logarithm of 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 (and hence summary statistics) underpinning $p_\varepsilon(\theta|\eta(\mathbf{y}))$ has little impact on the nature of the resultant predictive, with all three auxiliary models generating approximate predictives that are extremely close to the exact.[13]

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 MCMC 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, Chou, & Kroner, 1992); with more recent work focusing also on random 'jump' processes, both in the asset price itself and its volatility (Bandi & Renò, 2016; Broadie, Chernov, & Johannes, 2007; Maneesoonthorn, Forbes, & Martin, 2017). Empirical regularities documented in the option pricing literature (see Garcia, Lewis,

[13] In this example the computing times for the calculation of the exact and approximate predictives are very similar. However, we remind the reader that we have only implemented a naive version of the ABC algorithm, which could easily be optimized to yield a faster implementation of ABF. In contrast, our MCMC implementation uses cutting-edge sparse-matrix calculations to minimize the execution time required to produce the exact predictive.

Pastorello, & Renault, 2011 for a 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 & 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 (Aït-Sahalia, Cacho-Diaz, & Laeven, 2015; Bandi & Renò, 2016; Fulop, Li, & Yu, 2014; 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 & Scharth, 2012; Maneesoonthorn et al., 2017; Maneesoonthorn, Martin, Forbes, & Grose, 2012). To capture the possibility of extreme movements in volatility, and in the spirit of Lombardi and Calzolari (2009) and Martin et al. (2018), 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, Mallows, and Stuck (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 $BV_t = \frac{\pi}{2}\left(\frac{M}{M-1}\right)\sum_{i=2}^{M}\left|r_{t_i}\right|\left|r_{t_{i-1}}\right|$, with r_{t_i} denoting the ith, 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 & 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 Eqs. (27) and (28), h_t, Z_t and ΔN_t, evolve, respectively, according to

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

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

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

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 Aït-Sahalia et al. (2015), Fulop et al. (2014) and Maneesoonthorn et al. (2017). We estimate d (in (31)) indirectly via the unconditional intensity implied by this particular structure, namely, $d^0 = d/(1-\beta-\gamma)$.

Exact inference on the full set of static parameters,

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

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 relative accuracy of approximate predictives based on alternative choices of summaries, $\eta(\mathbf{y})$.[14] 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, in turn, required to construct $g(y_{T+1}|\mathbf{y})$ under any choice for $\eta(\mathbf{y})$. We adopt the bootstrap particle filter of Gordon, Salmond, and Smith (1993) for this purpose.[15] 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, as shown in Section 3.3. Hence, in choosing $\eta(\mathbf{y})$, from a set of alternatives, according to the accuracy of the associated predictive, we are - in spirit - choosing an approximate predictive that is as close as possible (in terms of predictive accuracy) 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, $JV_t =$

[14] In contrast to the previous numerical examples, where y_t was univariate, in this example our goal is to jointly forecast log-returns, r_t, and the logarithm of bi-power variation, $\ln BV_t$. Therefore, in what follows $y_{T+1} = (r_{T+1}, \ln BV_{T+1})'$.

[15] 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.

Table 3
Auxiliary model specifications for ABC posterior inference for the model in (27)–(31). The error terms, ε_t, ζ_t and u_t, in the second and third columns 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. The final column gives the set of supplementary summary statistics used in addition to the scores from each auxiliary model. The total number of summary statistics used in each specification is denoted by d_η in the first column.

Auxiliary model		Supplementary statistics
GARCH-N $d_\eta = 11$	$r_t = \sigma_t\varepsilon_t, \varepsilon_t \sim N(0,1)$ $\sigma_t^2 = \gamma_0 + \gamma_1 r_{t-1}^2 + \gamma_2\sigma_{t-1}^2$	$Mean(sign(r_t)\sqrt{JV_t})$, $Var(JV_t)$ $Corr(JV_t, JV_{t-1})$ Skewness($\ln BV_t$), Kurtosis($\ln BV_t$) Estimated regression coefficients from: $\ln BV_t = \kappa_0 + \kappa_1 \ln\hat{\sigma}_t^2 + \kappa_3\zeta_t$
GARCH-T $d_\eta = 12$	$r_t = \sigma_t\varepsilon_t, \varepsilon_t \sim t(\nu)$ $\sigma_t^2 = \gamma_0 + \gamma_1 r_{t-1}^2 + \gamma_2\sigma_{t-1}^2$	$Mean(sign(r_t)\sqrt{JV_t})$, $Var(JV_t)$ $Corr(JV_t, JV_{t-1})$ Skewness($\ln BV_t$), Kurtosis($\ln BV_t$) Estimated regression coefficients from: $\ln BV_t = \kappa_0 + \kappa_1 \ln\hat{\sigma}_t^2 + \kappa_3\zeta_t$
TARCH-T $d_\eta = 13$	$r_t = \sigma_t\varepsilon_t, \varepsilon_t \sim t(\nu)$ $\sigma_t^2 = \gamma_0 + \gamma_1 r_{t-1}^2 + \gamma_2 I_{(r_{t-1}<0)} r_{t-1}^2$ $+\gamma_3\sigma_{t-1}^2$	$Mean(sign(r_t)\sqrt{JV_t})$, $Var(JV_t)$ $Corr(JV_t, JV_{t-1})$ Skewness($\ln BV_t$), Kurtosis($\ln BV_t$) Estimated regression coefficients from: $\ln BV_t = \kappa_0 + \kappa_1 \ln\hat{\sigma}_t^2 + \kappa_3\zeta_t$
RGARCH $d_\eta = 12$	$r_t = \sigma_t\varepsilon_t, \varepsilon_t \sim N(0,1)$ $\ln\sigma_t^2 = \gamma_0 + \gamma_1 \ln BV_{t-1} + \gamma_2 \ln\sigma_{t-1}^2$ $\ln BV_t = \gamma_3 + \gamma_4 \ln\sigma_{t-1}^2 + \gamma_5\varepsilon_t$ $+\gamma_6\left(\varepsilon_t^2 - 1\right) + \gamma_7 u_t, u_t \sim N(0,1)$	$Mean(sign(r_t)\sqrt{JV_t})$, $Var(JV_t)$ $Corr(JV_t, JV_{t-1})$, Kurtosis($\ln BV_t$)

$\max(RV_t - BV_t, 0)$, where $RV_t = \sum_{i=1}^{M} r_{t_i}^2$ 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, Huang, and Shek (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. All ABC posteriors are produced by the nearest-neighbour version of Algorithm 1 described in Section 3.4.1, but with α_T and N_T depending on the sample size T as per Frazier et al. (2018) (see Section 3.4.3 for additional discussion.) We note that the ability of GARCH auxiliary models to yield summary statistics that guarantee posterior concentration has been numerically verified in similar models in Martin et al. (2018). However, we believe a formal verification of Bayesian consistency in the current context, as was done with the examples in Section 3.4, is beyond the scope of this paper.

5.2. Empirical forecasting results

In Table 5 we report the ABC marginal 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 noticeably 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 for 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

Table 4
Lower and upper bounds of 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	d	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	0.50
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	−0.02	(−0.47,0.47)	0.00	(−0.49,0.46)	−0.01	(−0.47,0.48)	−0.01	(−0.48,0.47)
ψ_1	1.26	(0.77,1.49)	1.25	(0.83,1.49)	1.20	(0.73,1.49)	0.96	(0.51,1.45)
σ_{BV}	0.45	(0.02,0.96)	0.47	(0.03,0.95)	0.48	(0.02,0.95)	0.55	(0.04,0.99)
ω	−0.04	(−0.68,0.38)	−0.10	(−0.34,0.20)	−0.17	(−0.48,−0.01)	0.19	(−0.95,0.97)
ρ	0.94	(0.81,0.99)	0.93	(0.83,0.99)	0.92	(0.81,0.98)	0.79	(0.52,0.99)
σ_h	0.20	(0.08,0.29)	0.21	(0.08,0.30)	0.20	(0.06,0.30)	0.13	(0.01,0.29)
α	1.76	(1.52,1.98)	1.76	(1.52,1.99)	1.77	(1.52,1.99)	1.80	(1.52,1.99)
d^0	0.11	(0.01,0.28)	0.11	(0.01,0.27)	0.10	(0.01,0.28)	0.10	(0.01,0.27)
β	0.69	(0.51,0.90)	0.69	(0.51,0.91)	0.69	(0.51,0.90)	0.69	(0.52,0.90)
γ	0.12	(0.02,0.20)	0.12	(0.02,0.20)	0.12	(0.02,0.20)	0.13	(0.03,0.20)
μ	0.07	(−0.87,0.94)	0.05	(−0.86,0.90)	0.12	(−0.81,0.88)	0.23	(−0.69,0.94)
σ_z	1.21	(0.52,2.57)	1.23	(0.53,2.72)	1.14	(0.53,2.49)	1.01	(0.53,2.15)

Table 6
Average predictive log score (LS), quadratic score (QS) and continuous rank probability score (CRPS) for the one-step-ahead approximate 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	−1.57	−1.28	**−1.20**	−1.95
	QS	0.38	0.47	**0.52**	0.27
	CRPS	−1.52	−1.05	**−0.99**	−2.10
$\ln BV_t$	LS	**−2.73**	−2.76	−2.93	−2.83
	QS	**0.10**	0.05	0.02	0.09
	CRPS	−2.04	−1.42	**−1.38**	−2.57

predictive scores in all but one case) 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 approximate predictive produced using 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.

6. Discussion

This paper explores the use of approximate Bayesian computation (ABC) in generating probabilistic forecasts and proposes 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 a short amount of time. Whilst the asymptotic results based on merging formally exploit the property of Bayesian consistency, numerical evidence suggests that lack of consistency for the ABC posteriors does not preclude the possibility of a close match to the exact predictive being produced in any given finite sample. The theoretical results presented regarding expected scores are 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, Robert, & Rousseau, 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.

Appendix. Proofs

Let $\{\mathcal{F}_t : t \geq 1\}$ 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 appendix 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 of Theorem 1 by first proving 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,

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

A.2. Theorem 1[16]

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

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

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

$$V_\epsilon := \{\theta \in \Theta : \rho_H^2\{F_{\mathbf{y}}, P(\cdot|\mathbf{y}, \theta)\} > \epsilon/4\}.$$

By convexity of $\rho_H^2\{F_{\mathbf{y}}, \cdot\}$, and Jensen's inequality,

$$\begin{aligned} \rho_H^2\{F_{\mathbf{y}}, P_{\mathbf{y}}\} &\leq \int_{\Theta} \rho_H^2\{F_{\mathbf{y}}, P(\cdot|\mathbf{y}, \theta)\} d\Pi[\theta|\mathbf{y}] \\ &= \int_{V_\epsilon} \rho_H^2\{F_{\mathbf{y}}, P(\cdot|\mathbf{y}, \theta)\} d\Pi[\theta|\mathbf{y}] \\ &\quad + \int_{V_\epsilon^c} \rho_H^2\{F_{\mathbf{y}}, P(\cdot|\mathbf{y}, \theta)\} d\Pi[\theta|\mathbf{y}] \\ &= \Pi[V_\epsilon|\mathbf{y}] + \frac{\epsilon}{4} \Pi[V_\epsilon^c|\mathbf{y}]. \end{aligned}$$

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:

$$\rho_H^2\{F_{\mathbf{y}}, P_{\mathbf{y}}\} \leq o_{\mathbb{P}}(1) + \frac{\epsilon}{4}. \tag{33}$$

Likewise, a similar argument yields

$$\begin{aligned} \rho_H^2\{F_{\mathbf{y}}, G_{\mathbf{y}}\} &\leq \int_{\Theta} \rho_H^2\{F_{\mathbf{y}}, P(\cdot|\mathbf{y}, \theta)\} d\Pi[\theta|\eta(\mathbf{y})] \\ &= \int_{V_\epsilon} \rho_H^2\{F_{\mathbf{y}}, P(\cdot|\mathbf{y}, \theta)\} d\Pi[\theta|\eta(\mathbf{y})] \\ &\quad + \int_{V_\epsilon^c} \rho_H^2\{F_{\mathbf{y}}, P(\cdot|\mathbf{y}, \theta)\} d\Pi[\theta|\eta(\mathbf{y})] \\ &= \Pi[V_\epsilon|\eta(\mathbf{y})] + \frac{\epsilon}{4} \Pi[V_\epsilon^c|\eta(\mathbf{y})] \\ &= o_{\mathbb{P}}(1) + \frac{\epsilon}{4}. \end{aligned} \tag{34}$$

Now, note that

$$\begin{aligned} \frac{1}{2}\left[\rho_H^2\{F_{\mathbf{y}}, P_{\mathbf{y}}\} + \rho_H^2\{F_{\mathbf{y}}, G_{\mathbf{y}}\}\right] &\geq \frac{1}{4}\left[\rho_H\{F_{\mathbf{y}}, P_{\mathbf{y}}\} + \rho_H\{F_{\mathbf{y}}, G_{\mathbf{y}}\}\right]^2 \\ &\geq \frac{1}{4}\left[\rho_H\{P_{\mathbf{y}}, G_{\mathbf{y}}\}\right]^2, \end{aligned}$$

where the first line follows from the Cauchy-Schwartz inequality and the second line from the triangle inequality.

[16] A previous version of the proof of this theorem contained an error. We thank Alex Cooper for bringing this to our attention.

Applying equations (33) and (34), we then obtain

$$\rho_H^2\{P_{\mathbf{y}}, G_{\mathbf{y}}\} \leq \epsilon + o_{\mathbb{P}}(1).$$

Recall that, for probability distributions P, G,

$$0 \leq \rho_{TV}^2\{P, G\} \leq 4 \cdot \rho_H^2\{P, G\}.$$

Applying this relationship between ρ_H^2 and ρ_{TV}^2, yields the stated result. ■

A.3. Theorem 2

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

$$\begin{aligned} P_{\mathbf{y}}(B) &= \int_{\Omega}\int_{\Theta} p(Y|\mathbf{y},\theta)d\Pi[\theta|\mathbf{y}]d\delta_Y(B) \\ &= \int_{\Omega}\int_{\Theta} p(Y|\mathbf{y},\theta)d\delta_\theta(\theta_0)d\delta_Y(B) \\ &\quad + \int_{\Omega}\int_{\Theta} p(Y|\mathbf{y},\theta)\{d\Pi[\theta|\mathbf{y}] - d\delta_\theta(\theta_0)\}d\delta_Y(B) \\ &= F_{\mathbf{y}}(B) + \int_{\Omega}\int_{\Theta} p(Y|\mathbf{y},\theta)d\delta_Y(B) \\ &\quad \times \{d\Pi[\theta|\mathbf{y}] - d\delta_\theta(\theta_0)\}. \end{aligned} \tag{35}$$

The second term in Eq. (35), $\int_{\Theta}\int_{\Omega} p(Y|\mathbf{y},\theta)d\delta_Y(B)\{d\Pi[\theta|\mathbf{y}] - d\delta_\theta(\theta_0)\}$, is a bounded and continuous function of θ for each $\mathbf{y}$. Therefore, from the posterior concentration of $\Pi[\theta|\mathbf{y}]$ to δ_{θ_0}, Assumption 1 part (4),

$$\int_{\Omega}\int_{\Theta} p(Y|\mathbf{y},\theta)d\delta_Y(B)\{d\Pi[\theta|\mathbf{y}] - d\delta_\theta(\theta_0)\} = o_{\mathbb{P}}(1)$$

and it follows that $P_{\mathbf{y}} = F_{\mathbf{y}} + o_{\mathbb{P}}(1)$. Applying this result to $\mathbb{M}(P_{\mathbf{y}}, F_{\mathbf{y}})$ we can conclude

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

The same derivations to the above yield that, under Assumption 1 part (4), $G_{\mathbf{y}} = F_{\mathbf{y}} + o_{\mathbb{P}}(1)$, and

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

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 $p(\mathbf{y}|\theta) = p(\mathbf{y})p(\eta_0|\theta)$. ■

A.4. Posterior consistency in the INAR(1) and MA(2) examples

Under the assumption of correct model specification, posterior consistency in ABC can be demonstrated by verifying the sufficient conditions given in Theorem 1 of Frazier et al. (2018), which we restate here for ease of exposition:

[A1] There exist a continuous, injective map $b : \Theta \to \mathcal{B} \subset \mathbb{R}^{k_\eta}$ and a function $\rho_T(\cdot)$ satisfying: $\rho_T(u) \to 0$ as $T \to \infty$ for all $u > 0$, and $\rho_T(u)$ monotone non-increasing in u (for any given T), such that, for all $\theta \in \Theta$,

$$P_\theta\left[d\{\eta(\mathbf{z}), b(\theta)\} > u\right] \leq c(\theta)\rho_T(u),$$
$$\int_{\Theta} c(\theta)d\Pi(\theta) < +\infty$$

where either of the following is satisfied:

- **(i)** *Polynomial deviations:* There exist a positive sequence $v_T \to +\infty$ and $u_0, \kappa > 0$ such that $\rho_T(u) = v_T^{-\kappa}u^{-\kappa}$, for $u \leq u_0$.
- **(ii)** *Exponential deviations:* There exists $h_\theta(\cdot) > 0$ such that $P_\theta[d\{\eta(\mathbf{z}), b(\theta)\} > u] \leq c(\theta)e^{-h_\theta(uv_T)}$ and there exists $c, C > 0$ such that

$$\int_{\Theta} c(\theta)e^{-h_\theta(uv_T)}d\Pi(\theta) \leq Ce^{-c(uv_T)^\tau}, \quad \text{for } u \leq u_0.$$

[A2] The prior $p(\theta)$ is absolutely continuous with respect to the Lebesgue measure and satisfies $p(\theta^0) > 0$.

A.4.1. INAR(1)

Recall the INAR(1) model

$$y_t = \sum_{j=0}^{y_{t-1}} B_j(\rho) + \varepsilon_t,$$

where $B_j(\rho)$ are *i.i.d* are Bernoulli random variables with probability ρ, and ε_t is *i.i.d* Poisson with intensity parameter λ. The summary statistics chosen for this example were the sample mean, $\bar{y}$, and the first three sample autocovariances.

The parameters are $\theta = (\rho, \lambda)'$ and our prior space is uniform over

$$\Theta := \{\theta \in \Theta : \rho \in [0, 1-\delta],\ \lambda \in [0, 10]\},$$

for some small $\delta > 0$. The uniform prior $p(\theta)$ over Θ automatically fulfills Assumption **[A2]** for θ_0 in this space.

For any $\theta \in \Theta$ and $\mathbf{z}$ simulated from (13), it follows that

$$\eta(\mathbf{z}) = \begin{pmatrix} \bar{z} \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \begin{pmatrix} \lambda/(1-\rho) \\ \rho \\ \rho^2 \\ \rho^3 \end{pmatrix} + o_P(1).$$

Define $b(\theta) = (\lambda/(1-\rho), \rho, \rho^2, \rho^3)'$ and note that $b(\theta)$ is continuous. The linear allocation of ρ as the second element of $b(\theta)$ ensures that $\theta \mapsto b(\theta)$ is injective in θ. From the structure of $\eta(\mathbf{z})$, $V(\theta) = \mathbb{E}[\{\eta(\mathbf{z}) - b(\theta)\}\{\eta(\mathbf{z}) - b(\theta)\}']$ satisfies $\text{tr}\{V(\theta)\} < \infty$ for all $\theta \in \Theta$. By Markov's inequality,

$$P_\theta\{\|\eta(\mathbf{z}) - b(\theta)\| > u\} = P_\theta\left\{\|\eta(\mathbf{z}) - b(\theta)\|^2 > u^2\right\} \leq \frac{\text{tr}\{V(\theta)\}}{u^2 T}.$$

As a result, Assumption **[A1]** is satisfied with $\rho_T(u) = 1/(T^{1/2}u)^2$.

A.4.2. MA(2)

We now verify the conditions in the moving average model of order two:

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} \quad (t = 1, \ldots, T),$$

where $\{e_t\}_{t=1}^T$ is a sequence of white noise random variables with variance σ^2 such that, for some $\delta > 0$, $\mathbb{E}[e_t^{4+\delta}] < \infty$. Our prior for $\theta = (\sigma^2, \theta_1, \theta_2)'$ is uniform over the following region,

$$\Theta := \left\{\theta \in \Theta : 0 \leq \sigma^2 \leq 3,\ 0 \leq \theta_1 \leq 1-\delta,\ 0 \leq \theta_2 \leq 1-\delta\right\},$$

for some small $\delta > 0$. The summary statistics for this exercise are given by the sample autocovariances $\eta_j(\mathbf{y}) = T^{-1}\sum_{t=1+j}^T y_t y_{t-j}$, for $j = 0, 1, \ldots, l$. For any $\theta \in \Theta$, $\eta_j(\mathbf{z}) = T^{-1}\sum_{t=1+j}^T z_t z_{t-j}$. Define $b_j(\theta) = \mathbb{E}_\theta(z_t z_{t-j})$ and take $b(\theta) = (b_0(\theta), b_1(\theta), b_2(\theta))'$. Each choice of the summary statistics in the MA(2) example that leads to a Bayesian consistent posterior has these components in common.

Firstly, note that $\theta \mapsto b(\theta) = (\sigma^2(1+\theta_1^2+\theta_2^2), (1+\theta_2)\theta_1, \theta_2)'$ is continuous in θ. In addition, from the linear allocation of θ_2 in $b_2(\theta)$, it follows that $\theta \mapsto b(\theta)$ is injective over Θ. Now, take $d\{\eta(\mathbf{z}), b(\theta)\} = \|\eta(\mathbf{z}) - b(\theta)\|$ for simplicity. Under the moment restriction on e_t above, $V(\theta) = \mathbb{E}[\{\eta(\mathbf{z}) - b(\theta)\}\{\eta(\mathbf{z}) - b(\theta)\}']$ satisfies $\text{tr}\{V(\theta)\} < \infty$ for all $\theta \in \Theta$. By Markov's inequality,

$$P_\theta\{\|\eta(\mathbf{z}) - b(\theta)\| > u\} = P_\theta\left\{\|\eta(\mathbf{z}) - b(\theta)\|^2 > u^2\right\} \leq \frac{\text{tr}\{V(\theta)\}}{u^2 T} + o(1/T),$$

where the $o(1/T)$ term comes from the fact that there are finitely many non-zero covariance terms due to the m-dependent nature of the series. As a result, Assumption **[A1]** is satisfied with $\rho_T(u) = 1/(T^{1/2}u)^2$.

The uniform prior $p(\theta)$ automatically fulfills Assumption **[A2]** for θ_0 in this space.

References

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

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

Barndorff-Nielsen, O. E., & 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., & Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, *96*(4), 983–990.

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

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

Blomstedt, P., & 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*(491), 1178–1187.

Blum, M. G. B., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, *28*(2), 189–208.

Bollerslev, T., Chou, R. Y., & 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., & Johannes, M. (2007). Model specification and risk premia: Evidence from futures options. *The Journal of Finance*, *62*(3), 1453–1490.

Canale, A., & Ruggiero, M. (2016). Bayesian nonparametric forecasting of monotonic functional time series. *Electronic Journal of Statistics*, *10*(2), 3265–3286.

Chambers, J. M., Mallows, C. L., & 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., & 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., & Kristensen, D. (2015). Bayesian indirect inference and the ABC of GMM. arXiv preprint, arXiv:1512.07385.

Creel, M., & 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., & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, *14*, 1–26.

Drost, F. C., van den Akker, R., & 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., & Lee, A. (2015). Bayesian indirect inference using a parametric auxiliary model. *Statistical Science*, *30*(1), 72–95.

Fearnhead, P., & 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., & Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika*, *105*(3), 593–607.

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

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

Gallant, R., & Tauchen, G. (1996). Which moments to match? *Econometric Theory*, *12*, 657–681.

Garcia, R., Lewis, M.-A., Pastorello, S., & 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., & Samanta, T. (1995). On convergence of posterior distributions. *The Annals of Statistics*, *23*(6), 2145–2152.
Ghosh, J. K., & Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer-Verlag New York, Inc.
Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B.*, *69*(2), 243–268.
Gneiting, T., & 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., & Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, *140*(2), 107–113.
Hansen, P. R., Huang, Z., & 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., & Has'Minskii, R. Z. (2013). *Statistical estimation: Asymptotic theory, Vol. 16*. Springer Science & Business Media.
Jaakkola, T. S., & Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, *10*(1), 25–37.
Jasra, A. (2015). Approximate Bayesian computation for a class of time series models. *International Statistical Review*, *83*(3), 405–435.
Joyce, P., & Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, *7*(1), 1–16.
Jung, R. C., & Tremayne, A. (2006). Binomial thinning models for integer time series. *Statistical Modelling*, *6*(2), 81–96.
Koopman, S. J., & 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., & Fearnhead, P. (2018a). Convergence of regression-adjusted approximate Bayesian computation. *Biometrika*, *105*(2), 301–318.
Li, W., & Fearnhead, P. (2018b). On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika*, *105*(2), 285–299.
Lombardi, M. J., & Calzolari, G. (2009). Indirect estimation of α-stable stochastic volatility models. *Computational Statistics & Data Analysis*, *53*(6), 2298–2308.
Maneesoonthorn, W., Forbes, C. S., & 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., & 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., & 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., & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, *22*(6), 1167–1180.
Marjoram, P., Molitor, J., Plagnol, V., & 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. M., Frazier, D. T., Maneesoonthorn, W., & Robert, C. P. (2018). Auxiliary likelihood-based approximate Bayesian computation in state space models. *Journal of Graphical and Computational Statistics* arXiv preprint, arXiv:1604.07949. Conditionally accepted.
Martin, V. L., Tremayne, A. R., & 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., & Harris, D. (2011). Efficient probabilistic forecasts for counts. *Journal of the Royal Statistical Society. Series B.*, *73*(2), 253–272.
Mengersen, K. L., Pudlo, P., & Robert, C. P. (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, *110*(4), 1321–1326.
Neal, P., & Rao, T. S. (2007). MCMC for integer-valued ARMA processes. *Journal of Time Series Analysis*, *28*(1), 92–110.
Petrone, S., Rousseau, J., & 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.
Price, L. F., Drovandi, C. C., Lee, A., & Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, *27*(1), 1–11.
Robert, C. P. (2016). Approximate Bayesian computation: A survey on recent results. In R. Cools, & D. Nuyens (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods. Springer Proceedings in Mathematics & Statistics, vol. 163*. Cham: Springer.
Sisson, S. A., & Fan, Y. (2011). *Likelihood-free MCMC*. New York: Chapman & Hall/CRC.
Sisson, S. A., Fan, Y., & Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *104*(6), 1760–1765.
Tran, M.-N., Nott, D. J., & Kohn, R. (2017). Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, *26*(4), 873–882.
van der Vaart, A. W. (1998). *Asymptotic statistics, Vol. 3*. Cambridge University Press.
Zhu, W., Marin, J. M., & Leisen, F. (2016). A bootstrap likelihood approach to Bayesian computation. *Australian and New Zealand Journal of Statistics*, *58*(2), 227–244.

David T. Frazier is Senior Lecturer, Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia. His research interests include statistical and econometric theory, simulation-based inference, Bayesian inference and financial econometrics.

Worapree (Ole) Maneesoonthorn is Lecturer, Melbourne Business School, University of Melbourne, Australia. Her research interests include Bayesian inference, simulation methods and financial econometrics.

Gael M. Martin is Professor, Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia. Her research interests include Bayesian econometrics, simulation methods, financial econometrics and forecasting methodology.

Brendan P.M. McCabe is Professor of Econometrics, School of Management, University of Liverpool, U.K. His research interests include statistical and econometric theory, Bayesian econometrics, count time series analysis and forecasting methodology.